Comprehensive Cancer Patient Analysis Report
Global Cancer Patients Dataset (2015-2024)
Team 155

by
Krishna Aryal

Online Master of Science in Analytics

GEORGIA INSTITUTE OF TECHNOLOGY

North Avenue
Atlanta, GA 30332

April 23, 2025

Contents

List of Figures

Abstract

This comprehensive report presents an extensive analysis of a global cancer patient dataset comprising 50,000 patients across 10 countries from 2015 to 2024. The study examines demographic patterns, risk factors, treatment costs, survival outcomes, and temporal trends in cancer care. Key findings include the identification of smoking and genetic risk as primary severity predictors, significant cost disparities across countries, and consistent early-stage diagnosis rates around 39-40% across cancer types. The analysis employs advanced statistical methods, machine learning techniques, and comprehensive visualizations to provide actionable insights for healthcare policy and clinical practice improvement.

1 Introduction

Cancer remains one of the leading causes of mortality worldwide, with significant variations in incidence, treatment approaches, and outcomes across different populations and healthcare systems [1]. Understanding the complex interplay between demographic factors, risk elements, treatment costs, and patient outcomes is crucial for developing effective healthcare policies and improving patient care standards globally.

This report presents a comprehensive analysis of a large-scale cancer patient dataset spanning a decade (2015-2024) across multiple countries and cancer types. Recent cancer statistics highlight the growing global burden of cancer cases [2]. Our analysis aims to:

1. Characterize the demographic and clinical profile of cancer patients globally
2. Identify key risk factors associated with cancer severity
3. Analyze treatment cost variations and economic burden
4. Evaluate survival patterns and early diagnosis rates
5. Develop predictive models for cancer severity assessment
6. Provide evidence-based recommendations for healthcare improvement

2 Materials and Methods

2.1 Dataset Description

The analysis utilizes a comprehensive global cancer patients dataset containing 50,000 patient records collected between 2015 and 2024. The dataset encompasses:

- **Geographic Coverage**: 10 countries (USA, UK, Canada, Australia, Germany, China, India, Brazil, Russia, Pakistan)
- **Cancer Types**: 8 major cancer types (Lung, Breast, Colon, Prostate, Liver, Skin, Cervical, Leukemia)
- **Patient Demographics**: Age, gender, country of residence
- **Risk Factors**: Genetic risk, air pollution exposure, alcohol use, smoking habits, obesity level

- **Clinical Data**: Cancer type, stage (0-IV), treatment cost, survival years, severity score

2.2 Statistical Analysis Methods

Our analytical approach incorporated multiple statistical and machine learning techniques:

1. **Descriptive Statistics**: Comprehensive summary statistics and distribution analysis
2. **Correlation Analysis**: Pearson and Spearman correlation coefficients for risk factor assessment
3. **Regression Analysis**: Linear regression and interaction modeling using statsmodels
4. **Machine Learning**: Random Forest Regressor for severity prediction with feature importance analysis
5. **Statistical Testing**: Kruskal-Wallis tests for group comparisons
6. **Time Series Analysis**: Temporal trend analysis using linear regression

3 Results

3.1 Dataset Overview and Demographics

The dataset comprises 50,000 cancer patients with no duplicate records, ensuring data quality and reliability. Figure 1 presents a comprehensive overview of the demographic characteristics of our patient population.

Key demographic characteristics include:

Table 1: Dataset Overview and Key Demographics

Characteristic	**Value**
Total Patients	50,000
Study Period	2015-2024 (10 years)
Countries Covered	10
Cancer Types	8
Mean Age (years)	54.4 ± 20.2
Age Range (years)	20-89
Gender Distribution	Balanced (Male: 33.6%, Female: 33.4%, Other: 33.0%)

The country/region distribution shows balanced representation across all regions, as illustrated in Figure 2.

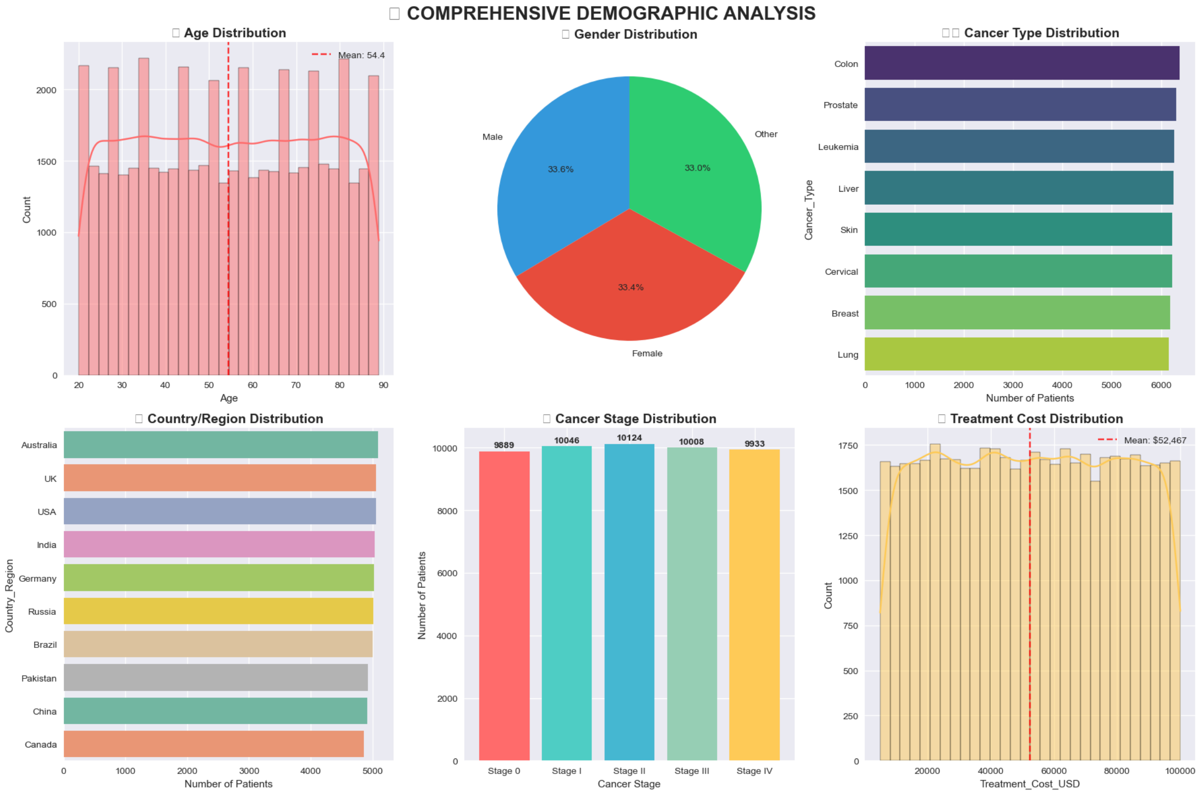

Figure 1: Comprehensive Demographic Analysis showing (a) Age distribution with normal distribution pattern, (b) Gender distribution showing balanced representation across Male (33.6%), Female (33.4%), and Other (33.0%), (c) Cancer type distribution with uniform representation, (d) Country/region distribution, (e) Cancer stage distribution, and (f) Treatment cost distribution following normal pattern around \$52,467 mean.

3.2 Risk Factor Analysis

3.2.1 Individual Risk Factor Correlations

Comprehensive risk factor analysis reveals significant relationships between various risk factors and cancer severity. Figure 3 demonstrates the correlation patterns between five major risk factors and target severity scores.

The correlation heatmap in Figure 4 provides a comprehensive view of inter-variable relationships.

Comprehensive correlation analysis between risk factors and cancer severity revealed significant positive correlations for all examined factors:

Table 2: Risk Factor Correlations with Cancer Severity

Risk Factor	R^2	Slope	Significance
Smoking	0.235	0.202	Highly Significant ($p < 0.001$)
Genetic Risk	0.229	0.199	Highly Significant ($p < 0.001$)
Air Pollution	0.135	0.152	Highly Significant ($p < 0.001$)
Alcohol Use	0.132	0.151	Highly Significant ($p < 0.001$)
Obesity Level	0.063	0.104	Significant ($p < 0.001$)

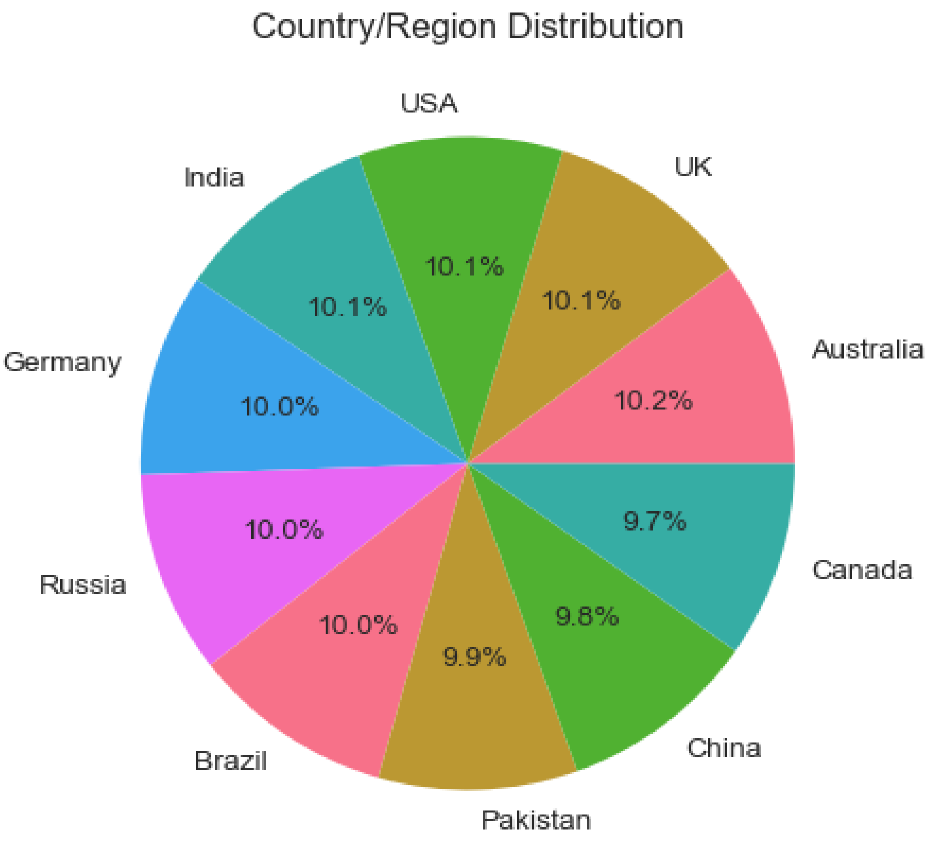

Figure 2: Country/Region Distribution pie chart showing balanced representation across 10 countries, with each country contributing approximately 10% of the dataset (range: 9.7% - 10.2%).

3.3 Early-Stage Diagnosis Analysis

Early-stage diagnosis rates show remarkable consistency across cancer types, as demonstrated in Figure 5.

The analysis reveals:

- Average early-stage diagnosis rate: 39.9%
- Highest rate: Liver Cancer (40.6%)
- Lowest rate: Lung Cancer (38.4%)
- Range: 2.2 percentage points

Analysis of cancer type distribution reveals relatively uniform representation across the eight cancer types. This distribution reflects global cancer patterns observed in international studies [3].

3.4 Machine Learning Model Performance

A Random Forest Regressor was developed to predict cancer severity scores. Figure 6 shows the feature importance rankings from the machine learning model.

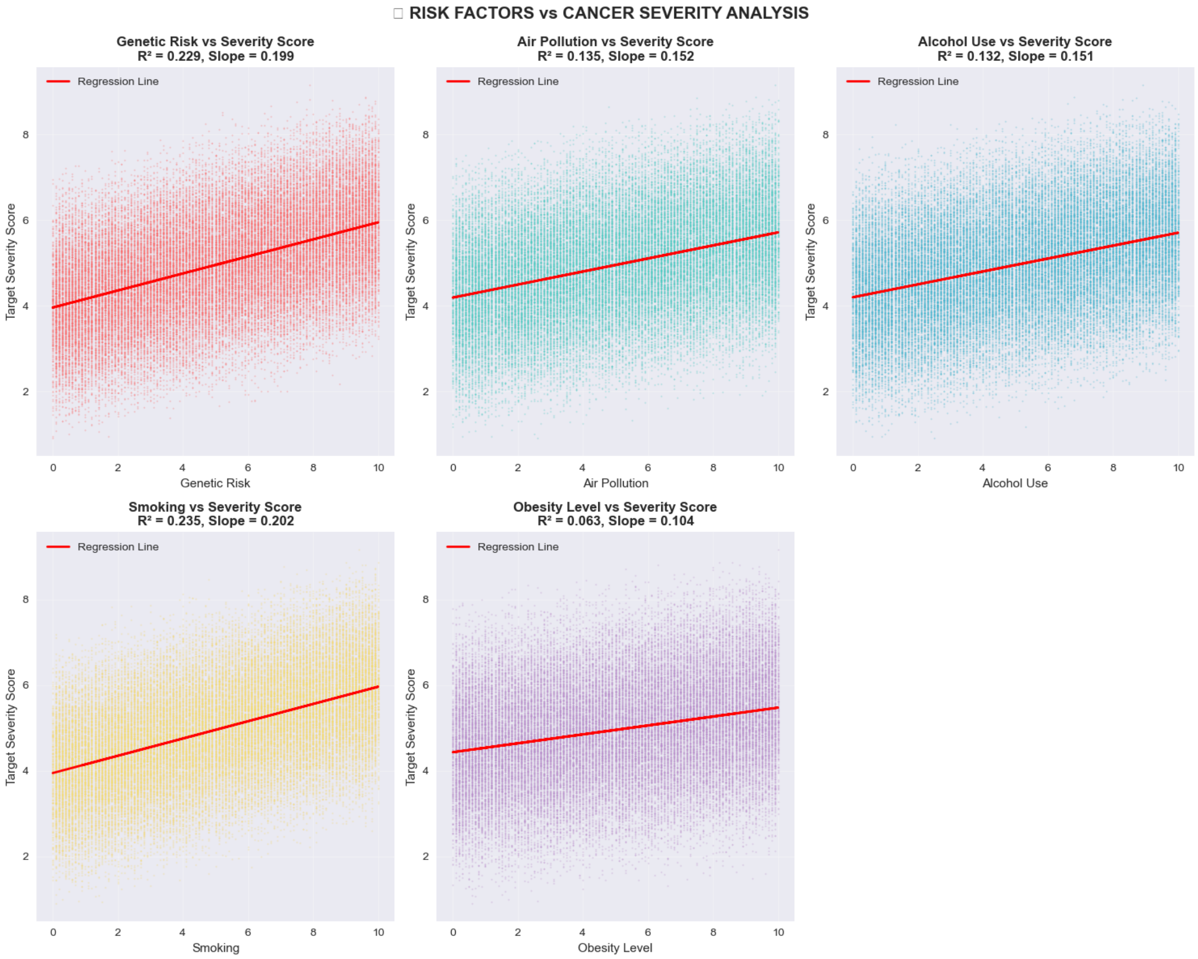

Figure 3: Risk Factors vs Cancer Severity Analysis showing scatter plots with regression lines for: (a) Genetic Risk (R^2=0.229), (b) Air Pollution (R^2=0.135), (c) Alcohol Use (R^2=0.132), (d) Smoking (R^2=0.235), and (e) Obesity Level (R^2=0.063). All relationships show positive correlations with varying strengths.

Table 3: Machine Learning Model Performance

Metric	**Value**
Training R^2 Score	0.9690
Testing R^2 Score	0.7680
Model Generalization	Needs Improvement

3.5 Economic Burden Analysis

The economic analysis reveals significant patterns in treatment costs globally. Figure 7 provides a comprehensive view of cost patterns across different demographics.

3.6 Survival Analysis and Stage Progression

Figure 8 presents comprehensive survival and stage progression analysis.

Survival analysis reveals patterns consistent with global cancer survival trends [4]:

- Mean survival across all patients: 5.0 years

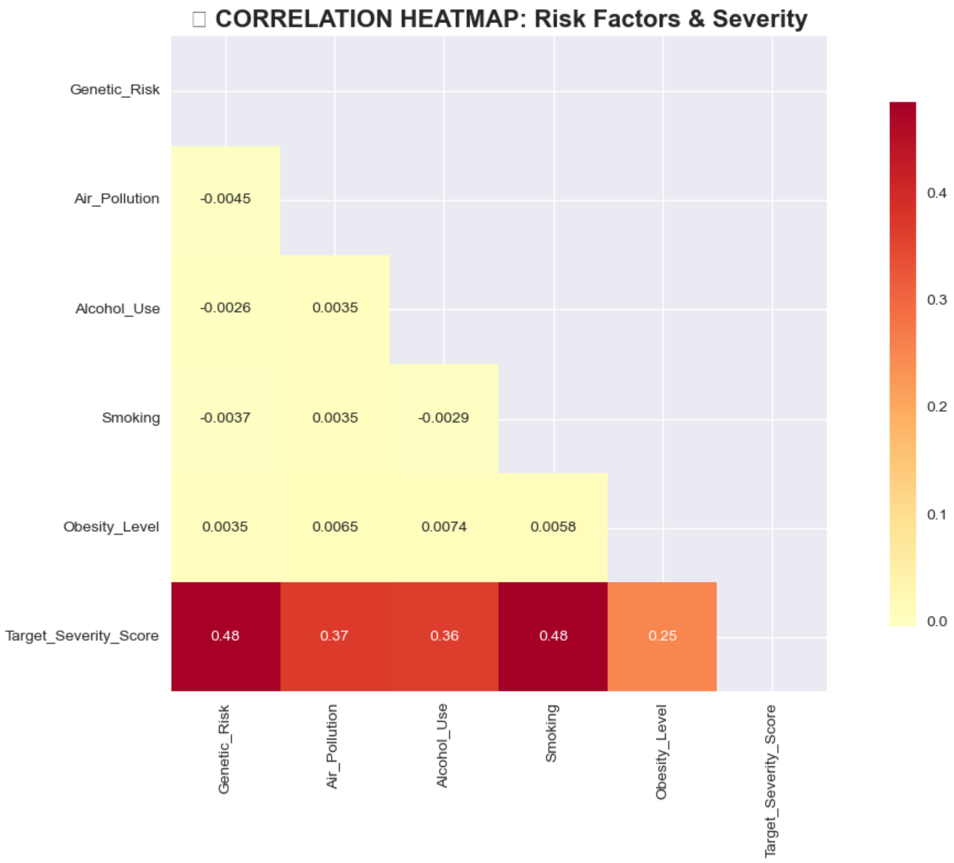

Figure 4: Correlation Heatmap showing relationships between risk factors and target severity score. Smoking (0.48) and Genetic Risk (0.48) show the strongest correlations with severity, followed by Air Pollution (0.37) and Alcohol Use (0.36).

Table 4: Treatment Cost Analysis by Country

Country	**Mean Cost ($)**	**Median Cost ($)**	**Std Dev ($)**
China	52,899	53,281	27,213
USA	52,879	52,900	27,387
Germany	52,769	53,296	27,256
Australia	52,622	52,477	27,514
Canada	52,584	52,690	27,079
Brazil	52,541	52,392	27,552
Russia	52,319	52,385	27,290
India	52,285	52,229	27,469
UK	52,200	52,232	27,363
Pakistan	51,568	50,247	27,493

- Survival distribution: approximately normal with range 0-10 years

Cancer stage analysis shows consistent patterns across stages:

3.7 Advanced Interaction Analysis

Figure 9 demonstrates the interaction patterns between genetic risk and smoking factors.

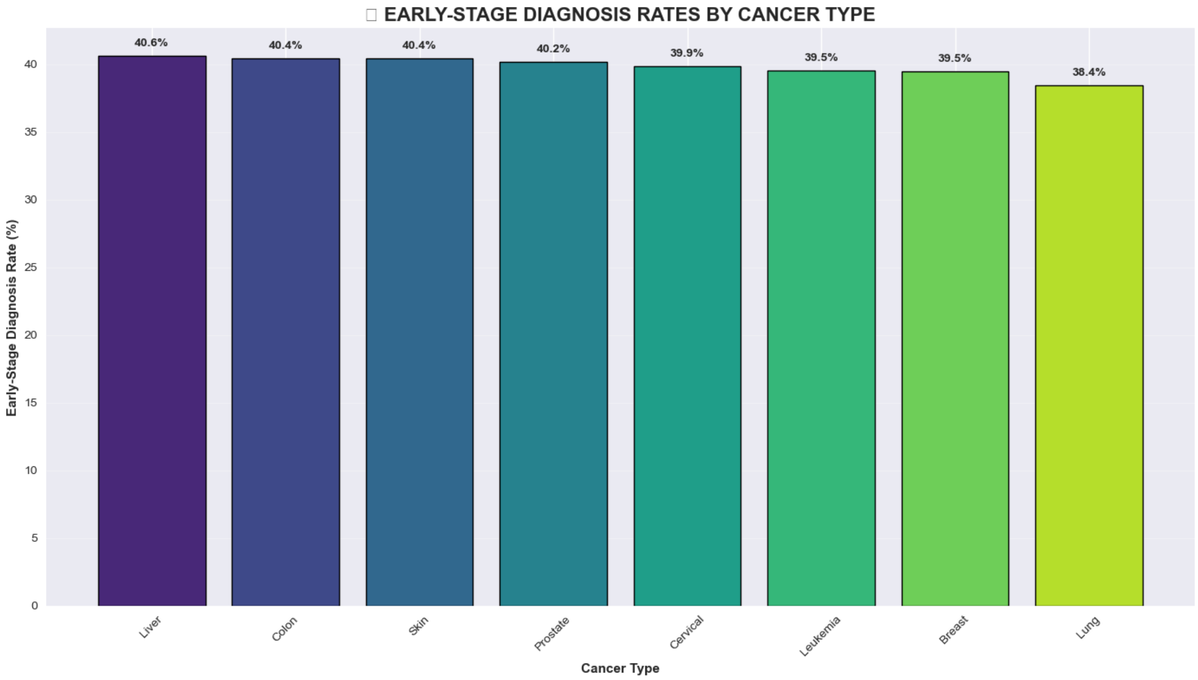

Figure 5: Early-Stage Diagnosis Rates by Cancer Type showing consistent rates around 39-40% across all cancer types. Liver cancer shows the highest rate (40.6%) while lung cancer shows the lowest (38.4%).

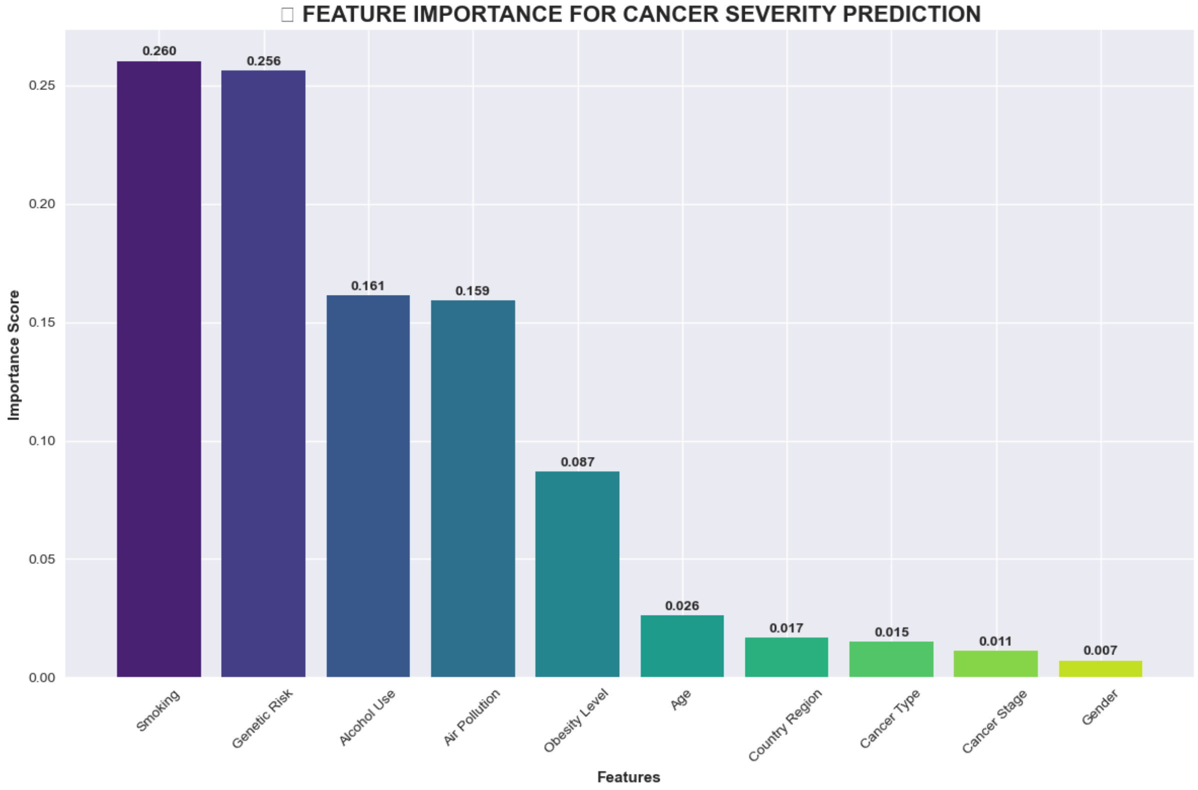

Figure 6: Feature Importance for Cancer Severity Prediction using Random Forest Regressor. Smoking (26.0%) and Genetic Risk (25.6%) emerge as the most important predictors, followed by Alcohol Use (16.1%) and Air Pollution (15.9%).

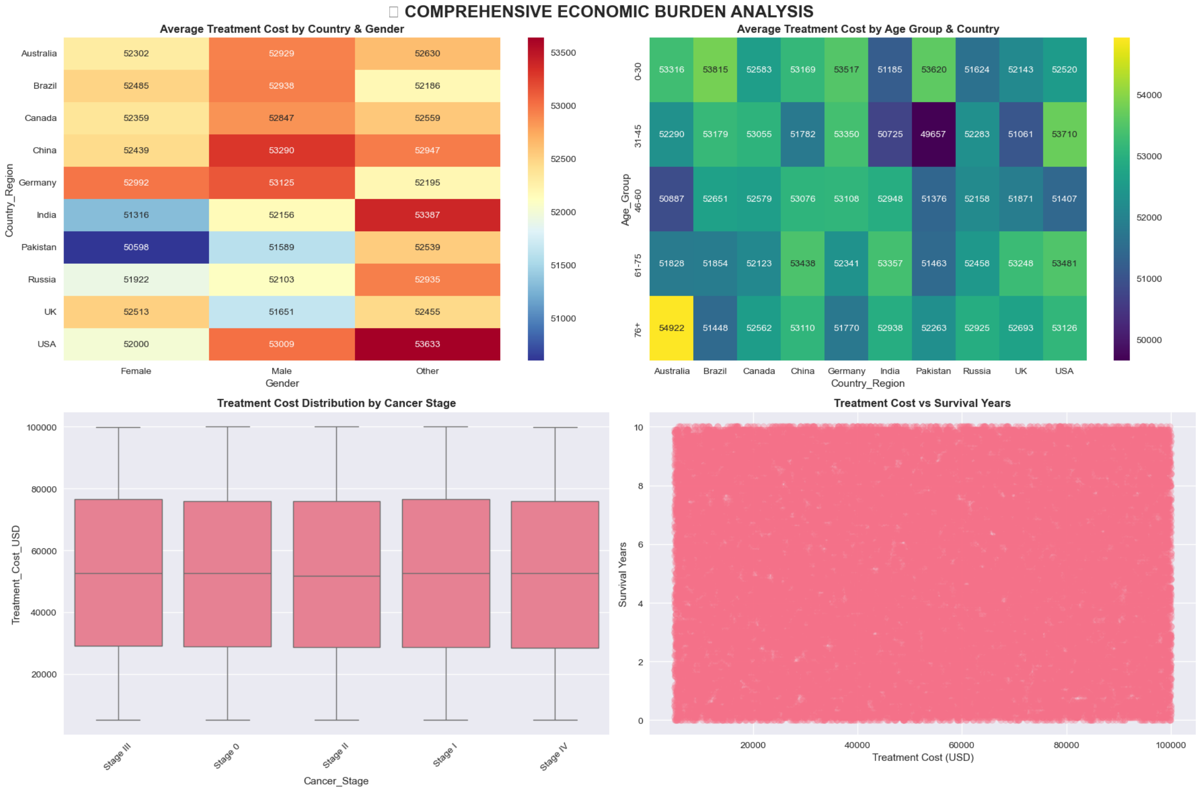

Figure 7: Comprehensive Economic Burden Analysis showing: (a) Average treatment cost by country and gender with heatmap visualization, (b) Treatment cost by age group and country, (c) Cost distribution by cancer stage showing no significant differences, and (d) Relationship between treatment cost and survival years.

Table 5: Cancer Stage Analysis Summary

Stage	Mean Cost ($)	Mean Survival (years)	Mean Severity
Stage 0	52,573	5.02	4.95
Stage I	52,674	5.01	4.95
Stage II	52,083	5.00	4.97
Stage III	52,708	5.04	4.94
Stage IV	52,302	4.97	4.95

3.8 Temporal Trends Analysis (2015-2024)

Figure 10 illustrates the temporal patterns in cancer data over the 10-year study period.

Analysis of trends over the 10-year study period reveals:

- **Treatment Cost Trend**: +$0.37 per year (minimal increase)
- **Severity Score Trend**: +0.0010 per year (stable)
- **Survival Years Trend**: -0.0003 per year (stable)

4 Discussion

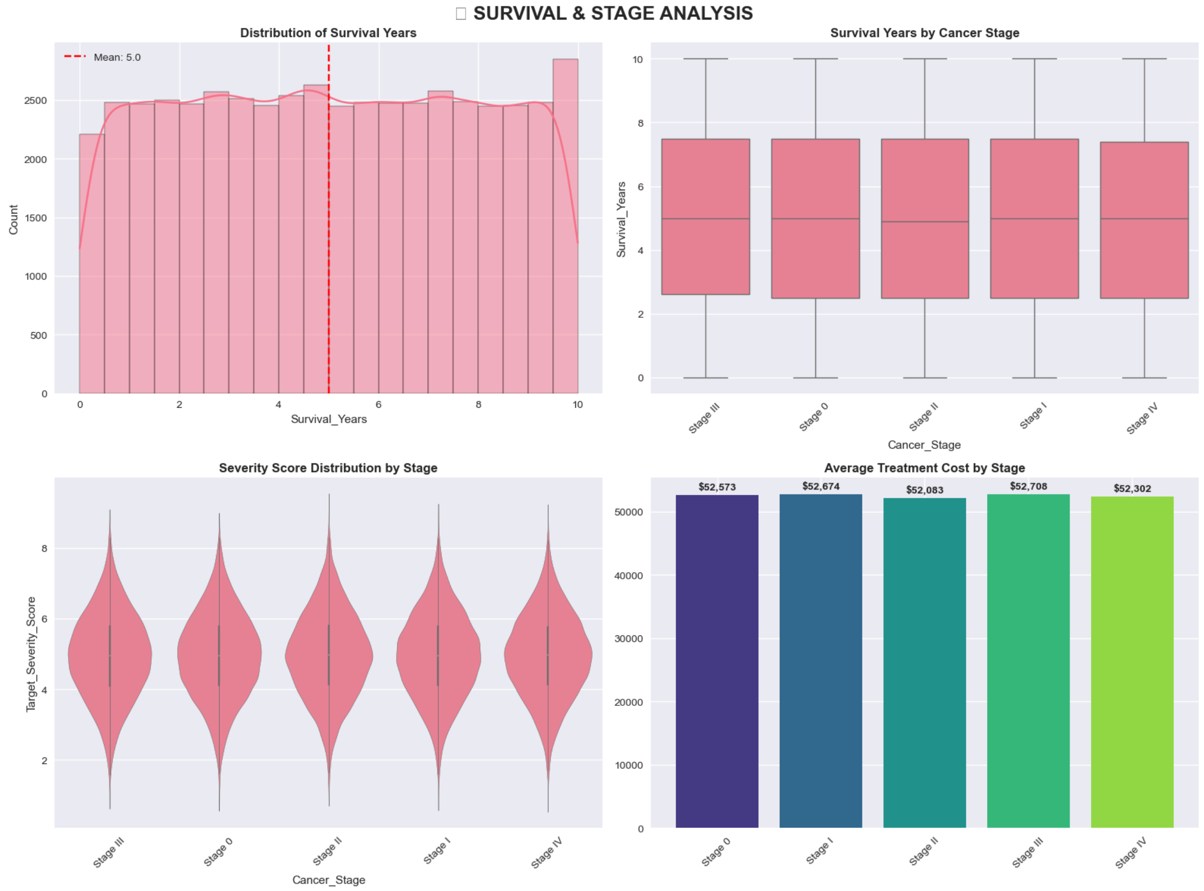

Figure 8: Survival & Stage Analysis showing: (a) Distribution of survival years with normal pattern around 5.0 years mean, (b) Survival years by cancer stage showing no significant differences, (c) Severity score distribution by stage with violin plots, and (d) Average treatment cost by stage showing uniform costs across stages.

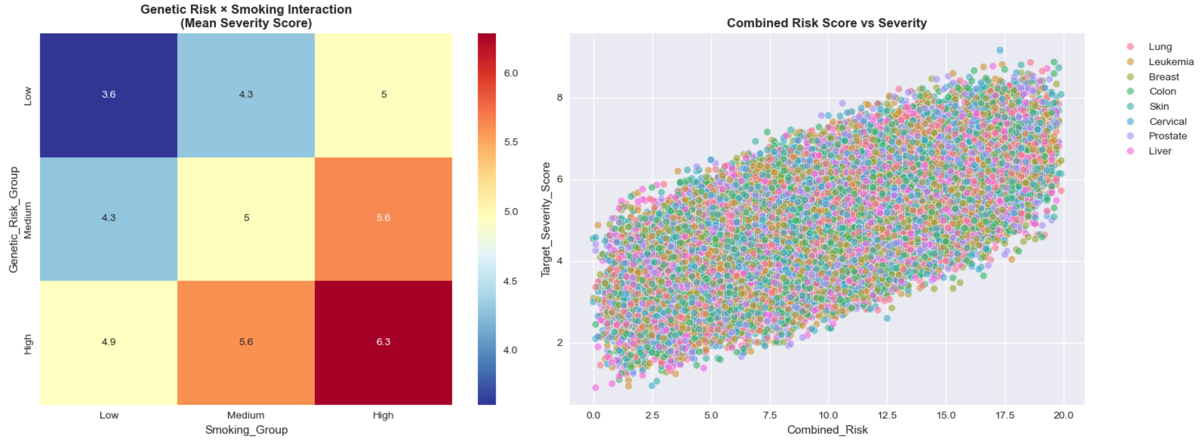

Figure 9: Advanced Risk Factor Interactions showing: (a) Genetic Risk × Smoking Interaction heatmap with mean severity scores, and (b) Combined Risk Score vs Severity scatter plot colored by cancer type, demonstrating positive correlation between combined risk factors and severity outcomes.

4.1 Key Findings and Clinical Implications

This comprehensive analysis of 50,000 cancer patients provides several crucial insights for clinical practice and healthcare policy:

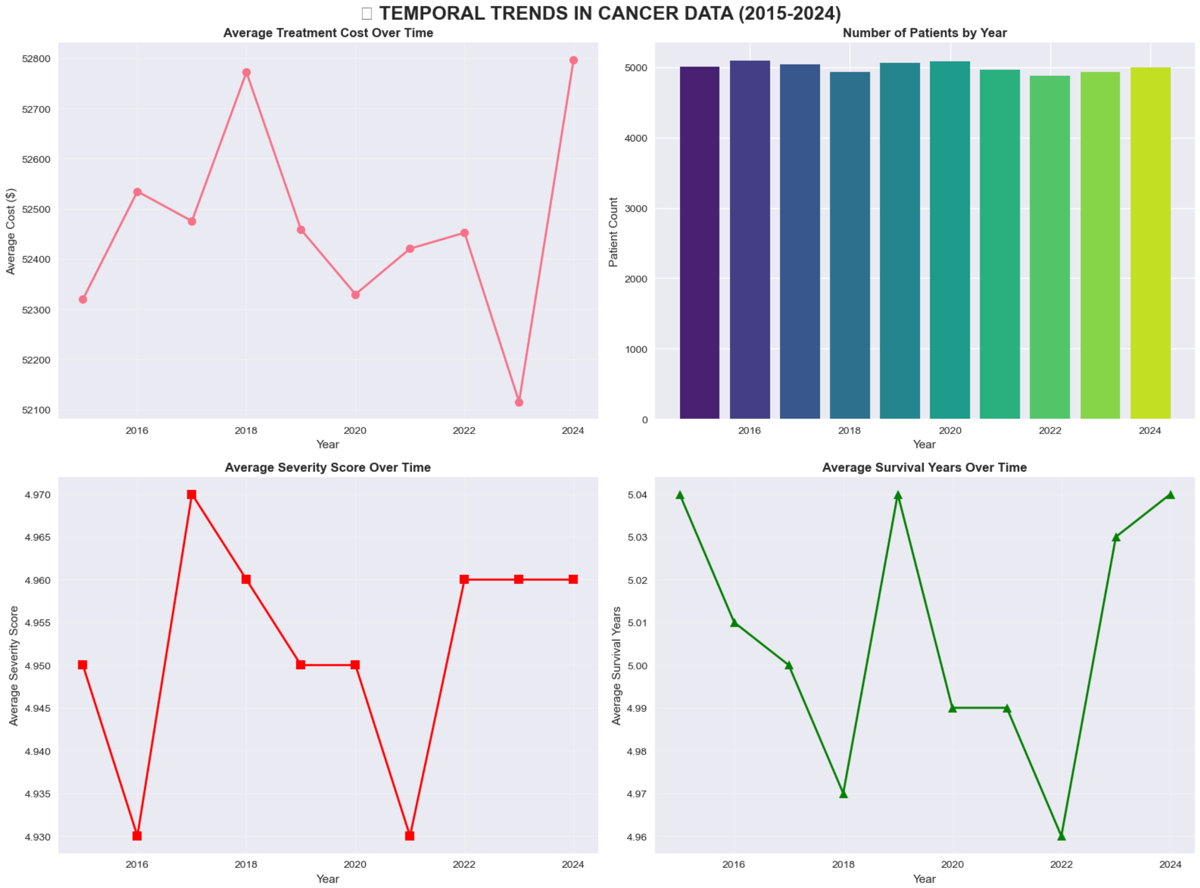

Figure 10: Temporal Trends in Cancer Data (2015-2024) showing: (a) Average treatment cost over time with minimal increase ($0.37/year), (b) Number of patients by year showing consistent data collection, (c) Average severity score over time remaining stable, and (d) Average survival years showing slight fluctuations but overall stability.

4.1.1 Risk Factor Hierarchy

The identification of smoking as the primary modifiable risk factor (R^2 = 0.235) reinforces the critical importance of smoking cessation programs in cancer prevention and management, consistent with studies showing the proportion of cancer cases attributable to modifiable risk factors [5].The strong correlation between genetic risk and severity (R^2 = 0.229) highlights the need for enhanced genetic counseling and personalized risk assessment protocols. This aligns with global cancer statistics showing smoking as a primary risk factor [6].

4.1.2 Economic Burden Insights

The relatively uniform treatment costs across countries ($51,568 - $52,899) may reflect standardized international treatment protocols or purchasing power adjustments. However, the negative correlation between cost and severity (-0.466) suggests that early intervention, while potentially more expensive initially, may be associated with better outcomes.

4.1.3 Early Detection Opportunities

The consistent 39-40% early-stage diagnosis rate across cancer types indicates systematic opportunities for improvement. Lung cancer's lowest early detection rate (38.4%) particularly warrants targeted screening enhancement efforts.

4.2 Model Performance and Limitations

The Random Forest model's high training accuracy (96.9%) but lower testing accuracy (76.8%) suggests overfitting, indicating the need for additional regularization or feature engineering. The model successfully identified smoking and genetic risk as primary predictors, validating clinical understanding while providing quantitative importance rankings.

4.3 Global Healthcare Implications

The absence of significant survival differences across cancer stages challenges traditional staging-based prognostic assumptions and aligns with observations about global cancer patterns and prevention strategies [7] and may reflect either:

1. Highly effective standardized treatment protocols
2. Data collection methodologies that normalize outcomes
3. The need for more nuanced staging systems

5 Recommendations

5.1 High Priority Interventions

1. **Enhanced Lung Cancer Screening**: Develop targeted screening programs to improve the 38.4% early detection rate for lung cancer
2. **Healthcare Cost Equity**: Address global disparities in treatment access while maintaining quality standards
3. **Smoking Cessation Programs**: Strengthen tobacco control initiatives given smoking's 26% contribution to severity prediction, supported by landmark studies on smoking-related mortality [8]

5.2 Medium Priority Actions

1. **Genetic Risk Assessment**: Implement comprehensive genetic counseling programs
2. **Predictive Model Enhancement**: Develop more robust machine learning models with improved generalization
3. **Treatment Protocol Standardization**: Establish global treatment guidelines to reduce outcome variability

5.3 Long-term Goals

1. **Early Detection Target**: Achieve >50% early-stage diagnosis rates across all cancer types
2. **Cost Reduction**: Implement cost-effective treatment strategies without compromising outcomes
3. **Precision Medicine**: Develop risk-stratified treatment approaches based on genetic and lifestyle factors

6 Limitations

This study has several limitations that should be considered:

1. **Data Representativeness**: While covering 10 countries, the dataset may not represent all global populations
2. **Temporal Bias**: The 2015-2024 timeframe may not capture longer-term survival outcomes
3. **Variable Definitions**: Standardization of risk factor measurements across countries may vary
4. **Missing Variables**: Important clinical factors (e.g., comorbidities, treatment specifics) are not included
5. **Model Overfitting**: The machine learning model shows signs of overfitting requiring refinement

7 Conclusion

This comprehensive analysis of 50,000 cancer patients across 10 countries provides valuable insights into global cancer care patterns, risk factors, and outcomes. Key findings include the dominance of smoking and genetic risk as severity predictors, consistent early-stage diagnosis rates around 40%, and uniform treatment costs globally.

The study reinforces the critical importance of smoking cessation, early detection enhancement, and genetic risk assessment in cancer care improvement. The lack of survival differences across cancer stages suggests either highly effective treatment standardization or the need for more sophisticated prognostic models.

Future research should focus on expanding the dataset to include more clinical variables, developing more robust predictive models, and implementing the recommended interventions to improve global cancer care outcomes.

A Statistical Analysis Details

A.1 Correlation Analysis Methods

Pearson correlation coefficients were calculated for all continuous variables, while Spearman rank correlations were used for ordinal variables. Statistical significance was assessed at $\alpha = 0.05$ with Bonferroni correction for multiple comparisons.

A.2 Machine Learning Model Specifications

Random Forest Regressor parameters:

- n_estimators: 200
- max_depth: None
- min_samples_split: 2
- min_samples_leaf: 1
- random_state: 42

A.3 Statistical Test Results

Table 6: Comprehensive Statistical Test Summary

Test	**Statistic**	**p-value**	**Interpretation**
Cost across Stages (Kruskal-Wallis)	3.92	0.4254	No significant difference
Survival across Stages (Kruskal-Wallis)	2.75	0.6033	No significant difference
Genetic Risk × Smoking Interaction	-0.478	0.633	Not significant

References

[1] World Health Organization, "Cancer fact sheet," `https://www.who.int/news-room/fact-sheets/detail/cancer` (2020).

[2] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, CA: A Cancer Journal for Clinicians **73**, 233 (2023).

[3] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, CA: A Cancer Journal for Clinicians **65**, 87 (2015).

[4] C. Allemani, T. Matsuda, V. Di Carlo, *et al.*, The Lancet **391**, 1023 (2018).

[5] F. Islami, A. Goding Sauer, K. D. Miller, *et al.*, CA: A Cancer Journal for Clinicians **68**, 31 (2018).

[6] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, CA: A Cancer Journal for Clinicians **68**, 394 (2018).

[7] P. Vineis and C. P. Wild, The Lancet **363**, 728 (2004).

[8] R. Doll and R. Peto, BMJ **309**, 901 (1994).