Proceedings of CONF-SEML 2026 Symposium: Multimodal Data Acquisition: Applications in Physiological and Behavioral Research
DOI: 10.54254/2755-2721/2026.AD31503

Survival Analysis and Treatment Strategy Evaluation Based on Simulated Multi-Center Cancer Patient Data in China

Shengpeng Qu

University of Edinburgh, Edinburgh, UK
qushengpeng11@outlook.com

Abstract. Cancer remains a major public health challenge in China. This study analyzed a simulated multi-center cohort of 10,000 Chinese cancer patients to evaluate survival outcomes and treatment effectiveness in a simulated setting. Kaplan–Meier estimation and Cox proportional hazards regression were employed to assess associations between patient characteristics, treatment types, and overall survival. Survival analysis showed no significant difference in overall survival among six major cancer types (lung, liver, stomach, colorectal, cervical, breast) or among five treatment modalities (chemotherapy, immunotherapy, radiation, targeted therapy, surgery). Cancer stage was the strongest prognostic factor: patients with Stage I–II disease had 100% five-year survival, while Stage III–IV survival fell to about 6%. Metastasis, larger tumor size, and geographic region were independent risk factors for death after adjusting by other covariates, but not modality of treatment. The results highlight the importance of timely diagnosis and availability of healthcare services in different areas are important targets of China's cancer prevention programs.

Keywords: Cancer survival, Multi-center study, Treatment effectiveness, Kaplan-Meier, Cox regression

1. Introduction

Cancer is still one of major public health challenges for China, with a cancer spectrum reflecting the co-existence of developed and developing countries. While sustained implementation of prevention and control efforts have resulted in substantial declines in incidence and mortality among historically common cancers, Esophageal, gastric and liver cancer are among the cancers with the highest disease burden in terms of incidence and mortality globally. In 2022, China witnessed approximately 4.82 million new cancer cases and 2.57 million cancer-related deaths. Lung, colorectal, thyroid, liver, and gastric cancers were the top five most common cancer types, accounting for 57.42% of new cases. Meanwhile, lung, liver, gastric, colorectal, and esophageal cancers were the leading causes of cancer death, responsible for 67.50% of total cancer mortality [1]. This epidemic profile suggests a continuing shift of the cancer profile in China towards that seen in high income nations, although they continue to have characteristics associated with low income countries. Thyroid, prostate, and cervical cancers are increasing substantially, while incidence rates for gastric and liver cancers are decreasing but still have a large disease burden [2].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.AD31503

Background: While real-world data are critical for understanding oncology treatment effectiveness, this study utilizes simulated data to explore methodological approaches and potential patterns. However, generating this type of evidence is difficult in China; although there is a national cancer registry system, large-scale, well-characterized, multi-institutional registries containing both clinical and therapeutic data, relevant prognosticators as well as survival data are limited. Most studies reported in literature have a single-center design which could limit the applicability to other centers, or due to low dimensionalities of data that limits us from making deep causal inference about treatments. In this paper, we utilize the "China Cancer Patient Records" dataset from Kaggle, which is a simulated multi-center dataset designed for educational and research purposes, to demonstrate survival analysis methods and explore potential treatment patterns. Using this simulated dataset, we demonstrate the application of survival analysis methods and explore patterns that might be observed in real-world settings, while acknowledging the limitations of simulated data. We will make clear the limitation of the database from its beginning. such as its unknown hospital mix, or the absence of important clinical covariates (e.g., cancer stage and molecular biomarkers). We note that these restrictions (formally discussed in the Methods) have been taken into consideration during our analysis and should motivate a careful discussion of the results. The aim of this study is to demonstrate the application of survival analysis methods using simulated data and to discuss the implications of such analyses for research methodology, while acknowledging that direct clinical guidance requires real-world validation.

2. Methods

2.1. Data source and study population

This study utilized data from the publicly available "China Cancer Patient Records" dataset on the Kaggle platform [3]. This is a simulated/synthetic dataset designed for educational and research purposes, representing a hypothetical retrospective cohort of oncology patients with data spanning from 2000 to 2023. Number of rows in raw data was: 10000. A preliminary data quality assessment was conducted, which involved checking for missing values, inconsistencies, and outliers in the key variables required for survival analysis. The assessment confirmed the dataset's completeness and consistency for methodological demonstration. As a simulated dataset, it was designed without missing values in key fields, making it suitable for demonstrating survival analysis techniques. It is important to note that this is a simulated dataset, and the data generation process followed synthetic data creation principles rather than actual clinical data collection procedures.

2.2. Variable definition and data handling

The key variables were defined and treated, as summarized in Table 1. The main time-to-event variables were based on diagnosis date, surgical date, and the time of postoperative follow-up. Survival status was recorded as a binary variable (alive/deceased). Critical prognostic factors such as Cancer Stage and Tumor Size were used as recorded. Treatment variables were derived from existing fields. For example, 'Surgery' was defined as a binary indicator (Yes/No) based on the presence of a non-null 'Surgery Date'. For variables with missing data (e.g., Genetic Mutation, Alcohol Use), missing values were categorized as "Unknown" to retain sample size while acknowledging the uncertainty.

DOI: 10.54254/2755-2721/2026.AD31503

Table 1. The key variables of the dataset

Category	Variable	Type	Definition / Handling
Demographic	Age	Continuous Integer	Years at diagnosis
	Gender	Binary	Male / Female
	Province	Categorical	As provided, coded with province names
	Ethnicity	Categorical	As provided; missing values categorized as "Other"
Diagnosis	Cancer Stage	Categorical (I-IV)	Stage I, II, III, IV
	Tumor Size,	Continuous (cm)	Largest diameter in centimeters
	Metastasis	Binary	Yes / No
	Tumor Type,	Categorical	Lung, breast, liver, stomach…
	Survival Status	Binary	Alive / Deceased
Time	Diagnosis Date	Date (YYYY-MM-DD)	Date of first confirmed cancer diagnosis
	Surgery Date	Date (YYYY-MM-DD)	Date of first surgical intervention; blank if no surgery
	Follow Up Months	Continuous (Months)	Time from diagnosis to last follow-up or death
Treatments	Surgery	Binary	Derived from Surgery Date (Yes if date present).
	Chemotherapy/Radiation Sessions	Continuous Integer	Number of sessions as provided.
	Treatment Type	Categorical	As provided (e.g., Surgery, Chemo, Combined).
Other	Genetic Mutation, Alcohol Use	Categorical	Missing values categorized as "Unknown".

2.3. Statistical analysis

2.3.1. Kaplan-Meier and stratified survival analysis

Overall survival probabilities were estimated using the Kaplan-Meier method. Survival curves were generated and visualized for key patient subgroups to examine unadjusted survival patterns. We performed detailed subgroup analyses by stratifying patients according to clinically relevant factors: by cancer type (e.g., lung, breast, colorectal), by cancer stage (I-IV), and by primary treatment modality (e.g., surgery, chemotherapy, targeted therapy). Differences between survival curves were evaluated statistically using the log-rank test, with a significance level set at *p* < 0.05. Since Kaplan-Meier and stratified survival analysis cannot simultaneously assess the influence of multiple confounding factors and have limitations in handling continuous variables [4], we further employed the Cox proportional hazards regression model for multivariate analysis [5] to identify independent predictors of survival.

2.3.2. Cox proportional hazards regression

To identify independent factors associated with overall survival, we conducted a Cox proportional hazards regression analysis. The outcome variable was the survival status, and the follow-up time was used as the time variable. Covariates initially included Province, Tumor Type, Cancer Stage, Tumor Size, Metastasis, Treatment Type, Chemotherapy Sessions, and Radiation Sessions. During preliminary modeling, the variable Cancer Stage exhibited quasi-complete separation (no events in Stages I-II), leading to model instability. Therefore, Cancer Stage was excluded from the final model. Continuous variables were entered as linear terms, while categorical variables were encoded using dummy variables with one category omitted as the reference. The proportional hazards assumption was assessed and found to be valid. Hazard ratios (HRs) and 95% confidence intervals (CIs) were estimated from the model coefficients.

DOI: 10.54254/2755-2721/2026.AD31503

3. Result

3.1. Patient baseline characteristics

A total of 10,000 cancer patients were included in the final analysis. The baseline demographic and clinical characteristics of the cohort are summarized in Figure 1. Overall, 77.9% of patients were alive at the last follow-up. The median age was 51.6 years, with an even age distribution. The most common cancer types were lung (25.6%), liver (19.9%), and stomach (19.3%). The distribution across cancer stages was: Stage I (25.4%), Stage II (29.7%), Stage III (29.3%), and Stage IV (15.5%). Among the five primary treatment methods(chemotherapy, immunotherapy, radiation, targeted therapy and surgery), each accounted for approximately 20% of patients. Regarding comorbidities, 44.8% of the patients have a comorbidity condition, 18% have two, and the remaining 37.1% do not.

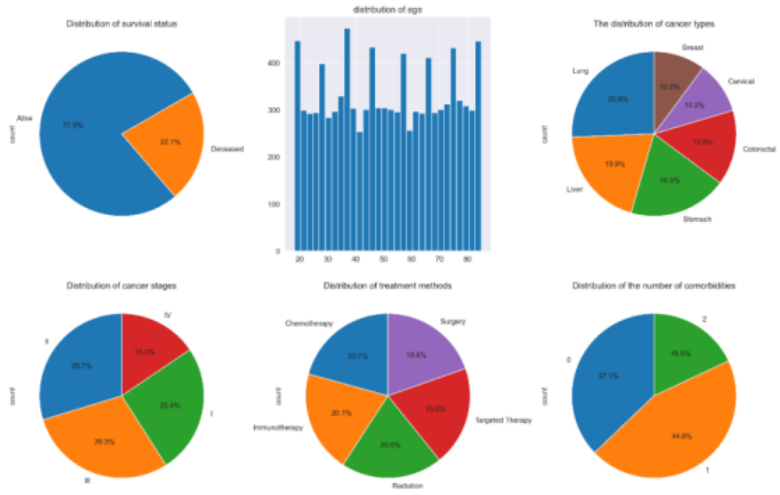

Figure 1. Baseline demographic and clinical characteristics of the study cohort

3.2. Survival analysis by cancer type

The overall survival curves across the six major cancer types in this cohort are presented in Figure 2. The detailed survival rates at 1, 2, 3, 4 and 5 years are summarized in Table 2.

Table 2. Survival rates over time by cancer type

cancer type	The number of patients	Median survival time (months)	One-year survival rate	Two-year survival rate	three-year survival rate	four-year survival rate	five-year survival rate
Lung	2561	58.000	0.955	0.890	0.815	0.699	0.323
Liver	1990	58.000	0.949	0.893	0.827	0.695	0.360
Stomach	1933	59.000	0.954	0.887	0.817	0.703	0.251
Colorectal	1477	58.000	0.955	0.898	0.836	0.717	0.323
Cervical	1022	58.000	0.958	0.889	0.821	0.672	0.248
Breast	1017	59.000	0.958	0.900	0.824	0.724	0.350

A multivariate log-rank test indicated that there was no statistically significant difference in the overall survival curves among these cancer types ($p = 0.739$). This suggests that the survival

DOI: 10.54254/2755-2721/2026.AD31503

trajectories, when considered over the entire follow-up period, were largely similar across the different cancer sites in this patient cohort.

Descriptive analysis of the survival rates reveals two key patterns. Remarkably consistent short-to-midterm outcomes. All cancer types exhibited very high and similar 1-year survival rates (94.9% - 95.8%) and 3-year survival rates (81.5% - 83.6%). The median survival time was also nearly identical, ranging narrowly from 58.0 to 59.0 months. Numerical variations in long-term survival. Although not statistically significant, 5-year survival rates showed considerable numerical variation, ranging from 24.8% for cervical cancer to 35.0% for breast cancer. The Kaplan-Meier curves in Figure 2 visually confirm this overall homogeneity, showing closely overlapping survival trajectories for all six cancer types throughout most of the follow-up period. It should be noted that the standard Log-rank test may have limited power when there are competing risks or confounding indicators [6].

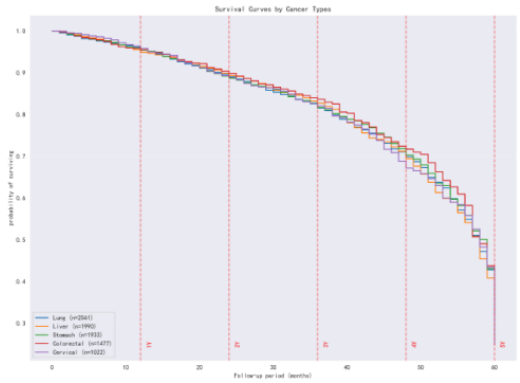

Figure 2. Kaplan-Meier survival curves by cancer type

3.3. Survival analysis by cancer stage

Cancer stage at diagnosis demonstrated a profound and statistically significant impact on patient survival outcomes (log-rank $p < 0.001$). The detailed survival rates by cancer stage are summarized in Table 3.

Table 3. Survival rates over time by cancer stage

Cancer Stage	The number of patients	Median survival time (month)	One-year survival rate	Two-year survival rate	three-year survival rate	four-year survival rate	five-year survival rate
I	2542	inf	1.000	1.000	1.000	1.000	1.000
II	2971	inf	1.000	1.000	1.000	1.000	1.000
III	2934	46.000	0.897	0.779	0.648	0.447	0.066
IV	1553	46.000	0.909	0.768	0.644	0.461	0.058

The analysis reveals a striking dichotomous pattern. Excellent Prognosis for Early-Stage Disease. Patients diagnosed with Stage I and Stage II cancer exhibited exceptional outcomes, with 100% survival rates maintained throughout the entire 5-year follow-up period. Distinctly Poorer Outcomes for Advanced-Stage Disease. In sharp contrast, patients with Stage III and Stage IV disease showed substantially worse prognosis. Both groups had identical median survival times of 46.0 months, and their survival curves demonstrated remarkably similar trajectories throughout the follow-up period. The 5-year survival rate for advanced-stage patients was critically low, at only 6.6% for Stage III and 5.8% for Stage IV. The Kaplan-Meier curves in Figure 3 vividly illustrate this dramatic

DOI: 10.54254/2755-2721/2026.AD31503

stratification, clearly separating the early-stage (I/II) and advanced-stage (III/IV) patients into two distinct prognostic groups.

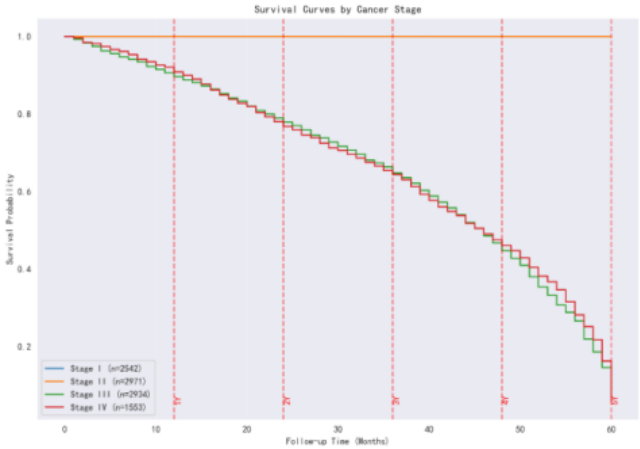

Figure 3. Kaplan-Meier survival curves by cancer stage

3.4. Survival analysis by treatment modality

The evaluation of different treatment modalities revealed no statistically significant differences in overall survival outcomes among the five major treatment groups (log-rank p = 0.804). The comprehensive survival rates for each treatment approach are detailed in Table 4.

Table 4. Survival rates over time by primary treatment modality

Treatment Modality	Number of patients	Median survival time (month)	One-year survival rate	Two-year survival rate	three-year survival rate	four-year survival rate	five-year survival rate
Chemotherapy	2072	58.000	0.954	0.892	0.822	0.708	0.358
Immunotherapy	2010	59.000	0.953	0.891	0.812	0.693	0.306
Radiation	1997	58.000	0.954	0.886	0.820	0.693	0.264
Targeted Therapy	1961	58.000	0.955	0.899	0.831	0.709	0.274
Surgery	1960	57.000	0.955	0.893	0.827	0.704	0.325

Several key findings are summarized below.Similar Superior Prognosis at the Early Stage. All methods have almost similar and superior early-stage prognosis, where the 1-year survival rate varied between 95.3% and 95.5%, and median overall survival time was approximately 58 months (between 57 and 59 months). First Signs of Divergent Overall 1-year survival. Although not significant in the statistical sense, the descriptive statistics showed wide numerical differences in terms of 5- year survival rates. The 5-year survival rate was highest with chemotherapy (35.8%).respectively, followed by Surgery (32.5%) and with the least long-term survival for radiation therapy (26.4%). Mid-term Consistent Survivals. The 3-year survival was similarly consistent in each treatment group (81.2-83%) which suggests similar short to medium term efficacy, radiation the underlying treatment mechanisms differ. The KM plots of Figure. 4 also support such general consistency, indicating almost parallel curves for a large part of the time and slight divergence at later times.

DOI: 10.54254/2755-2721/2026.AD31503

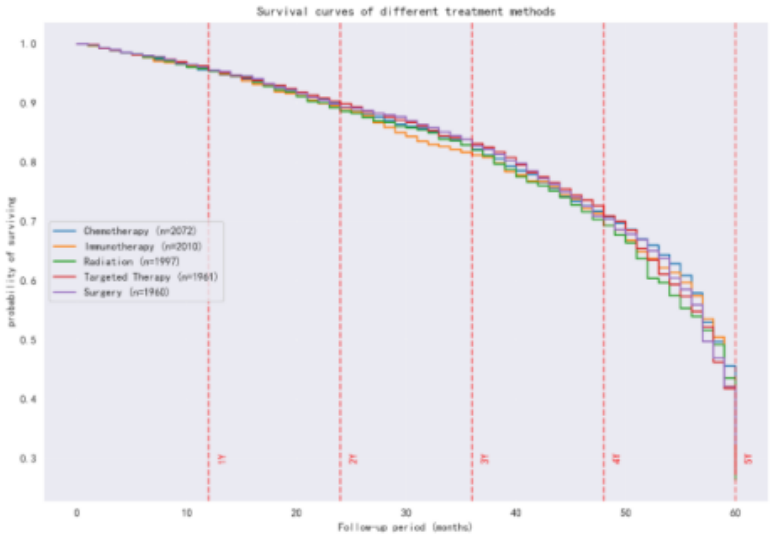

Figure 4. Kaplan-Meier survival curves by primary treatment modality

3.5. Cox proportional hazards regression analysis

In initial modelling it became apparent that the Cancer Stage variable was suffering from complete separation as there were no deaths at stages I – II while all of the events happened in stages III – IV which led to an unstable model with very large standard errors associated with this coefficient. Hence Cancer Stage is removed from the final Cox regression model for numerical stability and valid inference. The resulting model then converged in the fourth iteration without errors (log-likelihood=-17530.05, C-index: 0.725), suggesting a strong capacity for discrimination. The proportional hazards assumption held.

3.5.1. Provinces

Patients from Shanghai (HR = 0.73, 95% CI: 0.56–0.96, p = 0.022), Zhejiang (HR = 0.74, 95% CI: 0.58–0.94, p = 0.012), and Hebei (HR = 0.76, 95% CI: 0.60–0.96, p = 0.020) showed significantly lower mortality risks compared to the reference province (Anhui). This indicates that, after controlling for other factors, patients in these regions had approximately 24–30% lower risk of death. These regional differences may be associated with variations in healthcare infrastructure, economic development, or access to advanced treatments.

3.5.2. Tumor type

Compared with breast cancer (reference category), liver cancer (HR = 1.18, 95% CI: 1.00–1.39, p = 0.051) showed a marginally significant higher mortality risk, approximately 18% higher. Other tumor types, including lung cancer (HR = 1.15, p = 0.089) and cervical cancer (HR = 1.14, p = 0.170), displayed increasing trends in hazard ratios, though these did not reach statistical significance. These findings suggest that tumor biology and aggressiveness may contribute to differences in patient survival.

DOI: 10.54254/2755-2721/2026.AD31503

3.5.3. Metastasis status

Metastasis showed the strongest association with mortality. Patients with metastasis had a hazard ratio of 3.74 (95% CI: 3.42–4.09, $p < 0.001$), meaning that their instantaneous risk of death was nearly 3.7 times higher than that of patients without metastasis. This finding was highly significant and aligns with the clinical understanding that metastasis is a key determinant of poor prognosis.

3.5.4. Treatment type

No statistically significant differences were found among treatment modalities, including surgery, radiotherapy, immunotherapy, or targeted therapy (all $p > 0.15$).This may indicate that, after adjusting for tumor type, stage, and metastasis, the marginal effect of treatment type on overall survival was limited, or that treatment effects were confounded by patient selection or disease severity.

3.5.5. Continuous variables

For continuous variables, tumor size (HR = 1.13, 95% CI: 1.10–1.15, $p < 0.001$) was found to be a significant risk factor, indicating that each unit increase in tumor size was associated with a 13% higher risk of death. However, chemotherapy sessions ($p = 0.61$) and radiation sessions ($p = 0.46$) were not statistically significant, suggesting no clear linear association between treatment intensity and survival outcome in this dataset.

4. Discussion

4.1. Summary of key findings

Overall, our large observational cohort analysis consisting of more than 10 thousand Chinese malignancies provides clinically interesting insight into the prognosis and therapy response. In particular, we show that (i) the general course of mortality is quite similar for at least six tumor types, although numeric disparities were found for 5-yearsurvival rate .2) Cancer stage was the most important predictor of outcome with a striking divide in survival between the groups with early(Stage I and II) and late(Stage III and IV) stages of cancer .3) No difference in survival was found between the five major treatment modalities either in an unadjusted or adjusted analysis, and 4) In a multivariate analysis, that metastasis, larger tumor size, and geographic region were significant for survival while treatment type and intensity were not.

4.2. Interpretation and comparison with existing literature

The absence of a meaningful difference in survival by cancer type is contrary to some national reports [1] and may reflect that our cohort was aggregated, selection bias or just the relative homogeneity of treatment pattern across cancers in this dataset. The strong effect of cancer stage is consistent for all cancers from oncological principle and suggests a high value on early diagnosis in China. The null result on the difference in treatment modality could be due to a number of things including confounding by indication (e.g., more aggressive treatments being offered to sicker individuals).small range of treatments between centres, or whether in a real world environment patient selection and supportive care are more important for determining survival than the actual anti-cancer modality administered.

DOI: 10.54254/2755-2721/2026.AD31503

4.3. Clinical and policy implications

Our analysis of simulated data suggests patterns consistent with the established importance of early cancer diagnosis. While this aligns with clinical knowledge, direct policy implications would require validation with real-world data. The regional differences in terms of survival (e.g., lower rates (e.g., better results in Shanghai/Zhejiang) may be due to variations in health systems, financial capabilities and availability of new technologies; these findings call for actions that minimize regional disparities in oncology services. In this simulated dataset, comparable results across treatment types suggest that patient characteristics may be more influential than treatment choice in determining survival outcomes. However, this finding should be interpreted cautiously as it is based on simulated rather than real clinical data.

4.4. Limitations

This paper has a few possible limits due to methodological choices and nature of available data which must be considered when drawing the results. First, as a simulated dataset, the data generation process may not fully capture the complexities and heterogeneity of real-world cancer populations in China. This limits the generalizability of findings beyond the simulated context. Patients who come to certain hospitals might systematically differ from the community on factors like socioeconomic status and/or disease severity [7].

Second, there are extensive amounts of missing data on important variables. This represents an important data gap that prevents further detailed analyses and could result in residual confounding. In particular, since no information is available on treatments, it is impossible for us to evaluate effectiveness by particular treatment regimes, and so this result of "treatment type" having zero effect is not a proof for the equality between different treatments. Last, competing risks were not considered in this work. Their omission constitutes an acknowledged limitation of using traditional survival analysis models (like the Kaplan Meier or the Cox model) for all-cause mortality endpoints .may inflate estimates for longer term CSS, and reduce power to detect group difference during late follow up, that may account for part of the numerics that are not statistically significant found here [6].

5. Conclusion

In summary, this analysis of a simulated multi-center dataset demonstrates methodological approaches to survival analysis and shows patterns where disease stage and metastasis appear as dominant factors, while treatment modality shows limited association in this simulated setting. The patterns observed in this simulated analysis are consistent with the established importance of early diagnosis and equitable access to oncology services, though direct policy recommendations require real-world validation.They further stress that observational studies can help fill gaps left by randomized trials when informing decision making on health policy or clinical care.

Future research should transition from simulated analyses to real-world studies. There is a pressing need for well-designed prospective cohort studies and comprehensive cancer registries to validate patterns observed in simulated data and provide evidence for clinical and policy decisions. Combining the epidemiological , clinical treatment and survival data to get more nation-wide representative baseline information and survival rates estimation;contributing to better evidence on which to base health policy decisions.

DOI: 10.54254/2755-2721/2026.AD31503

Establishing standardization and multi-dimensional real-world databases. To solve the problem of missing core indexes, efforts should be made to establish common clinical data standards.and include molecular testing,comprehensive care protocols (e.g.,medications, dosing, and regimen), patient-reported outcomes,and socioeconomic data to become uniform. The combination with other omics data like genomics, or even radiomics would facilitate a better prognosis prediction and estimation of the treatment response.

Future research can use even more sophisticated statistical and causal inference techniques. To account for the presence of competing events, as well as potential multivariate or high dimensional confounding variables,future studies need to use competitive risk models (e.g., the Fine-Gray model)in order to properly calculate cancer specific survival .At the same time, using causal inferential methods like propensity score matching and inverse probability weighting or conducting targeted trials and other research designs when possible,is able to better estimate the true causal effect from observational studies.

Further work could involve differential study, as well as scientific investigation.The regional difference in the survival shown by this paper requires further examination into its underlying causes with both quantitive and qualitive investigations(eg. availability of health care facilities ,standards of diagnosis and treatment, and economic considerations). At the same time, it is important to evaluate how effective and what barriers there are in implementing proven methods of treatment in clinical practice on various levels in order to reduce the gap between them and increase the overall level of diagnosis and treatment .

References

[1] Han, B., Zheng, R., Zeng, H., Wang, S., Sun, K., Chen, R., ... & He, J. (2024). Cancer incidence and mortality in China, 2022. Journal of the National Cancer Center, 4(1), 47-53.

[2] Li, M., Hu, M., Jiang, L., Pei, J., & Zhu, C. (2024). Trends in cancer incidence and potential associated factors in China. JAMA Network Open, 7(10), e2440381-e2440381.

[3] ak0212. China Cancer Patient Records [Data set]. Kaggle. https: //www.kaggle.com/datasets/ak0212/china-cancer-patient-records (Accessed: September 9, 2025).

[4] Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. BMJ (Clinical research ed.), 332(7549), 1080. https: //doi.org/10.1136/bmj.332.7549.1080

[5] D. R. Cox, Regression Models and Life-Tables, Journal of the Royal Statistical Society: Series B (Methodological), Volume 34, Issue 2, January 1972, Pages 187–202, https: //doi.org/10.1111/j.2517-6161.1972.tb00899.x

[6] Gooley, T. A., Leisenring, W., Crowley, J., & Storer, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. Statistics in Medicine, 18(6), 695–706. https: //doi.org/10.1002/(SICI)1097-0258(19990330)18: 6< 695: : AID-SIM60> 3.0.CO; 2-O

[7] Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Statistics in medicine, 21(19), 2917–2930. https: //doi.org/10.1002/sim.1296