Intelligent Health Systems – From Technology to Data and Knowledge
E. Andrikopoulou et al. (Eds.)

doi:10.3233/SHTI250379

Lung Cancer Survival Estimation Using Data from Seven German Cancer Registries

Sebastian GERMER[a,1], Christiane RUDOLPH[b], Alexander KATALINIC[b,c], Natalie RATH[d], Katharina RAUSCH[d], and Heinz HANDELS[a,e]
[a] *German Research Center for Artificial Intelligence (DFKI), Lübeck*
[b] *Institute for Cancer Epidemiology, University of Luebeck*
[c] *Institute for Social Medicine and Epidemiology, University of Luebeck*
[d] *Cancer Registry Saarland, Saarbrücken*
[e] *Institute of Medical Informatics, University of Luebeck*
ORCiD ID: Sebastian Germer https://orcid.org/0000-0001-8439-872X

Abstract. Predicting the survival of cancer patients is of high importance for the medical community, e.g. for evaluating therapy strategies. This study is based on lung cancer data retrieved from seven German cancer registries according to the German basic oncology dataset. After data integration and preprocessing, we predicted the survival for 6, 12, 18 and 24 months respectively using a gradient boosting algorithm. To gain insight into the decision process of the models, we identified the features that have a high impact on patient survival using permutation feature importance scores as explainability metric. They show that age at diagnosis as well as the presence of distant metastases are key factors for long-term survival. The found factors can be used in a next step for multi-variate survival analysis.

Keywords. Lung Cancer, Explainable AI, Survival Analysis

1. Introduction

Lung cancer is the most common malignant neoplasm worldwide in 2022, and is associated with the highest mortality rate [1]. In Germany, 22,892 women and 34,572 men were newly diagnosed with lung cancer in 2019, and the absolute five-year survival rate is 25.0% for women and 20.6% for men [2].

Cancer cases are reported to the cancer registries of the federal states. The registries are committed to a mandatory German basic oncology dataset, which contains around 130 variables [2]. Since 2023, a majority of this dataset is submitted to the German Center for Cancer Registry Data in a modified format (oBDS-RKI, [3]). For this study, we requested lung cancer data from all 15 federal state cancer registries with dates of diagnosis between 2014 and 2022.

The key factors for long-term survival can vary depending on the type of cancer. Common factors affecting survival include the stage of cancer at diagnosis, age, sex, general health and treatment [4]. Identifying factors that are critical for short- and long-term survival could improve patient outcomes and treatment strategies. Using survival

[1] Corresponding Author: Sebastian Germer. Ratzeburger Allee 160, 23562 Lübeck, Germany; E-mail: sebastian.germer@dfki.de.

analysis methods combined with explanatory algorithms such as permutation feature importance, it is possible to obtain clues about these important variables.

In this study, we retrieve and integrate lung cancer data from German registries to predict survival outcome. Furthermore, we plan to assess the generalization ability of the models, using a "leave-one-registry-out" approach: As we have received data from seven registries, we can train models on six registries and externally validate them on the seventh registry.

2. Methods

2.1. Survival Analysis

Survival Analysis is a common method for assessing the outcome of cancer care. Survival data are usually divided into predictor variables for the patient (X_i), the time to an event ($T_i \in [0, T_{max}]$), and the event indicator ($D_i \in \{0,1\}$). If $D_i = 1$, T_i corresponds to the total survival time between the first diagnosis of lung cancer and the death of the patient, else, it is the time delta between the diagnosis and the last known follow-up. These cases are called right censored and prevent the usage of normal regression methods to estimate the event time. However, methods developed for this kind of data, such as Cox Regression [5] and Random Survival Forests [6], have problems to adapt to datasets with many (non-ordinal) categorical data columns, and their performance is usually hard to differentiate quantitatively [7]. There are multiple studies transferring this regression problem to multiple binary classification problems [8,9]. This allows us to train classifiers based on the non-censored subset of the data at specific time steps.

In this work, we aim to binary classify the patients' survival after 6, 12, 18 and 24 months, respectively, and predict the feature importance at each step. For this, we used CatBoost, which is a gradient boosting algorithm [10]. Its predictor is built by iteratively combining weaker tree classifiers into a strong learner. It is specialized on dealing with categorical data through an ordinal target encoding strategy, which is important as most variables in our requested data are categorical.

2.2. Data Retrieval and Preprocessing

As of August 2024, seven of the requested registries have delivered data for analysis. These are the cancer registries of Hamburg (HH), Lower Saxony (NDS), North Rhine-Westphalia (NRW), Saarland (SL), Saxony (SN), Saxony-Anhalt (SNA) and Schleswig-Holstein (SH). We requested the data in oBDS-RKI format. This XML scheme is supported by all German cancer registries and has both syntactical and semantical constraints. Thus, using it eases data integration. The retrieved data was transformed into tabular data using XQuery and R. For our study, we exclude cases which where only observed after death. To simplify data handling, only the first tumor, surgery, radiotherapy and systemic therapy item of each patient was used. For each binary classification step, we separated the patients into three groups:

1. Patients who live longer than x months ($T_i > x$),
2. Patients who decease within x months ($T_i \leq x$ & $D_i = 1$) and
3. Patients who are right censored within x months ($T_i \leq x$ & $D_i = 0$).

Only the first and second group are used as training and test data for each time step.

3. Results

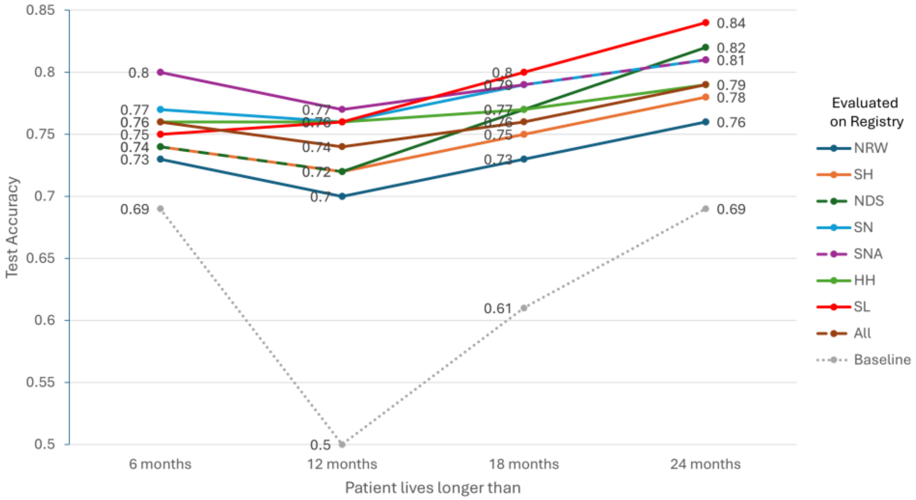

Figure 1. Test accuracy on each left-out registry dataset for 6, 12, 18 and 24 months. For ‘All’, a random 80% split of all registry data at this timestep was used for training, while 20% were used for validation. The ‘Baseline’ represents a classifier that always returns the most frequent class in the respective ‘All’ dataset.

3.1. Retrieved Data

We retrieved data of a total of 198,111 lung cancer patients with 201,637 tumours. These split into 99,066 patients (100,302 tumours) from NRW, 22,244 (22,656) from SH, 21,624 (22,204) from NDS, 21,311 (21,587) from SN, 13,215 (13,589) from SNA, 12,017 (12,438) from HH and 9,642 (9,888) from SL. Men were diagnosed with lung cancer more often (56.3% to 70.0% of the registries data) than women. The mean age at diagnosis lies between 68.5 in NDS and 70 in HH. The proportion of uncensored patients decreases over time: At 6 months, 64.6% to 87.2% of patients are uncensored (mean 81.5%), at 24 months, 58.9% to 81.2% (mean 74.0%). 39.0% to 61.2% of patients live (confirmed) longer than 6 months after diagnosis (mean 55.0%), while 11.7% to 26.5% (mean 21.3%) live longer than 24 months.

3.2. Binary Classifications

For each classification time step, the models were trained on all data except for one cancer registry, which was used for evaluation. In a further ablation, training and evaluation data are a stratified random 80/20 split of the combined data from all registries (‘All’). The classification results can be seen in Figure 1. For the classification whether a patient lives longer than 6 months, the model accuracy is between 0.73 and 0.8 (mean 0.76), decreases to 0.7-0.77 (mean 0.74) for 12 months, and afterwards increases to 0.76-0.84 (mean 0.8) for 24 months. All models performed better than baseline classifiers that return the most frequent class label. For the F1 score of the class of deceased patients, we observe a mean value of 0.59 (6 months), 0.76 (12 months), 0.83 (18 months) and 0.87 (24 months) across all CatBoost models.

3.3. Feature Importance

For the ‘All’ ablation, we performed a permutation feature importance analysis [11], which can be seen in Figure 2. Across all classification models, the age at diagnosis is the most important feature. The presence of distant metastases, reflected by the features “Localisation Distant Metastases”, “cTNM_M” (clinical metastases assessment) and “TNM_M” (final metastases assessment), is consistently one of the top features, which is consistent with our previous study [7]. Also, information about the therapy in form of the kind of systemic therapy, used substances, and its intention are weighted relatively high.

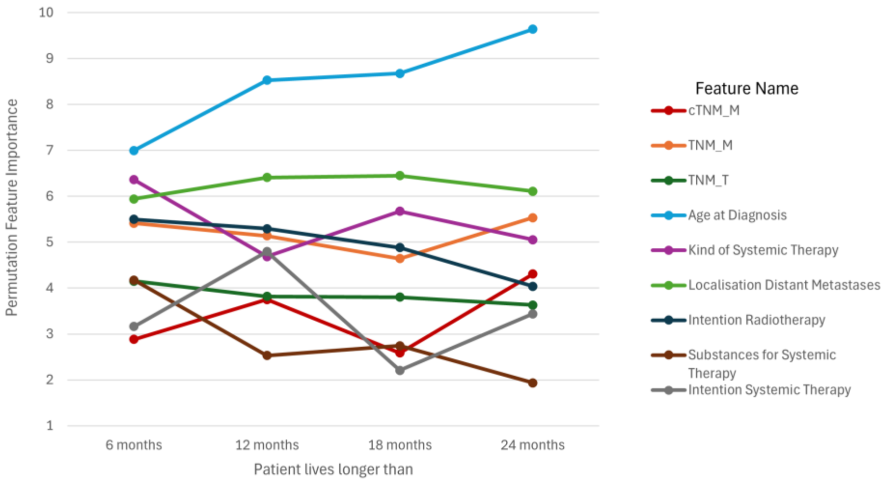

Figure 2. Permutation Feature Importance of the nine most important features.

4. Discussion

Survival regression models are complicated to train due to the presence of censored data. We have trained multiple binary classifiers and could show that this approach circumvents the censoring problem.

It is possible to combine the output of the binary classifiers to a more powerful multi-class ordinal classifier or regressor. This can be done via formula [12] or by using another classification algorithm using the predicted probabilities of the classifiers as input. We leave this for future research. The found features by the importance analyses can be used for survival analysis models in a next step, for example to differentiate the survival based on the chosen kind of systemic therapy and used substances.

We could observe differences of up to 20% in the proportion of therapies administered, though lung cancer treatment in Germany is usually standardized by guidelines. This suggests a possible under-reporting to the registries. The differences in censoring and survival rates between the registries also need to be investigated in future.

5. Conclusions

We requested lung cancer data in the form of the oBDS-RKI from the state cancer registries in Germany. Seven cancer registries provided data, allowing us to evaluate classification models using individual cancer registries as a validation set. We found that the classification models performed better than baseline across all different training registries and time steps. We were also able to see the progression of feature importance over these time steps, providing valuable information for future research. The source code for our models and analyses is available at https://github.com/AI-CARE-Consortium/AI-CARE-Survival-Classification.

Acknowledgements

This work was done as part of the AI-CARE project, funded by the German Federal Ministry of Health (BMG), grant number ZMI5-2522DAT13J. We would like to thank all project members (https://ai-care-cancer.de/info#workinggroup) for their support. Also, we would like to thank all cancer registries that provided data for this study.

References

[1] Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, et al. Global Cancer Observatory: Cancer Today. n.d. Accessed August 13, 2024. https://gco.iarc.who.int/today/.

[2] Katalinic A, Halber M, Meyer M, Pflüger M, Eberle A, Nennecke A, et al. Population-Based Clinical Cancer Registration in Germany. Cancers 2023;15:3934. doi:10.3390/cancers15153934.

[3] Meisegeier S, Imhoff M, Berg K, Kraywinkel K. Bundesweiter klinischer Krebsregisterdatensatz - Datenschema und Klassifikationen 2023. doi:10.5281/zenodo.10022040.

[4] Brierley J, Gospodarowicz MK, Wittekind C, editors. TNM classification of malignant tumours. Eighth edition. Chichester, West Sussex, UK ; Hoboken, NJ: John Wiley & Sons, Inc; 2017.

[5] Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological) 1972;34:187–220.

[6] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The Annals of Applied Statistics 2008;2:841–60. doi:10.1214/08-AOAS169.

[7] Germer S, Rudolph C, Labohm L, Katalinic A, Rath N, Rausch K, et al. Survival Analysis for Lung Cancer Patients: A Comparison of Cox Regression and Machine Learning Models. International Journal of Medical Informatics 2024:105607. doi:10.1016/j.ijmedinf.2024.105607.

[8] Mourad M, Moubayed S, Dezube A, Mourad Y, Park K, Torreblanca-Zanca A, et al. Machine Learning and Feature Selection Applied to SEER Data to Reliably Assess Thyroid Cancer Prognosis. Sci Rep 2020;10:5176. doi:10.1038/s41598-020-62023-w.

[9] Doppalapudi S, Qiu RG, Badr Y. Lung cancer survival period prediction and understanding: Deep learning approaches. International Journal of Medical Informatics 2021;148:104371. doi:10.1016/j.ijmedinf.2020.104371.

[10] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc.; 2018. doi:10.48550/arXiv.1706.09516.

[11] Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 2007;8:25. doi:10.1186/1471-2105-8-25.

[12] Frank E, Hall M. A Simple Approach to Ordinal Classification. In: De Raedt L, Flach P, editors. Machine Learning: ECML 2001, Berlin, Heidelberg: Springer; 2001, p. 145–56. doi:10.1007/3-540-44795-4_13.