Understanding Risk Factors for Mortality Among Older Individuals

(A Comprehensive Analysis using Classical, Ensemble and Deep Learning Models)

Srimanta Ghosh

University Roll No.: E1773U234013

Supervisors:

Dr. Rahul Ghosal[1]

Dr. Sudhakar Sahoo[2]

[1]Assistant Professor, Arnold School of Public Health
University of South Carolina

[2]Associate Professor
Institute of Mathematics and Applications, Bhubaneswar

Outline

Outline

The Problem

- As populations around the world continue to age, understanding the determinants of healthy aging and longevity becomes increasingly vital.
- While physical activity has been shown to reduce mortality risk, the complex interplay between lifestyle behaviors, demographic characteristics, and health conditions remains poorly understood.
- Leveraging advancements in statistical modeling and machine learning, this project aims to create a comprehensive framework for predicting mortality risk—enabling more precise identification of high-risk individuals and informing public health strategies for aging populations.

The Problem

- As populations around the world continue to age, understanding the determinants of healthy aging and longevity becomes increasingly vital.
- While physical activity has been shown to reduce mortality risk, the complex interplay between lifestyle behaviors, demographic characteristics, and health conditions remains poorly understood.
- Leveraging advancements in statistical modeling and machine learning, this project aims to create a comprehensive framework for predicting mortality risk—enabling more precise identification of high-risk individuals and informing public health strategies for aging populations.

The Problem

- As populations around the world continue to age, understanding the determinants of healthy aging and longevity becomes increasingly vital.
- While physical activity has been shown to reduce mortality risk, the complex interplay between lifestyle behaviors, demographic characteristics, and health conditions remains poorly understood.
- Leveraging advancements in statistical modeling and machine learning, this project aims to create a comprehensive framework for predicting mortality risk—enabling more precise identification of high-risk individuals and informing public health strategies for aging populations.

Outline

Research Questions

- How are physical activity, demographic factors, and health indicators related to mortality in older adults?
- What are the key risk factors that contribute to mortality in older adults?
- How do classical statistical models, ensemble methods, and deep learning approaches compare in their effectiveness at predicting mortality risk?

Research Questions

- How are physical activity, demographic factors, and health indicators related to mortality in older adults?
- What are the key risk factors that contribute to mortality in older adults?
- How do classical statistical models, ensemble methods, and deep learning approaches compare in their effectiveness at predicting mortality risk?

Research Questions

- How are physical activity, demographic factors, and health indicators related to mortality in older adults?
- What are the key risk factors that contribute to mortality in older adults?
- How do classical statistical models, ensemble methods, and deep learning approaches compare in their effectiveness at predicting mortality risk?

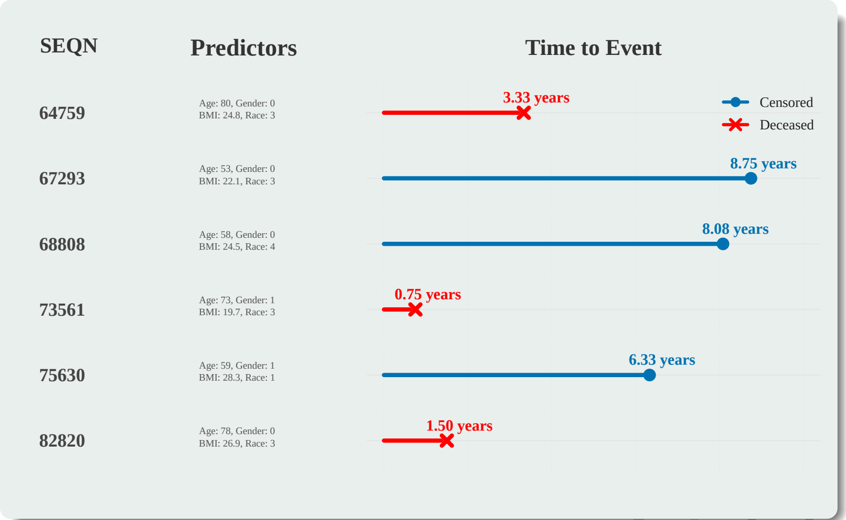
Outline

Data Source

The study will use data from the National Health and Nutrition Examination Survey (NHANES) 2011-2014, "***https://wwwn.cdc.gov/nchs/nhanes/Default.aspx***", which includes demographic, lifestyle, and health-related variables. The mortality information is linked to the National Death Index (NDI).

Outline

The Survival Data

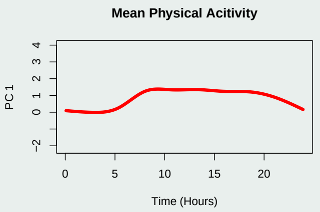


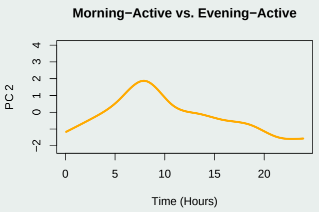
Outline

Preprocessing

FPCA

$$X_i(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t) + \epsilon_i(t)$$





Outline

1. Introduction
   - The Problem
   - Research Questions
2. The Data
   - Data Source
   - The Survival Data
   - Preprocessing (Physical Activity Data)
3. The Models
   - Structures
   - Output
4. Conclusion and Future Work

Models

Model	Type	Loss Function
Cox	Classical	$\mathcal{L}(\beta) = -\sum_{i:\delta_i=1}\left(x_i^\top\beta - \log\sum_{j\in R(T_i)}\exp(x_j^\top\beta)\right)$
Penalized Cox	Classical	$\mathcal{L}(\beta) = -\sum_{i:\delta_i=1}\left(x_i^\top\beta - \log\sum_{j\in R(T_i)}\exp(x_j^\top\beta)\right) + \lambda\|\beta\|_1$
GAM Cox	Classical	$\mathcal{L}(f,\beta) = -\sum_{i:\delta_i=1}\left(\eta_i - \log\sum_{j\in R(T_i)}\exp(\eta_j)\right) + \lambda\sum_{j=1}^{p}\int\left(f_j''(x)\right)^2 dx$ where $\eta_i = \sum_{j=1}^{p} f_j(X_{ij}) + \sum_{k=1}^{q}\beta_k Z_{ik}$
DeepSurv	Deep Learning	$\mathcal{L}(\theta) = -\frac{1}{N_E}\sum_{i:\delta_i=1}\left[h_\theta(x_i) - \log\sum_{j\in R(T_i)}\exp(h_\theta(x_j))\right] + \lambda\|\theta\|_2^2$

Model	Type	Split Rule
RSF	Ensemble	Log-rank test statistic: $L(x,c) = \sum_{i=1}^{N}\frac{\left(d_{i,1} - \frac{Y_{i,1}d_i}{Y_i}\right)}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i,1}}{Y_i}\left(1-\frac{Y_{i,1}}{Y_i}\right)\frac{(Y_i-d_i)}{(Y_i-1)}d_i}}$

Models

Model	Type	Loss Function
Cox	Classical	$\mathcal{L}(\beta) = -\sum_{i:\delta_i=1}\left(x_i^\top\beta - \log\sum_{j\in R(T_i)}\exp(x_j^\top\beta)\right)$
Penalized Cox	Classical	$\mathcal{L}(\beta) = -\sum_{i:\delta_i=1}\left(x_i^\top\beta - \log\sum_{j\in R(T_i)}\exp(x_j^\top\beta)\right) + \lambda\|\beta\|_1$
GAM Cox	Classical	$\mathcal{L}(f,\beta) = -\sum_{i:\delta_i=1}\left(\eta_i - \log\sum_{j\in R(T_i)}\exp(\eta_j)\right) + \lambda\sum_{j=1}^{p}\int\left(f_j''(x)\right)^2 dx$ where $\eta_i = \sum_{j=1}^{p} f_j(X_{ij}) + \sum_{k=1}^{q}\beta_k Z_{ik}$
DeepSurv	Deep Learning	$\mathcal{L}(\theta) = -\frac{1}{N_E}\sum_{i:\delta_i=1}\left[h_\theta(x_i) - \log\sum_{j\in R(T_i)}\exp(h_\theta(x_j))\right] + \lambda\|\theta\|_2^2$

Model	Type	Split Rule
RSF	Ensemble	Log-rank test statistic: $L(x,c) = \sum_{i=1}^{N}\frac{\left(d_{i,1} - \frac{Y_{i,1}d_i}{Y_i}\right)}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i,1}}{Y_i}\left(1-\frac{Y_{i,1}}{Y_i}\right)\frac{(Y_i-d_i)}{(Y_i-1)}d_i}}$

Models

Model	Type	Loss Function
Cox	Classical	$\mathcal{L}(\beta) = -\sum_{i:\delta_i=1} \left(x_i^\top \beta - \log \sum_{j\in R(T_i)} \exp(x_j^\top \beta) \right)$
Penalized Cox	Classical	$\mathcal{L}(\beta) = -\sum_{i:\delta_i=1} \left(x_i^\top \beta - \log \sum_{j\in R(T_i)} \exp(x_j^\top \beta) \right) + \lambda \lVert\beta\rVert_1$
GAM Cox	Classical	$\mathcal{L}(f,\beta) = -\sum_{i:\delta_i=1} \left(\eta_i - \log \sum_{j\in R(T_i)} \exp(\eta_j) \right) + \lambda \sum_{j=1}^{p} \int \left(f_j''(x) \right)^2 dx$ where $\eta_i = \sum_{j=1}^{p} f_j(X_{ij}) + \sum_{k=1}^{q} \beta_k Z_{ik}$
DeepSurv	Deep Learning	$\mathcal{L}(\theta) = -\frac{1}{N_E} \sum_{i:\delta_i=1} \left[h_\theta(x_i) - \log \sum_{j\in R(T_i)} \exp(h_\theta(x_j)) \right] + \lambda \lVert\theta\rVert_2^2$

Model	Type	Split Rule
RSF	Ensemble	Log-rank test statistic: $L(x,c) = \sum_{i=1}^{N} \frac{\left(d_{i,1} - \frac{Y_{i,1} d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^{N} \frac{Y_{i,1}}{Y_i} \left(1 - \frac{Y_{i,1}}{Y_i} \right) \frac{(Y_i - d_i)}{(Y_i - 1)} d_i}}$

Models

Model	Type	Loss Function
Cox	Classical	$\mathcal{L}(\beta) = -\sum_{i:\delta_i=1}\left(x_i^\top \beta - \log \sum_{j\in R(T_i)} \exp(x_j^\top \beta)\right)$
Penalized Cox	Classical	$\mathcal{L}(\beta) = -\sum_{i:\delta_i=1}\left(x_i^\top \beta - \log \sum_{j\in R(T_i)} \exp(x_j^\top \beta)\right) + \lambda \\|\beta\\|_1$
GAM Cox	Classical	$\mathcal{L}(f,\beta) = -\sum_{i:\delta_i=1}\left(\eta_i - \log \sum_{j\in R(T_i)} \exp(\eta_j)\right) + \lambda \sum_{j=1}^{p} \int \left(f_j''(x)\right)^2 dx$ where $\eta_i = \sum_{j=1}^{p} f_j(X_{ij}) + \sum_{k=1}^{q} \beta_k Z_{ik}$
DeepSurv	Deep Learning	$\mathcal{L}(\theta) = -\frac{1}{N_E}\sum_{i:\delta_i=1}\left[h_\theta(x_i) - \log \sum_{j\in R(T_i)} \exp(h_\theta(x_j))\right] + \lambda \\|\theta\\|_2^2$

Model	Type	Split Rule
RSF	Ensemble	Log-rank test statistic: $L(x,c) = \sum_{i=1}^{N} \frac{\left(d_{i,1} - \frac{Y_{i,1} d_i}{Y_i}\right)}{\sqrt{\sum_{i=1}^{N} \frac{Y_{i,1}}{Y_i}\left(1 - \frac{Y_{i,1}}{Y_i}\right)\frac{(Y_i - d_i)}{(Y_i - 1)} d_i}}$

Models

Model	Type	Loss Function
Cox	Classical	$\mathcal{L}(\beta) = -\sum_{i:\delta_i=1}\left(x_i^\top\beta - \log\sum_{j\in R(T_i)}\exp(x_j^\top\beta)\right)$
Penalized Cox	Classical	$\mathcal{L}(\beta) = -\sum_{i:\delta_i=1}\left(x_i^\top\beta - \log\sum_{j\in R(T_i)}\exp(x_j^\top\beta)\right) + \lambda\|\beta\|_1$
GAM Cox	Classical	$\mathcal{L}(f,\beta) = -\sum_{i:\delta_i=1}\left(\eta_i - \log\sum_{j\in R(T_i)}\exp(\eta_j)\right) + \lambda\sum_{j=1}^{p}\int\left(f_j''(x)\right)^2 dx$ where $\eta_i = \sum_{j=1}^{p} f_j(X_{ij}) + \sum_{k=1}^{q}\beta_k Z_{ik}$
DeepSurv	Deep Learning	$\mathcal{L}(\theta) = -\frac{1}{N_E}\sum_{i:\delta_i=1}\left[h_\theta(x_i) - \log\sum_{j\in R(T_i)}\exp(h_\theta(x_j))\right] + \lambda\|\theta\|_2^2$

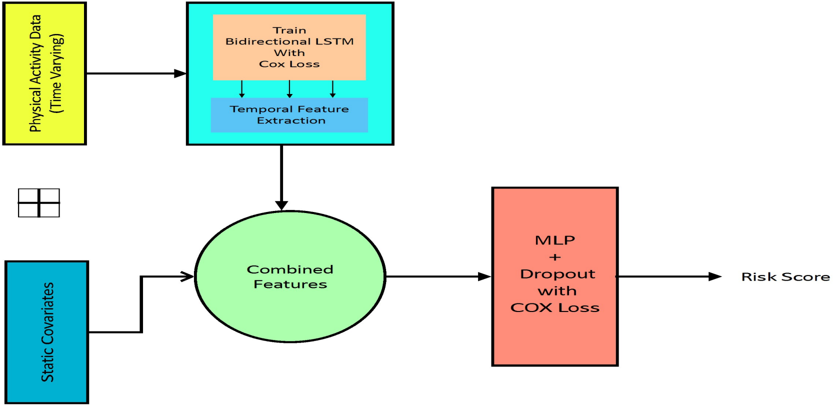
Model	Type	Split Rule
RSF	Ensemble	Log-rank test statistic: $L(x,c) = \sum_{i=1}^{N}\frac{\left(d_{i,1} - \frac{Y_{i,1}d_i}{Y_i}\right)}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i,1}}{Y_i}\left(1-\frac{Y_{i,1}}{Y_i}\right)\frac{(Y_i-d_i)}{(Y_i-1)}d_i}}$

Models

Model	Type	Loss Function
Cox	Classical	$\mathcal{L}(\beta) = -\sum_{i:\delta_i=1}\left(x_i^\top\beta - \log\sum_{j\in R(T_i)}\exp(x_j^\top\beta)\right)$
Penalized Cox	Classical	$\mathcal{L}(\beta) = -\sum_{i:\delta_i=1}\left(x_i^\top\beta - \log\sum_{j\in R(T_i)}\exp(x_j^\top\beta)\right) + \lambda\|\beta\|_1$
GAM Cox	Classical	$\mathcal{L}(f,\beta) = -\sum_{i:\delta_i=1}\left(\eta_i - \log\sum_{j\in R(T_i)}\exp(\eta_j)\right) + \lambda\sum_{j=1}^{p}\int\left(f_j''(x)\right)^2 dx$ where $\eta_i = \sum_{j=1}^{p} f_j(X_{ij}) + \sum_{k=1}^{q}\beta_k Z_{ik}$
DeepSurv	Deep Learning	$\mathcal{L}(\theta) = -\frac{1}{N_E}\sum_{i:\delta_i=1}\left[h_\theta(x_i) - \log\sum_{j\in R(T_i)}\exp(h_\theta(x_j))\right] + \lambda\|\theta\|_2^2$

Model	Type	Split Rule
RSF	Ensemble	Log-rank test statistic: $L(x,c) = \sum_{i=1}^{N}\frac{\left(d_{i,1} - \frac{Y_{i,1}d_i}{Y_i}\right)}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i,1}}{Y_i}\left(1-\frac{Y_{i,1}}{Y_i}\right)\frac{(Y_i-d_i)}{(Y_i-1)}d_i}}$

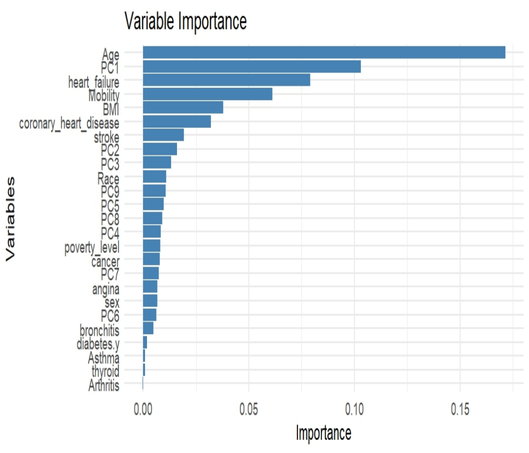
BiLSTM-Deepsurv Architecture



Outline

Output

Variable	exp(coef.)	Pr(>|z|)
Age	1.06	< 2e-16
BMI	0.96	4.67e-07
Mobility2	0.58	2.03e-08
povertylevel	0.92	0.00517
heartfailure1	1.87	7.23e-08
PC1	0.86	3.04e-14

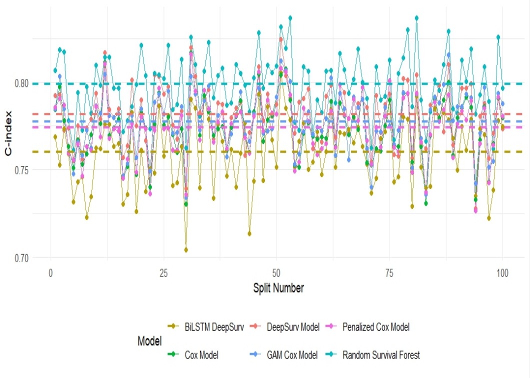


C-Index Comparison

$$C = \frac{\sum 1(h(X_i) > h(X_j)) \cdot 1(T_i < T_j)}{\sum 1(T_i < T_j)}$$

where $h(X)$ is the model's predicted risk score, and T is the observed survival time.

Model	Average C-Index
Cox	0.77421
GAM Cox	0.0.77763
Penalized Cox	0.77424
RSF	0.79895
Deepsurv	0.78173
BiLSTM Deepsurv	0.76041



Conclusion and Future Work

- If your primary goal is achieving high accuracy, Random Survival Forest (RSF) is the best-performing model.
- If your primary goal is interpretability, the Generalized Additive Cox model is one of the best reliable choices.
- DeepSurv is a strong deep learning-based alternative that shows good result in prediction.
- The BiLSTM-DeepSurv model appears promising, but it still requires improvement. With more computational resources and further tuning, it may perform better in future experiments.

Conclusion and Future Work

- If your primary goal is achieving high accuracy, Random Survival Forest (RSF) is the best-performing model.
- If your primary goal is interpretability, the Generalized Additive Cox model is one of the best reliable choices.
- DeepSurv is a strong deep learning-based alternative that shows good result in prediction.
- The BiLSTM-DeepSurv model appears promising, but it still requires improvement. With more computational resources and further tuning, it may perform better in future experiments.

Conclusion and Future Work

- If your primary goal is achieving high accuracy, Random Survival Forest (RSF) is the best-performing model.
- If your primary goal is interpretability, the Generalized Additive Cox model is one of the best reliable choices.
- DeepSurv is a strong deep learning-based alternative that shows good result in prediction.
- The BiLSTM-DeepSurv model appears promising, but it still requires improvement. With more computational resources and further tuning, it may perform better in future experiments.

Conclusion and Future Work

- If your primary goal is achieving high accuracy, Random Survival Forest (RSF) is the best-performing model.
- If your primary goal is interpretability, the Generalized Additive Cox model is one of the best reliable choices.
- DeepSurv is a strong deep learning-based alternative that shows good result in prediction.
- The BiLSTM-DeepSurv model appears promising, but it still requires improvement. With more computational resources and further tuning, it may perform better in future experiments.

References

Rahul Ghosal, Marcos Matabuena, and Sujit K Ghosh.
Functional time transformation model with applications to digital health.
Computational Statistics & Data Analysis, page 108131, 2025.

Hemant Ishwaran and Udaya B Kogalur.
Random survival forests for r.
R news, 7(2):25–31, 2007.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger.
Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network.
BMC medical research methodology, 18:1–12, 2018.