Predicting Postoperative Complications for Gastric Cancer Patients using Data Mining

Hugo Peixoto[1], Alexandra Francisco[2], Ana Duarte[2], Márcia Esteves[2], Sara Oliveira[2], Vítor Lopes[3], António Abelha[1] and José Machado[1]

[1] Algoritmi Research Center, University of Minho, Campus Gualtar, 4470 Braga, Portugal
[2] University of Minho, Campus Gualtar, 4470 Braga, Portugal
[3] Tâmega e Sousa Hospital Center, Av. Padre Américo, 4564 Penafiel, Portugal
jmac@di.uminho.pt

Abstract. Gastric cancer refers to the development of malign cells that can grow in any part of the stomach. With the vast amount of data being collected daily in healthcare environments, it is possible to develop new algorithms which can support the decision making processes in gastric cancer patients treatment. This paper aims to predict, using the CRISP-DM methodology, the outcome from the hospitalization of gastric cancer patients who have undergone surgery, as well as the occurrence of postoperative complications during surgery. The study showed that, on one hand, the RF and NB algorithms are the best in the detection of an outcome of hospitalization, taking into account patients' clinical data. On the other hand, the algorithms J48, RF, and NB offer better results in predicting postoperative complications.

Keywords: Data Mining, Clinical Decision Support Systems, CRISP-DM, Gastric Cancer, WEKA.

1 Introduction

Gastric cancer (GC) is the development of malign cells that can grow in any part of the stomach. It has the potential to invade local and distant organs, the liver, oesophagus, and lungs. There are several histological subtypes of gastric cancer, with adenocarcinoma being the most common one – it can be further divided into intestinal and diffuse type, according to the Lauren Classification. Other possible and less frequent subtypes are squamous, adenosquamous, medullary, and undifferentiated carcinomas. Patients are usually asymptomatic, as symptoms often correspond to an advanced stage disease [1, 2, 3, 4]. Epidemiologically, GC is the second most frequent cause of mortality related to cancer and it is the fourth most common cancer in the world - it has been noticed a decreasing incidence in the past few years [3, 5, 6].The highest incidences of this disease are in Eastern Europe, Central and South America, and in East Asia. Lower rates occur in North America, Northern Europe, Australia, New Zealand, and most parts of Africa. It is more prevalent in males and older people. The survival rate within 5 years is below 30%, except in Japan, where 70% of these can-

cers are diagnosed as stages I and II of the TNM classification. A possible explanation for the higher survival rate in Japan is the existence of screening programs, which leads to an early diagnosis of this cancer [3, 5, 6].Several risk factors have been identified, such as tobacco consumption, poor diet and Helicobacter pylori infection. The decreasing incidence of gastric cancer can be related to a better control of these infections – with the improved hygiene conditions and antimicrobial treatment. Nevertheless, it has been demonstrated that populations with high predominance of these type of infections have low GC rates, which indicates that there are other significant factors for the development of this disease. Therefore, having a family history of GC, low standards of hygiene, being exposed to radiation or even smoking are also factors that may increase the risk [3, 4].

This paper has the main purpose of implementing Data Mining techniques in order to develop predictive models that are capable of predicting the outcome of hospitalization of patients with GC who have undergone surgery and the occurrence of postoperative complications. This article is divided into five sections, which include the introduction, proceeded by the state-of-the-art and the methodologies, materials, methods, and results. Finally, it is presented the discussion of the results and the conclusions and ideas for future work related to this paper.

2 State-of-the-Art

2.1 Data Mining

Nowadays, vast amounts of data are being collected daily in diverse industries and healthcare environments are not the exception. In fact, healthcare data has suffered an exponential growth throughout the years due to the vast quantity of transactions that are performed daily and the digitalization and computerization of healthcare. Therefore, Data Mining (DM) emerged in response to the overwhelming increase of data in medical facilities as a way to transform this data into useful and relevant information for healthcare organizations [7]. From a technical point of view, DM can be defined as a set of techniques and methods used to analyse and explore large sets of data with the intention of discovering previously unknown and meaningful tendencies or patterns. Thus, the goal of DM is to learn from data by extracting new and useful knowledge. Subsequently, this information can be used to build predictive models, which use past information to determinate what might happen in the future, i.e. to give an outcome [8, 9]. Therefore, DM includes descriptive techniques, e.g. clustering techniques that are responsible for discovering information hidden in data, and predictive techniques, e.g. classification and regression techniques, that are used to retrieve new information from existing data [8, 9, 10]. This paper focuses on predictive techniques, more specifically, on classification techniques. Undeniably, DM has become essential in healthcare, namely because of the information acquired by the analysis and exploration of medical data, which can help healthcare organizations and their caregivers to provide more accurate decisions and, therefore, improve the quality of the delivered care [7]. Despite the benefits DM techniques provide to the healthcare industry, they have some limitations. In fact, healthcare data has limited accessibility

due to its dispersion in different systems whereby medical data must be gathered and combined before the DM process. Furthermore, ethical and legal problems may occur if the ownership and privacy of healthcare data is not guaranteed. The lack of quality of the medical data, which includes missing and inconsistent data, is also another limitation since it directly affects DM results [7].

2.2 Decision Support Systems and Clinical Decision Support Systems

Decision Support Systems (DSS) are computer-based systems which are capable of supporting problem solving as well as all stages of the decision-making process allowing the decision maker to control the process. However, in order to help the decision-making process, these systems need knowledge and useful information which can be extracted through DM techniques. Thus, as mentioned before, these techniques are used to analyse and explore data with the aim of discovering patterns that might be helpful for decision-making [11, 12]. In order to health professionals make more accurate decisions, DSS are incorporated into healthcare organizations being known as Clinical Decision Support Systems (CDSS). These systems were specifically designed to aid caregivers in the clinical decision-making process. For this purpose, health providers must enter the data of a specific patient in the system. Once entered, the data must be processed and then linked and compared to knowledge present in the system so that it can give back useful information and suggestions to the caregivers and, therefore, improve the quality of the delivered care [13]. CDSS can perform different actions having, nonetheless, the same goal which is to improve the quality and efficiency of treatments delivered by healthcare organizations and, therefore, lead to higher patient safety and reduce the incidence of adverse events. Firstly, CDSS can be used to alert health providers about problems or irregularities that are occurring as well as remind them about certain actions that must be performed. Additionally, these systems are widely used to give a more accurate diagnosis and to help predicting outcomes based on patient-specific clinical data provided to the system. Many of these systems are also used to assist health professionals, e.g. to calculate the appropriate doses of medications, thus reducing the risk of occurring medical errors. Moreover, CDSS may also offer suggestions to caregivers giving them guidelines or recommendations in order to perform appropriate care and reduce the likelihood of adverse events [13, 14, 15].

2.3 Data Mining Techniques

As mentioned in the previous subsection, the CRISP-DM process was the reference model followed in this paper for the development of the DM project. Moreover, Waikato Environment for Knowledge Analysis (WEKA), which is a ML software, was used in the Modelling phase to analyse and explore the available data and to create the intended models. Thus, a total of five modelling techniques were used with the referred software in order to induce the DM models: Random Forest (RF), Naïve Bayes (NB), Sequential Minimal Optimization (SMO), J48, and JRIP. In a simplified way, RF initially selects a bootstrap sample from the training data, which is a random sam-

ple obtained with replacement, to induce a Decision Tree (DT). Subsequently, this step is repeated until an ensemble of DT is created, having each one of them its own prediction value. Lastly, the final output, i.e. the prediction, is obtained by combining the output from all trees, which corresponds to the most frequent output obtained by the ensemble. RF is known as being very efficient as well as one of the most accurate classifiers. Moreover, the performance attained by RF is usually improved comparatively to single DT [16, 17, 18, 19]. On the other hand, NB is a probabilistic classifier that assumes that all variables are equally independent from the value to be predicted regardless of the correlations that may exist between them. Thus, this is the reason why NB is considered a naïve classifier. This classifier uses the Bayes theorem to predict new instances which, given a set of input values, chooses the most probable output value as the prediction. NB is known as being one of the most effective algorithms for certain domains, namely to classify text documents. Additionally, NB is easily applied and learn fast [19, 20]. SMO corresponds to a new and improved Support Vector Machine (SVM) algorithm, which has the aim of finding the best function that can classify the instances of the training data into the different classes, and emerged as a solution for the quadratic programming problem of SVM. Therefore, SMO, which is used for training SVM, breaks the referred problem into a set of smaller ones that can be resolved analytically, thus being much more simple, faster, and easier to use. SVM is commonly used since it is a highly accurate and reliable method [21, 22]. The J48 classifier implemented by WEKA corresponds to an enhanced implementation of the C4.5 algorithm, which is a DT classifier. In order to create the DT for a certain dataset, the referred algorithm recursively divides the data generating, in each step, several tests. Subsequently, the test that offers the best information gain is then chosen. It must be mentioned that J48 is considered one of the most powerful and commonly used DT classifier [23, 24, 24]. Lastly, JRIP is a rule classifier and corresponds to WEKA's implementation of the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm.

2.4 Related Work

Undeniably, DM techniques are essential tools to aid the decision-making process in medical environments and to improve the quality of healthcare facilities as well as the care delivered by them. Thus, in the interest of having a better understanding of the potential of these tools, some existing works are presented in this subsection. Delen et al. (2005) used three DM techniques along with 10-fold cross-validation to develop predictive models that could predict breast cancer survivability. More specifically, they used two ML algorithms (Artificial Neural Networks (ANN) and DT) as well as one statistical method (Logistic Regression (LR)) on a dataset provided by the Surveillance, Epidemiology, and End Results (SEER) program which contained over 200,000 cases. Additionally, sensitivity analysis was performed on the ANN model with the aim of discovering the variables, which influence the prediction of breast cancer survivability [26]. The results revealed that the DT model was the best predictor (accuracy of 93.6%), followed by the ANN model (accuracy of 91.2%) and, lastly, by the LR model (accuracy of 89.2%), thus proving that DM techniques successfully

create predictive models with a high accuracy. Lastly, the sensitivity results showed that the grade of the cancer was, without a doubt, the most important variable to predict breast cancer survivability [27]. Khalilia et al. (2011) resorted to RF, SVM, boosting, and bagging on a National Inpatient Sample (NIS) dataset with the intent of predicting the risk of not only one but eight chronic diseases. In addition, since the dataset was class-imbalanced, these methods were also performed using random subsampling in order to solve this problem. Overall, the RF model gave the best results having an average accuracy of 88.79%. Moreover, the results obtained with subsampling were better, thus proving that sub-sampling was able to solve the imbalance problem. It must be mentioned that one of the advantages of this study was the fact that they successfully predicted the risk of 8 different diseases and not only one [27].

3 Methodologies, Material, Methods, and Results

The dataset that serves as the basis for the study project was provided by a Portuguese hospital and is constituted of data from GC patients who have undergone surgical procedures. The data contain a set of clinical indicators that are associated to each patient. Given that the purpose of this study is to predict two parameters - the result of hospitalization and whether the patient will or will not have postoperative complications, the dataset was divided in two, considering only the attributes that are associated with each type of goal. The CRISP-DM methodology was implemented in this work in the way described in the following points, where each of them corresponds to a phase.

3.1 Business Understanding

Among the different surgeries related to GC, many of them originate postoperative complications and have very low success rates. Understanding anticipated results of a surgery could assist doctors in the decision of choose the best option for each patient. Thus, this work intends to evaluate different DM techniques in order to determine the ones that provide the best results at the predictive capacity level.

3.2 Data Understanding

At this stage, it was necessary to extract and analyze the dataset provided by the hospital. It were identified 65 attributes and 154 instances. The number of existing attributes is adequate for what is intended, but the number of instances is low, taking into account the characteristics of the DM. The individual meanings of each attribute were verified to understand what kind of influence they might have in the analysis. The existing data are in the nominal type and, for each of them, the frequency of each label was characterized. Some attributes have conditions of compatibility between them such as, for example, the number of invaded glands must be always lower than the number of resected glands, and there is only lymphatic resection if exists resected glands. At first sight, one of the attributes in the dataset that appears to have influence

in the results is the cancer stage, which indicates the severity of the disease. This phase of Data Understanding aims to detect the inconsistencies and verify the duplicated data. Furthermore, it is also important to find the redundant data, observe if the data are in agreement with each other, and check the missing values.

3.3 Data Preparation

In this step, the data are prepared in order to: - Eliminate those that are repeated and those that have relations of dependence; - Eliminate those that have many attributes without filling; - Remove inconsistencies; - Create groups such as age groups; - Add new necessary attributes. It should be noted that the data are treated in two distinct datasets, one for each of the two analyses considered. For the dataset that is related to the complications, a new column had to be created to predict this result. In each dataset, the instances that were removed were those in which there were at least four null values, in which the attribute "Surgery Performed" was null, and still those that presented null values in the columns that were intended to predict the columns of the result of the hospitalization and the occurrence of postoperative complications. At this stage, it was still necessary to correct some errors and make some changes to the dataset. As an example, since the label “99” of the variable “Degree Differentiation” was wrong, it was replaced by a null value, as happened with the label “3” in the variable “Reconstruction”. It was also necessary to create a new label in the column of attribute “Lymphatic Resection” for cases in which no glands have been resected.

3.4 Modeling

In this phase, the Data Mining Model (DMM) is constructed according to a target variable (T), the choices regarding the scenarios considered (S), the DM techniques chosen (DMT), the approaches followed (DA), and the sampling methods used (SM):

- T = {Hospitalisation Result, Surgery Complications}
- S = {S1, S2, S3, S4, S5}
- DMT = {JRIP, J48, RandomForest, SMO, NaiveBayes}
- DA = {With Oversampling, Without Oversampling}
- SM = {Cross-validation 10 Folds, Percentage Split 66%}

Where for DM Hospitalisations Result (DM1):

- S1 = {all attributes}
- S2 = {Resected Glands, Cancer Stage, Surgery Performed, Lymphatic Resection, Performed Surgery Aim, Num Surgery Complications, ASA, Hospitalisation Result}
- S3 = {Sex, Provenance, Motive, Age Group, Hospitalisation Result}
- S4 = {HPreOp, HPosOp, Resected Glands, Degree Differentiation, Invaded Glands, Lymphatic Resection, Hospitalisation Result}
- S5 = {Local P T, Surgery Performed, Reconstruction, Operating Room Discharge, Performed Surgery Aim, Num Surgery Complications, Access Surgery, ASA, Hospitalisation Result}

And for DM Surgery Complications (DM2):

- S1 = {all attributes}
- S2 = {Sex, Surgery Complications}
- S3 = {Sex, Provenance, Motive, Age Group, Surgery Complications}
- S4 = {HPreOp, Resected Glands, Degree Differentiation, Invaded Glands, T, Cancer Stage, Surgery Complications}
- S5 = {Local P T, Surgery Performed, Reconstruction, Performed Surgery Aim, Access Surgery, ASA, Surgery Complications}.

The choice of the various samples was based on sample S1. To obtain the sample S2, it had been applied the filter "AttributeSelection" in WEKA. This filter considers only the most relevant attributes, reducing significantly the number of attributes analysed. Scenarios S3, S4, and S5 contain the attributes that were chosen by group elements, taking into account the patient's own data and the surgery and tumour data. This methodology was used in both cases, for DM1 and DM2. In the case of the DM1, there were only six patients who died after the hospitalization and two who maintained their own state. Thus, as these data were insufficient, it was necessary to do oversampling in these instances, in order to increase the number of data of "died" and "same state". Oversampling is a technique of duplication of data that should be used when there are few instances for analysis. On the other hand, in the case of DM2, there was no oversampling done because the result presents 23 patients who had surgical complications and 63 who did not, which corresponds to a sufficient number of data for analysis. Thus, the DMM that is given by DMM = {T, S, DMT, DA, SM} represents 2 x 5 x 5 x 1 x 2 = 100 simulations (two targets, five scenarios, five DM techniques, one approach by target, and two sampling methods).

3.5 Evaluation

After the construction of the models, they are analyzed and evaluated to see if they adequately fulfil the intended objectives. The verification of the models is carried out with the results of the simulations, considering the accuracy, sensitivity, specificity, and execution time of the technique as important parameters of the analysis. Tables 1 and 2 show the best values obtained for accuracy, sensitivity, and specificity according to each of the techniques used. For each value found, the execution time of the respective DM technique is associated.

Table 1. DM1 – Hospitalization Result.

DM Technique	Scenario	Sampling Method	Accuracy	Time	Sensitivity	Time	Specificity	Time
JRIP	S5	Cross-validation	91.4	0.00	91.1	0.00	78.9	0.00
J48	S2	Cross-validation	89.2	0.06	89.3	0.06		
	S5	Percentage Split					88.6	0.00
RandomForest	S1	Cross-validation	97.8	0.49	98.1	0.49	99.8	0.49
SMO	S1	Cross-validation	91.4	0.2	90.6	0.20		
	S1	Percentage Split					71.3	0.16
NaiveBayes	S1	Cross-validation	95.7	0.02	95.7	0.02	89.3	0.02

Table 2. DM2 – Surgery Complications.

DM Technique	Scenario	Sampling Method	Accuracy	Time	Sensitivity	Time	Specificity	Time
JRIP	S5	Cross-validation	72.6	0.00				
	S2, S5	Percentage Split			83.3	0.00		
	S2, S4	Cross-validation					26.7	0.00
J48	S5	Percentage Split	83.2	0.00			64.0	0.00
	S2, S4	Percentage Split			83.3	0.01		
RandomForest	S5	Percentage Split	82.1	0.01	83.3	0.01	48.7	0.01
SMO	S1	Percentage Split	69.0	0.00				
	S2, S3, S4, S5	Percentage Split			83.3	0.01		
	S1	Cross-validation					35.9	0.00
NaiveBayes	S1	Percentage Split	82.1	0.00				
	S2, S3, S4, S5	Percentage Split			83.3	0.00		
	S1	Percentage Split					48.7	0.00

As seen from the tables above, there are methods that provide high accuracy results that can be used for future work. In general, it is observed that the execution times of the algorithms are low, and the values of accuracy are high.

4 Discussion

From the analysis of tables 4 and 5, it is observed that the results of DM1 present values of accuracy, sensitivity, and specificity higher than those of DM2. This worse result is possible associated with the considered attributes that may not have a great influence on the postoperative complications. Another of the differences between DM1 and DM2 is related to the method that presents better results in the classification. In DM1, cross-validation yields better results whereas in DM2 it is the percentage split that has the best results. Regarding the selected scenarios, S1 and S5 have a high predictive character in DM1 and scenarios S2, S4, and S5 are practically equivalent and yield better results in DM2. It should be noted that in DM2, with scenario S2, that only has the attribute “sex” and the classification attribute, it is possible to obtain good values of predictive capacity. However, using, for example, JRIP, the returned rule is simply "no complications", without being associated with the sex attribute. This result indicates that this scenario is not suitable for the intended objectives. The RF and NB techniques present values above 95% in accuracy and sensitivity when used with sample S1, through the cross-validation method for DM1 analysis. Thus, these techniques are the most suitable to be used for DM because they present the best values. Between RF and NB techniques, there is one fundamental difference: RF is much slower in its execution, which can have considerable effects on processing da-

tasets with thousands of instances. It can still be observed that the values of specificity are low in DM2, which removes credibility to these results. Despite this, DM2 presents good values for accuracy and sensitivity, especially when using the sample S5 and the J48, RF or NB techniques.

5 Conclusions and Future Work

In conclusion, it was verified that the dataset of DM1 produced good results for the construction of predictive models, at the level of gastric cancer surgeries. In this case, the RF technique and the cross-validation method should be used. On the other hand, it was also found that not all data allow to obtain good results and that, therefore, they should be reanalysed, such as the DM2 data. Thus, it can be concluded that the RF and NB algorithms are good options in the detection of the result of hospitalization from the clinical data of the patient and that the algorithms J48, RF, and NB offer good predictions for the existence of postoperative complications. To complement the developed work, a new study could be done, with a larger dataset and a bigger number of instances, in order to determine if the results would be maintained for the RF and NB techniques and if the execution time associated with these algorithms would be acceptable.

Acknowledgments

This work has been supported by Compete: POCI-01-0145-FEDER-007043 and FCT within the Project Scope UID/CEC/00319/2013.

References

1. Biglarian A, Hajizadeh E, Kazemnejad A, Zali MR. Application of Artificial Neural Network in Predicting the Survival Rate of Gastric Cancer Patients. Iranian Journal of Public Health 2011; 40(2):80-86.
2. Rugge M, Fassan M, Graham DY. Epidemiology of Gastric Cancer. In: Strong V, editors. Gastric Cancer. Springer, Cham; 2015. p. 3-34.
3. Brenner H, Rothenbacher D, Arndt V. Epidemiology of Stomach Cancer. In: Verma M, editors. Methods of Molecular Biology. Springer; 2009. p. 467-477.
4. Sitarz R, Skierucha M, Mielko J, Offerhaus GJA, Maciejewski R, Polkowski W. Gastri Cancer: epidemiology, prevention, classification, and treatment. Cancer Management and Research 2018; 10:239-248.
5. Roder DM. The epidemiology of gastric cancer. Gastric Cancer 2002; 5(Suppl 1):5-11.
6. Karimi P, Islami F, Anandasabapathy S, Freedman ND, Kamangar F. Gastric Cancer: Descriptive Epidemiology, Risk Factors, Screening, and Prevention. Cancer Epidemiology, Biomarkers & Prevention 2014; 23(5):700-713.
7. Koh HC, Tan G. Data mining applications in healthcare. Journal of healthcare information management 2011; 19(2):64-72.

8. Witten I, Frank E. Data Mining: Practical machine learning tools and techniques. 2nd ed. San Francisco, California: Morgan Kaufmann; 2005.
9. Tuffery S. Data mining and statistics for decision-making. 1st ed. Oxford: Wiley-Blackwell; 2011.
10. Fonseca F, Peixoto H, Miranda F, Machado J, Abelha A. Step Towards Prediction of Perineal Tear. Procedia Computer Science 2017;113:565-570.
11. Bâra A, Lungu I. Improving decision support systems with data mining techniques. Advances in data mining knowledge discovery and applications. INTECH Open Access Publisher; 2012. p. 397-418.10
12. Shim J, Warkentin M, Courtney J, Power D, Sharda R, Carlsson C. Past, present, and future of decision support technology. Decision Support Systems. 2002; 33(2):111-126.
13. Beeler P, Bates D, Hug B. Clinical decision support systems. Swiss Med Wkly 2014;144:w14073.
14. Trowbridge R, Weingarten S. Chapter 53. Clinical Decision Support Systems [Internet]. United States Department of Health & Human Services Agency for Healthcare Research and Quality website. 2001 [cited 6 May 2018]. Available from: https://archive.ahrq.gov/clinic/ptsafety/chap53.htm
15. Morais A, Peixoto H, Coimbra C, Abelha A, Machado J. Predicting the need of Neonatal Resuscitation using Data Mining. Procedia Computer Science. 2017; 113:571-576.
16. Svetnik V, Liaw A, Tong C, Culberson J, Sheridan R, Feuston B. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. Journal of Chemical Information and Computer Sciences. 2003; 43(6):1947-1958.
17. Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. 2004.
18. Zhang C, Liu C, Zhang X, Almpanidis G. An up-to-date comparison of state-of-the-art classification algorithms. Expert Systems with Applications. 2017;82: 128-150.
19. Khoshgoftaar T, Golawala M, Hulse J. An Empirical Study of Learning from Imbalanced Data Using Random Forest. 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007). 2007.
20. Mitchell T. Machine Learning. New York: McGraw-Hill; 1997.
21. Platt J. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.
22. Wu X, Kumar V, Quinlan J, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowledge and information systems. 2009; 14(1):1-37.
23. Zhao Y, Zhang Y. Comparison of decision tree methods for finding active objects. Advances in Space Research. 2008; 41(12):1955-1959.
24. Rajput A, Aharwal R, Dubey M, Saxena S, Raghuvanshi M. J48 and JRIP rules for e - governance data. International Journal of Computer Science and Security (IJCSS). 2011;5(2):201-207.
25. Mohamed W, Salleh M, Omar A. A comparative study of reduced error pruning method in decision tree algorithms. In: 2012 IEEE International Conference on Control System, Computing and Engineering. 2012. p. 392-397.
26. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine. 2005; 34(2):113-127.
27. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision-making. 2011; 11(1):51.