**Key Drivers of Cancer Treatment Costs: Leveraging Machine Learning for Financial Planning**

Esti Stern

July 28, 2025

**Abstract**

Accurately predicting cancer treatment costs is a growing priority in healthcare, as it enables better financial planning for patients and more efficient resource allocation for providers.

The goal of this study is to identify key predictors for predicting the costs of cancer treatment.

To estimate cancer treatment costs, a supervised learning framework was implemented using Random Forest and XGBoost regression models on a deidentified Kaggle dataset of 50,000 patient records with 15 predictive features. After preprocessing, feature selection, and thorough model evaluation – including LazyPredict benchmarking – Bayesian Ridge Regression was chosen to be evaluated for its strong performance, with modeling conducted in Python using Scikit-learn and grid search optimization.

Advanced machine learning models like XGBoost achieved high predictive accuracy for cancer treatment costs, with an R2 of 0.98 and RMSE of 3156. Bayesian Ridge Regression showed signs of overfitting, highlighting the challenges in cost modeling.

Machine learning – especially XGBoost – proved highly effective in predicting cancer treatment costs by uncovering nonlinear patterns and identifying indirect yet impactful predictors. Future models enriched with variables like treatment type and insurance coverage could further sharpen cost forecasts and support more personalized healthcare strategies.

**Introduction**

Financial toxicity, which includes both direct out-of-pocket expenses for treatment and indirect costs such as travel, time, and changes to employment, places a significant burden on cancer patients and their families. This financial strain affects access to necessary care as well as their overall well-being, influencing treatment decisions (Abrams et al., 2021). While the high costs of treatments are well known, current guidelines lack detailed recommendations for physician-patient discussions about financial challenges, limiting the patient’s ability to exert some control over treatment expenses (Agarwal et al., 2021). Although, financial toxicity is seen across cancer types, countries, and healthcare systems, some are at a higher risk of experiencing it than others (Leighl et al., 2021). A key factor in managing these costs is how insurance companies assess and allocate coverage. By improving cost prediction models, insurers can offer more accurately priced premiums, and better coverage plans tailored to patient needs, reducing the burden of treatment expenses. This study aims to develop a predictive model that enables insurance companies to estimate healthcare costs more accurately based on diverse patient profiles.

Cancer drug prices vary widely, and dot not always correlate with value or clinical benefit (Leighl et al., 2021). Drug development costs on average $1.2 billion to $1.3 billion per approved biopharmaceutical. Only 16% to 19% of products that enter clinical trials successfully reach the market. When new versions of an approved drug come out, the older drug is viewed as substandard and the cheaper generic versions become outdated and obsolete sustaining the monopoly (Siddiqui & Rajkumar, 2012).

Beyond drug pricing, the stage at which the cancer is diagnosed plays a critical role in determining treatment expenses and patient outcomes. Costs of cancer care vary by cancer site and phase of care (de Oliveira et al., 2016) (Mariotto et al., 2020). In a study of women with breast cancer, younger patients (under 45 years old) were found to face a higher prevalence of late-stage disease and higher cost within each stage compared to older women (45 – 64 years old) (Allaire et al., 2017). While patients diagnosed at an earlier stage incur higher lifetime payments due to longer survival time, their annual average costs are substantially lower than those at advanced stages. Early detection of cancer can result in more effective treatment and lower management costs (Sun et al., 2018) (Reddy et al., 2022).

In 2009, the highest costs per hospital day were for cancer of the prostrate ($4,600 per day), cancer of the breast ($4,100 per day), and cancer of the thyroid ($3,500 per day) (Price et al., 2012). In pediatric cancers, leukemias and brain tumors are the most common and costly (Merrill et al., 2007). In Ontario, Canada, lifetime costs ranged from less than $55,000 for lung and liver cancers to over $110,000 for leukemia, multiple myeloma, lymphoma, and breast cancer (de Oliveira et al., 2016). These cost variations underscore the economic challenges of cancer treatment, highlighting the need for strategic healthcare planning.

While cancer drug prices vary across countries and may not correlate directly with the country’s GDP, prices are highest in the United States (Leighl et al., 2021). Unlike the UK, where pharmaceutical prices are negotiated with manufacturers, the US generally accepts the prices set by drug companies. This absence of price negotiation in the US results in higher costs for cancer treatments (Prasad & Mailankody, 2016). In countries without universal healthcare, patients are responsible for treatment costs (Leighl et al., 2021). In the United States, the share of costs of cancer paid for by private insurance and Medicaid have increased (Tangka et al., 2010). To better understand cancer care costs worldwide, it is important to compare them in detail. Standardizing cost data, especially for medicines, would make it easier to compare expenses and treatment outcomes between countries (Torkki et al., 2022).

High costs of cancer drugs are driven by various factors, impacting patients, insurers, and healthcare systems (Siddiqui & Rajkumar, 2012). Understanding the cumulative costs of cancer treatment from diagnosis is critical for resource allocation and policy making. Accurate cost measurement requires detailed data on services used, unit costs, and adjustments for missing data, patient deaths and follow-ups. Variations in costs arise from factors such as stage at diagnosis, treatment regiments, patient demographics, and geographical differences (Chirikos, 2002). By standardizing cost data and improving financial transparency, healthcare policymakers can create more equitable treatment pathways and alleviate financial toxicity for patients.

**Methods**

To predict cancer treatment costs, a supervised learning approach was employed using random forest regression models applied to patient data. The analysis utilized a publicly available dataset from Kaggle, comprising 50,000 entries with 15 features each. These features included patient demographics, environmental and genetic risk factors, disease severity, and treatment costs across various countries and years. The dataset was deidentified to maintain patient privacy.

Before modeling, exploratory data analysis was conducted to better understand the structure and distribution of the data. This included examining feature distributions and visualizing relationships between variables using plots and correlation heatmaps.

The dataset contained no missing values. Categorical variables such as ‘Country’ and ‘Cancer Type’ were transformed using one-hot encoding, while ‘Gender’ and ‘Cancer Stage’ were encoded using ordinal encoding. Feature importance scores from the Random Forest model in Scikit-learn were used to identify the most influential predictors. Features with minimal impact were removed to enhance model efficiency and accuracy.

Both Random Forest Regressor and XGBoost were evaluated. Random Forest was initially selected due to its ability to provide feature importance, its resistance to overfitting, and its independence from feature scaling or normalization. XGBoost was later introduced for comparison, given its computational efficiency and superior handling of overfitting.

To explore whether additional models might outperform the initial selections, LazyPredict was applied after evaluating Random Forest and XGBoost. This tool provided a quick comparison of multiple regression algorithms using consistent preprocessing and evaluation metrics. Bayesian Ridge Regression emerged as a strong contender, demonstrating competitive performance in terms of R2 and RMSE. Based on these results, a Bayesian Ridge Regression model was implemented and evaluated.

The dataset was divided into training and testing sets in a 75:25 ratio. Hyperparameters for the Random Forest and XGBoost models were optimized using grid search with 3-fold cross-validation. Model performance was assessed using root mean squared error (RMSE) and R2, appropriate metrics for regression tasks.

All modeling was conducted in Python using Scikit-learn and Pandas within a Jupyter notebook environment. Random seeds were set to 23 to ensure reproducibility.

**Results**

The main demographic and clinical features of the study participants are shown in **Figure 1**. The gender breakdown was nearly equal, with 33.6% male, 33.4% female, and 33.0% as other, suggesting that the study included a wide range of individuals. The distributions of cancer type and cancer stage also show relatively balanced proportions, with colon (12.8%) and prostate (12.6%) being the most common diagnoses among specified cancer types, and the highest cancer stage representation at Stage II (20.2%). These relatively even distributions across categories suggest that no single group is overrepresented. The geographic distribution reflects a similar pattern, with participants from ten countries and the largest shares from Australia (10.2%) and the UK, USA, and India each at 10.1%.

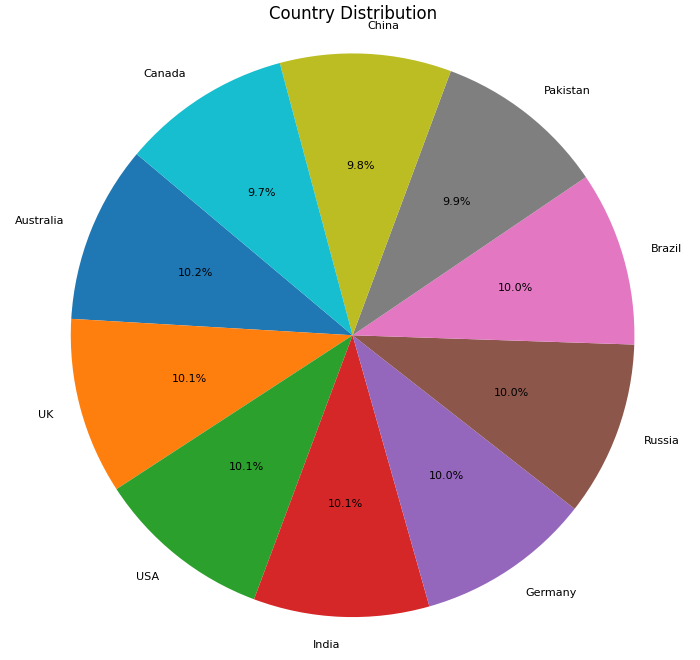
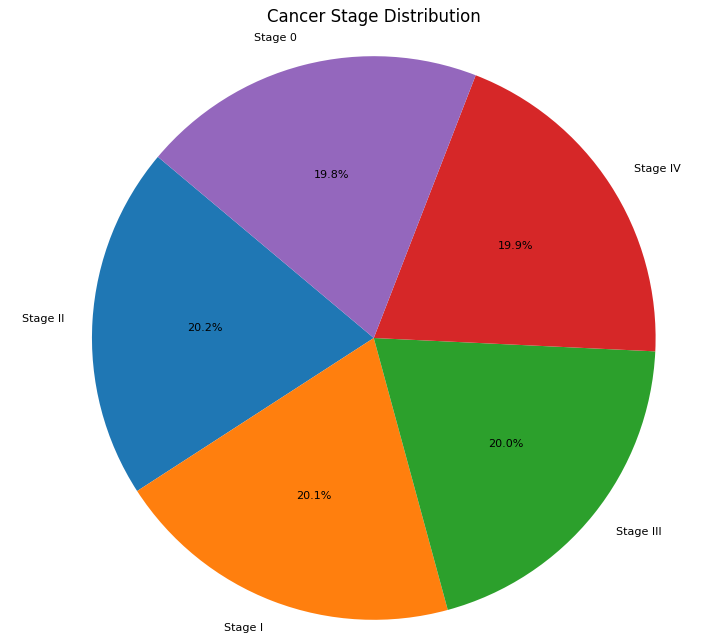
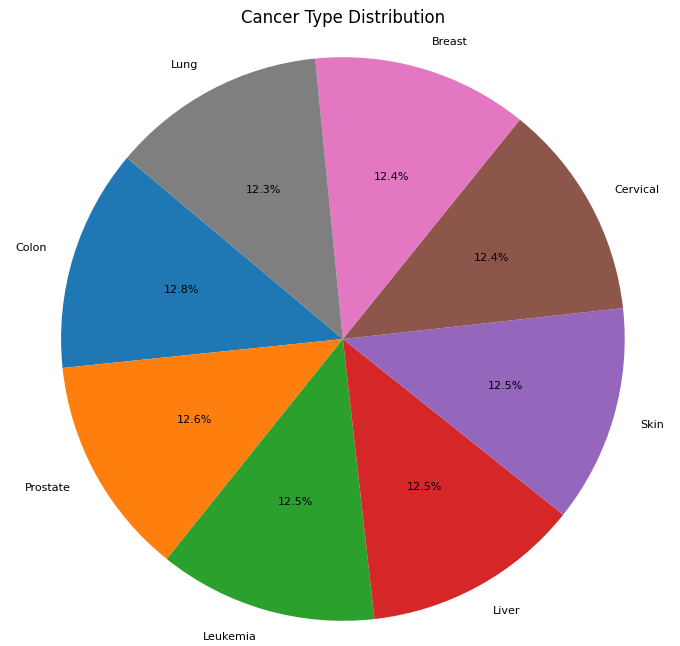
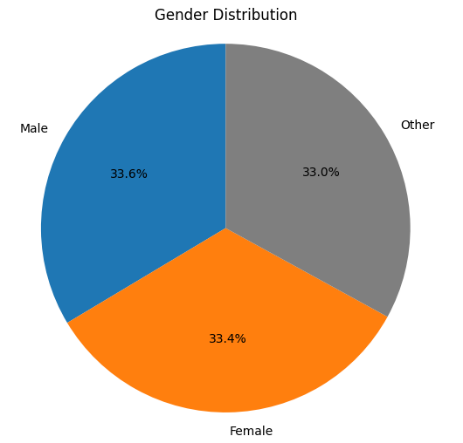


Figure 1: Distribution of Gender, Cancer Type, Stage and Geographic Location

The overall average treatment cost was $52,467 (**Table 1**). Although the average treatment costs across countries were relatively similar, there were notable deviations from the global mean (**Figure 2**). None of the features, apart from severity score, had a significant linear correlation with the target variable, treatment cost, as shown in **Figure 3**.

Table 1: Summary Statistics

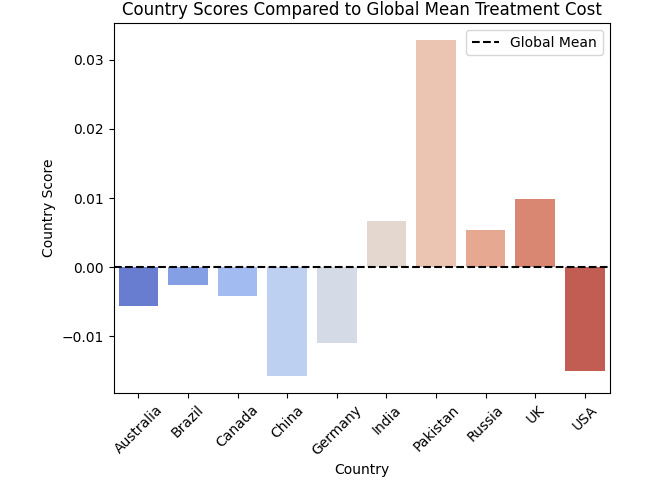
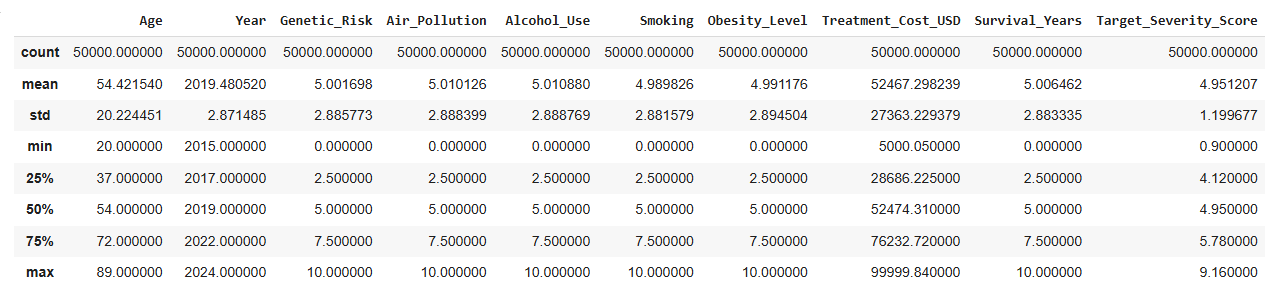


Figure 2: Global Treatment Cost Deviations by Country

Each bar shows a country’s score relative to the global mean: higher bars (in red) mean lower-than-average treatment costs, and lower bars (in blue) mean higher-than-average costs

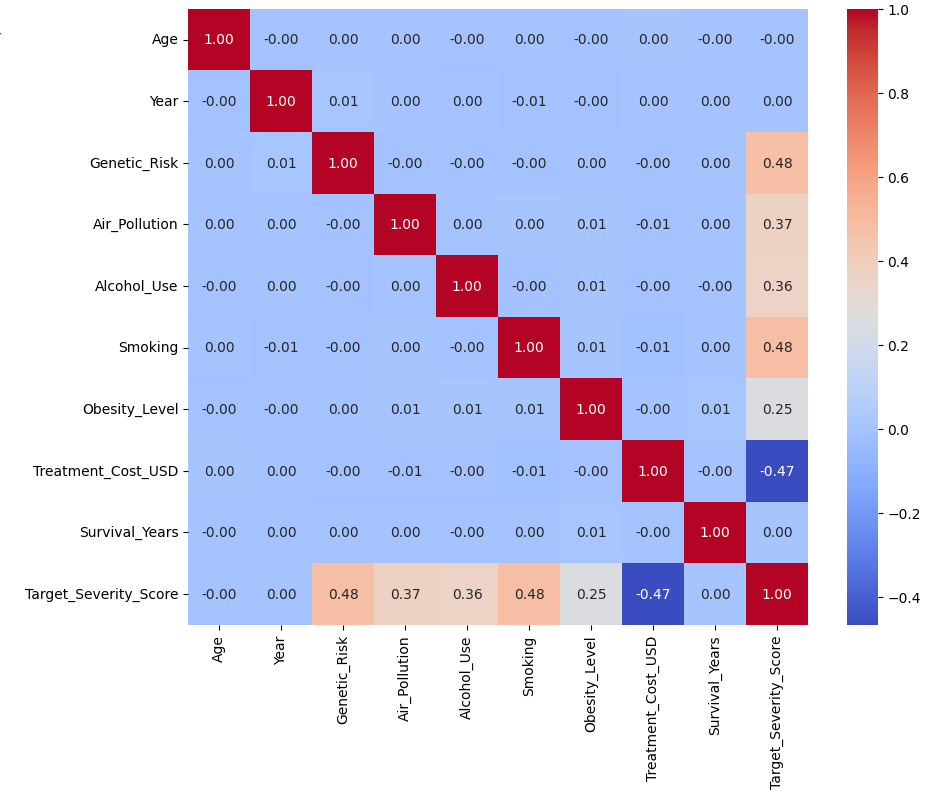


Figure 3: Correlation Between Features and Treatment Cost

Following the training of both a Random Forest Regressor and an XGBoost model, feature importance was analyzed, and variables with minimal contribution were removed, as illustrated in **Figure 4**. The XGBoost model demonstrated superior performance—achieving lower RMSE and higher R²—when all features were retained, indicating that even low-importance features contributed to predictive accuracy. Because the difference was very slight, it was decided to remove features as a measure to prevent overfitting. Model optimization was conducted using grid search in combination with cross-validation to determine the optimal hyperparameters. A comparison of model performance metrics before and after optimization is provided in **Table 2**.

Bayesian Ridge Regression was identified as a top-performing candidate by LazyPredict. Despite the features exhibiting very low linear correlation with the target variable (**Figure 3**), the model achieved an unexpectedly high R2 value of 0.9999. This anomaly is likely attributable to overfitting.

|  |  |  |
| --- | --- | --- |
|  | Random Forest | XGBoost |
| RMSE | 8946.15 | 5042.69 |
| R2 | 0.8936 | 0.9662 |

|  |  |  |
| --- | --- | --- |
|  | Random Forest | XGBoost |
| RMSE | 8862.98 | 3156.17 |
| R2 | 0.8956 | 0.9867 |

Table 2: Model Metrics Before and After Optimization



Table 3: Model Comparison

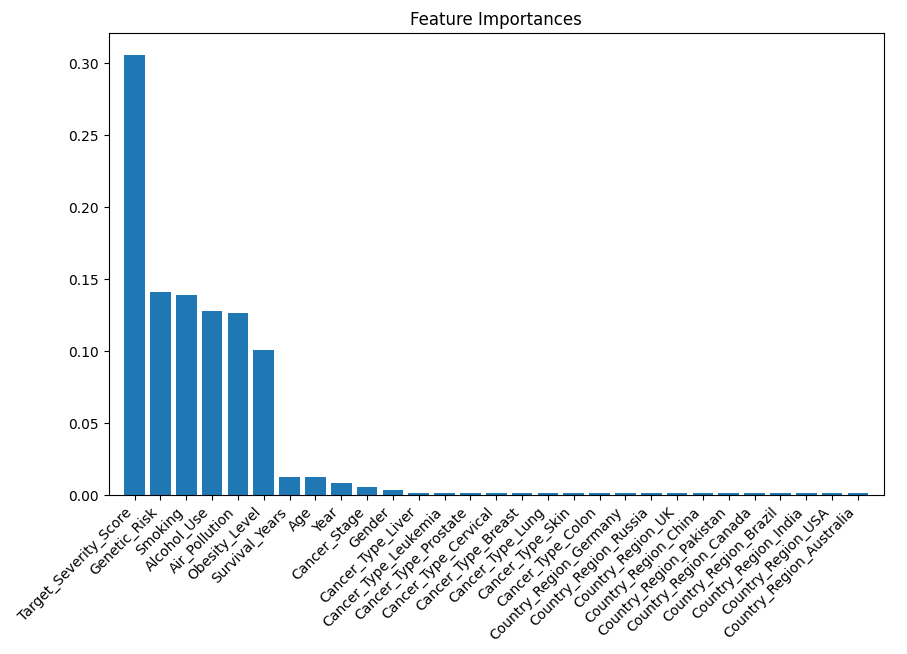
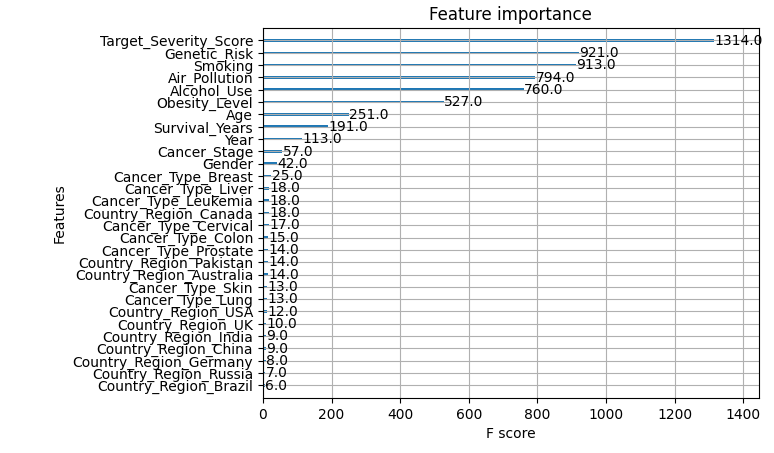


Figure 4: Feature Importance from Random Forest (top) and XGBoost (bottom)

**Discussion**

Despite weak linear correlations between most clinical and demographic factors and treatment cost, advanced models – particularly XGBoost – demonstrated strong predictive accuracy. The study found XGBoost to be the top-performing model, achieving an R2 of 0.98 and an RMSE of 3156. The exceptionally high R2 from Bayesian Ridge Regression likely shows overfitting, highlighting how complex it is to predict cancer treatment costs. These insights reinforce the importance of integrating predictive analytics into healthcare planning to support fair and data-informed financial policies.

Genetic and environmental factors appear to play an important role in predicting treatment costs. While they may not have a direct effect on cost, they influence the severity of cancer, which in turn correlates strongly with cost. As a result, these factors emerge as key predictors in the overall cost modeling.

These findings highlight the importance of recognizing and addressing modifiable risk factors – such as smoking and alcohol consumption – through preventative healthcare policies. Doing so can not only improve patient outcomes but also help manage long-term expenditure more effectively.

While tree-based models offer notable advantages they also present specific limitations, especially when working with categorical variables. One such limitation is that they generally perform better with numerical data compared to one-hot encoded features as numerical features allow flexible splitting while one-hot encoded features increase dimensionality making splits less informative. This can lead to lower feature importance scores, even if the original categorical variable was meaningful. In this dataset, the ‘Country’ variable shows no significant differences in costs across the ten included countries, suggesting that geographical location is not a meaningful predictor in this context, as reflected in its low importance score. Although tree-based models may carry inherent assumptions or biases, the feature importance scores still provide a broadly reliable indication of variable relevance.

Tree-based algorithms also often sacrifice interpretability, as it’s difficult to trace the exact decision path that leads to a prediction. However, they typically offer stronger performance and provide useful insights through feature importance scores, helping identify which variables most influence the model’s output.

Insurance providers can leverage this model to design more personalized policies based on patients’ individual risk profiles and anticipated treatment costs. Similarly, healthcare practitioners can use the model to promote transparency, helping patients better understand potential expenses and make more informed decisions.

For future work, incorporating additional variables – such as treatment type, duration of therapy, hospital stay length, and insurance coverage – could enhance the model’s predictive accuracy and a provide a more nuanced understanding of cost drivers in cancer care.

**Conclusion**

This study demonstrates the value of machine learning models – particularly XGBoost – in accurately predicting cancer treatment costs. Despite weak linear relationships, tree-based methods excelled by capturing nonlinear patterns in the data. Key predictors like genetic and environmental influences, although indirect, surfaced as critical in shaping disease severity and ultimately costs.

While tree-based models present challenges in interpretability and handling categorical variables, their strong performance and insightful feature importance scores underscore their utility in healthcare analytics. Integrating these tools into clinical and policy planning can support more personalized and transparent patient care.

Looking ahead, incorporating variables such as treatment type, therapy duration, hospital stay, and insurance coverage promises to elevate predictive accuracy and enrich understanding of cost dynamics. Ultimately, this work contributes to a smarter, more preventative healthcare system

**Bibliography**

Abrams, H. R., Durbin, S., Huang, C. X., Johnson, S. F., Nayak, R. K., Zahner, G. J., & Peppercorn, J. (2021). Financial toxicity in cancer care: origins, impact, and solutions. *Translational Behavioral Medicine*, *11*(11), 2043–2054. https://doi.org/10.1093/tbm/ibab091

Agarwal, A., Livingstone, A., Karikios, D. J., Stockler, M. R., Beale, P. J., & Morton, R. L. (2021). Physician-patient communication of costs and financial burden of cancer and its treatment: a systematic review of clinical guidelines. *BMC Cancer*, *21*(1), 1036. https://doi.org/10.1186/s12885-021-08697-5

Allaire, B. T., Ekwueme, D. U., Poehler, D., Thomas, C. C., Guy, G. P., Subramanian, S., & Trogdon, J. G. (2017). Breast cancer treatment costs in younger, privately insured women. *Breast Cancer Research and Treatment*, *164*(2), 429–436. https://doi.org/10.1007/s10549-017-4249-x

Chirikos, T. N. (2002). Cancer Economics: on Variations in the Costs of Treating Cancer. *Cancer Control*, *9*(1), 59–66. https://doi.org/10.1177/107327480200900110

de Oliveira, C., Pataky, R., Bremner, K. E., Rangrej, J., Chan, K. K. W., Cheung, W. Y., Hoch, J. S., Peacock, S., & Krahn, M. D. (2016). Phase-specific and lifetime costs of cancer care in Ontario, Canada. *BMC Cancer*, *16*(1). https://doi.org/10.1186/s12885-016-2835-7

Leighl, N. B., Nirmalakumar, S., Ezeife, D. A., & Gyawali, B. (2021). An Arm and a Leg: The Rising Cost of Cancer Drugs and Impact on Access. *American Society of Clinical Oncology Educational Book*, *41*, e1–e12. https://doi.org/10.1200/edbk\_100028

Mariotto, A. B., Enewold, L., Zhao, J., Zeruto, C. A., & Yabroff, K. R. (2020). Medical Care Costs Associated with Cancer Survivorship in the United States. *Cancer Epidemiology, Biomarkers & Prevention*, *29*(7), 1304–1312. https://doi.org/10.1158/1055-9965.EPI-19-1534

Merrill, C. T., Nagamine, M., & Hambrick, M. M. (2007). Pediatric Hospital Stays for Cancer, 2005. *Healthcare Cost and Utilization Project (HCUP)*.

Prasad, V., & Mailankody, S. (2016). The UK Cancer Drugs Fund Experiment and the US Cancer Drug Cost Problem: Bearing the Cost of Cancer Drugs Until It Is Unbearable. In *Mayo Clinic Proceedings* (Vol. 91, Issue 6, pp. 707–712). Elsevier Ltd. https://doi.org/10.1016/j.mayocp.2016.04.028

Price, R. A., Stranges, E., & Elixhauser, A. (2012). Cancer Hospitalizations for Adults, 2009. *Healthcare Cost and Utilization Project (HCUP)*.

Reddy, S. R., Broder, M. S., Chang, E., Paydar, C., Chung, K. C., & Kansal, A. R. (2022). Cost of cancer management by stage at diagnosis among Medicare beneficiaries. *Current Medical Research and Opinion*, *38*(8), 1285–1294. https://doi.org/10.1080/03007995.2022.2047536

Siddiqui, M., & Rajkumar, S. V. (2012). The high cost of cancer drugs and what we can do about it. In *Mayo Clinic Proceedings* (Vol. 87, Issue 10, pp. 935–943). Elsevier Ltd. https://doi.org/10.1016/j.mayocp.2012.07.007

Sun, L., Legood, R., Dos-Santos-Silva, I., Gaiha, S. M., & Sadique, Z. (2018). Global treatment costs of breast cancer by stage: A systematic review. In *PLoS ONE* (Vol. 13, Issue 11). Public Library of Science. https://doi.org/10.1371/journal.pone.0207993

Tangka, F. K., Trogdon, J. G., Richardson, L. C., Howard, D., Sabatino, S. A., & Finkelstein, E. A. (2010). Cancer treatment cost in the United States. *Cancer*, *116*(14), 3477–3484. https://doi.org/10.1002/cncr.25150

Torkki, P., Leskelä, R.-L., Bugge, C., Torfadottir, J. E., & Karjalainen, S. (2022). Cancer-related costs should be allocated in a comparable way—benchmarking costs of cancer in Nordic countries 2012–2017. *Acta Oncologica*, *61*(10), 1216–1222. https://doi.org/10.1080/0284186X.2022.2124883