International Journal of
Health Sciences
(IJHS)

Unveiling the Impact of Demographic Factors on Disease Survival: A Multifaceted Examination across Diverse Medical Conditions

CARI Journals

International Journal of Health Sciences

ISSN: 2710-2564 (Online)

Vol. 7, Issue No. 2, pp 1 - 14, 2024

www.carijournal.org

Unveiling the Impact of Demographic Factors on Disease Survival: A Multifaceted Examination across Diverse Medical Conditions

Lakshmi Sahitya Cherukuri [1*], Rohan Singh Rajput [2], Shantanu Neema [3]

[1*] Independent Researcher, Dallas, TX, USA

[2] Headspace Inc., Los Angeles, CA USA

[3] Syntelli Solutions Inc., Charlotte, NC USA

https://orcid.org/0009-0006-7442-6591

Accepted: 17th Feb 2024 Received in Revised Form: 3rd Mar 2024 Published: 17th Mar 2024

Abstract

Purpose: In this research, we studied the intricate interplay between demographic indicators and survival rates across various diseases, aiming to address the gap in comprehensive analyses across multiple conditions.

Methodology: Drawing from a dataset encompassing 9105 critically ill patients from five medical centers in the United States [1], admitted between1989-1991 and 1992-1994, our analysis spans eight disease categories. Leveraging techniques such as Cox-proportional hazard models and machine learning algorithms, we explore the influence of socio-economic status, gender, race, and education on survival outcomes.

Findings: Our findings underscore significant demographic disparities in disease survivability, with ethnicity, gender, and education level showing varying impacts across different medical conditions. Notably, Asians exhibit lower hazards for certain diseases but higher hazards for others, while females demonstrate better survival probabilities compared to males. Moreover, individuals with higher education levels tend to have slightly increased hazards for certain conditions.

Unique Contribution to Theory, Practice, and Policy: The call for comparative analyses across multiple diseases using comprehensive datasets marks a pivotal shift in research strategy. It aims to highlight the interplays and shared risk factors across diseases, contributing significantly to the advancement of theoretical frameworks, the refinement of healthcare practices, and the shaping of informed public health policies. This approach seeks to bridge a critical gap in the literature, offering a foundation for interventions designed to enhance disease management and improve population health outcomes comprehensively.

Keywords: *Survival Analysis, Scikit-Survival, Concordance Index, Disease Survival*

1 Introduction

Demographic factors play a pivotal role in shaping individuals' lives and significantly influencing their ability to withstand various diseases. Extensive research has highlighted the profound impact of socio-economic status, gender, and education on survival rates across a spectrum of medical conditions, including cancer and AIDS [2]. These studies reveal persistent disparities in survivability, even in countries with universal health coverage like England [3] and Australia [4] with universal health coverage. This study endeavors to bridge this gap by exploring the complex relationship between demographic indicators and survival rates for diverse diseases. Leveraging methodologies such as Cox-proportional hazard models and machine learning algorithms tailored for survival analysis, we aim to gain deeper insights into these dynamics. This study seeks to provide insights from a population health perspective, facilitating informed decision-making for public health policy initiatives and interventions aimed at improving overall health outcomes across different demographic groups. Despite the extensive literature on survivability factors, the current study aims to fill this void by comprehensively examining the impact of socio-economic indicators on survival rates [5] across a spectrum of diseases, thus contributing to a deeper understanding of population health and disease management. The factors affecting survivability are backed by several studies done on survival rates of diseases like cancer [6] and AIDS [7]. The impact of these factors was also prevalent in rich countries like England and Australia, despite universal health coverage [5, 8]. Other than race and socio-economic groups, gender also plays a key role in the survivability of diseases like heart failure [9].

Additionally, some studies have also suggested that the level of education affects survivability in the case of Alzheimer’s disease [10]. This study is a summary of such factors across various diseases like acute respiratory failure, chronic obstructive pulmonary disease (COPD), congestive heart failure (CHF), liver disease, coma, colon cancer, lung cancer, multiple organ system failure (MOSF) with malignancy and MOSF with sepsis. Available literature does address different diseases and often evaluates survival time and has noted the coefficients of a range of factors on survival times using techniques such as cox-proportional hazard models [11] and other machine learning models [12]. Popular libraries such as scikit-survival and lifelines could be used to evaluate survivability. In this study, our aim is from a population health perspective when studying multiple diseases to allow assessment of the overall burden of disease and the effect of socioeconomic indicators on survivability. This information could be useful to make informative decisions for public health policy initiatives and interventions to improve overall health outcomes by addressing a range of conditions in different socio-economic groups.

Despite the wealth of research on individual diseases, there remains a notable gap in the literature regarding comprehensive analyses across multiple diseases, particularly in assessing the impact of demographic factors. Despite the increasing recognition of the importance of studying multiple diseases simultaneously, existing literature predominantly focuses on individual diseases or specific disease pairs, neglecting the potential insights that could be gained from comparative

analyses across a broader spectrum of conditions. While numerous studies have contributed valuable insights into the epidemiology, risk factors, and treatment strategies for various diseases, there remains a notable gap in the literature regarding the utilization of datasets containing information on multiple diseases for comprehensive analysis. This gap represents a missed opportunity to understand the complex interactions between diseases, identify common risk factors, and develop more effective healthcare approaches. Consequently, there is a pressing need for research that leverages comprehensive datasets to conduct comparative analyses across multiple diseases, filling this identified gap in the literature and advancing our understanding of population health and disease management.

2 Data Description

The dataset encompasses 9105 critically ill patients from five medical centers in the United States, admitted between 1989-1991 and 1992-1994. Each entry pertains to hospitalized patients meeting specific criteria across nine disease categories: acute respiratory failure, chronic obstructive pulmonary disease, congestive heart failure, liver disease, coma, colon cancer, lung cancer, multiple organ system failure with malignancy, and multiple organ system failure with sepsis. The objective is to ascertain the 2- and 6-month survival rates of these patients, leveraging various physiological, demographic, and disease severity metrics. This research holds significance as it addresses the pressing national concern surrounding patients' end-of-life care, facilitating timely decision-making, and planning to mitigate the occurrence of prolonged, distressing, and mechanized dying processes.

3 Data Pre-processing

Data Preparation step is necessary to ensure data is in the right format for use by machine learning algorithms. The size of the datasets for each disease is more than 500, so all the diseases are included in the analysis [Table 1]. Some of the features which were computed from the patients' biomarker profile were removed as they are derived from other data points from the dataset. The time of the event occurrence is estimated using the 'Age' feature. Death and Age are paired to formulate target variable y = [(Age, Death)], death event is recorded as 1 if the data is not censored and death event is recorded during the study. This brought down the columns to 31. All the readings from individuals are grouped by disease as the intent is to estimate the time of event happening. Continuous variables are normalized using the Standard Scalar model and missing values are filled using KNN-Imputer with 4 nearest neighbors. One-hot encoding is a common strategy in machine learning that transforms each categorical data level into a separate binary variable. Race and Sex features are one-hot encoded to better differentiate the effect of these variables on the survival time.

Disease	Size
ARF/MOSF w/ Sepsis	3515
CHF	1387
COPD	967
Cirrhosis	508
Colon Cancer	512
Coma	596
Lung Cancer	908
MOSF w/ Malignant	712

Table 1. Number of patients for each disease type

4 Methodology

For this study, we utilized the scikit-survival library [12] in Python, which incorporates the Cox-proportional hazard model along with other machine learning survival models based on random forest, gradient-boosted trees, and support vector machines. To compare the performance of various models from scikit-survival across different diseases, we employed the concordance index or C-Index [13] to assess their performance on validation sets. Upon comparing scores from different models for various diseases, it was observed that Cox-proportional hazard models with L2 regularization [Table 2] could be preferred, despite exhibiting slightly lower average performance compared to tree-based models such as gradient boosted or random forest. This preference was attributed to the interpretability of Cox-proportional hazard models, as well as their better generalization indicated by similar scores in both training and validation sets, which mitigates overfitting and instills greater confidence in our results.

5 Related Work

The dataset utilized in this study encompasses eight distinct diseases, for which survival estimates were computed and are presented in Table 1. To conduct the survival analysis, we employed models from the scikit-survival module available on pypi. Specifically, we compared various scikit-survival models including Cox proportional hazards (CoxPH), Random Survival Forest, and Gradient Boosted Survival. Examination of the coefficients derived from the CoxPH models provided insights into the effects of demographic factors on survivability. Notably, while there was a noticeable difference in the training scores across these models, the testing scores exhibited similarity. Leveraging these scores, we determined the most suitable method for a generalized

machine learning model. Within the CoxPH framework, we applied Lasso, Ridge, and Elastic Net regularization techniques to identify the best-performing model. Ridge (L2) regularization surpassed Lasso and Elastic Net [14, 15]. The consistent test scores among all models suggest comparable generalization to unseen data, aligning with the primary goal in most modeling scenarios of achieving good generalization. Models with lesser discrepancy between training and test scores tend to offer enhanced reliability and interpretability, especially in critical applications, as they are less likely to overfit the training data. Thus, we concluded that the Cox PH model with Ridge regularization best suited the studied data. This model is expected to be more robust, potentially simpler, less overfitted, and possibly more interpretable.

model	ARF/MOSF w/Sepsis	CHF	COPD	Cirrhosis	Colon Cancer	Coma	Lung Cancer	MOSF w/Malig
Cox	0.694	0.671	0.656	0.745	0.57	0.64	0.59	0.655
Cox_Ridge	0.694	0.672	0.658	0.742	0.572	0.642	0.592	0.658
Cox_Lasso	0.576	0.63	0.591	0.689	0.499	0.541	0.506	0.656
Cox_Elasticnet	0.576	0.63	0.591	0.689	0.499	0.541	0.506	0.656
RandomForest	0.692	0.685	0.634	0.73	0.638	0.674	0.6	0.663
GradientBoosted	0.7	0.692	0.655	0.759	0.576	0.65	0.573	0.675

Table 2. Test scores of C-indices for various survival models

For disease-related predictions, various health data points show a correlation. It becomes complex to calculate the parameters of each feature's contribution to the survival estimates. To address this problem, we considered the Cox model with added penalty to the coefficients. Cox Proportional Hazards model and Cox Net Survival model introduce regularization to the model. To improve the feature selection, we tested Elastic Net and Lasso models with L1 ratio of 0.9 and L1 ratio of 1.0 for deriving correlativity among features. Alpha at 0.5 fit best for the L2 penalty Cox model, it is chosen from alphas [0, 1]. Random Survival Forest model, which is an ensemble of decision trees, is employed to balance the accuracy and generalization. Random survival forest model is configured with '*n_estimators = 100*', offering a balance between model complexity and computational efficiency, and '*in_samples_split = 10*', '*min_samples_leaf = 10*' ensure that a node must have at least 10 samples to split, and a leaf must have at least 10 samples, respectively, helping to prevent overfitting by avoiding overly complex trees. In a Gradient Boosting model, '*n_estimators = 100*' sets the number of boosting stages (trees) to 100, optimizing performance and complexity; '*learning_rate = 0.1*' adjusts the contribution of each tree to the outcome, controlling the speed of convergence; '*max_depth = 4*' limits each tree's maximum depth to prevent overfitting by simplifying the model. Of all the models evaluated, Cox PH model with Ridge penalty has better concordance index score [Table 2]. The hyperparameters chosen are the default ones for Random survival forest and Gradient boosted models. Although tuning the parameters

might have improved the scores, in this research we focused on comparing the models as is from the scikit-survival module.

The data is exploited to make sense of the occurrence of the disease and their correlation with demographics of the individuals.

Feature	Coefficients for various disease							
	ARF/MOSF w/Sepsis	CHF	COPD	Cirrhosis	Colon Cancer	Coma	Lung Cancer	MOSF w/Malig
Education	0.02	0.04	0.01	0.04	0.02	0.05	0.02	0.00
Female	-0.03	-0.27	-0.05	0.00	-0.01	-0.10	-0.01	0.01
Male	0.03	0.27	0.05	0.00	0.01	0.10	0.01	-0.01
Asian	-0.42	-0.38	-0.12	0.37	-0.19	0.48	-0.19	-0.51
Black	0.05	-0.05	-0.29	0.40	-0.20	-0.32	-0.20	0.27
Hispanic	0.10	-0.29	-0.08	-0.12	0.60	-0.08	0.60	1.21
Other Demographics	0.30	-0.25	0.10	0.63	-0.48	0.66	-0.48	0.90
White	-0.39	-0.77	-0.48	-0.41	-0.48	-0.69	-0.48	-0.02

Table 3. Coefficients for various diseases based on demographic factors.

The data is exploited to make sense of the occurrence of the disease and their correlation with demographics of the individuals. The interpretation of the coefficients indicated a correlation between ethnicity, gender, and the timing of the event's occurrence. Asians generally have a lower hazard of ARF/MOSF w/Sepsis, CHF, COPD, and Lung Cancer compared to other ethnicities. However, they have a higher hazard of Cirrhosis, Colon Cancer, and MOSF w/Malig. Blacks have a lower hazard of COPD and Colon Cancer but a higher hazard of Cirrhosis, Lung Cancer, and MOSF w/Malig. Hispanics have a higher hazard of several medical conditions including ARF/MOSF w/Sepsis, Cirrhosis, Colon Cancer, Coma, Lung Cancer, and MOSF w/Malig. Whites generally have a lower hazard for most conditions except for Cirrhosis and Coma. Females generally have a lower hazard of CHF, COPD, Coma, and Lung Cancer compared to males. Males generally have a higher hazard of CHF, COPD, Coma, and Lung Cancer compared to females.

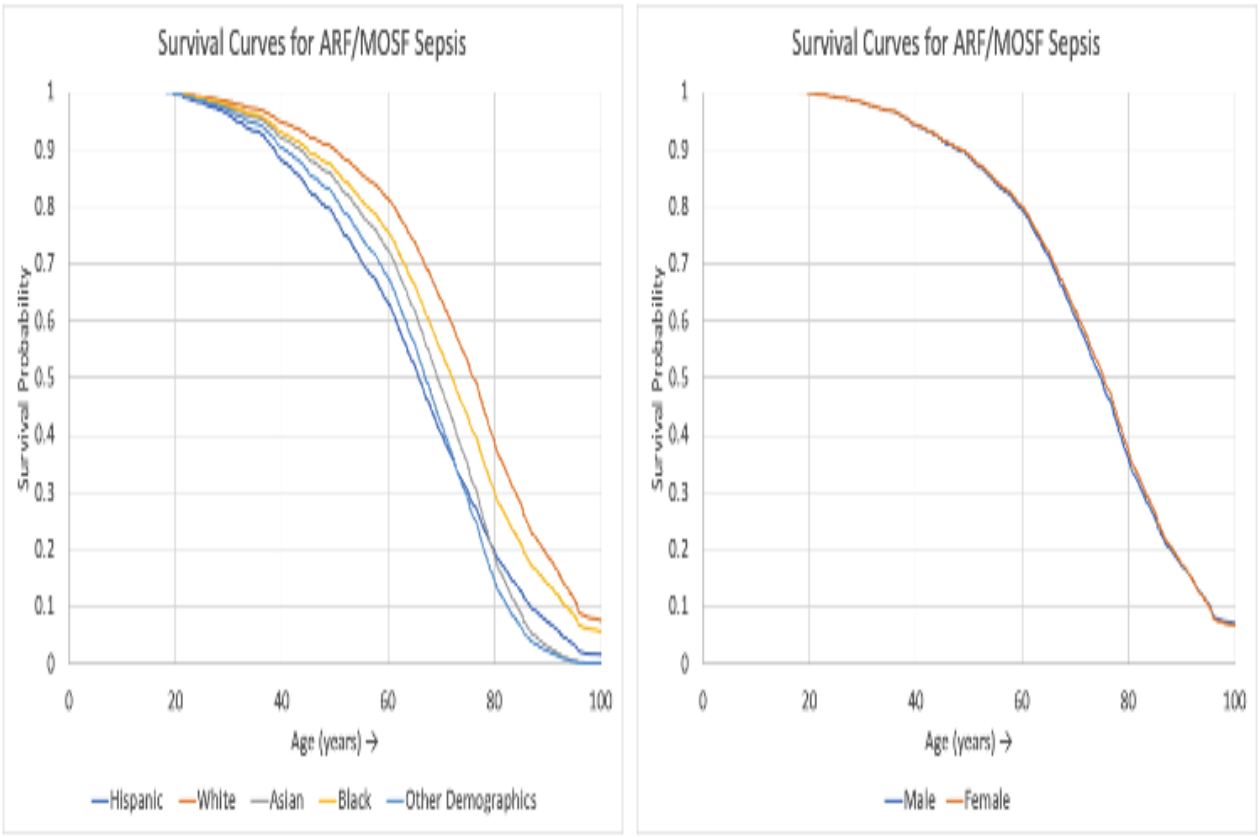

Figure 1. Survival Curves for ARF/MOSF Sepsis

In this section, we examine the survival curves for six varied diseases, segmented by race and gender. Figure 1, The Cox PH model coefficients show the impact of various data points from clinical profile on the survival probabilities for sepsis disease. These coefficients indicate the relative hazard for each variable, with a positive coefficient indicating a higher risk of the occurrence of Sepsis event and a negative coefficient indicating a protective effect against the event. “Other” race has the highest coefficient (0.30), suggesting that individuals in this group have a higher risk of the disease. Hispanics (0.10) and Blacks (0.05) also have positive coefficients but are lower than “other” races, indicating higher risk relative to other demographics. Whites and Asians have negative coefficients (-0.39 and -0.42, respectively), indicating that individuals in these groups are likely to have a higher survival probability. Inferring from the survival curves, the age at which the “other” race reaches a 50% survival probability would be younger than the other races included in the study, due to the higher positive coefficient. The number of comorbidities coefficient also has a high positive (0.17). It is likely that people with simultaneous presence of multiple medical conditions tend to have less chance of survival. Conversely, Whites and Asians, with their negative coefficients, would reach the 50% survival probability at the latter ages of 70s.

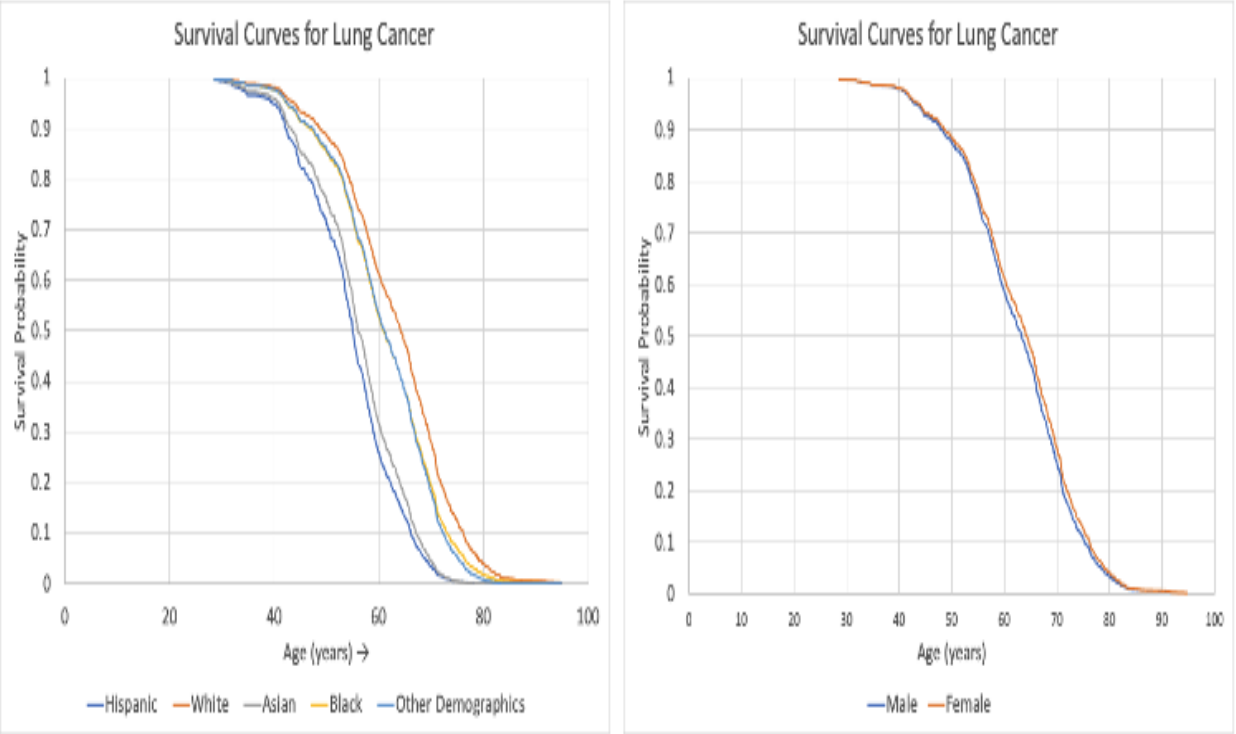

Figure 2. Survival Curves for Lung Cancer

Probabilities for occurrence of death due to Lung Cancer are drawn in Figure 2. Hispanic and Asian populations have a slightly higher chances of disease occurrence compared to White and Black populations. Differences in genetic factors, access to healthcare, or lifestyle choices between the racial groups potentially have an impact. Black males have shown higher rates of lung cancer incidence and mortality, highlighting a significant disparity in this group [16]. Females with a coefficient of –0.01 are slightly better than males whose coefficient is at 0.01, hinting at possible gender-related biological differences in disease progression or response to treatment.

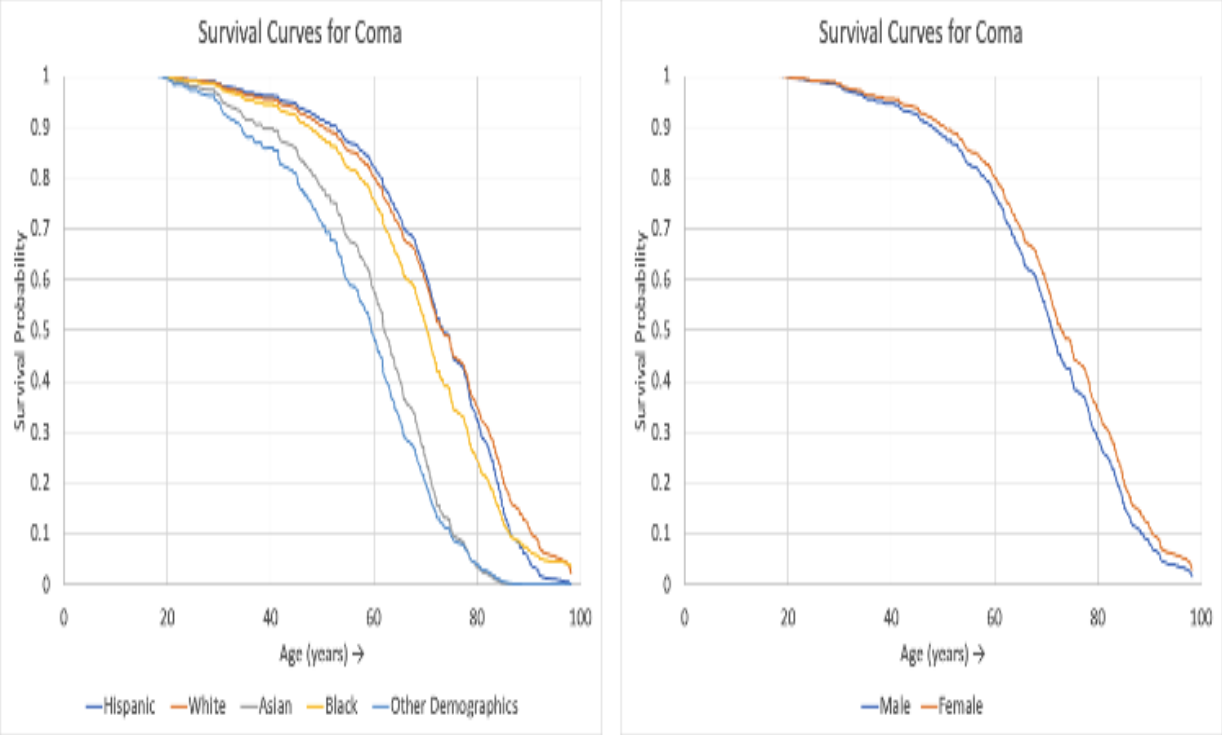

Figure 3. Survival Curves for Coma

Coma can stem from a number of comorbidities. Figure 3, "Other Demographics" have the highest coefficient of 0.66, implying the lowest survival probability, followed by Asians (0.48). Coma can also be a result of complications from conditions like diabetes, this is observed from the data, patients with diabetes have contributed (0.29) to the higher chances of death from the condition. Males have a slightly lower survival probability than females, as indicated by the respective positive (0.1) and negative coefficients (-0.1) for gender. Hispanics, Blacks, and Whites, in that order, show a protective effect against the risk of death from coma, with Whites benefiting the most.

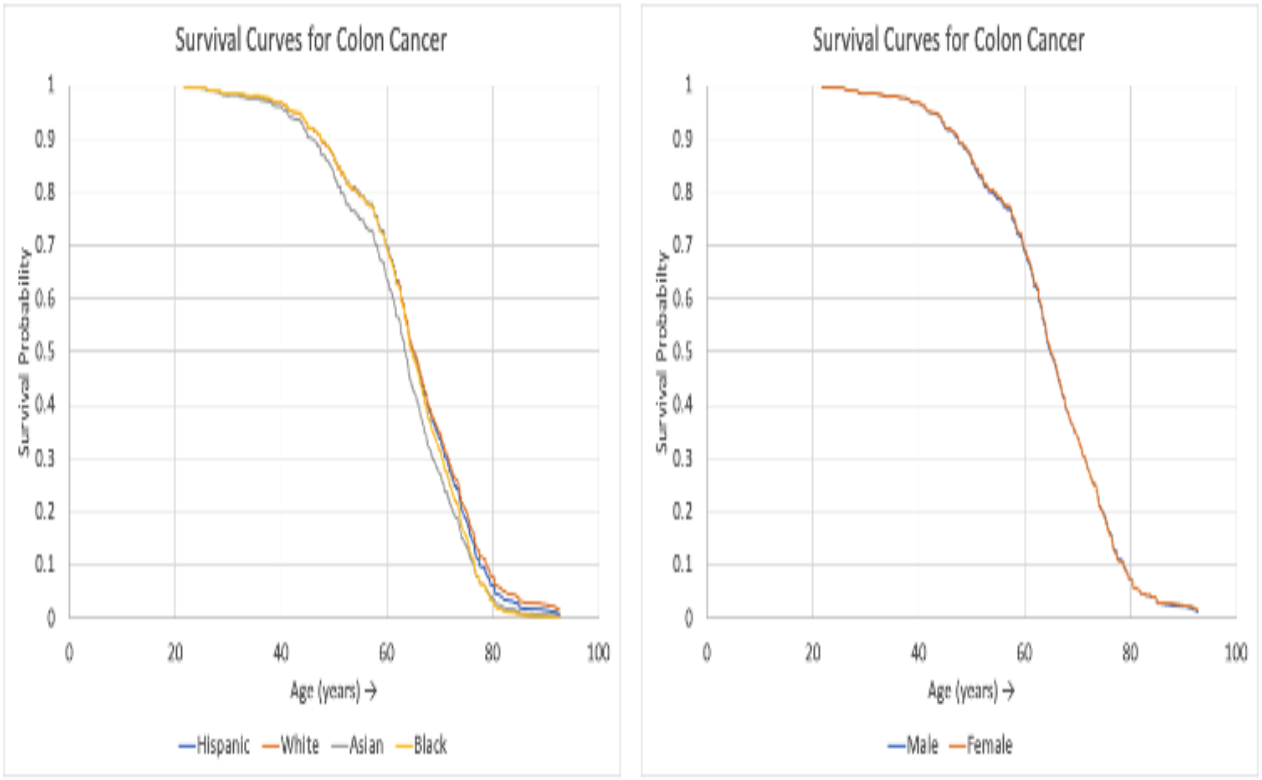

Figure 4. Survival Curves for Colon Cancer

Examining the colon cancer survival curves, Figure 4, reveals that racial and gender factors are significant determinants of survival. With the most substantial protective coefficient of -0.71, the Asian demographic exhibits the highest survival probability, which aligns with their curve's slower descent. The coefficients for other races indicate a lesser but still significant protective effect in the order of Whites (-0.62), Others (-0.55), Hispanics (-0.53), and Blacks (-0.44). The gender-based survival probability is slightly in favor of females, with a negative coefficient of -0.003, compared to males who have a positive coefficient of the same magnitude. The curves reflect this subtle difference, with the female curve staying just above the male curve. When interpreting the age at which there is a 50% probability of survival, it is observed that Asians would reach this median survival age later than other races, followed by Whites, Others, Hispanics, and Blacks in that order.

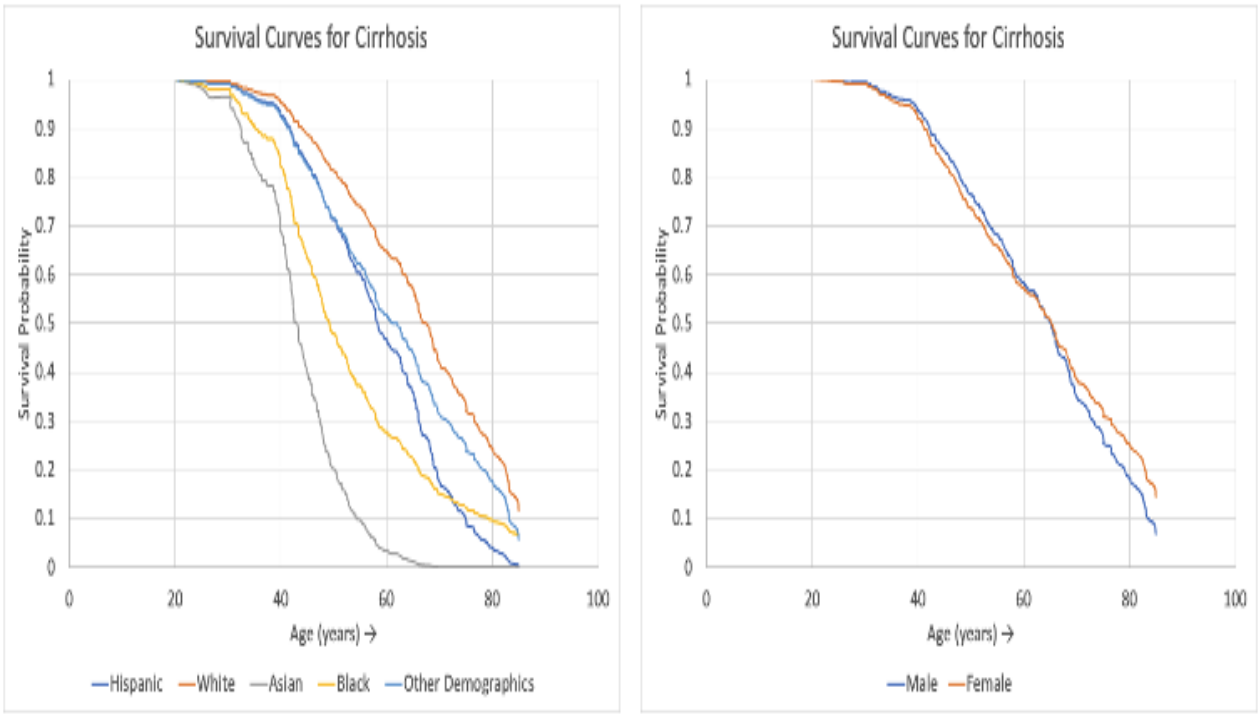

Figure 5. Survival Curves for Cirrhosis

Cirrhosis is a condition characterized by severe scarring of the liver, leading to impaired liver function. The chart (Figure 5) illustrates the differences in survival probabilities for cirrhosis between males and females over time. Until the age of 65, males tend to have better chances of survival after which females did better. Asians are at higher risk of the fatal outcome, followed by Blacks, Hispanics and Whites. Excessive alcohol consumption and Hepatitis are some of the causes of the disease. The clear distinction between survival probabilities indicates the demographic disparities. Among Asians, India had the highest number of deaths and Disability-Adjusted Life Years (DALYs) related to liver cirrhosis and other chronic liver diseases in 2019, with more than 40% of total deaths attributed to hepatitis C [17].

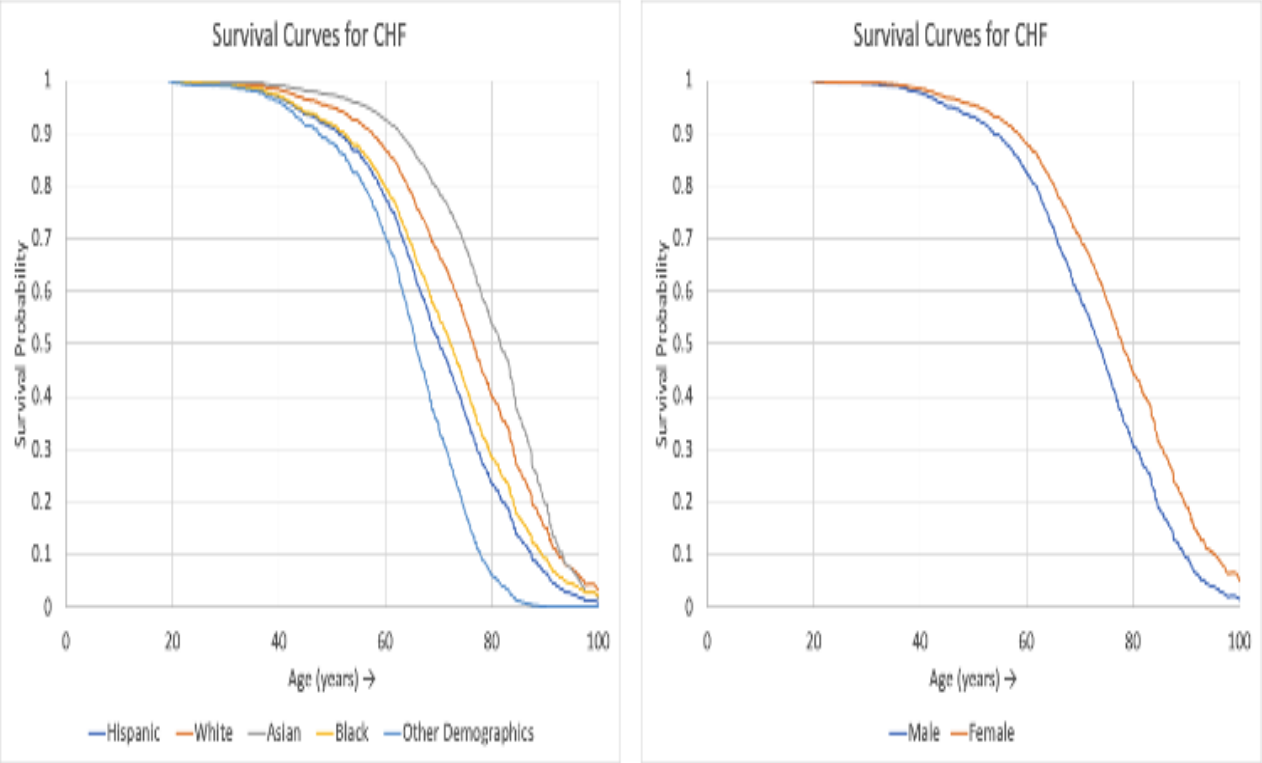

Figure 6. Survival Curves for CHF

Figure 6. Males, at a coefficient of 0.27, are contributing the most to the occurrence of the CHF disease compared to females. Observing the survival probabilities, females have high chances of survival compared to males. Black individuals have a mortality rate and are more likely to have an incidence of heart failure at younger age compared to White adults, with significant geographic variation as observed in Figure 6. [18]

6 Conclusion

1. In conclusion, the research illuminates the intricate interplay between demographic indicators and survival rates across diverse diseases, aiming to fill a crucial gap in comprehensive analyses across multiple conditions. Through rigorous examination of a dataset encompassing 9105 critically ill patients from various US medical centers, admitted between 1989-1991 and 1992-1994, spanning eight disease categories, significant disparities in disease survivability have been uncovered, influenced by ethnicity, gender, and education level. Notably, Asians exhibit varying hazards for different diseases, females generally demonstrate better survival probabilities than males, and individuals with higher education levels tend to face slightly increased hazards for certain conditions. These findings underscore the complex nature of disease outcomes and emphasize the importance of considering demographic factors in healthcare planning and policymaking.

7Recommendation

The study contributes valuable insights for informing public health interventions and initiatives aimed at reducing health disparities and enhancing overall health outcomes. By elucidating the nuanced impacts of socio-economic status, gender, race, and education on disease survivability, the research provides a foundation for designing targeted interventions to address population health challenges effectively. Moving forward, leveraging comprehensive datasets and employing advanced analytical techniques will remain crucial in advancing the understanding of disease management and promoting health equity across diverse demographic groups.

7 References:

[1] Vanderbilt University Department of Biostatistics, Professor Frank Harrell 2022, url: https://hbiostat.org/data/

[2] Simard EP, Pfeiffer RM, Engels EA. Mortality due to cancer among people with AIDS: a novel approach using registry-linkage data and population attributable risk methods. AIDS. 2012 Jun 19;26(10):1311-8. doi: 10.1097/QAD.0b013e328353f38e. PMID: 22472857; PMCID: PMC3377813.

[3] Pickwell-Smith BA, Spencer K, Sadeghi MH, et alWhere are the inequalities in colorectal cancer care in a country with universal healthcare? A systematic review and narrative synthesisBMJ Open 2024;14:e080467. doi: 10.1136/bmjopen-2023-080467

[4] Stanbury JF, Baade PD, Yu Y, Yu XQ. Cancer survival in New South Wales, Australia: socioeconomic disparities remain despite overall improvements. BMC Cancer. 2016 Feb 1;16:48. doi: 10.1186/s12885-016-2065-z. PMID: 26832359; PMCID: PMC4736306.

[5] Yu, X.Q., O'Connell, D.L., Gibberd, R.W. et al. Assessing the impact of socio-economic status on cancer survival in New South Wales, Australia 1996–2001. Cancer Causes Control 19, 1383–1390 (2008). https://doi.org/10.1007/s10552-008-9210-1

[6] Catherine Lejeune, Franco Sassi, Libby Ellis, Sara Godward, Vivian Mak, Matthew Day, Bernard Rachet, Socio-economic disparities in access to treatment and their impact on colorectal cancer survival, *International Journal of Epidemiology*, Volume 39, Issue 3, June 2010, Pages 710–717, https://doi.org/10.1093/ije/dyq048

[7] Mitchell H. Katz, Ling Hsu, Michael Lingo, Greg Woelffer, Sandra K. Schwarcz, Impact of Socioeconomic Status on Survival with AIDS, *American Journal of Epidemiology*, Volume 148, Issue 3, 1 August 1998, Pages 282–291, https://doi.org/10.1093/oxfordjournals.aje.a009637

[8] Aimilia Exarchakou, Bernard Rachet, Aurélien Belot, Camille Maringe, Michel P Coleman, Impact of national cancer policies on cancer survival trends and socioeconomic inequalities in England, 1996-2013: population based study, the British Medical Journal, (March 2018) https://doi.org/10.1136/bmj.k764

[9] Faisal Maqbool Zahid,Shakeela Ramzan,Shahla Faisal ,Ijaz Hussain, Gender based survival prediction models for heart failure patients: A case study in Pakistan, Feb 2019, https://doi.org/10.1371/journal.pone.0210602

[10] A. Bruandet; F. Richard; S. Bombois; C.A. Maurage; I. Masse; P. Amouyel; F. Pasquier, Cognitive Decline and Survival in Alzheimer's Disease according to Education Level, Dement Geriatr Cogn Disord (2007) 25 (1): 74–80. https://doi.org/10.1159/000111693

[11] Sameer Sundran, James Lu, Computing the Hazard Ratios Associated With Explanatory Variables Using Machine Learning Models of Survival Data, JCO Clinical Cancer Informatics, Vol 5, Issue 5, https://doi.org/10.1200/CCI.20.00172

[12] Pedregosa F, Varoquaux, Ga"el, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12(Oct):2825–30.

[13] Brentnall AR, Cuzick J. Use of the concordance index for predictors of censored survival data. Statistical Methods in Medical Research. 2018;27(8):2359-2373. doi:10.1177/0962280216680245

[14] Bøvelstad, H. M. *et al.* Predicting survival from microarray data—A comparative study. *Bioinformatics* **23**, 2080–2087 (2007). https://doi.org/10.1093/bioinformatics/btm305

[15] Van Wieringen, W. N., Kun, D., Hampel, R. & Boulesteix, A. L. Survival prediction using gene expression data: A review and comparison. *Comput. Stat. Data Anal.* **53**, 1590–1603 (2009). https://doi.org/10.1016/j.csda.2008.05.021

[16] Cranford HM, Koru-Sengul T, Lopes G, Pinheiro PS. Lung Cancer Incidence by Detailed Race-Ethnicity. Cancers (Basel). 2023 Apr 5;15(7):2164. doi: 10.3390/cancers15072164. PMID: 37046824; PMCID: PMC10093016.

[17] Wu, XN., Xue, F., Zhang, N. *et al.* Global burden of liver cirrhosis and other chronic liver diseases caused by specific etiologies from 1990 to 2019. *BMC Public Health* 24, 363 (2024). https://doi.org/10.1186/s12889-024-17948-6

[18] Lewsey SC, Breathett K. Racial and ethnic disparities in heart failure: current state and future directions. Curr Opin Cardio. 2021 May 1;36(3):320-328. doi: 10.1097/HCO.0000000000000855. PMID: 33741769; PMCID: PMC8130651.