Diffuse Large B-cell Lymphoma Survival Prediction using Encoding Clinical Features

Sy-Phuc Pham*,[1], Sae-Ryung Kang*,[2], Hyung-Jeong Yang**,[1], Deok-Hwan Yang**,[2], Sudarshan Pant[1], Soo-Hyung Kim[1], and Guee-Sang Lee[1]

[1]Chonnam National University, Gwangju, South Korea
[2]Chonnam National University Hwasun Hospital, Gwangju, South Korea
{phamsyphuc123,sudarshan.pant}@gmail.com,
{srkang,hjyang,drydh,shkim,gslee}@jnu.ac.kr

Abstract. Diffuse Large B-cell Lymphoma (DLBCL) is a type of blood cancer that has a high mortality rate. Accurately predicting the survival time of DLBCL patients is crucial for guiding treatment decisions and developing new therapies. In this study, we proposed Encoding Clinical Features (ECF) and used CoxCC and DeepSurv to predict the survival time of DLBCL patients. We experimented with our dataset, which was provided by Chonnam National Hwasun Hospital. We applied ECF to the dataset using a set of dimensions for categorical variables and evaluated the performance of the model using the C-index and the Integrated Brier Score (IBS). Our results showed that the ECF technique had a high C-index and a low IBS, indicating good performance in predicting the survival time of DLBCL patients. We also found that the selected dimensions for embedding categorical variables were suitable for our dataset. Our study demonstrates the potential of ECF with CoxCC and DeepSurv for improving the prediction of DLBCL patient outcomes and for identifying important prognostic factors that can be used to guide treatment decisions.

Keywords: Diffuse Large B-cell Lymphoma · Survival analysis · Clinical information.

1 Introduction

Diffuse Large B-Cell Lymphoma (DLBCL) is a type of non-Hodgkin lymphoma, which is a cancer that affects the lymphatic system [1]. The lymphatic system is a network of vessels and organs that help to fight infection and disease in the body. DLBCL is the most common type of non-Hodgkin lymphoma and it is characterized by the rapid growth of abnormal B-cells, a type of white blood cell that is involved in the immune response [2]. The cancer cells can spread to various parts of the body, including the lymph nodes, bone marrow, liver, spleen and other organs. Symptoms of DLBCL can include swollen lymph

* These authors contributed equally to this work.
** Corresponding author.

nodes, fever, weight loss, night sweats, and fatigue. The exact cause of DLBCL is unknown, but certain risk factors have been identified such as age, certain infections, and a weakened immune system [3]. DLBCL can be treated with a combination of chemotherapy, radiation therapy, and immunotherapy. The prognosis varies depending on the stage of the cancer, the patient's overall health, and the effectiveness of the treatment. With early diagnosis and appropriate treatment, the survival rate of DLBCL can be quite high.

Survival analysis is a statistical method used to study the time to an event of interest, such as death or failure. In the context of DLBCL, survival analysis is used to estimate the probability of survival in patients with this type of lymphoma. A common method used in survival analysis is the Kaplan-Meier estimator, which is used to estimate the survival probability over time. The Kaplan-Meier [4] estimator is based on the idea of censoring, which means that some patients may not have experienced the event of interest (death) at the time of the analysis. These patients are considered "censored" and their survival time is not included in the estimate. Another method used in survival analysis is the Cox proportional hazards model [5], which is used to estimate the effect of various factors (such as age, stage of the cancer, and treatment) on the risk of death. The Cox model allows for the estimation of hazard ratios, which indicate the relative risk of death for a particular group of patients compared to another group. Overall, Survival analysis is a useful tool for assessing the prognosis of DLBCL patients and for identifying factors that may influence the risk of death. These techniques can help physicians and researchers to identify patient subgroups that have a higher or lower risk of death, and to develop more effective treatment strategies for DLBCL.

DLBCL is a type of non-Hodgkin lymphoma (NHL), which is a cancer of the lymphatic system. DLBCL is the most common type of NHL and represents approximately 30% of all NHL cases [6]. DLBCL is typically diagnosed through a combination of clinical examination, laboratory tests, and imaging studies. A biopsy of the affected tissue is usually performed to confirm the diagnosis and to determine the stage of the disease. The treatment of DLBCL depends on the stage of the disease and the patient's overall health. The standard treatment for DLBCL is a combination of chemotherapy and radiation therapy. In some cases, immunotherapy or targeted therapy may also be used. Prognosis of DLBCL varies depending on the stage of the disease and the patient's overall health. With early detection and appropriate treatment, the overall survival rate for DLBCL is around 70%. However, the prognosis is poorer for patients with advanced-stage disease and those who do not respond well to treatment. The treatment of DLBCL typically involves a combination of chemotherapy and radiation therapy, and the prognosis varies depending on the stage of the disease and the patient's overall health. Clinical information is important to every DLBCL patient and this information is recorded by the physician throughout the course of treatment.

Artificial Intelligence (AI) has been increasingly used in survival analysis, which is a statistical method used to study the time to an event of interest, such as death. AI-based methods have been developed to improve the prediction of patient outcomes and to identify important prognostic factors that can be used to guide treatment decisions. AI-based methods have also been used to analyze large datasets and to identify patterns and trends that can help to advance biomedical research and to develop new drugs and therapies. For example, AI-based methods have been used to analyze clinical trial data to identify important predictors of patient outcomes and to identify potential new treatments for diseases such as cancer. Despite the advancements, AI in survival analysis is still in its early stages and there are many challenges that need to be addressed. One of the main challenges is the lack of standardization and regulation in the development and deployment of AI-based survival analysis tools, which can lead to inconsistency in the quality of the tools and in the results obtained. Additionally, there is a need for more clinical validation of the AI-based tools to ensure their safety and effectiveness before they are widely adopted in the healthcare system. Overall, AI has the potential to revolutionize survival analysis and to improve patient outcomes, but it is important to continue to invest in research and development to address the challenges and to ensure that the benefits of AI are realized in the healthcare system.

In this paper, we perform survival analysis tasks based on recent methods. In the first section, we introduce overall the DLBCL, the survival task, the clinical information in DLBCL, and the AI in the survival task. We present the recent method for survival tasks in the second section. In the third section, we introduce our approach for survival task. In the next section, we summarize the dataset and experiment detail for this work. The results of our experiment are presented in the next section. We include the conclusion in the last section.

2 Related work

2.1 Cox Proportional Hazards Model

The Cox Proportional Hazards Model (CoxPH) [7] is a statistical model used to estimate the effect of various factors on the risk of an event, such as death. The model is named after its creator, Sir David Cox. The model is widely used in survival analysis, which is a statistical method used to study the time to an event of interest, such as death. The CoxPH model is a semi-parametric model, which means that it makes some assumptions about the shape of the hazard function, but it does not specify the exact form of the function. The model assumes that the hazard ratio is constant over time. This assumption is known as the proportional hazards assumption. The CoxPH model estimates the effect of various factors on the hazard ratio. The model estimates the hazard ratio as a function of the values of the predictor variables. The hazard ratio is a measure of the relative risk of an event for a particular group of patients compared to another group. The CoxPH model can be used to estimate the effect of various factors on the risk of death in DLBCL patients, allowing to identify patient subgroups that have a higher or lower risk of death and to develop more effective treatment strategies for DLBCL.

2.2 DeepSurv

DeepSurv [8] is a deep learning-based survival analysis method that uses artificial neural networks to predict the risk of an event, such as death. It is an adaptation of a traditional Cox proportional hazards model, which is a widely used statistical model in survival analysis. DeepSurv uses a neural network to model the hazard function and to estimate the effect of various factors on the hazard ratio. The neural network is trained on a dataset of patients with the event of interest, and the model learns the complex non-linear relationships between the predictor variables and the hazard ratio. One of the main advantages of DeepSurv is that it is able to handle high-dimensional and non-linear data, which can be difficult to analyze using traditional survival analysis methods. Additionally, DeepSurv can handle missing data, which is a common problem in survival analysis. DeepSurv has been applied to various medical fields, such as oncology, and has shown to improve the prediction of patient outcomes and to identify important prognostic factors that can be used to guide treatment decisions.

3 Proposed method

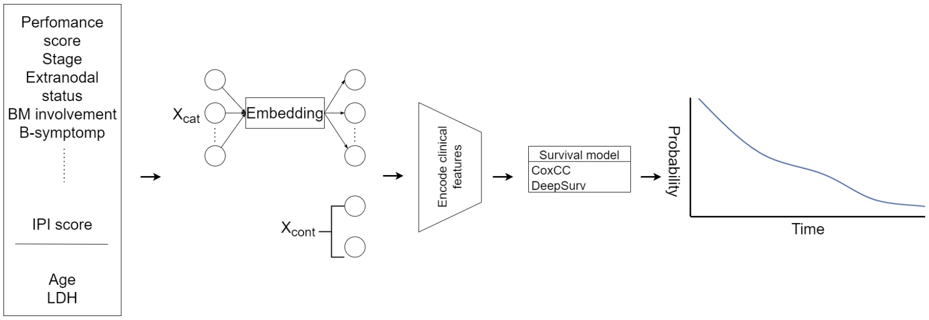

Fig. 1. Perspective of the proposed survival based encoding clinical feature for DLBCL survival prediction.

In this section, we present our method to deal with our dataset for use in the survival task. Our method has two main parts: encoding clinical features and survival prediction. The overall architecture has been shown in Figure 1. First, we introduce the encoding clinical features part. Linear analysis, machine learning, and deep learning are the core components of most approaches, however these can only be used with numerical data. Since that the data may be in categorical form, it is important to process the input data to transform the non-numerical fields into numerical fields before feeding the data into the deep network. It has been demonstrated in some research that the embedding method is superior to the one-hot vector approach, and it has also been utilized in some studies to accomplish the encoding [9,10]. For categorized information, Mikolov et al. [11] presented an embedding method. Equation 1 shows our definition of clinical input, which incorporates both continuous and categorical data.

$$\chi_{clinical} = \chi_{continuous} \oplus \chi_{category} \tag{1}$$

We set the dimensions to represent the embedding in the continuous fields using a variable from the categorical fields $\chi_i \in \chi_{category} \rightarrow \nu_i \in \mathbb{R}^{k_i}$. The k_i dimension of space is a function of the characteristics of the variables in the category. The clinical features were obtained by appending the vectors representing the continuous variables and the embeddings together.

$$Z_{clinical} = f(\chi_{continuous}, f(g(x_1, ..., x_n))) \tag{2}$$

where

- $Z_{clinical}$ is the clinical feature after after feature extraction part.
- $\chi_{continuous}$ is the numeric fields.
- $x_1, ..., x_n \in \chi_{category}$ with n is the number of categories in the clinical data.
- f is the concatenate operator.
- g is the embedding operator.

Second, we apply methods based on CoxCC [12] and DeepSurv to predict the survival hazard rate. For the purpose of determining whether or not encoding clinical information is effective, we used two standard models in the survival task.

4 Experimentals

4.1 Dataset

Clinical information is essential for the management of DLBCL patients as it helps to determine the patient's diagnosis, stage of the disease, treatment options and prognosis. The diagnosis of DLBCL is made on the basis of clinical data. This includes symptoms, physical examination, laboratory test results, and imaging studies. The extent to which cancer has spread throughout the body is quantified by the disease's stage, which is in turn determined by the available clinical data. This is important in determining the appropriate treatment plan and in estimating the patient's prognosis. Clinical data are analyzed to establish the best course of treatment for the patient. This includes the type of chemotherapy, radiation therapy, and immunotherapy that may be appropriate for the patient, as well as the potential side effects of these treatments. Clinical data is analyzed to determine the patient's prognosis, which is a prediction of the patient's chance of remission. This includes factors such as age, overall health, and the stage of the disease, as well as the patient's response to treatment. Clinical data is tracked over time to assess a patient's response to treatment and make any necessary adjustments. This can include regular blood tests, imaging

Table 1. Statistics of each clinical feature of 602 DLBCL patients in the data set of CNUHH.

Data field	Characteristics	Value	CNUHH (n=602)
numeric	Age	min-max	17-92
numeric	LDH	min-max	144-8402
categories	Performance	1	201 (33.39%)
		2	322(53.49%)
		3	65 (10.8%)
		4	14 (2.33%)
categories	B symptom	0	504 (83.72%)
		1	98 (16.28%)
categories	Extranodal status	0	456 (75.75%)
		1	146 (24.25%)
categories	Stage	1	118 (19.6%)
		2	195 (32.39%)
		3	137 (22.76%)
		4	152 (25.25%)
categories	Spleen involvement	0	579 (96.18%)
		1	23 (3.82%)
categories	Bone marrow involvement	0	554 (92.03%)
		1	48 (7.975%)
categories	IPI score	0	81 (13.46%)
		1	152 (25.25%)
		2	134 (22.26%)
		3	131 (21.76%)
		4	76 (12.62%)
		5	28 (4.65%)
categories	IPI risk	1	233 (38.7%)
		2	134 (22.26%)
		3	131 (21.76%)
		4	104 (17.28%)
categories	R-IPI	1	81 (13.46%)
		2	286 (47.5%)
		3	235 (39.04%)

studies, and physical examinations. The medical expert at Chonnam National University Hwasun Hospital kindly provided the dataset that was used for the experiments that were conducted for this work. In this dataset, we divided the features into two types: numeric and categorical. We detailed the clinical feature in the Table 1.

4.2 Experiment set up

We carried out the experiment using the CoxCC, DeepSurv models as primary methods. The dataset includes 602 patients. We separated the dataset into 5-fold for the training process. To evaluate the performance of the model, we performed on independent test data sets. In total, we used 481 patients for the training process, and we used 121 patients for testing and evaluating the performance of the model. In each of the methods, the model parameter was configured in a similar way.

5 Experimental results

Our work was implemented with the Pytorch 1.11 library. We utilized two distinct metrics in order to evaluate the performance of the models. The first was the concordance index [13], or C-index for short. This evaluation approach is the one that is employed the most commonly. It measures the ability of a model to correctly rank the event times of a population, with higher values indicating better performance. We also utilized the Integrated Brier Score (IBS) [14] as a secondary metric. IBS is a measure of the accuracy of a model's predictions over time. Encoding clinical features was performed to improve CoxCC and DeepSurv. Experimental results in Table 2 show that our proposed model suitable for clinical tabular. Although our proposed method gives better results than only using CoxCC and DeepSurv, it lacks outstanding at C-index. That is something we need to improve in the future.

Table 2. Comparison results of the encoding clinical features and without encoding clinical features

Method	Survival model	IBS	C-index
W/o encoding clinical features	CoxCC	0.164	0.545
	DeepSurv	0.146	0.700
Encoding clinical features	CoxCC	**0.140**	**0.654**
	DeepSurv	**0.143**	**0.728**

6 Conclusions

This paper presented Encoding clinical features for using in traditional architecture in survival task such as CoxCC and DeepSurv. We defined the dimension for embedding categorical features by experiment and selected the best dimension for the CNUHH dataset. DeepSurv deep learning network architecture has somewhat better results than CoxCC statistical model. Clinical data with categorical data types are efficiently fed into deep learning network architectures and statistical models. We all know that patient data includes not just clinical information but also imaging data like PET scans and CT scans. In future work, we will integrate medical imaging into the model to conduct a survival analysis experiment of DLBCL patients.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT). (NRF-2020R1A2B5B01002085)

References

1. Sehn, Laurie H., and Gilles Salles. "Diffuse large B-cell lymphoma." New England Journal of Medicine 384.9 (2021): 842-858.
2. Pileri, Stefano A., et al. "Predictive and prognostic molecular factors in diffuse large B-cell lymphomas." Cells 10.3 (2021): 675.
3. Li, Shaoying, Ken H. Young, and L. Jeffrey Medeiros. "Diffuse large B-cell lymphoma." Pathology 50.1 (2018): 74-87.
4. Goel, Manish Kumar, Pardeep Khanna, and Jugal Kishore. "Understanding survival analysis: Kaplan-Meier estimate." International journal of Ayurveda research 1.4 (2010): 274.
5. Cox, David R. "Regression models and life-tables." Journal of the Royal Statistical Society: Series B (Methodological) 34.2 (1972): 187-202.
6. Armitage, James O., et al. "Non-hodgkin lymphoma." The lancet 390.10091 (2017): 298-310.
7. Fisher, Lloyd D., and Danyu Y. Lin. "Time-dependent covariates in the Cox proportional-hazards regression model." Annual review of public health 20.1 (1999): 145-157.
8. Katzman, Jared L., et al. "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network." BMC medical research methodology 18.1 (2018): 1-12.
9. Guo, Cheng, and Felix Berkhahn. "Entity embeddings of categorical variables." arXiv preprint arXiv:1604.06737 (2016).
10. Wang, Peng, et al. "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification." Neurocomputing 174 (2016): 806-814.
11. Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
12. Kvamme, Håvard, Ørnulf Borgan, and Ida Scheel. "Time-to-event prediction with neural networks and Cox regression." arXiv preprint arXiv:1907.00825 (2019).
13. Harrell, Frank E., et al. "Evaluating the yield of medical tests." Jama 247.18 (1982): 2543-2546.
14. Gerds, Thomas A., and Martin Schumacher. "Consistent estimation of the expected Brier score in general survival models with right-censored event times." Biometrical Journal 48.6 (2006): 1029-1040.