DOI: 10.23977/tranc.2024.050112
ISSN 2523-6482 Vol. 5 Num. 1

Pan-cancer Research Based on Survival Analysis and Multivariate Analysis of Variance

Fanrui Sun

College of Mathematical Science, Yangzhou University, Yangzhou, 225009, China

Keywords: Cancer Survival Rate, Exploratory Analysis, Multifactor Analysis of Variance, Tukey Test

Abstract: Cancer has become a significant public health and disease challenge worldwide, and there is an urgent need to improve the treatment levels for cancer patients. This study focuses on the survival analysis of cancer, using historical data on patient characteristics including year of diagnosis, gender, ethnicity, and cancer type to create box plots and line graphs for an initial exploratory analysis of the differences between various categories of these factors and how they have changed over time. Further employing multifactor analysis of variance to investigate the differences in various categories of factors and their interactive combinations, the article compared and selected an interaction model that presents the disparities among factors. Notably, cancer type, ethnicity, and year significantly impact survival rates, while gender differences are less pronounced, with four significant interactive combinations identified. The Tukey test is then used to analyze the differences between different categories within gender and ethnicity. Based on the factors with significant differences identified in this study, healthcare professionals can tailor personalized treatment plans according to patient characteristics, which also aids medical institutions in optimizing the allocation of healthcare resources, thereby improving patients' quality of life and survival rates.

1. Introduction

1.1 Research Background

In the realm of modern medicine, cancer has evolved into a paramount public health and disease concern on a global scale. According to data furnished by the World Health Organization (WHO), in excess of 20 million individuals are diagnosed with cancer annually, with a staggering mortality rate of 77%[1]. Despite the continuous advancements in medical technology, the incidence and mortality rates of cancer persist at elevated levels. Therefore, research on the treatment of cancer patients is of urgent necessity.

1.2 Research Significance

Cancer survival rates, as a crucial metric for assessing treatment effectiveness and disease burden, enable healthcare professionals to tailor treatment plans for individual patients by comparing the impact of different cancer types and patient characteristics with historical survival data, aiding patients and their families in making informed decisions. Biological differences among various races and genders lead to variations in cancer susceptibility, pathogenesis, and treatment responses. Disparities in socioeconomic status, education levels, and healthcare resources also affect the quality and availability of cancer screening, diagnosis, and treatment. Epidemiological studies have consistently shown significant disparities in cancer survival rates across races, genders, and types of cancer.

Multifactor analysis of variance plays a significant role in the study of cancer survival rates. It allows us to consider multiple related factors simultaneously, helping to reveal how they collectively affect survival rates. Through multifactor analysis of variance, researchers can identify key influencing factors, providing precise guidance for cancer treatment.

Therefore, survival analysis for cancer patients is vital for formulating more effective strategies for cancer prevention and treatment. This study aims to reveal differences in gender, ethnicity, and cancer type in the five-year survival rate tables of cancer patients in the United States in recent years, offering valuable insights and assistance for future cancer treatment research and public health practice.

2. Data

2.1 Data Description

This article utilizes a dataset from the Our World in Data platform, which provides five-year cancer survival rates for randomly selected years between 1963 and 2013 in the United States. The dataset is considered highly reliable.

The dataset encompasses survival rates, year of data collection, gender, ethnicity, and 17 distinct types of cancer, including pancreatic cancer, liver cancer, esophageal cancer, lung cancer, stomach cancer, brain cancer, myeloma, bladder cancer, breast cancer, cervical cancer, leukemia, ovarian cancer, colorectal cancer, prostate cancer, skin cancer, thyroid cancer, and an unspecified category of cancer. Each cancer type includes 111 data sets, with equal numbers of data sets across different genders and ethnicities, indicating that the data are random and unbiased. The temporal distribution is relatively uniform, with data collected approximately every three years, except for some older intervals that are spaced at ten-year intervals, and each time point has roughly equivalent amounts of data.

2.2 Data Preprocessing

To facilitate accurate analysis of the sample data, it is necessary to preprocess the data. Among the 1887 samples, there are 11.71% missing survival rate data, as shown in Figure 1. The presence of missing values is not conducive to subsequent analysis and visualization, and we cannot extract valid information from them. Moreover, considering that the selected sample size is sufficiently large and to prevent errors that may arise from imputing missing values, we have chosen to delete samples with missing values entirely. Subsequently, the data will be categorized and analyzed according to the required research types or time periods.

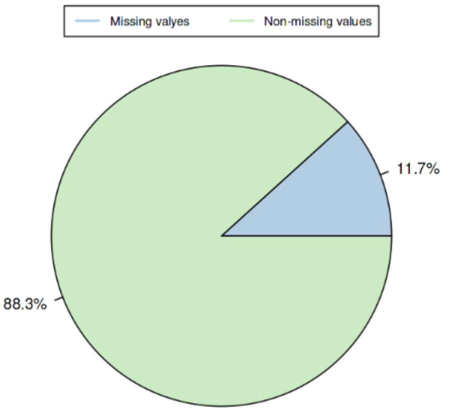


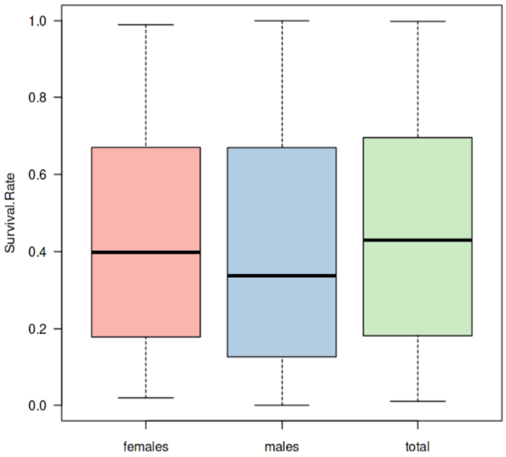
Figure 1: Proportion of missing values

3. Data Analysis

3.1 Exploratory data analysis

Survival analysis is a statistical method used to predict the probability of an event occurring within a certain period, taking into account time variations and handling censored data. It is widely applied in fields such as medicine, biology, engineering, economics, etc., for instance, studying disease prognosis, product lifespan, customer churn, etc.

Initially, this study aims to investigate whether cancer survival rates are influenced by gender. Therefore, box plots are created for the cancer survival rates categorized by gender, comparing females, males, and an unspecified gender type.



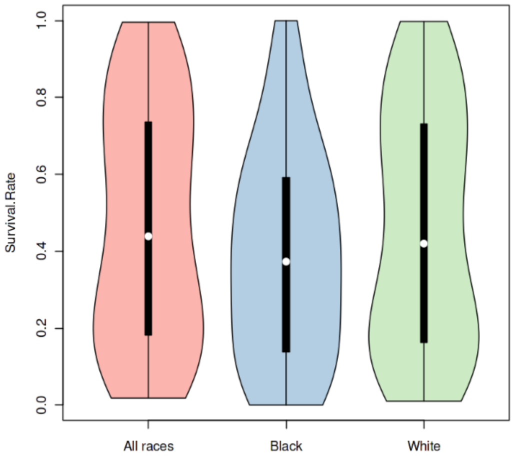


Figure 2: Survival Rate by Genders Figure 3: Survival Rate by Races

As shown in Figure 2, considering the median and lower quartiles, the cancer survival rate for females is greater than that of males, while the upper quartiles are roughly the same for both genders.

As shown in Figure 3, it can be observed that the survival rates for African Americans are lower than those of Whites, whether we compare the medians, upper quartiles, or lower quartiles.

Within the same gender and race, different types of cancer clearly have varying impacts on survival rates. To specifically compare the differences in survival rates across various types of cancer, box plots for the survival rates of 17 different types of cancer were created.

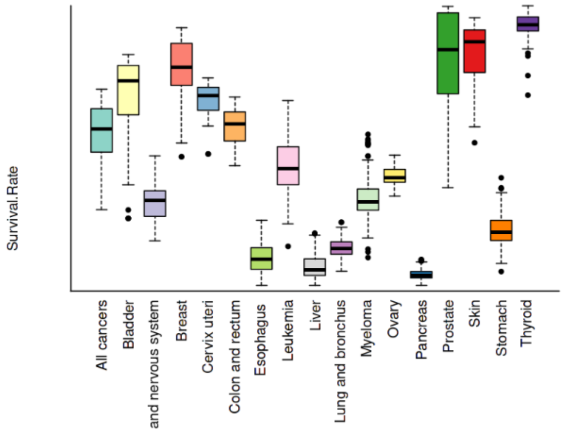


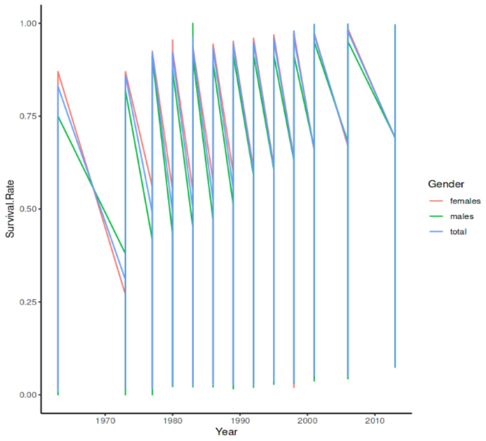
Figure 4: Survival Rate by Cancer Types

As shown in Figure 4, it can be observed that there are a few outliers outside the upper and lower quartiles in some of the box plots. The identification of outliers is usually based on the interquartile range (IQR), which is the difference between the upper quartile (Q3) and the lower quartile (Q1). If a data point exceeds Q3 + 1.5/QR or falls below Q1 - 1.5/QR, it is considered an outlier[2]. The presence of outliers is normal but represents exceptional cases within the sample and does not accurately represent the overall inference; therefore, we typically ignore them and only analyze the main body of the box plot during our analysis. Observing the box plots and disregarding the non-specified cancer type, the median survival rates, ranked from highest to lowest, are as follows: thyroid cancer > skin cancer > prostate cancer > breast cancer > bladder cancer > cervical cancer > colorectal cancer > leukemia > uterine cancer > brain cancer > myeloma > stomach cancer > lung cancer > esophageal cancer > liver cancer > pancreatic cancer.

3.2 Survival Analysis

In addition to gender, race, and type of cancer, survival rates may also change over time, and the extent and rate of change under different factors may vary. The following section will conduct a specific study.

Firstly, to explore whether gender differences have an impact on the change in cancer survival rates over time, we categorized the table data by different gender types and created line charts for the change in cancer survival rates over time.



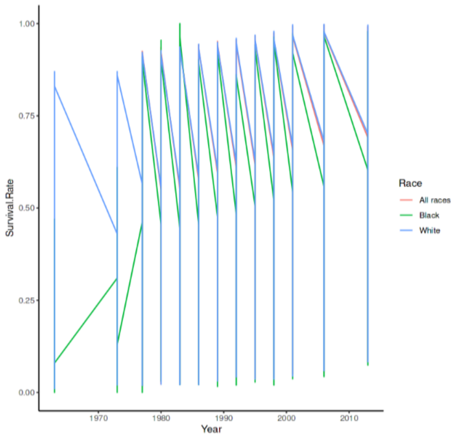


Figure 5: Survival Rate by Genders

Figure 6: Survival Rate by Races

As shown in Figure 5, the overall survival rate for all categories has been continuously increasing over time. Prior to 1990, there was a disparity in survival rates between men and women, with women having a higher survival rate than men. However, after 1990, the survival rates for both genders became nearly identical, yet both continued to steadily improve. Further analysis reveals that the increase in cancer survival rates for men during this period was greater than that for women. Although the growth for women was smaller, their survival rate remained higher than men's for most of the time period.

Next, to investigate the differences in the patterns of change over time in cancer survival rates among different races, we classified the data according to the table's racial categories and created line graphs illustrating the changes in cancer survival rates over time.

As shown in Figure 6, the cancer survival rate for all races has been continuously rising over time. However, the survival rate for Black people was notably low in 1963 and only returned to roughly the same level as White people around 1976.

Finally, to analyze the recent changes in survival rates for different types of cancer, line graphs illustrating the changes in survival rates over time were created based on 17 distinct types of cancer.

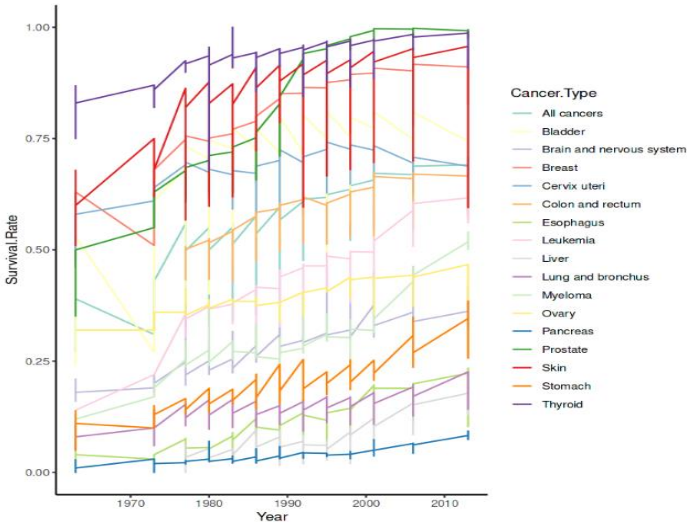


Figure 7: Survival Rate by Cancer Types Over Time

As shown in Figure 7, in addition to the previous analysis of the varying survival rates across different types of cancer, it is also evident that the survival rates for all cancer types have been continuously improving with the advancement of society and medical standards. Notably, the survival rates for leukemia, prostate cancer, multiple myeloma, skin cancer, and esophageal cancer have seen the most significant growth. In contrast, the growth rate of survival rates for cervical cancer, brain cancer, stomach cancer, thyroid cancer, liver cancer, and esophageal cancer has been slower. Moreover, the survival rate for cervical cancer has mostly declined from 1979 to 2013. The changes in cancer survival rates may be influenced by various factors, including advancements in medical technology, early detection and diagnosis, innovations in treatment methods, the scope of health insurance coverage, and population behaviors.

3.3 Multivariate Analysis of Variance (MANOVA)

In descriptive analysis, preliminary differences in survival rates due to year, gender, race, and type of cancer can be initially identified through box plots. However, the intuitive description of the graphs cannot determine whether there is a significant difference in the impact of these four factors on survival rates. To further qualitatively analyze the differences among these four factors, this study employs Multivariate Analysis of Variance (MANOVA), treating survival rate as the dependent variable and the four influencing factors as independent variables for specific calculation and analysis[3].

The principle of multivariate analysis of variance (MANOVA) is to investigate whether a dependent variable is affected by multiple independent variables, also known as factors, and their different levels of combination. It can analyze not only the effects of individual factors but also the interactions between factors. Considering that the results of main effects and interactions differ, to compare which presentation is better given the dataset, this article utilizes the calculation of AICc (corrected Akaike Information Criterion) values for each model to make a judgment. AICc is an information criterion that takes into account both the goodness of fit and complexity of the model, helping to select the most appropriate model[4].

Table 1: The table of AICc results comparing two models

Names	k	ICc	Delta-AICc	ModelLik	AICcWt	LL	Cum.Wt
interaction	83	6553.1	0	1	1	3617.7	1
mul-way	3	4734.8	1818.3	0	0	2390.8	1

As shown in Table 1, the 'mul_way' model has a higher AICc value and a larger ΔAICc compared to the optimal model, indicating that its goodness of fit is relatively poor. In summary, based on the AICc values and AICc weights, the 'interaction' model may better explain the observed data in this dataset, making it the optimal model given the data.

We can further analyze the interactions between different factors and create a table.

Table 2: The interaction effects results table

Type	Df	Sum Sq	Mean Sq	F value	P
Cancer.Type	16	134.25	8.391	9159.542	<2e-16
Gender	2	0.07	0.033	36.293	4.34e-16
Race	2	1.18	0.588	641.381	<2e-16
Year	1	5.11	5.113	5581.862	<2e-16
Cancer.Type:Gender	28	0.33	0.012	12.972	<2e-16
Cancer.Type:Race	32	1.85	0.058	63.094	<2e-16
Gender:Race	4	0.00	0.000	0.169	0.954
Cancer.Type:Year	16	1.67	0.105	114.154	<2e-16
Gender:Year	2	0.02	0.008	8.648	0.000
Race:Year	2	0.01	0.006	6.796	0.001
Cancer.Type:Gender:Race	56	0.05	0.001	0.936	0.609
Cancer.Type:Gender:Year	28	0.07	0.002	2.701	4.54e-06
Cancer.Type:Race:Year	32	0.23	0.007	7.830	<2e-16
Gender:Race:Year	4	0.00	0.000	0.120	0.975
Cancer.Type:Gender:Race:Year	56	0.03	0.001	0.602	0.991
Residuals	1384	1.27	0.001		

As shown in Table 2, by comparing the P-values of each combination with 0.05, it can be observed that the factors such as Year, Race, Gender, and Cancer. Type have a very significant impact on survival rates, although the significance of Gender is somewhat lower than that of the other factors. In addition to this, the interactions between Cancer. Type and Gender, Cancer. Type and Race, Cancer. Type and Year, Gender and Year, Cancer. Type with Gender and Year, and Cancer. Type with Race and Year are very significant. The interaction between Race and Year is also relatively significant, whereas there are no significant interactions between Gender and Race, Cancer. Type with Gender and Race, or among these four factors.

After quantitatively analyzing the impact of the four factors on survival rates, we aim to further investigate whether there are significant differences in the impact of different categories of single factors on survival rates. Therefore, this paper uses Tukey's test to explore the differences between each pair of factor levels and calculates the adjusted P-values based on these differences to determine which differences are statistically significant [5].

Firstly, this paper seeks to explore whether there is a difference in survival rates between different genders, summarizing the results in a table.

Table 3: The Tukey test table for gender.

Type	diff	lwr	upr	P
male-females	-0.015	-0.026	-0.004	0.003
total-females	-0.010	-0.021	-0.000	0.045
total-males	0.005	-0.006	0.002	0.572

As shown in Table 3, the P-value for the comparison between 'males' and 'females' is 0.0035, and the P-value for 'total' versus 'females' is 0.0454, while the P-value for 'total' versus 'males' is 0.5716. This suggests that the differences between 'females' and 'males', as well as 'females' and 'total', are statistically significant (with a significance level of 0.05), whereas the difference between 'total' and 'males' is not significant.

Next, we investigate whether there are any significant differences in survival rates among different races, summarizing the results in a table.

Table 4: The Tukey test table for races

Type	diff	lwr	upr	P
Black-All races	-0.06	-0.073	-0.052	0.000
White-All races	-0.014	-0.025	-0.004	0.004
White-Black	0.048	0.038	0.058	0.000

As shown in Table 4, it can be observed that there are very significant differences among the three categories, with the P-values for 'Black-All races' and 'White-Black' being 0, indicating an extremely significant difference.

4. Conclusion

After exploratory statistical analysis and multifactor analysis of variance, this study found that except for the combinations of Gender with Race, Cancer. Type with Gender and Race, and Gender with Race and Year, as well as the interactions among these four factors, all other single factors or factor combinations show significant differences, meaning they indeed have an impact on cancer survival rates. Furthermore, it uncovered disparities in specific cancer types across different ethnicities and time periods.

Medical professionals can infer survival rates based on these characteristic factors of patients to provide personalized treatment plans. These findings actively promote the prevention of cancer, the prediction of patient survival time, and the formulation of clinical treatment plans, carrying significant implications.

References

[1] Wang Jinsong, Wei Jiayan, Peng Min. Interpretation and enlightenment of 2023 American cancer statistics report and the latest global cancer statistics[J]. Journal of Practical Oncology, 2023, 38(6): 523-527.

[2] Liu, James Y., et al. "High-throughput screening of respiratory hazards: Exploring lung surfactant inhibition with 20 benchmark chemicals [J]." Toxicology 504 (2024): 153785.

[3] Zhang, Leilei, et al. "Exogenous curcumin mitigates as stress in spinach plants: a biochemical and metabolomics investigation [J]." Plant Physiology and Biochemistry (2024): 108713.

[4] Zuo, Yutong. "Research on Head Shape Optimization of High-speed Trains Based on Multifactor ANOVA and Streamlined Head Optimization CST Methods [J]." International Journal of Frontiers in Engineering Technology 6.1 (2024).

[5] Akinbobola, Olawale, et al. "Outcomes of resected lung cancer diagnosed through screening and incidental pulmonary nodule programs in a Mississippi Delta cohort [J]." JTO Clinical and Research Reports (2024): 100684.