I. Introduction

Code R : https://github.com/saliou-ds/survival_analysis_project

Problem Statement

Wilms' tumor, a type of kidney cancer primarily affecting children, remains a significant health concern despite advances in treatment. While the survival rates have improved over the years, a substantial number of patients still experience relapse after initial remission. Understanding the factors that influence relapse, and the effectiveness of new treatments is crucial for improving long-term outcomes for these patients.

Current State of Research

Research on Wilms' tumor has led to many advances, particularly through the efforts of the National Wilms Tumor Study Group (NWTSG) and the Children's Oncology Group (COG). These organizations have conducted extensive studies to improve treatment protocols and survival rates. Current research focuses on genetic factors, treatment optimization, and reducing long-term side effects. Despite these efforts, challenges remain in predicting relapse and tailoring treatments to individual patients' needs.

Gap in the Field

One of the critical gaps in current research is the lack of comprehensive survival analysis that accounts for various prognostic factors such as age, stage of the disease, and histology. While existing studies have provided valuable insights, there is a need for more detailed analysis to understand the time until relapse and the effectiveness of new treatments. This gap is particularly evident in the context of two-phase studies, where initial screening is followed by detailed analysis of a subset of patients.

Present Research as a Solution

The present research aims to address this gap by conducting a survival analysis on data from the National Wilms Tumor Study Group. By analyzing the time until tumor relapses and considering factors such as age, stage of the disease, and histology, this study seeks to provide a better understanding of the effectiveness of new treatments. The use of two-phase studies allows for a more efficient and detailed analysis, helping to identify key prognostic factors and improve treatment protocols.

Hypotheses

1. **Hypothesis 1**: Age at diagnosis significantly influences the time until tumor relapses in patients with Wilms' tumor.
2. **Hypothesis 2**: The stage of the disease at diagnosis is a critical factor in predicting relapses.
3. **Hypothesis 3**: Histological classification of the tumor (favorable vs. unfavorable) affects the likelihood and timing of relapse.

II. Methods

Study Population

The study was conducted on a cohort of 4028 patients diagnosed with Wilms' tumor. The patients ages ranged from 0 to 191 months (15 years old). The study population was divided into two groups: 1857 patients in one group and 2171 patients in the other. Additionally, a subcohort of 669 patients was selected for more detailed analysis.

Study Design

This research utilized a two-phase study design. In the first phase, a large cohort of patients was screened to gather initial data. In the second phase, a subset of the cohort was selected for more detailed analysis. This approach allowed for efficient use of resources while obtaining comprehensive information on a subset of the population.

Data Collection

Data were collected on various factors that influence the prognosis of Wilms tumor, including:

- Age: Numeric variable representing the age of the participant.
- Stage of the Disease: Both numeric (num_*stage) and categorical (fac*_stage) representations of the disease stage.
- Histology: Information on tumor histology was collected from both the local institution (fac_*instit) and the central lab (fac*_histol). Histology was categorized as 'Favorable' or 'Unfavorable' based on prognosis.

Variables

- pid: Participant ID, a unique identifier for each individual in the dataset.
- event: Indicator of whether the event of interest (e.g., relapse) occurred.
- time: The time duration until the event or censoring (e.g., time to relapse).
- num_age: Numeric variable representing the age of the participant.
- num_stage: Numeric variable representing the stage of the disease.
- fac_stage: Factor variable representing the stage of the disease, but as a categorical variable.
- fac_study: Factor variable indicating the study group the participant belongs to.
- fac_*in*_subcohort: Factor variable indicating whether the participant is in the subcohort.
- fac_instit: Histology factor variable based on the local institution's categorization.

- fac_histol: Histology factor variable based on the central lab's categorization.

Statistical Analysis

The primary focus of the analysis was on survival data, specifically the time until tumor relapse. The following statistical methods were employed:

- Survival Analysis: Techniques such as Kaplan-Meier estimates and Cox proportional hazards models were used to analyze the time until relapse and identify significant prognostic factors.
- Handling Censoring: Methods were applied to appropriately handle censored data, where the event of interest (relapse) had not occurred for some patients during the study period.

Equipment and Software

The data analysis was conducted using R, a statistical software environment. Key packages used included:

- survival: For creating survival objects and fitting survival models.
- survminer: For visualizing survival curves and model diagnostics.

Procedure

1. Data Preparation: The dataset was imported into R, and variables were formatted appropriately.
2. Survival Object Creation: A survival object was created using the Surv() function.
3. Kaplan-Meier Estimate: Kaplan-Meier survival curves were generated using the survfit() function.
4. Cox Proportional Hazards Model: A Cox model was fitted using the coxph() function to assess the impact of prognostic factors.
5. Model Diagnostics: Proportional hazards assumption and other diagnostics were checked to ensure model validity.
6. Modification has been made to optimize the cox model, transforming it into a weighted cox model.
7. Visualization: Survival curves and other relevant visualizations were plotted to interpret the results.

By following these methods, the study aimed to provide a comprehensive analysis of the factors influencing the prognosis of Wilms tumor and the effectiveness of new treatments. The detailed methodology ensures that the study can be replicated by other researchers in the field.

III. Results

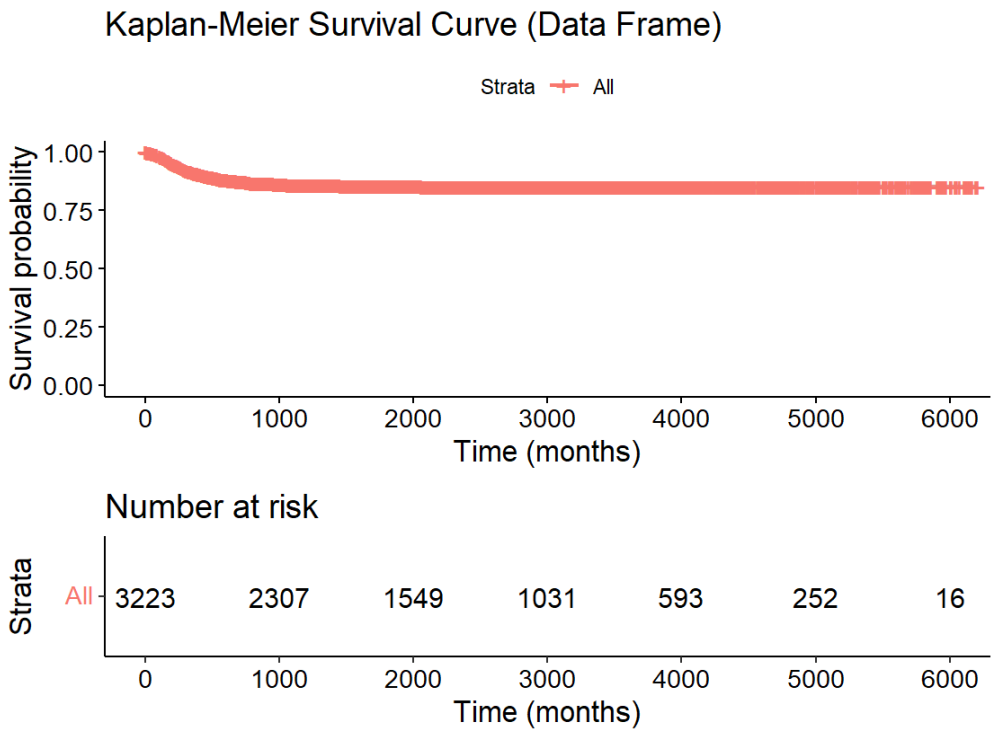

Figure 1. Kaplan-Meier Survival Curve of the Train Data

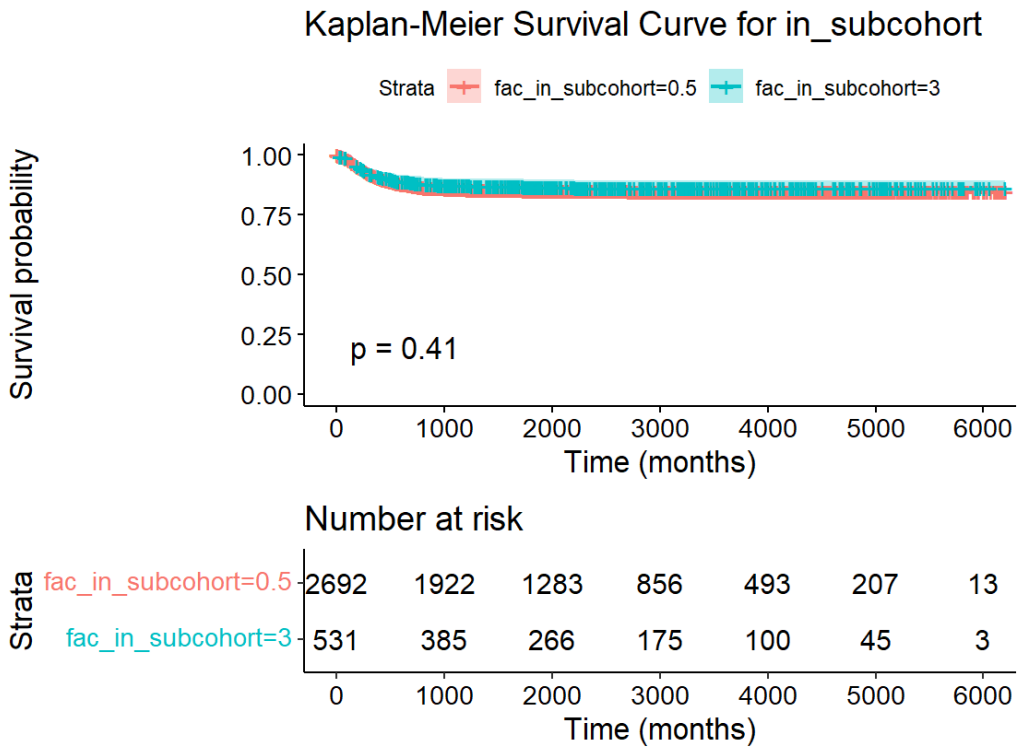

Figure 2. Kaplan-Meier Survival Curve for Subcohort

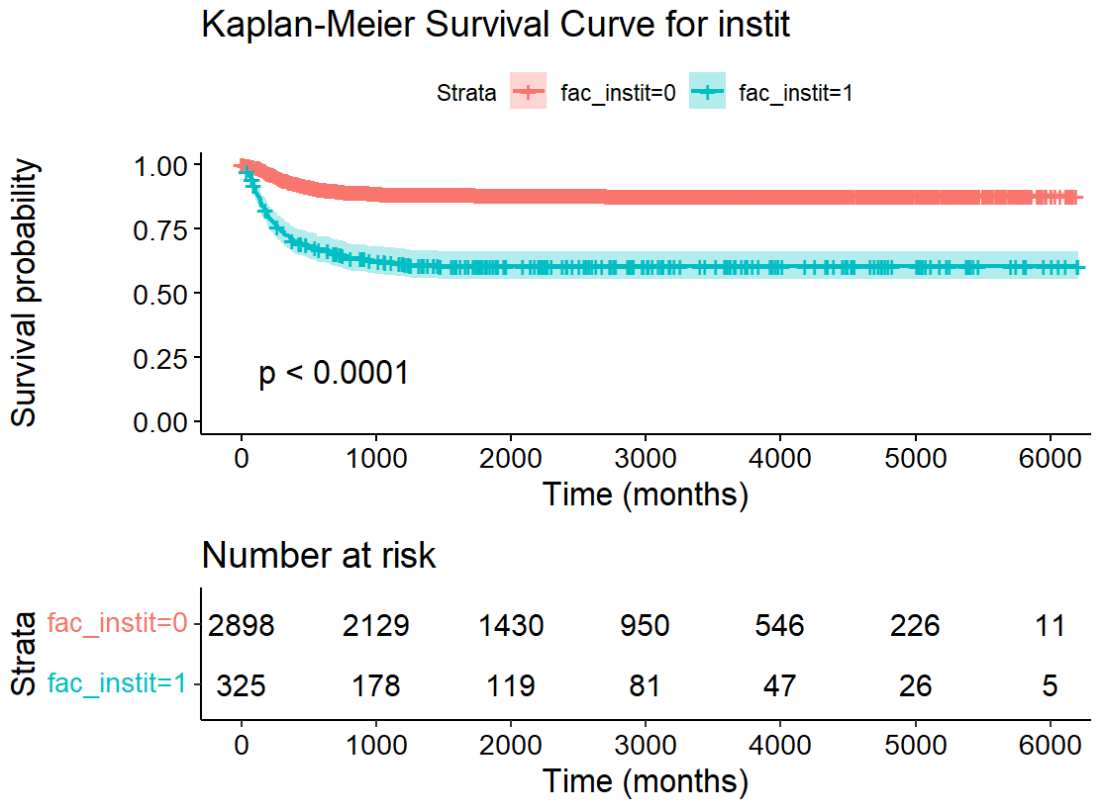

Figure 3. Kaplan-Meier Survival Curve based on Local Institution Histological Prognosis

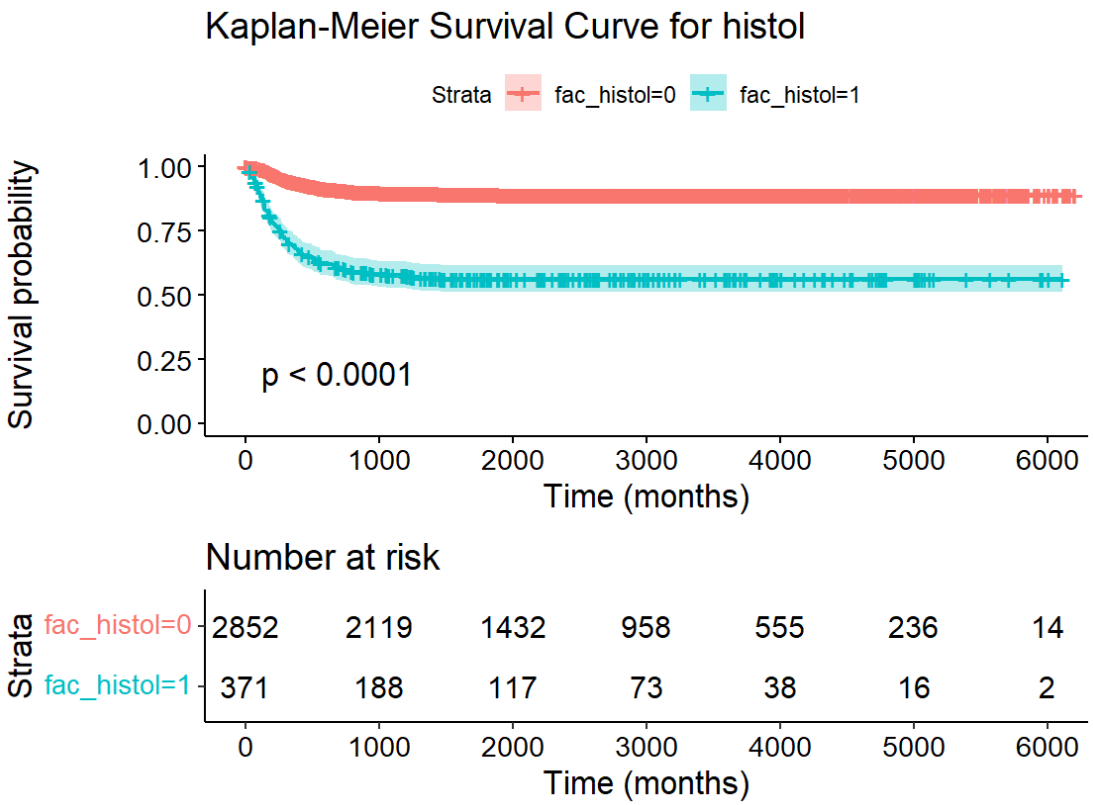

Figure 4. Kaplan-Meier Survival Curve on the Central Lab Histological Prognosis

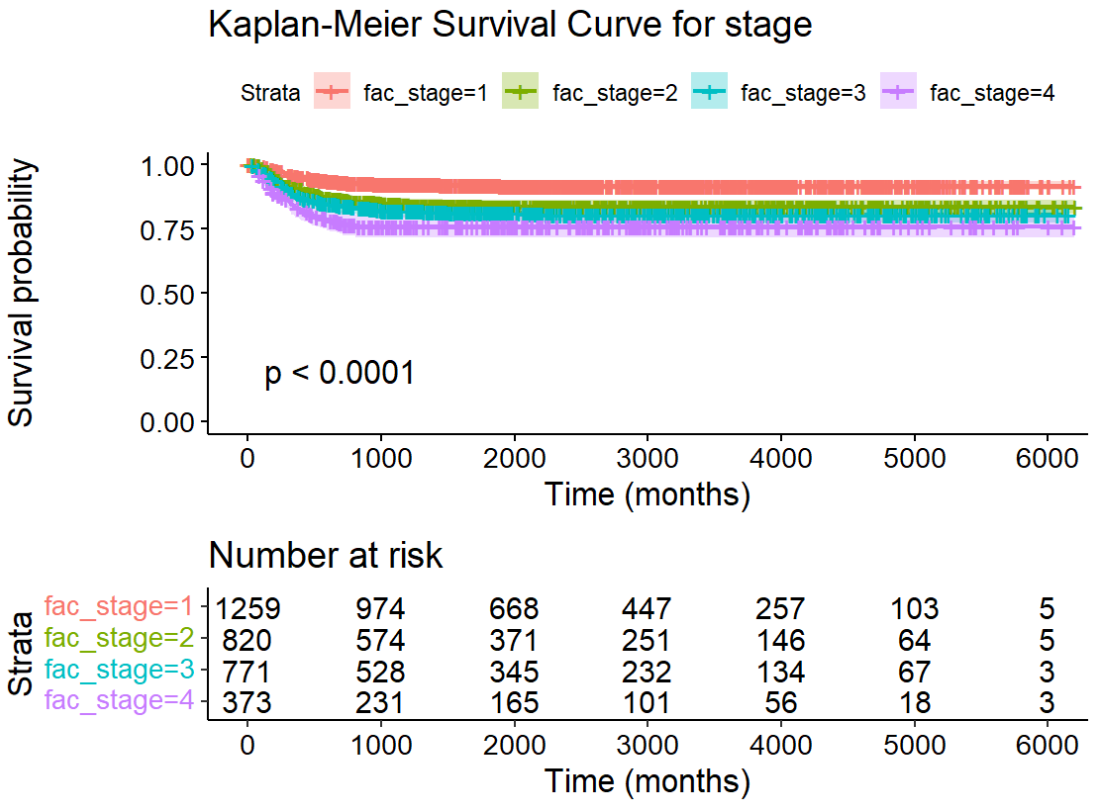

Figure 5. Kaplan-Meier Survival Curve by Tumor Stage (from I to IV)

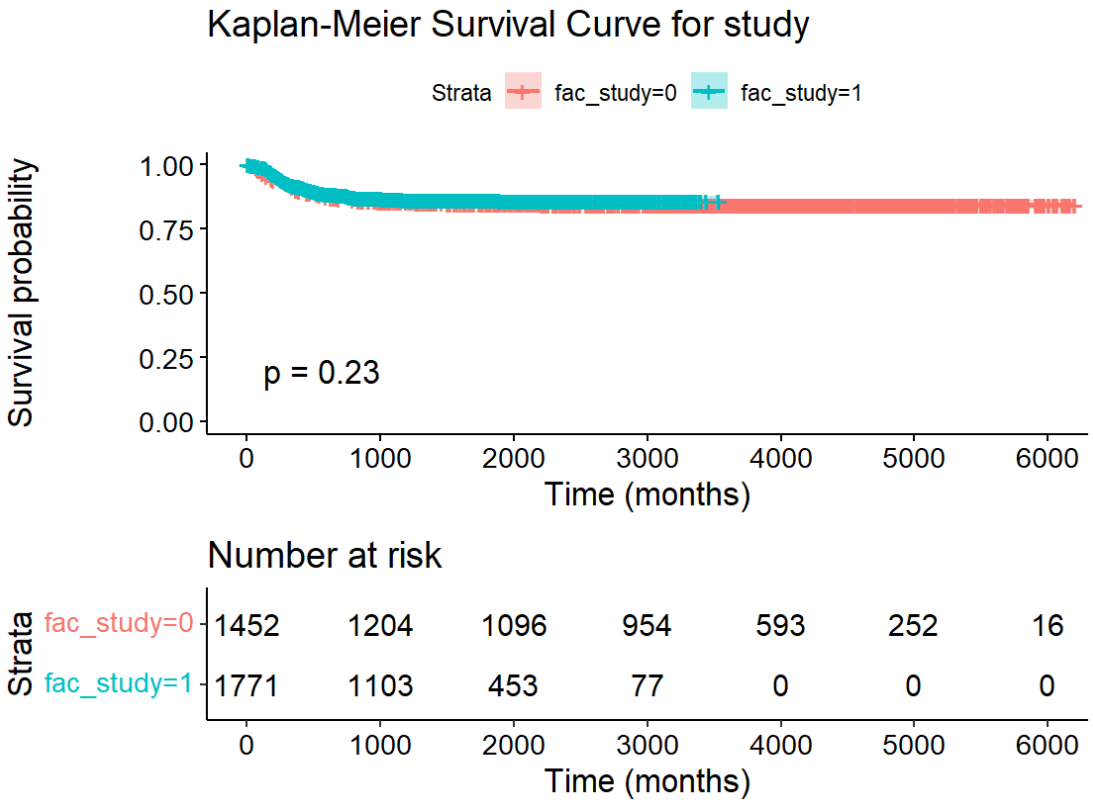

Figure 6. Kaplan-Meier Survival Curve by Study Group

Table 1. Correlation matrix of my significative variables

	num_age	fac_stage	fac_instit	fac_histol
num_age	**1**	**0.267718**	**-0.00436**	**0.018683**
fac_stage	**0.267718**	**1**	**0.142343**	**0.085159**
fac_instit	**-0.00436**	**0.142343**	**1**	**0.715412**
fac_histol	**0.018683**	**0.085159**	**0.715412**	**1**

Table 2. Concordance summary of weighted Cox-model

```
Call:
concordance.formula(object = surv_obj_test ~ test$predicted_risk)

n= 805
Concordance= 0.6235 se= 0.0237
concordant discordant     tied.x     tied.y    tied.xy
     52839      31878        168          5          0
```

IV. Discussion

Main Findings

The Kaplan-Meier survival curve for the training dataset shows that the survival probability decreases over time but stabilizes slightly above 0.75 between 1000 and 2000 months. This trend can be explained by the initial decline in survival probability followed by a period of relative stability.

The Kaplan-Meier survival curve for the in_subcohort variable shows a similar pattern to the overall curve, with no significant difference between the groups (p-value from log-rank test = 0.41). This indicates that subcohort membership does not significantly impact survival probability.

The Kaplan-Meier survival curve for the instit variable reveals a much lower survival probability for individuals deemed "Unfavourable" by this variable. The difference between the groups is significant (p-value from log-rank test = 0.0001). The survival probability stabilizes slightly above 0.5 for the "Favourable" group, while it remains lower for the "Unfavourable" group.

The Kaplan-Meier survival curve for the histol variable shows a similar pattern to the instit variable, with the same p-value and curve shape. This suggests that histological prognosis is closely related to institutional prognosis.

The Kaplan-Meier survival curve for the stage variable shows a clear degradation in survival probability with advancing tumor stage. The log-rank test is significant (p-value = 0.0001), indicating that tumor stage significantly impacts survival. The survival probability stabilizes around 0.9 for Stage I and slightly below 0.75 for Stage IV.

The Kaplan-Meier survival curve for the study variable is not significant (p-value from log-rank test = 0.23). One study group shows no patients at risk around 3500 months, while the initial number of patients is similar across study groups.

Correlation Analysis

The correlation matrix of significant variables shows that fac_histol and fac_instit are correlated. Based on our knowledge of the study, we chose fac_histol for further analysis, as central lab histology is generally more accurate than local institutions.

Model Performance

The concordance index of the final weighted Cox model is 0.6235 when applied to the test dataset. This indicates a moderate level of predictive accuracy. Adding the subcohort as a weight significantly improved the performance of our model, as evidenced by a higher concordance index, while including age as a time-dependent covariate helped explain the dynamic effect of age on survival probabilities.

V. Conclusion

Comparison to Other Research

Our findings align with previous studies that have demonstrated the significance of tumor stage, histological prognosis, and age in predicting survival outcomes. Conversely, the study variable did not show significant impact, indicating that study group differences may be less influential in this context. However, the subcohort created through a second phase of grouping proved to be more informative.

Limitations and Future Research

One limitation of our study is the potential for measurement error in the histological and institutional prognoses. Future research should focus on improving the accuracy of these measurements and exploring additional variables that may impact survival.

These findings overall suggest that tumor stage, age and histological prognosis are significant predictors of survival, while one phase study group differences are not. The moderate concordance index of our model indicates that there is room for improvement in predictive accuracy. Future research should aim to address the limitations identified and explore additional factors that may influence survival outcomes.

References

What's New in Wilms Tumor Research? | American Cancer Society

National Wilms Tumor Study Group - Wikipedia

Current treatment for Wilms tumor: COG and SIOP standards

Norman E Breslow and Nilanjan Chatterjee. "Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis". In: Journal of the Royal Statistical Society: Series C (Applied Statistics) 48.4 (1999), pp. 457–468.