CLASSIFICATION OF THE SURVIVAL MONTHS AS A FUNCTION OF THE AGE GROUP IN NON-HODGKIN LYMPHOMA:

Esti Vaknin, Sapir Machluf,

Department of Software Engineering,
Intelligent Systems
Afeka Tel Aviv Academic College of Engineering
Mivtza Kadesh 38, Tel Aviv
{ester.vaknin & sapir.machluf }@ s.afeka.ac.il

Keywords: Cancer, patient, multilabel, classification

ABSTRACT - This article addresses the problem of classifying lymphoma patients into groups that represent the length of time they will survive from the moment the disease is discovered, for different age groups. In the article we examine whether a model for all age groups will give at least as good results as dedicated models for each age group. In addition, we also examine the stability of the models.

Because these are groups with a hierarchy between them, we used multi-label classification models to examine the problem.

I. INTRODUCTION

In medicine, identifying relationships between groups of patients, can influence the nature of the treatment of certain patients, and help physicians make different decisions for patients with different characteristics.

For cancer, which affects each person differently, and conventional medicine has difficulty finding a cure, it is of paramount importance to tailor the best treatment to the type of patient. Various studies in the field of data science (Sunil Gupta, 2014) have shown that predictions and classification can be made and thus improve the predictability of physicians.

This paper describes a machine learning (ML) study based on data from cancer patients collected in the United States between 1973 and 2018 (SEER database). Different for different age groups. Also, examine whether the stability of the models improves in age-specific models.

This paper makes use of models of the classification of the months of survival of patients of different age groups.

The lymphatic system is one of the body's natural defence systems against infections and diseases and it is spread all over the human body. Lymphoma is a cancer of the lymphatic system and is divided into two main types: The first, Hodgkin's lymphoma, is characterized by the presence of abnormal cells called the reed-Stenberg. The second is a non-Hodgkin's lymphoma characterized by damaged lymphocytes that multiply and form a lump called a tumour.

The treatment the patient will receive depends on the type of lymphoma he is suffering from and its severity. Lymphoma treatment may include chemotherapy, immunotherapy drugs, radiation therapy, bone marrow transplant, surgery, or a combination of all of these.

According to the UK Cancer Research Centre (Hodgkin lymphoma survival statistics, 2021) 75% of Hodgkin's cancer patients survive at least 10 years from the time the disease is detected, compared to non-Hodgkin's patients where only 55% of patients survive over 10 years from the time the disease is detected.

In addition, patients aged 15-39 have the highest chances of survival relative to other age groups in both types of lymphoma.

Also, according to studies, in the last 40 years there has been a significant increase in survival with the disease from 47% to 80% for Hodgkin's lymphoma and from 22% to 63% for non-Hodgkin's lymphoma.

According to clinical studies, there is a difference in the effect of age groups on survival for different types of cancer. For example, in a study by (Maryska L.G. Janssen-Heijnen, 2010), a comparison was made for 13 different types of cancer, including non-Hodgkin's lymphoma, on the patient's survival for a period of 5 to 10 years from the time the disease was diagnosed and for 4 different age groups. According to this study, there are types of cancer (e.g., testicular cancer) in which the conditional survival was similar for the different age groups, and over the years became identical to that of the general population.

In contrast, there are types of cancer (e.g., ovarian cancer), in which the conditional survival was different between the different age groups, and over the years the differences between the age groups narrowed.

And finally, there were cancers (e.g., non-Hodgkin's lymphoma), which are the patients' conditional survival that differed between different age groups, and the difference was maintained over the years (Figure 1).

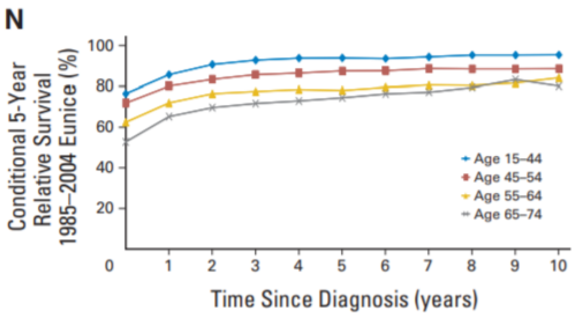

Figure 1
Non-Hodgkin's lymphoma

Studies in machine learning have shown that learning systems can in most cases predict cancer survival, at least as well as experts in the field. Also, when using social, economic data in addition to medical information about the patient, these predictions improve significantly (Sunil Gupta, 2014).

In this article we will present a comparison between machine learning models on a database containing all age groups (unified) and dedicated models for different age groups for lymphoma patients.
The models we will represent are models of classification, for each patient we will examine whether he survived up to a year, between one year and 5 years, between 5 and 10 years or over 10 years. Since there is a relationship between the groups, we used multi label classification models, which we will explain in more detail below.

II. The Research Goal

The aim of this study is to predict monthly survival for lymph patients in different age groups. We will make a separation between the main types of the disease: Hodgkin's lymphoma and non-Hodgkin’s lymphoma.

In our study we will examine the prediction of monthly survival by two approaches:

1. Building a unified prediction model for all age groups.
2. Building different separated models for different age groups.

We will compare the two different approaches and examine the prediction results to decide whether a unified model provides good enough prediction or whether different models divided into different age groups improves the stability and performance of the prediction model. We will examine which approach provides the best results for this data.

At this point, we will explain and focus our research question. In our research question, we answered a task of classification. We would like to predict for each patient the survival period according to the features we have chosen whether the patient will survive a period of one year, five years or ten years. A classification for these specific groups was chosen following studies published in the field which show this accepted division. In addition, the division of age groups starts from the age of 15 and is detailed in the chapter of the experiment and results. This division was also made based on the literature review and examination of the data. In our study we will focus on non-Hodgkin’s lymphoma mainly because most of the data are of this type.

III. Methodology

As we have seen in other studies, our work methodology has been applied as depicted Figure 2.

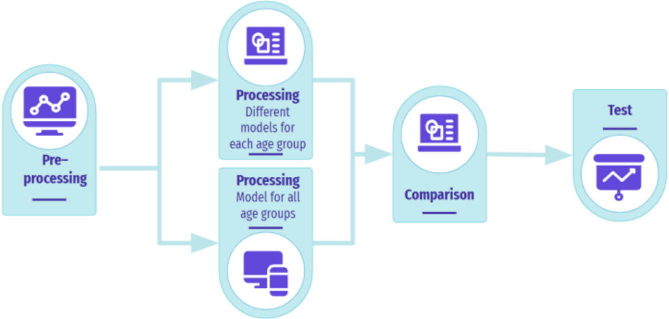

Figure 2

In the first stage, we pre-processed the data, examined which columns are relevant to our research question, what missing values we have and how they can be supplemented. This step is detailed in Chapter IV.
We then divided our data set into training data and validation data. We repeated this step for both a unified model and a model that is broken down into different age groups. We performed training for each model separately and in addition we performed cross-validation on the training set to test the stability of the model. The models we used for training are:
Gaussian Naive Base, Logistic Regression, SVM, XGboost.
All models were applied according to a multi-label method.
This step is detailed in chapter V.
We then examined each model on the exam set to examine our model results and understand which model gives the best results. This step is detailed in Chapter VI.

IV. Preprocessing

The pre-processing included a set of actions that allowed us to arrive at data that could be used for the model we built and tailored to the research we were conducting to arrive at good predictive results.
First, we reviewed all the data columns and understood their relationships to our research. The purpose of this action was to understand the meanings of the data and which columns are critical to performing the specific prediction we are focusing on.
After this necessary step we started to handle each column individually and performed a sequence of actions to get orderly and appropriate information for performing our prediction.

In the 'Year of diagnosis' column, we defined data from 1981 to 2008. A review of the literature we did in the previous step made us understand that the data must be treated within a period of up to 40 years back.

In the 'Patient Age' column we decided to removal all patients under the age of 15. This insight is a formula from the literature review we performed.

In addition, we removed all patients diagnosed with Hodgkin's lymphoma. Based on the focus of our research question, we arranged the 'Type of lymphoma' column so that the remaining data are only non-Hodgkin's lymphoma.

also, we removed all patients who died not because of non-Hodgkin's lymphoma in 'Cause of death' column.

Another step in this process focused on finding non-informative columns, we removed duplicate columns, columns with missing values or single value.

Moreover, the process included a step where we noticed columns with missing data and performed completions. This is after examining your necessity and significance to our research topic. The completion of the data is also done out of great thought and a desire to complete the missing data most correctly in order to maintain the reliability of the data.

In the 'Race recode' column we completed 17% missing values by the most common value.

The 'Median household income' column we supplemented according to OECD data to the average wage in the United States in those years.

After completing the step of handling the missing data, we performed actions that helped us arrange the data so that we could use it when running the model. The 'record numbers' column we changed to a binary value, whether the patient had one or more.

V. The process

After preparing our data for processing, we received 7 sets of training data and 7 sets of test data: a data set that includes all age groups, and data sets for each age group see Table 1

dataset	Train size	Test size
Unified model	83,833	35,929
Age 15-39	7543	3234
Age 40-49	9179	3935
Age 50-59	13512	5792
Age 60-69	18098	7757
Age 70-79	21149	9065
Age 80+	14348	6150

Table 1

The purpose of our model was to perform a classification for each patient, and to predict how long the patient would survive from the moment they discovered the disease. The groups into which we divided the patients are up to one year, between one and 5 years, between 5 and 10 years and over 10 years.

Because of the existing hierarchy between survival times, we used Multi label classification models. To test the accuracy of our models we used binary accuracy. For example: if a patient survived a period of between 5 and 10 years, and the model gave a classification of up to one year, and between 5 and 10 years, he is accurate in 2 out of 3 wards, which is the accuracy determined for the model.

We compared 4 multi label classification models: Gaussian Naive Base, Logistic Regression, SVM, XGboost with different adjustments.

The first approach to testing the model is OneVsRest:

Traditional two-department and multi-department problems can be moulded into multiple problems by limiting each instance to just one label. On the other hand, the generality of multi-label problems necessarily makes learning difficult. An intuitive approach to solving a multi-label problem is to break it down into several independent binary classification problems (one for each category).

In the "one-to-rest" strategy, it is possible to construct several independent classifiers, and in the invisible case, to choose the department for which confidence is increased (Nooney, 2018). In this approach we used the Gaussian Naive Base model.

The second approach by which we examined the model is Binary Relevance:

In this case an ensemble of single-label binary classifiers is trained, one for each class. Each classifier predicts the membership or non-membership of one class. The union of all the predicted classes was taken as a multi-label output. This approach is popular because it is easy to implement, however it also ignores the possible correlations between class labels. In this approach we used the Logistic Regression and XGboost models.

The third approach by which we examined the model is Classifier Chains:

A chain of binary classifiers C0, C1, . . . , Cn is constructed, where a classifier Ci uses the predictions of all the classifier Cj, where j <i. This way the method, also called classifier chains (CC), can consider label correlations.

The total number of classifiers needed for this approach is equal to the number of classes, but the training of the classifiers is more involved.

In this approach we used the SVM model.

The fourth and final approach we have examined is Label Powerset:

This approach does consider possible correlations between class labels. It considers each of the members of the power set of the labels in the training set as one label. It requires a lot of computational power.

In this approach we used the Logistic Regression model.

classifier	Accuracy (training)	CV- Score 1	CV- Score 2	CV- Score 3	CV- Score 4	CV- Score 5	Accuracy (test)
GaussianNB	0.408	0.404	0.406	0.406	0.406	0.407	0.408
LogisticRegression-Classifier Chains	0.358	0.362	0.360	0.363	0.370	0.368	0.356
SVC	0.355	0.350	0.353	0.354	0.354	0.355	0.354
LogisticRegression- Label Powerset	0.232	0.232	0.234	0.232	0.232	0.232	0.226
XGBClassifier- binary	0.576	0.523	0.523	0.524	0.523	0.523	0.526
XGBClassifier- multi	0.572	0.523	0.522	0.524	0.524	0.523	0.527

Table 2

classifier	Accuracy (training)	CV- Score 1	CV- Score 2	CV- Score 3	CV- Score 4	CV- Score 5	Accuracy (test)
Age 15-39	0.796	0.641	0.635	0.631	0.634	0.636	0.656
Age 40-49	0.773	0.626	0.628	0.621	0.622	0.621	0.631
Age 50-59	0.704	0.554	0.560	0.561	0.559	0.561	0.567
Age 60-69	0.624	0.459	0.459	0.459	0.464	0.466	0.468
Age 70-79	0.589	0.409	0.406	0.405	0.405	0.405	0.413
Age 80+	0.678	0.497	0.502	0.502	0.500	0.500	0.504

Table 3

VI. RESULTS

As can be seen in Table 2, for the consolidated data set, the model that yielded the best results is the XGB Classifier-multi which resulted in a 52.7% accuracy on the exam data for 4 classes.

However, all the models showed stability throughout the iterations of Cross-Validation.

For the data sets for each age group different results were obtained, the model that resulted in the best results is XGB Classifier- binary, the model results for each age group are shown in Table 3.

However, for some age groups, the accuracy on the exam set was very low relative to the training set, this model may have caused an over-fitting of the data.

In addition, the cross-validation results showed lower stability for this model

VII. DISCUSSION

Following the study, we concluded that for a unified model for age groups, performance is impaired, but the stability of the model is maintained.

For a model designed for any age group, performance increases but the stability of the model decreases, and, in some cases, there may be over-fitting to the data, these problems may be solved with a larger data set.

In addition, this article did not touch on data analysis for patients with lymphoma Hodgkin, for this disease further research is required, which may be similar in its characteristics to this study.

REFERENCES:

Ahmad, I' K 8) .'April 2020 .(Age-specific survival in prostate cancer using machine learning .*Emerald Insight ,* ,(2)*54* page 234-215

Hodgkin lymphoma survival statistics(2021) . CANCER RESERCH UK: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/hodgkin-lymphoma/survival

Maryska L.G. Janssen-Heijnen, A' G'-W 20) .'may 2010 .(Clinical Relevance of Conditional Survival of Cancer Patients in Europe: Age-Specific Analyses of 13 Cancers .*JOURNAL OF CLINICAL ONCOLOGY ,* *28*, page .2528-2520

Nooney, K 8) .'JUN 2018 .(*Deep dive into multi-label classification* .towardsdatascience: https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff

Sunil Gupta, T' T .(2014) .'Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry *.BMJ Open*.

U.S. Department of Health and Human Services *National Cancer Institute* .Surveillance, Epidemiology, and End Results: https://seer.cancer.gov/