Analysis of the SEER Data set for Lung Cancer Diagnosis for Stage Classification and Survival Analysis

V. Deepa and S. K. B. Sangeetha

Abstract The basic usage of survival analysis is to perform a medical observation and the amount of time needed for the analysis. Survival analysis is analysed mainly in the field of biological engineering. The tumours are diagnosed using a density function. Our study involves the evaluation of the parametric analysis of the probability survival models. Our proposed work is to develop models for the stage classification of lung cancer using the region-based SEER data set. The mortality rate is calculated at the end of the cancer's progression. The classification challenge can be seen via the viewpoint of survival analysis. The survival statistics helps in analysing the cancer stages and the mortality rate.

Keywords Mortality rate · Survival analysis · Cancer stages

1 Introduction

When a patient is diagnosed with lung cancer, clinicians attempt to determine the spread of the disease which is termed as staging. The stages of cancer is required for radiologist to determine the spread of the disease throughout the body. The survival statistics is generated using the stages of cancer. The term "survival time" refers to the time between the diagnosis of an illness and the patient's death. At the end of the cancer's progression, when the disease is most aggressive, a survival analysis is projected.

V. Deepa (✉) · S. K. B. Sangeetha
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, Chennai 600026, India
e-mail: dv1019@srmist.edu.in

S. K. B. Sangeetha
e-mail: sangeets8@srmist.edu.in

A. Kumar et al. (eds.), *Proceedings of the 4th International Conference on Data Science, Machine Learning and Applications*, Lecture Notes in Electrical Engineering 1038,
https://doi.org/10.1007/978-981-99-2058-7_3

1.1 Patients with Stage I Cancer

When a patient of over 50 years is diagnosed with stage I cancer, chemotherapy is done to reduce the survival rate of the patients. The treatment is mandatory to reduce the hazard. The hazard rate can be reduced with stage I cancer.

1.2 Patients with Stage II Cancer

The patient is given surgery and radiation when they are diagnosed with stage II cancer. When the age of the patient is more, the surgery and radiation is required at higher level to reduce the risk factor of the survival rate of the patient.

1.3 Patients with Stage III Cancer

When a patient below fifty age is diagnosed with stage III cancer, radiation is required. The radiation helps the patient to reduce the hazard. The survival rate reduces with the increase in age of the patients. The risk factor increases with age of the patient. The tumours spreads gradually if proper treatments are not given.

1.4 Patients with Stage IV Cancer

The patient is given cancer therapy and radiation when they are diagnosed with stage IV cancer. When the patient age increases, the mortality rate also increases. When the age of the patient is more, the surgery and radiation is required at higher level to reduce the risk factor of the survival rate of the patient. Lung cancer is divided into two stages, namely limited stage and the extended stage. A sponge like structure available in the chest area helps the breathing easier. The air enters through the mouth and the windpipe is called as alveoli. Lung cancer begins in the bronchial lining of the cells. Lung cancer is divided into three categories: localised, regional, and distant. The TNM stage is used to indicate the cancer stages such as tumour, nodule, and metastasis.

Table 1 Cancer stages and lymph node

S.no	Stages	Description
01	Stage I	The malignancy in the lungs has not migrated to the lymph nodes
02	Stage II	The malignancy has migrated to the lymph nodes adjacent
03	Stage III	The malignancy has migrated to the nearby lymph nodes denoted by III A,B
04	Stage IV	The fluid has spread around all parts of the organs which is called as advanced stage of lung cancer.

2 Risk Factors of Lung Cancer

when the patient is diagnosed with lung cancer the cancer cells are identified using the radiotherapy. The Lung cancer is classified into two types namely small cell and non-small cell adrenal lung cancer. The different symptoms of the lung cancer include fatigue, chest pain, loss of appetite, and shortness of breath. The CT scan was used to determine the disease's diagnosis. The scan helps in indicating the spread level of the tumour which helps the radiologist and medical analyst to examine the disease (Table 1).

3 Materials and Methods

3.1 Patient Population

The Seer data set has 18 cancer registries across United States. The data is extracted from the SEER*Stat version software. The SEER data set helps in analysing the TNM stages of the disease. The system helps in analysing the cases in the local, regional, and distant throughout the body. The different clinical characteristics under examination were the gender, age, year of diagnosis, surgery, and radiation therapy. A multivariate observation was done based on the different set of attributes. Cancer-specific survival was analysed for each sub-group using the Kaplan–Meier method.

4 Statistical Analysis

The SEER data contains different information on patient information such as primary tumour size, tumour morphology information, stage at diagnosis, and first course of treatment. The SEER registries keep track of cancer patients' survival rates. The analysis of the SEER data was done using the cohort selection. Cohort selection for our experiments is performed with SEER*Stat, a dedicated statistical software for

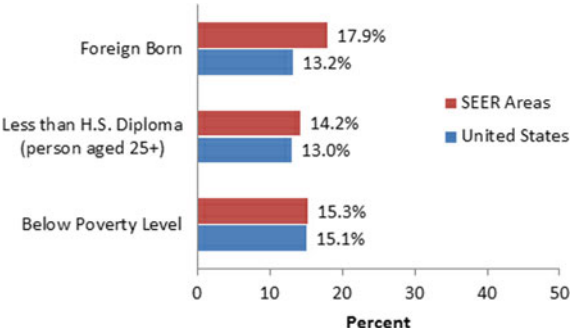

Fig. 1 Survey on the SEER population

the analysis of the SEER data. SEER*Stat configuration can be stored into a session file and re-used by others for the cohort analysis data. The cohorts consist of 229,011 cases for lung cancer which is called as the SEER*Stat session files. The SEER data included the cases of the malignant tumours and the benign tumours. The American Survey is taken between the years 2016–2021.The analysis of the lung cancer data set was analysed based on the different set of cancer survival based on the each survival group (Fig. 1).

4.1 Tumour Analysis

The SEER data set consists of the lung cancer data which performs stage classification based on the size of the tumour. The nodules present in the contralateral lungs is called as M1a, and the tumours in the extra-thoracic organ are called as metastases (Fig. 2).

Fig. 2 Tumours in lung CT image

The SEER data indicates that the chemotherapy causes 10% increase in the survival rate. Two types of treatment are given namely chemotherapy and radiotherapy. These treatments reduce the risk of the disease. Surgery is the most effective treatment for patients with early deduction of the stages of lung cancer. Surgical removal of tumour is faster and more efficient method compared to the other type of treatments. Different kinds of survival analysis include clinical trials, cohort studies, and statistical analysis of the lung cancer patients.

5 Parametric Analysis

The survival analysis is used to predict the mortality rate of the patients. The different survival analysis metrics are parametric and non-parametric analysis. The Kaplan–Meier estimator is used to compute the mortality rate of the patients for a limited set of groups.

Table Data

Surveillance, epidemiology, and end results program: unique analyses and critical insights.

Analyses	Critical insights
1. Population-based cancer rates	1. Absolute risk of cancer occurrence
2. Rare cancer rates	2. Precise and comprehensive description
3. Cancer rates in minority groups	3. Healthy disparity assessments
4. Birth cohort effect	4. Risk factor exposure assessments
5. Calendar periodic effect	5. Benefits/harms of screening

6 Experimental Setup

6.1 Online Lung Cancer Outcome Calculator

An online tool is used to predict the lung cancer. The analysis of the tool was done using the five outcome variables to remove the redundant attributes. In order to compute the mortality rate, 13 variables are used.

The description of the variables is given as follows:

1. **Patient age**: The age of the patient is given as numeric value during the diagnosis of the lung cancer.
2. **Birth place**: The birth place of the patient is given as character value. There are totally 198 options available in the SEER database to select for the attribute.

3. **Cancer grade**: It is represented as attributes such as well-grown, poor, and well differentiable. It is the description of how the cancer cells grows from the initial level.
4. **Diagnostic confirmation**: The most commonly used attribute for the confirmation of the lung cancer is denoted by the laboratory test results such as positive histology and negative histology.
5. **Tumour extension**: The tumour spreads from the local region to the metastasis region is called as spread. There are 20 options available which is called as localised or the lymphatic region. The attribute name is represented as 'EOD extension'.
6. **Lymph node**: The most commonly used attribute for the lymph node is denoted by EOD lymph node. There are eight options available.
7. **Surgery Type**: It is the description of the technique used to remove the cancerous tissue from the lung. There are totally 25 options available in the SEER database to select for the attribute.
8. **No surgery attribute**: The reason should be represented as character type. The different options available are surgery performed (yes/no) and reasons.
9. **Surgery and radiation therapy**: It is denoted as the sequential procedure for the different operations such as surgery and radiation.
10. **Lymph node surgery**: It is the description of the surgical procedures used to remove the lymph nodes at the time of surgery and biopsy. There are totally eight options available in the SEER database to select for the attribute.
11. **Cancer stage**: The stage is denoted as the spread of cancer such as tumour, region analysis, and the detection of the disease spread.
12. **Malignant tumours**: It is denoted as the total number of tumours during the patient lifetime. It helps us in identifying the numeric, categorical tumours.
13. **Regional lymph nodes examination**: It is denoted as the total number of regional lymph nodes that were removed and examined by the pathologist. A total of the 63 attributes are available which provides a maximum accuracy of 91.4% (Fig. 3).

7 SEER Data Dictionary

The SEER data dictionary consists of different variables such as age, size, tumour size, T, N, M classifications. Age: The age is represented as a three-digit code for denoting the patient age in years.

Grade: The grades are represented in ranges ICD-O-2.The Grade 1 indicates the cell may look normal, and Grade 2 indicates abnormal growth of the cells. Tumour Size: The tumour size is measured in mm. The codes for representing the data are 991–995.

The SEER data dictionary uses the unlabelled data which uses data-driven process [1]. The most commonly used techniques are the clustering for lung cancer stage classification and prediction.

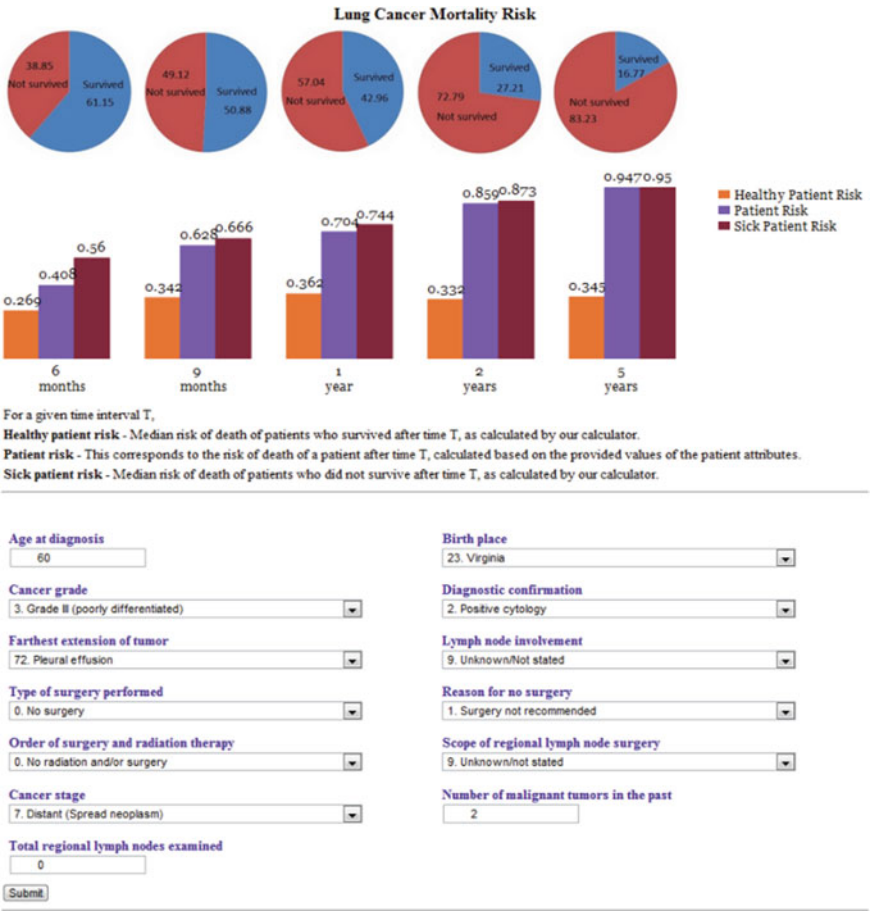

Fig. 3 Screenshot of the lung cancer outcome calculator

The ultimate goal of the SEER data model is the text classification which provides the best outcome for classification which operates on unlabelled data [2] for a given example.

The SEER data model refers the classification task having “true” or” false “value or “yes” or “no” [3].

The SEER data model is regarded as a class label for model prediction for lung cancer [4]. To predict the class for the given set of data points, it can be carried out in structures and unstructured data [5].

Another commonly used statistical model for lung cancer prediction and stage classification is the logistic regression (LR) [6].

7.1 Results

The SEER data helps in analysing the percentage of people affected with the lung cancer. It helps in calculating the mortality rate using the survival analysis. The statistical information can be computed from the cancer registry.

8 Conclusion and Future Work

The SEER population study has several limitations since it relies mainly on the cancer registry data. Our study of the SEER data also has several strengths. The patients are derived from population-based tumour registry, hence providing the accurate results. Because of the extensive data collected from the SEER program, we were able to analyse the cancer stage survival using the demographic information such as age, sex, and race. Lung cancer was one of the most spreading cancer type used for experimental analysis.

References

1. Magaji BA, Moy FM, Roslani AC, Law CW (20170 Survival rates and predictors of survival among colorectal cancer patients in a Malaysian tertiary hospital, pp 162–173
2. SEER* Stat Software. National Cancer Institute Surveillance, Epidemiology, and End Results Program (SEER). Available online: https://seer.cancer.gov/seerstat/. Accessed on 4 May 2018
3. Ung M, Rouquette I, Filleron T, Taillandy K, Brouchet L, Bennouna J et al (2016) Characteristics and clinical outcomes of sarcomatoid carcinoma of the lung. Clin Lung Cancer 234–245
4. Yendamuri S, Caty L, Pine M, Adem S, Bogner P, Miller A et al (2019) Outcomes of sarcomatoid carcinoma of the lung: a surveillance, epidemiology, and end results database analysis. Surgery 152(3):397–402
5. Imani F, Chen R, Tucker C, Yang H. Random forest mode ling for survival analysis of cancer recurrences, 9:183–192. Bentham publishers
6. Wongvibulsin S, Wu KC, Zeger SL (2020) Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF- SLAM) data analysis, 98:1–14
7. Pradeep KR, Naveen NC (2018) Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4.5 and Naive Bayes algorithms for healthcare analytics, 67:412–420
8. Nezhada MZ, Sadati N, Yanga K, Zhub D (2018) A deep active survival analysis approach for precision treatment recommendations: application of prostate cancer, 56:16–26
9. Fathima N, Liu L, Hong S, Ahmed H (2020) Prediction of breast cancer, comparative review of machine learning techniques, and their analysis, 8:173–174
10. Parikh RB, Manz C, Chivers, C, Regli SH (2019) Machine learning approaches to predict 6-month mortality among patients with cancer, 77, 1–7
11. Wongvibulsin S, Wu KC, Zeger SL (2020) Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis, 98:1–14
12. National Cancer Institute Surveillance, Epidemiology, and End Results Program (SEER) (2017) From electronic health-records, vol 56, Nov 2017, pp 37–56. Available online: https://seer.cancer.gov/. Accessed on 4 May 2018

13. Herbst RS, Morgensztern D, Boshoff C (2018) The biology and management of non-small cell lung cancer. Nature 553:446–454
14. Ettinger DS, Aisner DL, Wood DE et al (2018) NCCN guidelines insights: non-small cell lung cancer, version 5 2018. J Natl Compr Canc Netw 16:807–821
15. Roesel C, Terjung S, Weinreich G, Hager T, Chalvatzoulis E, Metzenmacher M et al (2016) Sarcomatoid carcinoma of the lung: a rare histological subtype of non-small cell lung cancer with a poor prognosis even at earlier tumour stages. Interact Cardiov Th 24(3):407–413