JOURNAL OF
EMERGING INVESTIGATORS

Article

Advancing pediatric cancer predictions through generative artificial intelligence and machine learning

Shiva K. Yadav[1]*, Amrit V. Yadav[1]*, Sandhya Viswanathan[2]

[1] Vista Ridge High School, Cedar Park, Texas

[2] QuantaHealth Research Institute, Austin, Texas

* These authors contributed equally to this work

SUMMARY

Pediatric cancers present unique challenges due to their rarity and the distinct biological factors involved, making early and accurate prediction of survival outcomes critical for guiding treatment. Recent advancements in artificial intelligence (AI) and machine learning (ML) have shown promise in enhancing predictive models for various diseases, including cancer. This study aimed to identify key factors influencing survival rates in pediatric cancer patients through the integration of generative AI and machine learning techniques, including the use of synthetic data. We hypothesized that age at diagnosis was an important predictor of survival outcomes, alongside other significant demographic and clinical variables. Our hypothesis is supported by our analysis, which includes 9184 pediatric cancer patients. Our results indicate that age at diagnosis, specific cancer types, and anatomical sites are significant predictors of survival. Stratification analyses and Kaplan-Meier survival curves consistently show that earlier diagnosis is associated with better survival outcomes, particularly for diagnoses such as neuroblastoma, B lymphoblastic leukemia/lymphoma with hyper diploidy, and osteosarcoma. Comparative analysis and sensitivity tests confirmed age at diagnosis as a critical factor. Classification models enhanced with synthetic data achieved an overall accuracy of 0.74, reflecting the potential of integrating AI-driven approaches with real and synthetic data to improve survival prediction. Broader impacts of this study include its potential to influence pediatric cancer treatment protocols by identifying high-risk groups early, thereby improving personalized treatment strategies. Additionally, this research demonstrates the utility of AI and synthetic data in healthcare, paving the way for more innovative applications across different medical fields.

INTRODUCTION

Pediatric cancer remains one of the most significant challenges in modern medicine, continuing to be a leading cause of disease-related death among children worldwide (1). Despite advancements in treatment, pediatric cancer presents a complex and heterogeneous disease landscape, necessitating the development of effective prognostic tools and personalized treatment strategies (2,3).

In pediatric oncology, early diagnosis is particularly crucial given the rapid progression often seen in pediatric cancers (4). Younger age at the time of diagnosis often allows for treatment at a stage when the disease is more localized and potentially more responsive to therapeutic interventions, leading to better survival rates across various types of cancer (5, 6). Studies have also shown that cancers diagnosed at younger ages often require less aggressive treatment, reducing the risk of severe side effects and improving the overall quality of life for patients (7). However, while the benefits of early diagnosis are well established, other factors may also play significant roles in survival outcomes.

Variables such as the type of cancer, its anatomical site, and the demographics of the patient population (including age, race, and sex at birth) must be considered in the development of accurate prognostic models. Certain cancer types may be inherently more aggressive, leading to poorer outcomes even when diagnosed early (8). Similarly, the anatomical site of the cancer can affect the ease of detection and the complexity of treatment, further complicating prognostic predictions. Moreover, demographic factors such as race and sex at birth have been shown to influence cancer outcomes, but their integration into predictive models is often inadequate (9).

Accurately predicting survival outcomes in pediatric oncology is particularly challenging due to the variability in disease presentation and response to treatment (10). Performance measures, such as those systematically analyzed by Sokolova and Lapalme, are critical when developing machine learning models for prognostic predictions (10). The importance of classification accuracy in these models is emphasized in the literature, as accurate predictions can significantly impact treatment decisions and outcomes. Although previous research has emphasized the significance of demographic and clinical factors such as age at diagnosis, race, ethnicity, sex at birth, diagnosis, and anatomical site in affecting survival rates, these variables have not been thoroughly incorporated into predictive models utilizing advanced machine learning techniques like Cox Proportional Hazards models, and Kaplan-Meier estimators. Additionally, the integration of generative AI, such as the CTGAN Synthesizer, has further enhanced the ability to model these factors more comprehensively (11).

Our study aims to fill this gap by applying generative AI and machine learning models to identify key predictors of survival outcomes. In this study, we utilized supervised learning models, including Random Forest, enhanced with generative AI techniques to predict survival outcomes in pediatric cancer patients.

This study leverages generative artificial intelligence

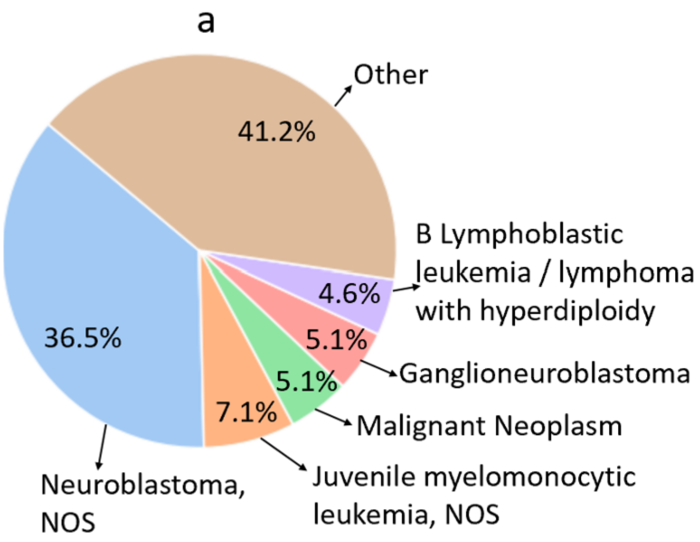

Figure 1. Cancer diagnosis distribution and age at diagnosis among pediatric cancer patients. a) Pie chart displaying the distribution of cancer diagnoses among pediatric patients, highlighting the top five most common diagnoses, including Neuroblastoma, NOS, and Malignant Neoplasm. The "Other" category consists of less frequent diagnoses. b) Histogram showing the age distribution of pediatric cancer patients at the time of diagnosis. The analysis was based on data from a pediatric patient cohort (n=9184), and the data were analyzed using Cox Proportional Hazards models.

(AI) and machine learning techniques, including the use of synthetic data, which means artificially created data that can simulate the patterns of real data in the world. Synthetic data is added to the existing dataset in an effort to overcome problems such as limited sample sizes and enhance the predictability of outcomes in pediatric cancer patients, ultimately aiming to enhance patient care and survival rates (11).

Advancements in pediatric oncology have improved survival rates, yet accurately predicting individual outcomes remains a significant challenge due to the complexity and variability of cancer progression. Traditional statistical methods, while useful, often struggle to account for the vast range of factors influencing prognosis. Recent developments in artificial intelligence and machine learning offer more sophisticated tools for handling such complexity. These technologies are capable of analyzing large, multidimensional datasets, uncovering patterns that might be missed by traditional methods, and generating more precise prognostic models. In this study, we apply advanced ML techniques to a dataset of 4592 pediatric cancer patients, expanded to 9184 participants through synthetic data generation, integrating early diagnosis data with variables like cancer type, anatomical site, and patient demographics. Our aim is to develop more accurate models that can inform personalized treatment strategies and ultimately improve survival rates in pediatric oncology.

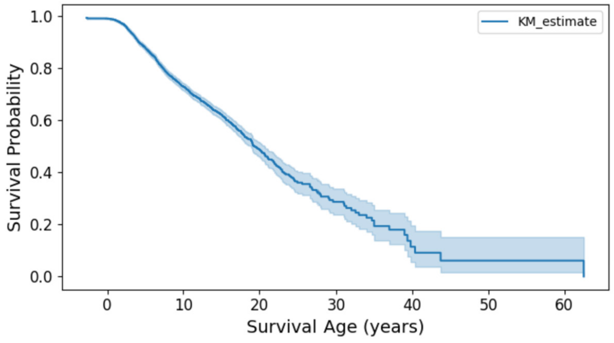

Figure 2. Kaplan-Meier survival curve for pediatric cancer patients. Kaplan-Meier survival curve illustrating the survival probability of pediatric cancer patients over time. The solid line indicates the estimated survival function, and the shaded region represents the 95% confidence interval around the estimate. The analysis was performed using the Kaplan-Meier estimator on a cohort of pediatric cancer patients (n=9184).

These predictive models highlighted the significant roles of variables such as age at diagnosis, specific cancer types, and anatomical sites in determining survival outcomes. The models achieved an accuracy of 0.74, demonstrating the potential of integrating AI-driven approaches into clinical practice to support personalized treatment strategies and improve patient care. These findings provide a compelling case for the adoption of AI in pediatric oncology, setting the stage for further exploration of the results in subsequent sections.

RESULTS

Patient demographics and characteristics

Descriptive statistical analysis was conducted on data from 4592 pediatric cancer patients, sourced from the Childhood Cancer Data Commons (CCDC) via the Clinical Commons portal at the National Cancer Institute. This analysis provided a foundation for predictive modeling. The predominant race in the dataset was White (~60%), and most patients were not Hispanic or Latino (~55%). A slight male predominance was observed, which is consistent with other similar studies. Neuroblastoma, Not Otherwise Specified (NOS), was the most frequent diagnosis, and the brain was the most common anatomic site. After expanding the dataset to 9184 participants through synthetic data generation,

https://doi.org/10.59720/24-175

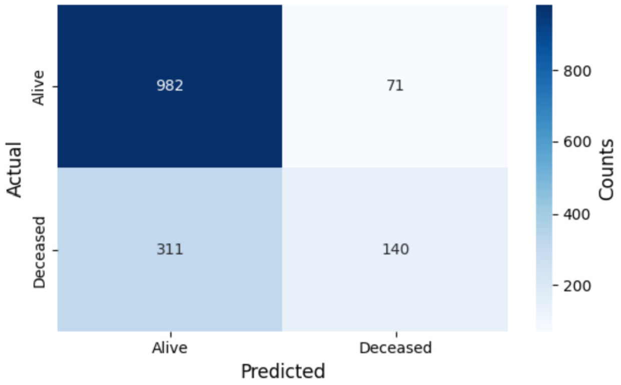

Figure 3. Confusion matrix of prediction patient survival outcomes with classification models. The confusion matrix illustrates the performance of the classification model in predicting survival outcomes (alive vs. deceased) among pediatric cancer patients. The analysis was based on data from a pediatric patient cohort (n=9184). The matrix compares actual outcomes (y-axis) with predicted outcomes (x-axis). The color intensity of each cell reflects the number of observations, with darker shades indicating higher counts.

additional demographic and clinical variations were included to enhance the robustness of the subsequent analyses.

Determining predictive power of patient characteristics

We then used comparative analysis to evaluate the predictive power of different variables using Cox Proportional Hazards models. This analysis was conducted on the original dataset of 4592 pediatric cancer patients. While age at diagnosis emerged as an important factor affecting survival outcomes, other variables such as ethnicity and specific cancer types also played significant roles. The histogram illustrates a higher concentration of diagnoses occurring in younger children, with a notable peak around ages 3 to 5 years. These results highlight the complexity of survival outcomes and underscore the need to consider multiple variables in prognostic models (**Figure 1**).

Determinants of survival

Kaplan-Meier survival curves were generated for different age groups at diagnosis to visualize their impact on survival outcomes (**Figure 2**). The survival probabilities were higher for patients diagnosed at younger ages compared to those diagnosed at older ages. However, the analysis also revealed significant survival differences based on other factors, such as specific diagnoses and ethnicity (log-rank test, $p < 0.001$), challenging the initial hypothesis that age at diagnosis is the most dominant predictor of survival outcomes. We conducted Cox Proportional Hazards regression analysis to identify significant predictors of survival, incorporating variables such as age at diagnosis, race, sex at birth, and specific cancer types. The hazard ratio (HR), calculated through this model, measures the relative risk of death at any point in time for one group compared to another. An HR greater than 1 indicates increased risk, while an HR less than 1 suggests decreased risk. In this analysis, age at diagnosis was an important factor in treatment outcomes. However, the HR of approximately 1.00 indicates that age at diagnosis did not significantly affect the daily mortality risk when controlling for other factors. The analysis of racial categories showed that race, can significantly impact survival outcomes, particularly for African American patients ($p<0.005$), but not Asian or native Hawaiian or other pacific islander individuals. The analysis indicated that sex at birth had minimal influence on survival outcomes, with HRs close to 1.00 and no substantial differences observed between different sexes. Certain cancer types, such as those categorized under tumor cells, malignant, also had high hazard ratios ($p< 0.05$), indicating a strong association with poorer survival outcomes.

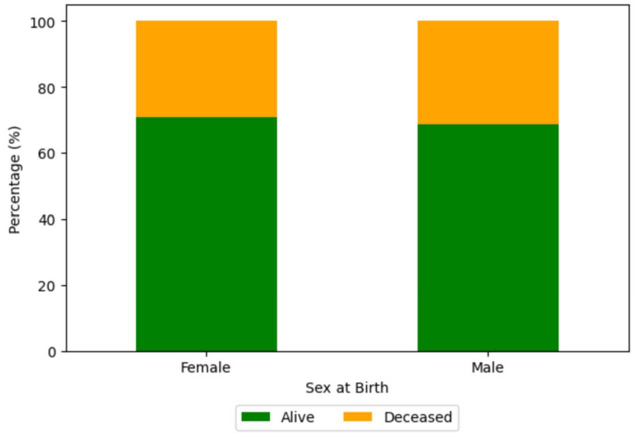

Figure 4. Survival rate by sex at birth. Bar graph showing the survival rates of pediatric cancer patients by sex at birth, with the percentage of patients who are alive (green) and deceased (orange) within each sex category (female and male). The analysis is based on data from a pediatric patient cohort (n=9184).

Impact of age of diagnosis on survival

We conducted a sensitivity analysis to explore the effect of age at diagnosis on survival predictions. The results showed that the predictive impact of decreasing the age at diagnosis by one year was not significant (**Figure 2**).

Classification models to predict patient status

Classification models, enhanced with generative AI techniques, were employed to predict survival outcomes and evaluate model performance in classifying patient statuses accurately. The integration of generative AI, particularly the CTGANSynthesizer, allowed the model to simulate and generate synthetic datasets that were used to augment the training data. This process improved the robustness of the models, enabling them to generalize better to unseen data. The generative AI techniques also facilitated the exploration of complex, non-linear relationships within the data, which traditional machine learning methods might not capture as effectively. By generating new data points that reflect the underlying distributions of the original dataset, the models could achieve higher accuracy and resilience against overfitting, particularly in scenarios with limited or imbalanced data. This enhancement is crucial for the model's ability to make more accurate and reliable predictions, ultimately leading to better-informed clinical decisions and improved patient outcomes. The true positives (patients correctly predicted to be alive) are 981, while the false positives (patients incorrectly predicted to be deceased) are 72. The false negatives (patients incorrectly predicted to be alive) are

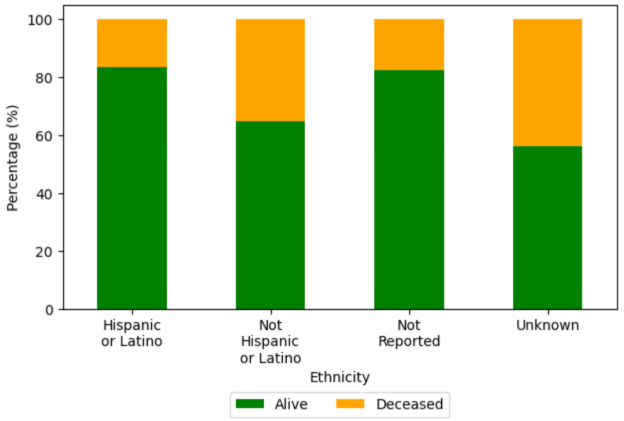

Figure 5. Survival rate by ethnicity. Bar graph displaying the survival rates of pediatric cancer patients across different ethnicities, including Hispanic or Latino, Not Hispanic or Latino, Not Reported, and Unknown. The bars represent the percentage of patients who are alive (green) and deceased (orange) within each ethnic category. The analysis is based on data from a pediatric patient cohort (n=9184).

312, and the true negatives (patients correctly predicted to be deceased) are 139. The models showed varying performance across different classes, with an overall accuracy of 0.74 (**Figure 3**).

Analyzing survival rates by sex at birth and ethnicity provided additional context. While these variables showed some influence on survival outcomes, ethnicity, particularly for Black or African American patients, emerged as a more significant factor than sex at birth. The bar graph illustrated that males had a slightly higher survival rate than females, and not Hispanic or Latino individuals had higher survival rates compared to Hispanic or Latino individuals (**Figures 4, 5**).

Impact of age at diagnosis on survival status

The boxplot compares the age at diagnosis between patients categorized as Alive or Deceased. The median age at diagnosis for deceased patients is lower than that of surviving patients, suggesting that younger children may face higher mortality risks in some cancer types. Surviving patients exhibit a wider spread in age, indicating better survival rates among older children. The presence of multiple outliers in both groups reflects variability in age at diagnosis and survival outcomes, emphasizing the importance of considering additional factors such as cancer type in survival predictions (**Figure 6**).

Factors that influence survival

Top 10 feature importances for predicting survival were identified using a Random Forest model. The bar chart shows that age at diagnosis (in days) had the highest relative importance in predicting survival outcomes, followed by cancer-related study titles and ethnicity. Other variables such as specific diagnoses (e.g., Neuroblastoma NOS) and race also played significant roles, indicating that these factors are critical in the survival prediction model (**Figure 7**).

Impact of cancer diagnosis and anatomic site on survival

Kaplan-Meier Survival Curves for Top Diagnoses and Sites: Kaplan-Meier survival curves for the top 10 cancer diagnoses and anatomic sites were analyzed to provide detailed survival probabilities over time. These curves demonstrated the varying survival outcomes associated with different cancer types and anatomical sites. Notably, Neuroblastoma, NOS (blue) and Juvenile myelomonocytic leukemia, NOS (orange) show significantly different survival patterns compared to other diagnoses. (**Figure 8**).

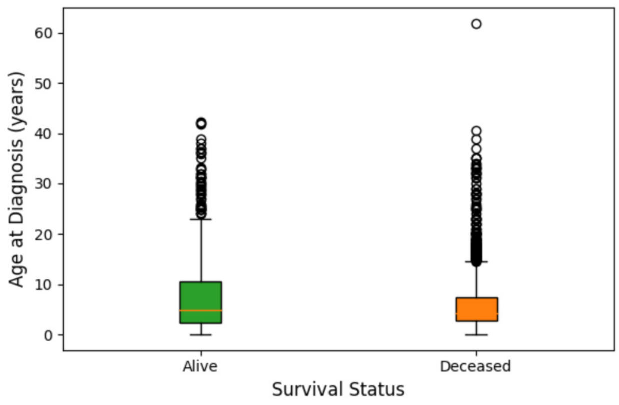

Figure 6. Age at diagnosis vs. survival status. Box plot comparing the age at diagnosis (in years) between pediatric cancer patients who are alive and those who are deceased. The plot indicates that age at diagnosis is an important predictor of survival outcomes, with the distribution showing that patients diagnosed at younger ages tend to have poorer survival rates, as indicated by the lower median age at diagnosis in the deceased group compared to the alive group. The median is shown by the line inside the box. The whiskers represent the range of ages, with outliers shown as individual points. The analysis is based on data from a pediatric patient cohort (n=9184).

DISCUSSION

This study aimed to identify key predictors of pediatric cancer survival, with a focus on early diagnosis, specific cancer types, anatomical sites, and demographic factors such as ethnicity. We initially hypothesized that age at diagnosis would be the most significant predictor of survival outcomes. However, our results suggest that while age at diagnosis is an important factor, ethnicity and specific cancer types have a stronger influence on survival rates. The dataset, consisting of 4592 entries, was expanded to 9184 through synthetic data generation, revealing a slight male predominance and a majority of white and not Hispanic or Latino patients. The most frequent diagnosis was neuroblastoma NOS, and the brain NOS was the most common anatomical site.

Our findings highlighted that early diagnosis improves survival in some cases but may not be the dominant predictor, as previously thought. Instead, the role of ethnicity and cancer type emerged as critical factors, with Kaplan-Meier survival curves by race, sex assigned at birth, and ethnicity providing valuable visual insights into the variability of survival outcomes. Cox Proportional Hazards models, enhanced with generative AI, identified significant predictors of survival, emphasizing the importance of considering multiple demographic and clinical factors when predicting outcomes.

Several factors could explain the variability in survival outcomes. Ethnicity played a prominent role, with black or African American patients exhibiting different survival trends compared to other groups, suggesting the need for a stratified analysis to understand how demographics and cancer type

https://doi.org/10.59720/24-175

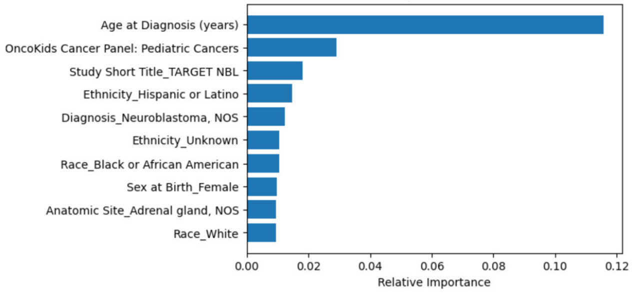

Figure 7. Top 10 feature importances for predicting survival. Bar chart displaying the relative importance of the top ten features in predicting survival outcomes for pediatric cancer patients, as determined by statistical method using a Random Forest model.

intersect to influence survival outcomes. The variability in cancer aggressiveness and response to treatment across demographic groups, particularly for cancers like acute lymphoblastic leukemia (ALL), may explain why earlier diagnosis improves survival in some cases but not all.

The inclusion of synthetic data expanded the dataset, offering significant advantages for statistical analysis and machine learning model development. However, this approach may also introduce artifacts that affect the accuracy of the analysis. Additionally, the presence of missing data and 'unknown' or 'not reported' categories may introduce bias, potentially affecting the survival predictions. While the classification models enhanced by generative AI showed an overall accuracy of 0.74, performance varied across different classes, highlighting areas for improvement. Future studies should aim to collect more complete and diverse datasets to address these limitations.

Future research should focus on validating the key predictors identified in this study, such as ethnicity, cancer type, and age at diagnosis, while also exploring the molecular and genetic mechanisms driving these associations. Larger datasets and stratified analyses by cancer type and demographic variables could provide more precise survival predictions and enable the development of personalized treatment strategies. Additionally, more sophisticated machine learning models that incorporate real-time data and adapt to evolving information could further improve prognostic accuracy.

Our study demonstrated the utility of generative AI in enhancing the predictive power of survival models, but future research should explore its application in unstructured data analysis, such as imaging or genomic data. Collaborative efforts across institutions will be essential to advance pediatric cancer research and improve the quality of life and survival of affected children.

In conclusion, the integration of machine learning and generative AI into pediatric cancer research has the potential to enhance prognostic models and support the development of personalized treatment plans. By identifying key survival predictors, clinicians can better tailor treatment strategies based on the unique characteristics of each patient's cancer, aligning with the principles of personalized medicine. This approach can optimize therapeutic outcomes by considering individual variability in genes, environment, and lifestyle, ultimately leading to better patient care and survival.

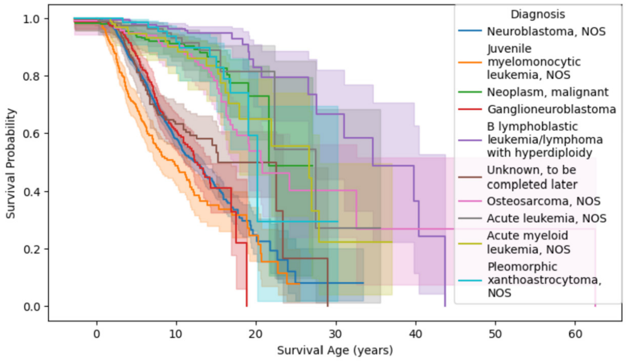

Figure 8. Kaplan-Meier Survival Curves by Top 10 Diagnoses. This plot presents the Kaplan-Meier survival curves for various pediatric cancer diagnoses. Each colored line represents a different diagnosis, with the corresponding shaded area indicating the 95% confidence interval. Data were analyzed using the Kaplan-Meier method, and survival differences between groups were assessed using the log-rank test.

MATERIALS AND METHODS

Data collection preparation and exclusion criteria

The data for this study were sourced from the Childhood Cancer Data Commons (CCDC), accessible through the Clinical Commons portal at the National Cancer Institute. The initial dataset comprised detailed demographic and clinical information for 6651 pediatric cancer patients.

To focus on the pediatric population, only patients aged 21 years or younger were included in the analysis. Patients were excluded from the dataset if they had incomplete data for critical variables necessary for survival analysis, such as age at diagnosis, diagnosis type, and survival status. This step ensured that the analysis was conducted on a robust and complete dataset, minimizing potential biases that could arise from data imputation. The final dataset, after applying these exclusion criteria, comprised 4592 patients who met all the necessary conditions for a thorough and reliable analysis.

Generative AI utilization

In this study, we utilized an Advanced Data Analysis Agent, a LLM developed by OpenAI, to support various computational tasks. The generative AI was employed primarily for data preprocessing, code generation, and preliminary analysis. The LLM played a crucial role in the generation of synthetic data to augment the dataset, ensuring that the model training and testing phases were conducted on a more comprehensive dataset. Additionally, the LLM assisted in organizing and structuring the dataset for subsequent analysis, generating starter code for implementing statistical models, such as the Kaplan-Meier method, Cox proportional hazards model, and Random Forest Classifier. These generated code snippets were then reviewed and refined by the research team to ensure accuracy and appropriateness for the study.

Statistical and machine learning methods

To identify significant predictors of survival, we employed Cox Proportional Hazards models, a widely used method in survival analysis for examining the relationship between survival time and predictor variables. The analysis was performed using Python (version 3.8) with relevant libraries such as Pandas (version 1.1.5) for data manipulation, Lifelines (version 0.25.9) for survival analysis, scikit-learn (version 0.24.2) for machine learning, along with specialized tools like NumPy (version 1.19.2) for numerical computations, Matplotlib (version 3.3.2) for plotting, Seaborn (version 0.11.0) for statistical data visualization, Statsmodels (version 0.12.1) for statistical modeling, Joblib (version 0.17.0) for job and data serialization, and Synthetic Data Vault (SDV, version 0.4.3) for generating synthetic datasets. For survival outcome prediction, we utilized classification algorithms including Random Forest, enhanced with the support of generative AI techniques. Model performance was evaluated using metrics such as precision, recall, F1-score, and accuracy. The confusion matrix was analyzed to gain insights into the model's predictions versus actual outcomes.

Data handling

The dataset was split into training (80%) and testing (20%) sets to build and validate the machine learning models. Preprocessing of data included the management of missing entries, normalization of numerical data, and transformation of categorical data into encoded formats.

Software and tools

All data analyses were conducted using Python (version 3.8) and associated libraries. The pandas library was used for data manipulation and preprocessing, while lifelines facilitated the survival analysis with Cox Proportional Hazards models. Classification algorithms and model evaluations were performed using scikit-learn, with generative AI techniques enhancing the data preprocessing and model training stages. The synthetic data generation was conducted using the CTGANSynthesizer from the Synthetic Data Vault (SDV). Generative AI, specifically an advanced data analysis assistant in the form of a large language model (LLM) in ChatGPT, played a pivotal role by organizing structured data, accelerating data cleaning, and facilitating visualization. The LLM also assisted in preliminary data analysis, identifying biases, and suggesting ethical considerations, which were critical for maintaining the study's integrity (6). The LLM, enhanced traditional techniques and expedited the implementation of models like the Kaplan-Meier method, Cox proportional hazards model, and Random Forest Classifier, which were integral to the study.

The complete code used for this study is available on GitHub at (https://github.com/sky-yng/Prognosis_Pediatric_Cancer), ensuring transparency and reproducibility of the analysis.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to the data providers and researchers who contributed to the pediatric cancer dataset used in this study. We also acknowledge the support from our respective institutions that made this research possible. Special thanks to the CCDC and the National Cancer Institute's Clinical Commons portal for access to the dataset, and to the developers of the SDV for their contributions to data augmentation techniques. We are also grateful to the OpenAI team for the development of the LLM in ChatGPT, which played a crucial role in data preprocessing and analysis. Finally, we thank Sarah Mechler, Science Teacher at Vista Ridge High School, for her guidance.

Received: June 18, 2024
Accepted: September 4, 2023
Published: December 21, 2024

REFERENCES

1. Blandin Knight, Sean et al. “Progress and prospects of early detection in lung cancer.” Open Biology, vol. 7, no. 9, 6 Sep. 2017, https://doi.org/10.1098/rsob.170070.
2. Rajkomar, A., et al. “Machine learning in medicine.” New England Journal of Medicine, vol. 380, no.14, 3 Apr. 2018, p.1347-1358, https://doi.org/10.1056/NEJMra1814259.
3. Ward, E., et al. Childhood and adolescent cancer statistics 2014. CA: A Cancer Journal for Clinicians, vol. 64, no.2, 31 Jan. 2014, , p. 83-103, https://doi.org/10.3322/caac.21219.
4. Harrell, F. E., et al. “Regression modeling strategies for improved prognostic prediction.”, vol. 3 no. 2, 04 Jun. 1984, p. 143-152, https://doi.org/10.1002/sim.4780030207.
5. Stehman, S. V. (1997). “Selecting and interpreting measures of thematic classification accuracy”. Remote Sensing of Environment, vol. 62, no.1, 1 Oct 1997, p. 77-89, https://doi.org/10.1016/S0034-4257(97)00083-7.
6. Bleyer, A., et al. “Cancer epidemiology in older adolescents and young adults 15 to 29 years of age”. National Cancer Institute, NIH Pub. No. 06-5767. ISBN: 978-1-58240-626-1. www.seer.cancer.gov/archive/publications/aya/index.html.
7. Anders, C. K., Johnson, R., Litton, J., Phillips, M., & Bleyer, A. “Breast cancer before age 40 years”. Seminars in Oncology, vol. 36, no.3, Jun 2009, p. 237–249, https://doi.org/10.1053/j.seminoncol.2009.03.001.
8. Kleinbaum, D. G., & Klein, M. (2012). “Survival analysis”. Springer. ISBN: 978-1-4419-6645-4. www.link.springer.com/book/10.1007/978-1-4419-6646-9.
9. Relling, M., Evans, W. “Pharmacogenomics in the clinic”. Nature Vol. 526, 14 Oct. 2015, p. 343–350, https://doi.org/10.1038/nature15817.
10. Sokolova, M., and Lapalme, G. (2009). “A systematic analysis of performance measures for classification tasks”. Information Processing & Management, vol. 45, no. 4, 4 Jul. 2009, p. 427-437, https://doi.org/10.1016/j.ipm.2009.03.002.
11. Cox, D. R. (1972). “Regression models and life-tables (with discussion)”. Journal of the Royal Statistical Society: Series B (Methodological), vol. 34, no.2, 5 Dec. 2018, p. 187-220, https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

structuring the dataset for subsequent analysis, generating starter code for implementing statistical models, such as the Kaplan-Meier method, Cox proportional hazards model, and Random Forest Classifier. These generated code snippets were then reviewed and refined by the research team to ensure accuracy and appropriateness for the study.