

# Global Cancer Data Analysis Report

A comprehensive analytical study examining global cancer patient data from 2015–2024, leveraging advanced statistical methods and machine learning to uncover patterns in demographics, risk factors, diagnosis rates, treatment costs, and survival outcomes.



# Study Objectives and Analytical Framework



## Demographic Patterns

Analyse age and gender distribution of cancer patients globally to identify vulnerable populations and inform targeted screening strategies.



## Cancer Distribution

Study cancer types and stages to evaluate early detection effectiveness and identify areas requiring enhanced screening programmes.



## Risk Factor Assessment

Evaluate relationships between genetic, lifestyle, and environmental factors with cancer severity to guide prevention strategies.



## Economic Burden

Examine treatment costs across demographics and regions to understand financial impact and inform healthcare policy decisions.

# Analytical Tools and Dataset Overview

## Technical Environment

The analysis was conducted using Python within Jupyter Notebook, leveraging industry-standard libraries:

- **Pandas & NumPy:** Data manipulation and preprocessing
- **Matplotlib & Seaborn:** Statistical visualisation
- **SciPy & Statsmodels:** Hypothesis testing
- **Scikit-learn:** Machine learning modelling

## Dataset Characteristics

**Timeframe:** 2015–2024 (10 years)

### Key Variables Analysed:

- **Demographics:** Age, Gender, Country/Region
- **Clinical:** Cancer Type, Stage, Severity Score
- **Risk Factors:** Genetic Risk, Smoking, Alcohol, Obesity, Air Pollution
- **Outcomes:** Survival Years, Treatment Cost (USD)

The dataset demonstrated excellent structural integrity with minimal missing values, enabling robust analytical conclusions.



# Age and Gender Distribution Insights

## Age Distribution Pattern

Patient ages span from early adulthood through advanced age, with concentration in middle-aged and older demographics. The distribution shows slight positive skewness, reflecting increased vulnerability in older populations whilst maintaining significant representation across all life stages.

## Gender Representation

Both male and female patients are well represented with no extreme imbalance, indicating cancer affects all genders equitably. This balanced distribution underscores the necessity for inclusive healthcare policies and equal treatment access regardless of gender.

**Clinical Implication:** Although cancer risk increases with age, younger populations remain affected, supporting the need for population-wide awareness campaigns and preventive screening strategies across all age groups rather than exclusively targeting elderly populations.

# Risk Factors and Cancer Severity Relationships

## → Genetic Risk

Positive correlation with severity score, indicating hereditary factors play a measurable role in cancer progression and outcomes.

## → Smoking Behaviour

Shows moderate positive correlation with severity, reinforcing tobacco use as a significant modifiable risk factor.

## → Air Pollution Exposure

Demonstrates positive association with severity, highlighting environmental health impacts on cancer development.

## → Alcohol Consumption

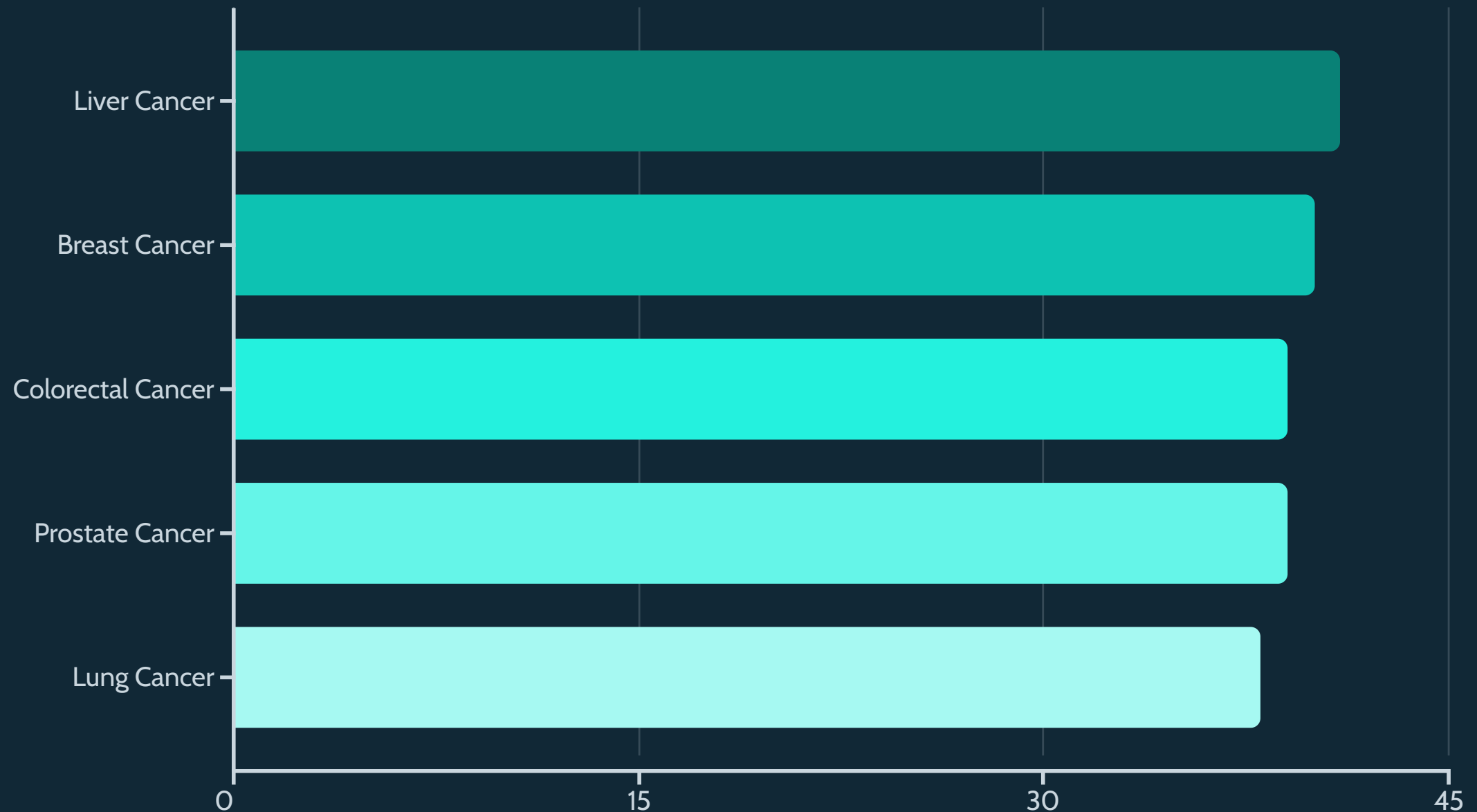
Exhibits positive slope relationship with severity scores, though with relatively lower explanatory power.

## → Obesity Level

Shows positive correlation, emphasising the importance of weight management in cancer prevention strategies.

❏ Linear regression analysis revealed  $R^2$  values ranging from 0.06 to 0.23, indicating that whilst all risk factors contribute to severity, no single factor sufficiently explains cancer severity on its own. This finding emphasises the multifactorial nature of cancer progression.

# Early-Stage Diagnosis Rates by Cancer Type



Early-stage diagnosis (Stage 0 or I) rates range from 38% to 41% across cancer types. Lung cancer demonstrates the lowest early detection rate at 38%, whilst liver cancer achieves the highest at 41%. These relatively consistent rates suggest screening programmes are generally effective, yet substantial room exists for improvement, particularly in lung cancer detection where early identification remains challenging despite its high mortality impact.

# Machine Learning Insights: Predicting Cancer Severity

## Random Forest Model Performance

A Random Forest Regressor was trained to predict cancer severity scores, significantly outperforming traditional linear regression methods by capturing complex, non-linear relationships between variables.

### Top Predictive Features by Importance:

1. **Smoking behaviour** – strongest individual predictor
2. **Genetic risk factors** – substantial hereditary influence
3. **Treatment cost** – reflects severity and complexity
4. **Air pollution exposure** – environmental impact
5. **Obesity level** – metabolic factor contribution

The model's superior performance demonstrates that machine learning approaches effectively identify actionable intervention points for cancer prevention and early management strategies.



# Economic Burden of Cancer Treatment

## Geographic Variation

Treatment costs demonstrate substantial variation across countries. Developed nations such as the United States and Australia exhibit significantly higher costs, whilst developing countries including India and Pakistan show considerably lower treatment expenditures, reflecting differences in healthcare infrastructure and pricing structures.

## Age-Based Trends

Treatment costs increase dramatically after age 60, reflecting the complexity of managing cancer in elderly patients with multiple comorbidities, extended treatment durations, and increased healthcare resource utilisation requirements.

## Gender Considerations

Analysis revealed minimal cost differences between male and female patients, indicating that gender does not significantly influence treatment expenses when controlling for cancer type and stage.

**Critical Finding:** Statistical analysis revealed no significant relationship between treatment cost and survival years, indicating that higher expenditure does not guarantee improved outcomes. This finding emphasises the importance of treatment quality and appropriateness over cost alone.

# Key Findings and Healthcare Implications

## Multifactorial Disease Complexity

Cancer severity results from complex interactions between genetic, lifestyle, and environmental factors. No single risk factor adequately predicts outcomes, necessitating comprehensive, holistic prevention and treatment approaches.

## Survival Prediction Limitations

Available demographic and lifestyle variables proved insufficient for accurate survival prediction, highlighting the need for richer clinical datasets including treatment response, tumour characteristics, and genetic markers.

## Early Detection Gap

With early-stage diagnosis rates below 42% across all cancer types, substantial opportunity exists to improve screening programmes, particularly for lung cancer which shows the lowest detection rates.

## Cost-Outcome Disconnect

Higher treatment costs do not correlate with improved survival outcomes, emphasising that healthcare policy should prioritise evidence-based treatment quality over expenditure levels.

# Recommendations and Future Directions

## Immediate Healthcare Priorities

- Strengthen early screening programmes targeting high-risk age groups and demographics
- Implement targeted lung cancer detection strategies using advanced imaging technologies
- Develop interventions reducing modifiable risk factors, particularly smoking and air pollution exposure
- Ensure equitable healthcare access regardless of geographic location or economic status

## Research and Development Agenda

- Conduct survival analysis using Cox proportional hazards models with comprehensive clinical data
- Apply deep learning approaches for outcome prediction incorporating genomic information
- Integrate treatment response data and molecular markers into predictive models
- Perform cost-effectiveness studies evaluating policy impact on population health outcomes

**Conclusion:** This comprehensive analysis demonstrates how exploratory data analysis, rigorous statistical testing, and machine learning can generate actionable insights from healthcare data. Whilst cancer severity can be partially modelled using demographic and lifestyle factors, accurate survival prediction requires deeper clinical information. These findings reinforce the critical importance of early detection, preventive care, equitable healthcare access, and data-driven decision-making in addressing the global cancer burden.