The Analysis of Genes, Geography, and the Clock: Untangling What Really Affects Cancer Patients Worldwide.

Introduction

This project explores how both biological traits and environmental conditions shape cancer outcomes. Using global patient data from 2015 to 2024, we investigate the influence of gender, diagnosis stage, day of diagnosis, and country-specific factors on survival rates, cancer types, and obesity trends. Through statistical methods including ANOVA, chi-square, and correlation analysis, we aim to understand whether it's our genetic makeup or the world around us that plays a bigger role in the cancer experience.

Dataset

Simple random sampling was used from the dataset. From a total of 33,600 cancer patient records (2015–2024), a random sample of 8,400 entries (25%) was selected. A further 6% sample (504 entries) was drawn from this subset for detailed analysis.

Research Question:

Do Genes Predict Cancer's Outcome, or Is Stage at Diagnosis the Real Game Changer?

Hypothesis:

1- : Is there a difference in cancer stage between males and females?

Null Hypothesis (H_0):

- There is no difference in the distribution of cancer stages between males and females.

Alternative Hypothesis (H_1):

- There is a difference in the distribution of cancer stage between males and females.

2- Is the stage at which a cancer is diagnosed has a stronger influence on survival and cancer severity than genetic predisposition?"

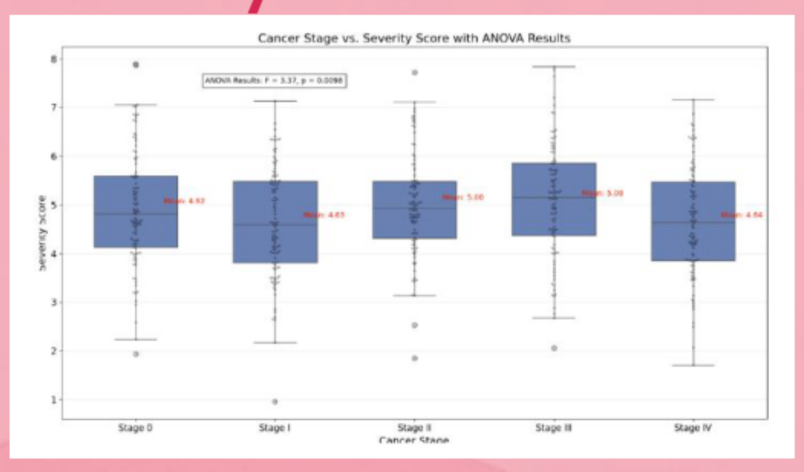
Null Hypothesis(H_0):

- There is no difference in the influence of genetic predisposition and cancer stage at diagnosis on survival and cancer severity.

Alternative Hypothesis (H_1):

- Cancer stage at diagnosis significantly influences survival and cancer severity more than genetic predisposition does.

Analysis



ANOVA Test – Severity Score by Cancer Stage:

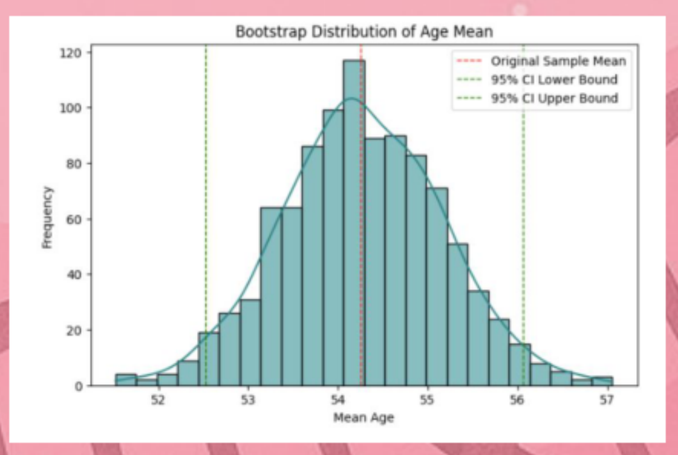
ANOVA results (F = 3.37, p = 0.0098) show a statistically significant difference in severity scores across stages. However, mean scores are similar (4.63–5.08) with no clear trend, and interquartile ranges largely overlap. This suggests that while stage affects severity statistically, the practical differences are modest and other factors likely contribute.

Correlation with Severity Score:

Genetic Risk shows the strongest positive correlation (0.83) with cancer severity. Smoking (0.52), Alcohol Use (0.47), and Treatment Cost (0.45) also show moderate positive correlations. Survival Years has a moderate negative correlation (-0.56), indicating higher severity links to shorter survival. Air Pollution and Obesity Level show weaker but notable contributions (0.2–0.3).

Genetic Risk vs. Target Severity Score:

The scatter plot shows a moderate positive linear relationship (r = 0.47) between genetic risk and severity score. Despite some scatter, data points generally trend upward, indicating that higher genetic risk is associated with higher severity. The distribution is fairly even, with mild clustering around mid-range values.



Bootstrap Distribution of Mean Age:

The histogram of 1,000 bootstrap samples shows a bell-shaped curve centered around 54.25 years, suggesting the average age is consistent across samples. The 95% confidence interval (52.54–56.03) indicates statistical stability, supporting the reliability of the sample mean. Similar bootstrapped analyses were conducted for Genetic Risk, Alcohol Use, Air Pollution, Obesity Level, and Smoking

Conclusion

This study shows that the stage of cancer at the time of diagnosis is very important in determining how severe the disease is (p = 0.0098). Although the differences between stages are not very large in real life, finding cancer early is still key to better treatment results.

On the other hand, genetic risk did not show a clear link with how long people survived (r = -0.04, p = 0.44). This suggests that simply having a genetic risk does not reliably predict how well someone will do. Overall, these results show that cancer outcomes depend on many factors, and knowing the cancer stage is more helpful for treatment than knowing about genetic risks.

The study also found a difference between men and women in how cancer stages were spread (p = 0.043). This may be due to biological differences, social factors, or how cancer is diagnosed. More research is needed to understand these gender differences in diagnosis and treatment.

Environmental and lifestyle factors like smoking, drinking alcohol, and air pollution were somewhat linked to cancer severity. This supports the idea that these factors can make cancer worse. However, other unmeasured factors like diet or access to healthcare might also play a role, so these results should be viewed carefully.

Interestingly, some cancers in more advanced stages (like Stage III) seemed to have better survival than those in earlier stages (like Stage 0). This surprising finding could be because of missing data, differences in treatment, or how the study was done. These problems show why it's important to have good data and follow patients over time in future studies

Details

Mira Micheal 23-101038
Mennatallah Amr23-101022
Maya Shalash 23-101004
Zeina Shaalan 23-101012

جامعة مصر للمعلوماتية
EGYPT UNIVERSITY OF INFORMATICS