Analyzing Demographic Influences on Frailty and Survival in UK Prostate Cancer Patients: A Synthetic Data Analysis

1 Introduction

Cancer remains a significant public health challenge in the UK and across other high-income countries [1]. As the leading cause of death among both men and women, it poses a critical area for medical research and public health interventions. Specifically in the UK, cancer is responsible for one in four premature deaths among individuals aged 30 to 69 years, marking its significant impact on society [2].

To address the complexities of cancer treatment and management, comprehensive and detailed data are essential. The Simulacrum dataset, developed by Health Data Insight (HDI) with support from AstraZeneca (AZ) and IQVIA, fulfills this need by providing a robust synthetic dataset that simulates the patient records held by the National Disease Registration Service (NDRS) of NHS England. Covering the years 2016 to 2019, this dataset includes synthetic records of cancer patients that encompass tumor diagnoses, treatments, and genetic information [1]. Crucially, these records replicate the appearance and structure of real patient data but contain no actual patient information, ensuring complete confidentiality and privacy.

The primary objective of this research is to delve into the Simulacrum data to analyze and understand the factors that influence frailty and overall survival rates of prostate cancer patients in the UK. This study employs statistical modeling techniques to identify key demographic and clinical predictors that significantly impact patient outcomes. The insights derived could be instrumental in guiding targeted interventions and shaping policy decisions, ultimately aiming to enhance patient management and resource allocation within the healthcare system. By leveraging the power of synthetic data, this research offers valuable insights intended to mitigate the impact of prostate cancer on public health, providing a basis for improved preventative measures and treatment strategies.

2 Literature Review

Cancer Research: Epidemiology and Treatment Outcomes Cancer remains a major health challenge in the UK, and its impact is growing as the population ages. According to projections in the "Cancer in the UK: Overview 2023" report[2], the number of new cancer cases is expected to rise significantly. Currently, about 384,000 people are diagnosed each year, and this is anticipated

[1] https://simulacrum.healthdatainsight.org.uk/using-the-simulacrum/requesting-data/
[2] https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk

to increase to over half a million by 2040, a one-third rise. The aging population largely drives this expected increase; the percentage of cancer cases in those aged 70 and older is predicted to rise from 50% to 60% by 2040 (see Figure 1).

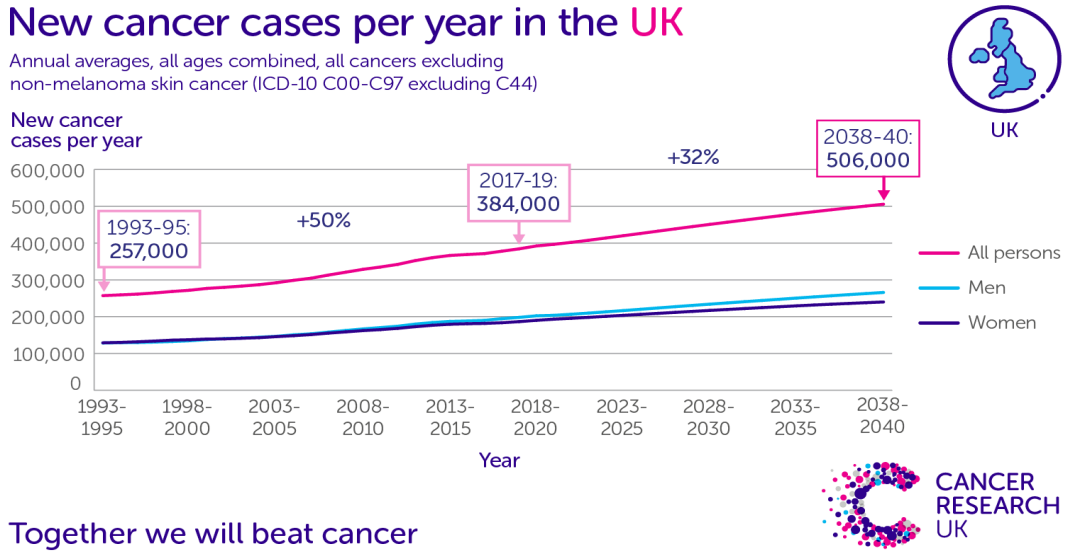


Figure 1: Projected annual increase in cancer cases in the UK, showing the growing challenge for public health.

2.1 Focus on Prostate Cancer

Prostate cancer is the most common type of cancer among men in the UK and many other parts of the world [3]. It primarily affects older men, with the median age at diagnosis being 66 years. In 2018, the UK's prostate cancer mortality rate was 40 per 100,000 men, which is higher than the European average of 33 deaths per 100,000 men, placing it ninth among 31 European countries (see Figure 2) [4].

Frailty is especially relevant in prostate cancer management due to its prevalence among older men. Our study classified patients based on a frailty score derived from age, comorbidity, and performance metrics [5]. This approach allows for personalized treatment plans that consider both the severity of the cancer and the patient's overall health [6].

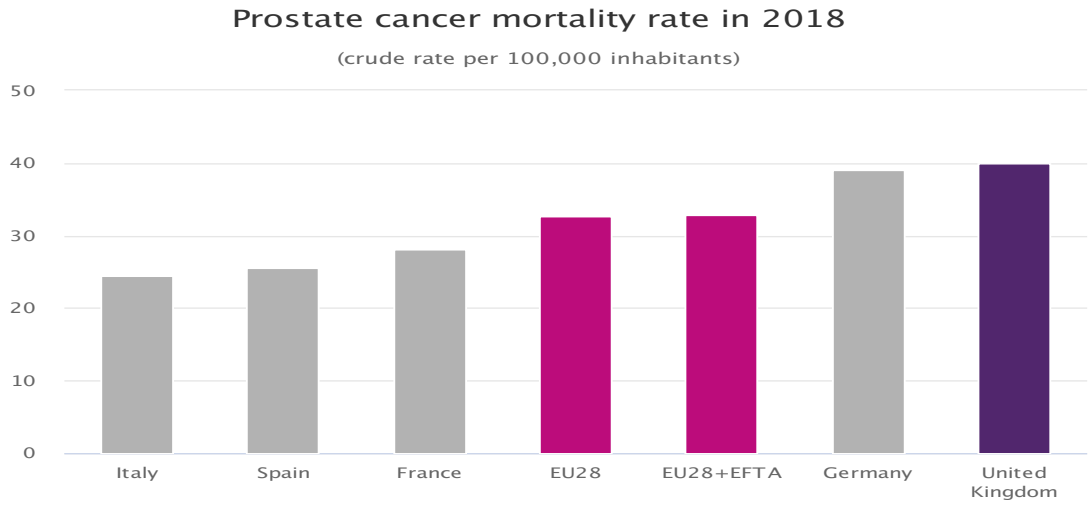


Figure 2: Prostate cancer mortality rates in the UK compared to the EU average, underlining the need for improved healthcare strategies.

2.2 Use of Synthetic Data in Research

The advent of synthetic data has revolutionized medical research, particularly in fields requiring extensive data but facing privacy constraints, like cancer research. The Simulacrum dataset provides realistic, anonymized data that mimics real patient records, enabling researchers to conduct detailed studies without compromising patient privacy. This dataset is invaluable for prostate cancer research, as it includes comprehensive information on patient demographics, tumor genetics, and treatment outcomes. Synthetic data allows researchers to simulate diverse clinical scenarios and treatment outcomes, providing insights crucial for enhancing patient care and advancing treatment methods.

3 The Dataset

This section details the structure, description, and preparation of the Simulacrum dataset used in our study. Figure 3 shows the flowchart of our data preparation and analysis process.

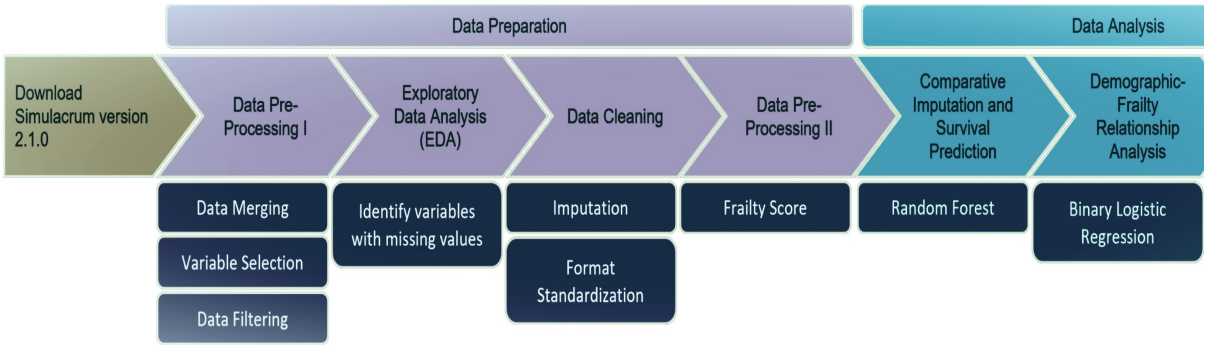


Figure 3: Flowchart of the data preparation and analysis in our study.

3.1 Data Structure and Description

The Simulacrum dataset, version 2.1.0, is a carefully crafted synthetic dataset designed to mimic real cancer patient records managed by NHS England's National Disease Registration Service (NDRS). It includes 11 distinct categories of data, such as patient demographics, tumor characteristics, genetic information, and treatment details. This diversity and complexity accurately reflect the real-world scenarios found in cancer patient data.

In this study, we primarily focus on two specific parts of the dataset, `SIM_AV_PATIENT` and `SIM_AV_TUMOUR`. These parts contain essential details such as patient registration and tumor data. We use these datasets in conjunction, linking them through a unique identifier, 'PATIENTID', which results in a comprehensive dataset encompassing 179,127 individuals across 13 different types of information. This dataset provides critical information including gender, ethnicity, and key clinical details such as diagnosis and vital status dates. Specifically, we concentrate on prostate cancer, utilizing ICD10 codes to identify relevant cases. This targeted approach allows us to deeply analyze how cancer impacts patients, evaluate the effectiveness of treatments, and assess overall patient outcomes based on a range of clinical and demographic factors.

3.2 Data Preparation

To prepare the dataset for in-depth analysis, we followed a structured approach that included several crucial steps:

1. **Data Merging and Variable Selection:** We combined the patient and tumor datasets using `PATIENTID` as the key. This allowed us to align records from both datasets. We specifically selected variables important for survival analysis, binary logistic regression, and random forest models, such as demographics, vital statuses, and diagnosis dates.

2. **Data Filtering:** We refined the dataset to only include records indicating whether patients were alive or deceased, marked as `A` (Alive) or `D` (Deceased). We also narrowed our focus to prostate cancer by selecting records with the ICD10 code `C61`.

3. **Creation of New Variables:** To further categorize patient outcomes, we introduced new variables:

   - `STATUS`: This identifies whether patients were alive or deceased as of a specific endpoint date.
   - `TIME`: This calculates the time from diagnosis to either the endpoint date or the date of death, expressed in years.

4. **Handling Missing Values and Data Imputation:** We addressed missing data in key fields like `ETHNICITY` and `PERFORMANCESTATUS`. Missing entries were filled based on logical imputation strategies to ensure data completeness and accuracy.

5. **Computation of Frailty Scores:** We calculated health and frailty scores using factors such as age, the Charlson Comorbidity Index, and performance levels. These scores were summed to determine a total frailty score, which classified patients as `Frail` or `Nonfrail`.

These steps ensured that the dataset was well-prepared for complex statistical analysis, enabling us to extract detailed insights about prostate cancer patient outcomes.

4 Data Analysis

This section examines the analysis conducted on the prostate cancer dataset, specifically focusing on age, vital status, and frailty. These factors are crucial for understanding the management and impact of prostate cancer among the elderly in the UK. We used a combination of descriptive statistics and visual tools to clarify the relationships and outcomes associated with prostate cancer in older adults.

Analysis Techniques: Our approach involved detailed descriptive statistics to outline the distribution and central tendencies of important variables such as age, deprivation, frailty scores, and survival status. To help interpret these characteristics more intuitively, we employed histograms and bar charts. These visuals effectively highlight the distribution and frequency of key data points in the dataset, as illustrated in Figure 4.

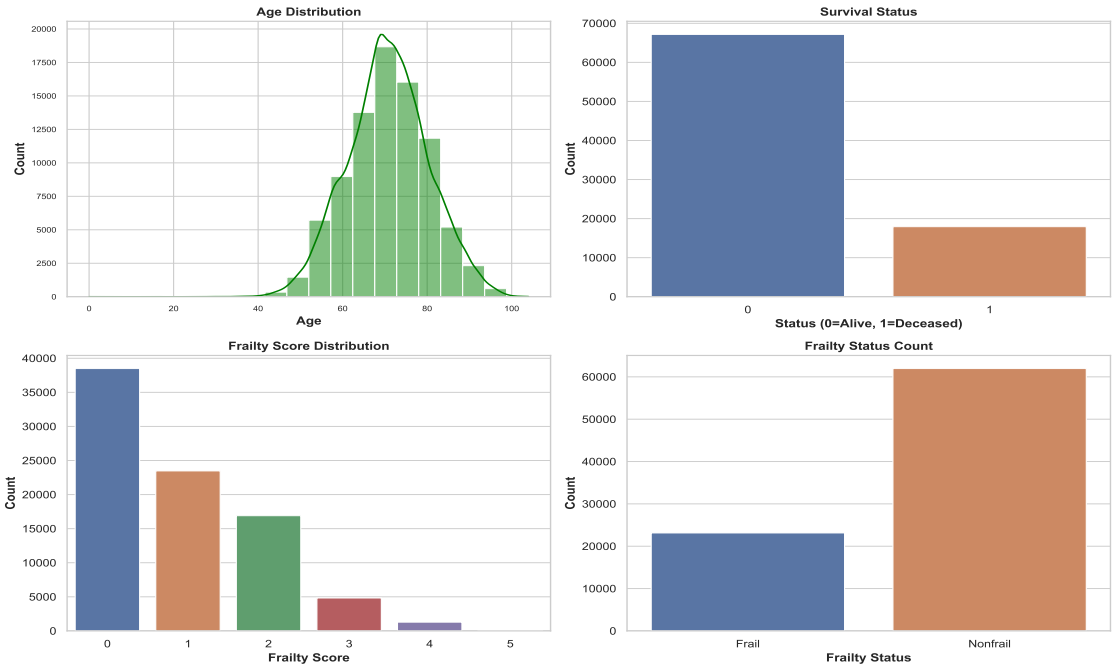


Figure 4: Visual summary of key metrics in the prostate cancer dataset: Age Distribution, Survival Status, Frailty Score Distribution, and Frailty Status Count.

Findings: The analysis provides several important insights:

- **Age Trends**: Age is a primary risk factor in prostate cancer incidence and progression. The average age of the cohort, around 70.74 years, underscores the disease's prevalence in older adults. This variable helps us examine how age influences treatment outcomes and survival rates, addressing the research question of how demographic factors impact patient health outcomes.
- **Survival Status**: Analyzing the survival status of patients provides insights into the effectiveness of current treatment protocols and the overall manageability of the disease. Understanding which factors contribute to higher survival rates can guide improvements in patient care and inform policy decisions regarding healthcare resource allocation.
- **Frailty Distribution**: Frailty scores offer a quantitative measure of a patient's health resilience and vulnerability. Investigating this distribution helps identify patients who might benefit from more aggressive or specialized treatment plans. It also contributes to the broader research question regarding the relationship between clinical characteristics (like frailty) and treatment outcomes. The prevalence of lower frailty scores and their impact on treatment options and health outcomes emphasizes the need for tailored treatment approaches based on individual health assessments.

These variables collectively provide a comprehensive view of how prostate cancer affects elderly patients and how different factors influence their treatment outcomes. By focusing on these specific aspects, our research can offer actionable insights that help refine clinical strategies for managing prostate cancer. This includes developing personalized treatment plans that consider each patient's unique health status and prognostic factors, ultimately aiming to improve survival rates and quality of life for older adults with prostate cancer. The findings stress the importance of integrating a holistic approach in clinical practice, where decisions are informed by a thorough understanding of how age and frailty interact with patient outcomes.

5 Modeling Section: Frailty and Survival Analysis in Prostate Cancer Patients

This section delves into the use of Binary Logistic Regression (BLR) and Random Forest (RF) models to analyze how demographic and clinical factors influence frailty and overall survival among prostate cancer patients in the UK. Utilizing the Simulacrum dataset, which replicates real patient data without compromising privacy, we explore two critical research questions.

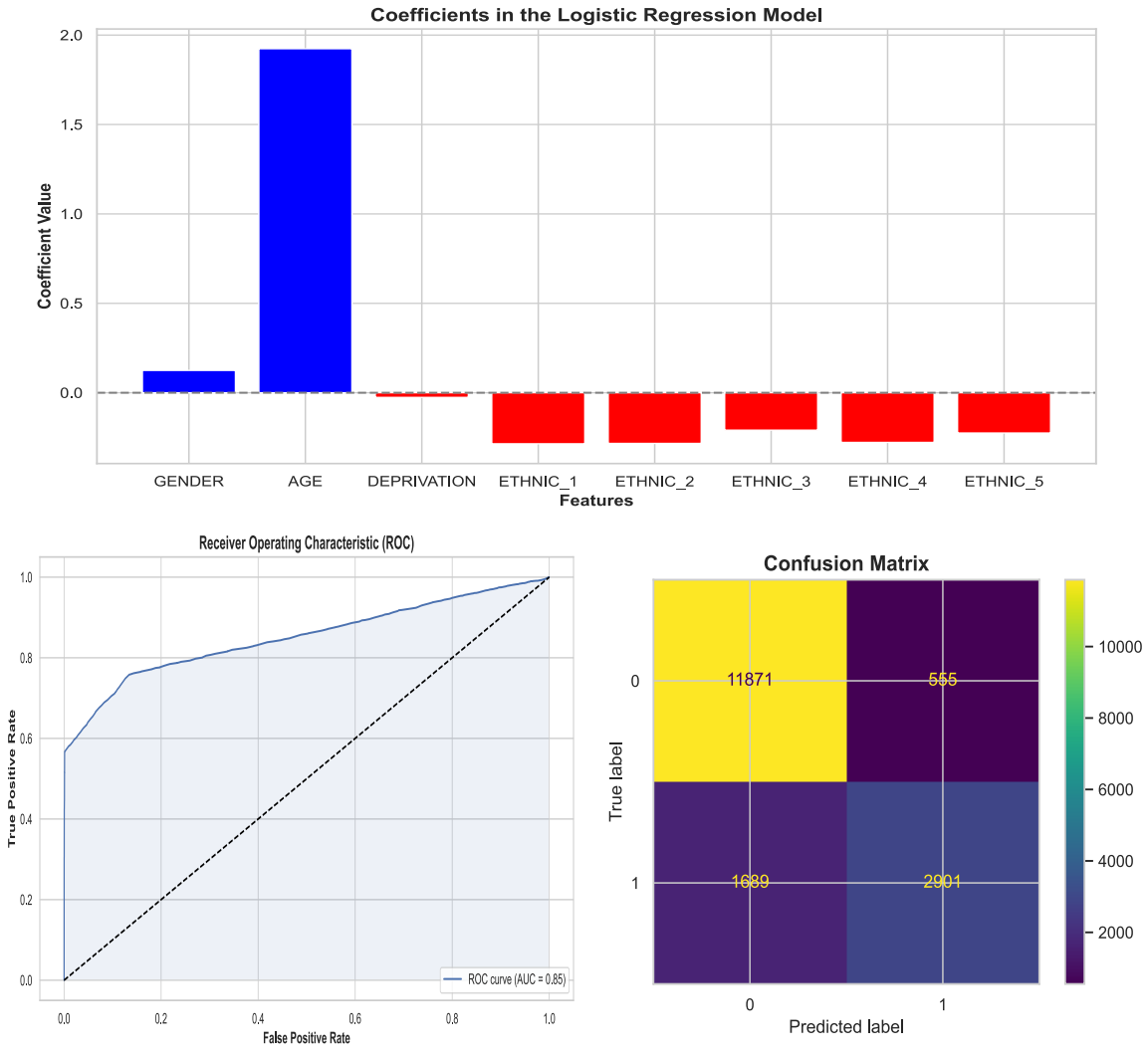
Research Question 1: What are the key demographic factors that influence the frailty score of prostate cancer patients?

Our analysis using Binary Logistic Regression (BLR) investigated the impact of age, gender, ethnicity, and deprivation on frailty scores among prostate cancer patients. Here are the key insights derived from the model:

- **Age as a Predictor**: The positive coefficient for age confirms that older patients are more likely to experience higher levels of frailty. This result is consistent with clinical expectations that frailty increases with age, underscoring the need for age-specific medical interventions.
- **Socio-economic and Ethnic Influences**: Negative coefficients for deprivation and certain ethnic categories suggest that higher socio-economic status and specific ethnic backgrounds may confer protective effects against frailty. These findings highlight the importance of considering socio-economic and ethnic factors in patient care strategies.
- **Model Performance and Accuracy**:
  - The ROC curve, depicted in Figure 5a, demonstrates an AUC of 0.85, indicating excellent model accuracy in discriminating between different frailty levels based on demographic factors.
  - The overall model accuracy is 87%, validating its effectiveness in clinical applications for predicting frailty.

- **Confusion Matrix Insights**: As shown in Figure 5b, the confusion matrix reveals a high number of true positives and true negatives. This suggests that the model is highly reliable in accurately classifying patients according to their frailty status, which is critical for planning appropriate interventions and managing healthcare resources efficiently.

These analytical outcomes are instrumental in enhancing our understanding of the demographic determinants of frailty in prostate cancer patients. They support the development of tailored healthcare interventions aimed at improving clinical outcomes and optimizing patient care, particularly for vulnerable groups identified through the model.



(a) (b)

Figure 5: Visualizations from the Binary Logistic Regression analysis showing the Coefficients Plot, ROC Curve, and Confusion Matrix.

Research Question 2: Can we predict the overall survival of prostate cancer patients based on

their frailty and demographic factors?

We utilized the Random Forest (RF) model to predict survival times, integrating frailty scores with demographic and deprivation indices. The analysis yielded significant insights:

- **Age as a Predominant Factor**: Age was identified as the most influential predictor in the model, highlighting its critical role in determining survival outcomes for prostate cancer patients. This finding aligns with clinical understandings that older age is often associated with reduced survival rates, emphasizing the need for age-specific interventions.
- **Model Accuracy and Effectiveness**: The Mean Squared Error (MSE) of the model was 2.43, which underscores the model's accuracy in forecasting survival times. This level of precision indicates that the model is highly effective at predicting how long patients might live following their diagnoses, which is essential for planning and optimizing treatment strategies.

The results from the Random Forest analysis are vividly displayed in Figure 6, where the feature importance graph shows the dominance of age and frailty in predicting survival outcomes. Additionally, the distribution of predicted survival times helps us understand the general prognosis for different patient groups based on their demographic characteristics and frailty levels.

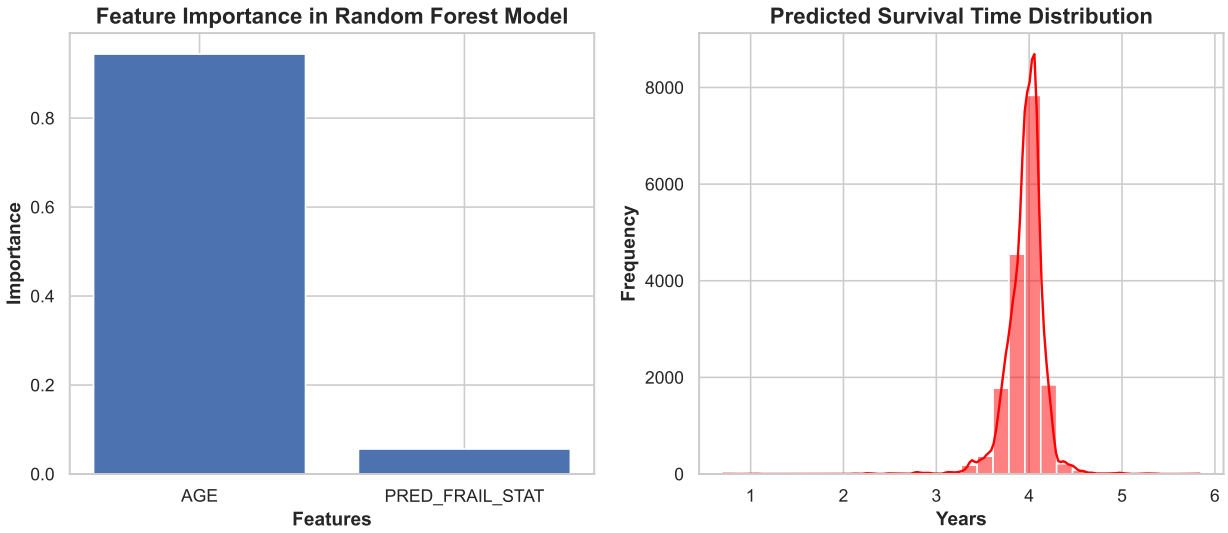


Figure 6: Random Forest model results, illustrating feature importance and the predicted survival time distribution, highlight how age and frailty critically influence survival predictions.

The insights derived from both the BLR and RF models are pivotal in understanding how demographic and clinical factors such as age and frailty impact the survival of prostate cancer patients. This comprehensive analysis aids in refining clinical strategies and enhancing patient care by providing robust evidence to tailor treatment plans according to individual patient profiles, ultimately improving treatment outcomes and resource allocation in healthcare settings.

5.1 Insights

The modeling analyses offer valuable insights into the management of prostate cancer, which can significantly influence clinical strategies and health policy:

- **Demographic Influence**: The BLR model highlights that age is a critical factor influencing both frailty and survival outcomes. This finding underscores the necessity for treatment strategies that are tailored to accommodate the physiological and socio-economic conditions of older adults. Age, as a determinant of frailty, suggests that older patients may require more personalized care approaches to manage their condition effectively.
- **Survival Predictions**: The Random Forest model's ability to accurately predict survival times demonstrates its utility in oncological settings. Such precise forecasting aids in better prognostication and informed treatment planning, allowing healthcare providers to adjust treatments based on predicted outcomes.
- **Strategic Healthcare Implementation**: The insights derived from these models can inform policy decisions and enhance individual patient care strategies. By understanding how demographic and clinical variables affect prostate cancer outcomes, healthcare systems can improve resource allocation and patient management, leading to better overall healthcare outcomes.
- **Integration of Frailty in Clinical Assessment**: The concept of frailty, as supported by recent research suggesting its significance in predicting disease progression and mortality among aging populations, aligns with our findings [7]. Incorporating frailty assessments into routine clinical practice can provide a more holistic view of a patient's health status, thereby facilitating more customized and effective treatment plans.

These findings enhance our understanding of prostate cancer management and highlight the importance of integrating comprehensive assessments of frailty and other demographic factors into clinical practice. This approach not only personalizes treatment but also optimizes outcomes for patients with prostate cancer.

6 Discussion

Our analysis with Binary Logistic Regression and Random Forest models has highlighted crucial associations between demographic factors, frailty, and survival in prostate cancer patients. Age emerged as the most significant predictor, corroborating clinical observations that prostate cancer disproportionately impacts older men, affecting their frailty scores and survival rates adversely.

This underscores the critical need for age-tailored patient management strategies in the UK's public health framework.

Significance of Socio-economic and Ethnic Factors: The observed protective effects against frailty for certain socio-economic statuses and ethnic backgrounds raise important questions. These factors appear to confer resilience against the severity of frailty, suggesting that socio-economic and cultural variables play a role in patient outcomes. Further investigation into these relationships could provide deeper insights into how to enhance patient care and optimize treatment strategies, ultimately leading to better health outcomes and more efficient resource utilization in healthcare.

Limitations: While our study provides valuable insights, it's essential to recognize the constraints associated with using the Simulacrum synthetic dataset:

- **Generalizability:** The synthetic nature of the data may not capture the full spectrum of real-world variability and the intricate interplay of clinical factors. This limitation may affect the applicability of our findings to actual patient populations.
- **Detail and Depth:** Although synthetic data allows for broad-scale analysis without compromising patient privacy, it might not include all the nuanced and rare conditions present in real clinical settings. This could lead to potential oversights in understanding complex patient scenarios.

These considerations underscore the need for careful interpretation of the results and suggest the importance of validating these findings with real-world data before they are used to inform clinical practice or policy changes. The insights gained, however, remain instrumental in guiding further research and developing hypotheses for future empirical studies that could validate and expand upon our findings.

7 Conclusion

This case study effectively leveraged Binary Logistic Regression and Random Forest models to uncover significant demographic and clinical predictors of frailty and survival among prostate cancer patients in the UK. Our findings confirmed that age is a critical determinant of both frailty and survival outcomes. Additionally, socio-economic and ethnic factors were found to influence frailty levels, suggesting that broader demographic contexts significantly impact patient health outcomes.

These insights are pivotal for crafting targeted interventions and enhancing healthcare strategies within the UK. By identifying which factors most significantly affect prostate cancer outcomes,

healthcare providers can better tailor treatments to individual patient needs, potentially improving overall survival rates and quality of life for patients.

Future Work: To build on the findings of this study, future research should aim to validate these results with real patient data. This step is crucial for confirming the patterns observed in the synthetic dataset and ensuring the applicability of our models to real-world settings. Suggested directions for further research include:

- **Longitudinal Studies**: These studies would track prostate cancer patients over time to gain more nuanced insights into the progression of frailty and its direct impact on survival. Such research could help in understanding the long-term effects of various demographic and clinical factors on patient outcomes.
- **Interaction Effects**: Investigating how demographic factors interact with specific clinical interventions could shed light on more complex dynamics that influence treatment efficacy. This research could identify potential synergies or conflicts between patient characteristics and treatment modalities.
- **Comparative Studies**: Conducting comparative analyses across different healthcare settings or even different countries could help identify best practices in prostate cancer management. These studies might reveal transferable strategies that could be adapted to improve healthcare delivery and policy in the UK and beyond.

The continuation of this research will not only deepen our understanding of the dynamics influencing prostate cancer management but also enhance the precision of predictive models. Ultimately, such efforts aim to optimize clinical practices and public health strategies, ensuring that prostate cancer patients receive the most effective and personalized care possible.

References

[1] W. H. Organization *et al.*, "Global health estimates: leading causes of death, cause-specific mortality, 2000–2019," 2020.

[2] P. H. England, "Chapter 2: major causes of death and how they have changed," 2017.

[3] L. S. Graham, J. K. Lin, D. E. Lage, E. R. Kessler, R. B. Parikh, and A. K. Morgans, "Management of prostate cancer in older adults," *American Society of Clinical Oncology Educational Book*, vol. 43, p. e390396, 2023.

[4] J. Ferlay, M. Colombet, I. Soerjomataram, T. Dyba, G. Randi, M. Bettio, A. Gavin, O. Visser, and F. Bray, "Cancer incidence and mortality patterns in europe: Estimates for 40 countries and 25 major cancers in 2018," *European journal of cancer*, vol. 103, pp. 356–387, 2018.

[5] J. K. Mak, R. Kuja-Halkola, Y. Wang, S. Hägg, and J. Jylhävä, "Can frailty scores predict the incidence of cancer? results from two large population-based studies," *GeroScience*, vol. 45, no. 3, pp. 2051–2064, 2023.

[6] M. A. Mafla-España, M. D. Torregrosa, and O. Cauli, "Analysis of frailty syndrome in men with metastatic prostate cancer: A scoping review," *Journal of Personalized Medicine*, vol. 13, no. 2, p. 319, 2023.

[7] Y.-Y. Pan, L.-C. Meng, H.-M. Chen, L.-K. Chen, and F.-Y. Hsiao, "Impact of frailty on survivals of prostate cancer patients treated with radiotherapy," *Archives of Gerontology and Geriatrics*, vol. 100, p. 104651, 2022.