International Conference on Actuarial Sciences (ICActS)
Available on-line at: **www.sicash.sliit.lk**

A Machine Learning Approach to Actuarial Life Table Estimation in Lung Cancer Patients

D. D. H. Tharushika[1]*, N. A. D. N. Napagoda[1]

[1]*Department of Mathematical Sciences, Faculty of Applied Sciences, Wayamba University of Sri Lanka, Kuliyapitiya, Sri Lanka.*

Corresponding author*: tharushikadampahala71@gmail.com

Abstract

Cancer-related mortalities worldwide are most caused by lung cancer, and one of the major causes of passing worldwide is still cancer. A dangerous disease is lung cancer, which requires accurate survival modelling to assist in actuarial evaluations, public health planning, and clinical decisions. Life expectancy and mortality risk across age groups are calculated using essential tools such as actuarial life tables, but complex real-world data is frequently struggled with by traditional methods. Actuarial life tables for patients with lung cancer are created using a data set of more than 500,000 patient records with 15 key variables from 2014 to 2024 across European countries, employing Extreme Gradient Boost Accelerated Failure Time (XGBoost AFT) based survival analysis. The main objective is to develop age-specific mortality rates and life expectancy for patients with lung cancer. In contrast to earlier research that was reliant on traditional models, the nonlinear learning capabilities of XGBoost AFT models are utilized in this study to allow for more accurate estimation of mortality trends. A data-driven, machine learning approach to actuarial life table development is contributed by this study, with information about lung cancer survival patterns being provided. The understanding of survival trends, treatment planning, efficient use of healthcare resources, and assessment of the results of initiatives is aided by physicians, researchers, and policymakers. Public health initiatives focused on early identification and prevention are also guided, as well as future healthcare requirements being forecast.

***Keywords*:** Actuarial life table, Extreme Gradient Boost, Lung cancer, Life expectancy, Mortality rates

Introduction

Cancer is stayed one of the primary causes of disease and mortality worldwide, with millions of deaths being accounted for each year. Among all cancer types, lung cancer is considered the most common cause of cancer-related dead worldwide, with an estimated 1.8 million deaths being accounted for annually (Wheless et al., 2013). A significant public health concern is stood by it because of its late detection, aggressive nature, and low long-term survival rates. Given this high mortality problem, survival trends and life expectancy among lung cancer patients are considered crucial for clinical decision-making, healthcare planning, and actuarial assessments. The life table is recognized as a fundamental tool in demography and actuarial science, with age-specific estimations of survival chance and life expectancy being provided. Traditionally, historical population data are used to create life tables, and uniform mortality risks are assumed.

A comprehensive data set of over 500,000 lung cancer patients in Europe from 2014 to 2024 is made use of in this study, which includes 15 critical variables. The primary analytical approach used to construct

https://doi.org/10.54389/XWTN8091

accurate and robust actuarial life tables is survival analysis, particularly through the Extreme Gradient Boosting Accelerated Failure Time (XGBoost AFT) model. This study aims to develop a reliable and precise actuarial life table for lung cancer patients by applying the XGBoost AFT survival analysis method. It involves a critical review and evaluation of existing research methodologies, emphasizing the potential of machine learning to enhance predictive accuracy and reliability across diverse patient-related factors.

Naing et al., (2017) conducted a reflective Cohort Analysis on primary lung cancer patients diagnosed in Brunei Darussalam between 1987 and 2012. Most patients were detected late, with a median survival of 6.1 months. Survival rates increased considerably between 1987-1999 and 2000-2012, despite a higher proportion of late-stage diagnoses in the later period. The study discovered that survival rates were comparable to those of other wealthy countries, but it also highlighted the significance of support early detection and anti-smoking programs. Park et al., (2020) conducted a retrospective Cohort Analysis in South Korea to estimate lifetime survival, Years of Life Lost (YLL), and medical expenses for lung cancer patients diagnosed between 2004 and 2010, and followed up until 2015. The findings revealed an average life survival of 4.5 years for patients and 14.5 years for controls, with an average YLL of 10 years. The study emphasized the importance of early identification, which enabled surgery to lessen disease burden. Veisani & Delpisheh, (2016) used actuarial life-table methods to studied survival rates and prognostic factors among 746 gastric cancer patients reported in Iran between 2001 and 2006. The median survival timed was 24.2 months, and 1 to 5 year survival rates fell from 73.6% to 29.7%. The studied demonstrated a miserable forecast for advanced-stage growths and identified surgery as an important option for enhanced survival. Age-specific mortality and life expectancy are forecasted by this work using a large-scale European dataset and the nonlinear, flexible XGBoost AFT technique. The accuracy of actuarial life tables is improved by this data-driven approach, and intricate mortality patterns are identified, in contrast to conventional cohort or parametric models. The major goal of this study is to provide reliable, disease specific actuarial life tables for lung cancer patients using a machine learning-based survival model. Nonlinear interactions and complicated risk factors hidden in clinical and demographic data are detected by the approach, allowing for more exact estimates of life expectancy. For practical uses including improving patient counselling, maximizing treatment planning, directing healthcare policy decisions, and boosting the accuracy of insurance and public health models customized to cancer outcomes, the essentially of this lung cancer-specific life tables is recognized.

Materials and Methods

Data description

Over 500,000 observations from the European region are contained in the data collection received through the web portal. A ten-year period, from 2014 to 2024, is covered, and a strong foundation for research is provided by 15 key variables, including both numerical and categorical data types.

Data pre-processing

Improvements to model performance and data quality are made through data preparation. Outliers are identified using the IQR method, and capping is applied to them. Categorical data is converted using label encoding, while features are adjusted for a balanced contribution through Z-score standardization. In order to assess model generalization, the dataset is finally divided into training and validation sets (80:20).

Extreme Gradient Boosting Accelerated Failure Time (XGBoost AFT) Survival Analysis

The Accelerated Failure Time (AFT) model is a parametric model that is used in survival analysis and that directly predicts the impact of factors on survival time Saikia & Pratim Barman, (2017). In contrast to the Cox model, it is suggested by AFT that a logarithmic transformation of survival time is linearly connected to the predictors. The logarithm of survival time can be stated as,

$$\log(T) = \beta x + \sigma\epsilon \tag{1}$$

Where x represents the vector of input features, β specifies the coefficients that quantify the importance of each feature, and σ is a scale parameter governing the spread of survival times, and ϵ is a random error term.

Concordance index (C-Index)

A performance statistic called the Concordance Index (C-Index) is used to assess how well survival models predict results. The percentage of all usable patient pairs in which the patient with the shorter survival time is accurately predicted to have a higher risk is determined, so that the match of the predicted survival times with the actual results can be assessed. A C-Index of 0.5 is denoted by random chance, whereas perfect prediction is denoted by 1.0(Blanche et al., 2019).

Results

Data pre-processing

Approximately 500,000 observations and 15 variables were included in the data set. Outliers were only found in the age variable, with 4,564 values recognized. The capping procedure was used to manage these outliers. After that, label encoding and standardization were performed, and the data set was divided into training and testing sets.

Extreme Gradient Boosting Accelerated Failure Time (xgboost AFT) Survival Analysis Life table

Survival patterns were evaluated by dividing patient data into 5-year age intervals. The probability of death (qx) was computed for each age group by dividing the number of deaths by the total population at risk within that group. To reduce random variations, the qx values were smoothed using a cantered moving average with a window size of three. Starting with a base of 100,000, predictable life table functions were computed iteratively: the number of survivors in each age group (lx), the number of deaths within the age group (dx), and the number of person-years lived in the interval (Lx). The total number of years remaining above each age group (Tx) was determined from the bottom up. Finally, the life expectancy at the start of each age interval (ex) was calculated by dividing the total remaining person-years (Tx) by the number of people alive at the beginning (lx). Table 1, an illustrated smoothed and ordered actuarial life table that shows the mortality of lung cancer patients stratified by age.

Table 1: *Lung Cancer Patient Life Table for Europe, 2014–2024*

Age group	qx	lx	dx	Lx	Tx	ex
0-4	0.145879	100000.00000	14587.90000	92706.05000	798031.71123	7.980317
5-9	0.146304	85412.10000	12496.10341	79164.04830	705325.66123	8.257913
10-14	0.125517	72915.99659	9152.19714	68339.89802	626161.61293	8.587438
15-19	0.111913	63763.79945	7136.01934	60195.78978	557821.71491	8.748251

Age group	qx	lx	dx	Lx	Tx	ex
20-24	0.088677	56627.78011	5021.60053	54116.97984	497625.92514	8.787664
25-29	0.079622	51606.17957	4109.00443	49551.67736	443508.94530	8.594105
30-34	0.078120	47497.17514	3710.46349	45641.94340	393957.26794	8.294330
35-39	0.083254	43786.71165	3645.43349	41963.99491	348315.32454	7.954818
40-44	0.092277	40141.27816	3704.13011	38289.21311	306351.32963	7.631828
45-49	0.092015	36437.14806	3352.76418	34760.76597	268062.11652	7.356836
50-54	0.085695	33084.38388	2835.16628	31666.80074	233301.35055	7.051706
55-59	0.078310	30249.21760	2368.80615	29064.81453	201634.54981	6.665777
60-64	0.071891	27880.41146	2004.34137	26878.24077	172569.73528	6.189641
65-69	0.073686	25876.07009	1906.69548	24922.72235	145691.49451	5.630356
70-74	0.073467	23969.37461	1760.95006	23088.89959	120768.77216	5.038462
75-79	0.088488	22208.42456	1965.17167	21225.83872	97679.87257	4.398325
80-84	0.098331	20243.25289	1990.53930	19247.98324	76454.03385	3.776766
85-89	0.118441	18252.71359	2161.86357	17171.78181	57206.05061	3.134112
90-94	0.120416	16090.85002	1937.59043	15122.05481	40034.26880	2.488015
95-99	0.124888	14153.25959	1767.57700	13269.47109	24912.21399	1.760175
100-104	0.119968	12385.68259	1485.87938	11642.74290	11642.74290	0.940016

A significant mortality rate is observed at young ages, with a death probability (qx) of 14.6% for those aged 0 to 4, implying that pediatric or early-start lung cancer cases may be particularly aggressive or frequently discovered late. Remarkably, the lowest death rates are recorded between the ages of 70 and 74, indicating a potential space of substantially better prediction, possibly due to improved therapy or earlier discovery during this time. Despite this, life expectancy (ex) is seen to fall steadily throughout all age categories. In senior age groups 70 and up, the risk of death is considered, underscoring the long-term impact of the disease.

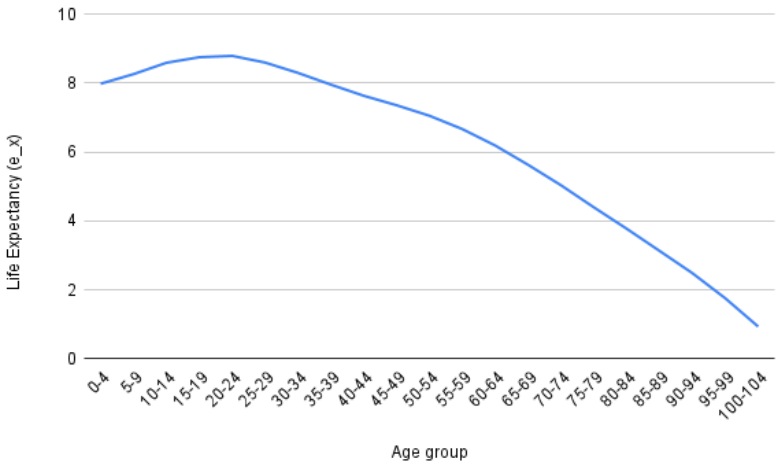

Figure 1: *The graph between life expectancy and age groups among lung cancer patients*

The link between life expectancy and age groups among lung cancer patients is showed by the figure 1. A declining trend is clearly demonstrated, with life expectancy decreasing as age increases. It is shown that a lesser anticipated survival time is associated with older lung cancer patients compared to younger patients.

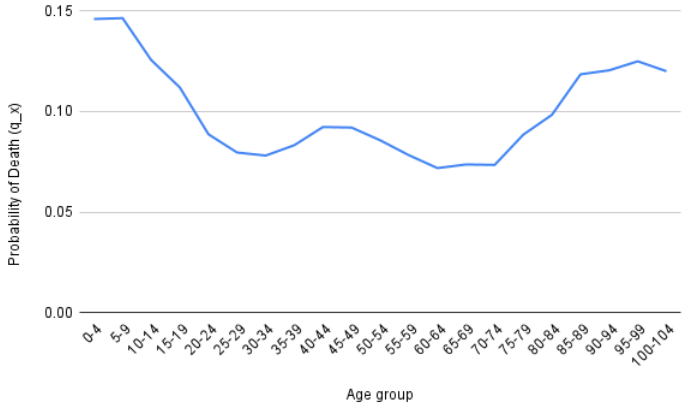

Figure 2: *Death probability curve*

U-shaped death probability curve is illustrated by Figure 2, which is had by lung cancer patients that varies with age, with the lowest risk being had by middle-aged individuals (55–59) and the highest risk being had by new-borns (0–4) and the elderly (90–94). The combined effects of increasing weakness later in life due to aging and other health difficulties, better treatment in maturity, and the severity of the disease in the very young.

Concordance Index (C-Index)

The XGBoost AFT survival model in this study is considered by a Concordance Index (C-Index) of 0.71, which suggests a high degree of predictive accuracy. It is indicated that a lower survival time for the patient is accurately predicted by the model in 71% of similar patient pairs.

Discussion

Previous research is done upon by this study through the application of XGBoost Accelerated Failure Time (AFT) survival modelling to a large, recent European lung cancer data set of over 500,000 patients from 2014 to 2024. The creation of detailed actuarial life tables and refined survival estimates that capture complex patient-level factors is allowed by this. Modern machine learning techniques are used in this study to improve predictive accuracy and reflect current practices, unlike the cohort analysis conducted by Naing et al., (2017) using traditional Kaplan-Meier methods, the semi-parametric treatment-group survival estimates provided by Park et al., (2020) and the gastric cancer life-table approach utilized by Veisani & Delpisheh, (2016) More clinical and genomic data should be included in future research, other machine learning survival models should be investigated, and findings should be validated across varied populations to improve actuarial life table accuracy and individualized prognosis in lung cancer management.

Conclusions

From 2014 to 2024, Extreme Gradient Boosting Accelerated Failure Time (XGBoost AFT) survival analysis was effectively utilized on a large data set of roughly 500,000 lung cancer patients in Europe. An actuarial life table was created through careful pre-processing of the stratifying of survival patterns into 5-year age intervals, revealing critical death trends in lung cancer patients. It is shown by the life table that significantly higher mortality risks are faced by young children and the elderly, whereas a relative survival advantage is held by the 60-69 age group, providing detailed insights into disease development and forecast. The primary contribution of this study is the use of advanced machine

learning survival modelling to build an improved, data-driven actuarial life table that is outperformed by existing methods in terms of accuracy and robustness. A more accurate tool for forecasting lung cancer patient survival, optimizing treatment strategies, and effectively allocating healthcare resources is provided to physicians, policymakers, and researchers by this model-based life table.

References

Naing, L., Abdullah, A., Abdullah, S., & Kifli, N. (2017). Survival of primary lung cancer patients in Brunei Darussalam, 1987–2012. *Annals of Translational Medicine, 5*(5), 98.

Park, H. Y., Hwang, J., Kim, D. H., Jeon, S. M., Choi, S. H., & Kwon, J. W. (2020). Lifetime survival and medical costs of lung cancer: a semi-parametric estimation from South Korea. *BMC cancer, 20*, 1-10.

Saikia, R., & Barman, M. P. (2017). A review on accelerated failure time models. *International Journal of Statistics and Systems, 12*(2), 311-322.

Saxena, K. (2025). *Lung Cancer Dataset*. Kaggle.com.https://www.kaggle.com/datasets/khwaishsaxena/lung-cancer-dataset

Thandra, K. C., Barsouk, A., Saginala, K., Aluru, J. S., & Barsouk, A. (2021). Epidemiology of lung cancer. *Contemporary Oncology/Współczesna Onkologia, 25*(1), 45-52.

Veisani, Y., & Delpisheh, A. (2016). Survival rate of gastric cancer in Iran; a systematic review and meta-analysis. *Gastroenterology and hepatology from bed to bench, 9*(2), 78.

Wheless, L., Brashears, J., & Alberg, A. J. (2013). Epidemiology of lung cancer. *In Lung Cancer Imaging*, 1-15. New York, NY: Springer New York.