Mortality research and forecasting: individual, population, ML, old-school

Koneoppiminen kuolevuustutkimuksessa – vahvistaa vanhaa vai tuo uutta tietoa?

Mikko Myrskylä

October 6, 2025

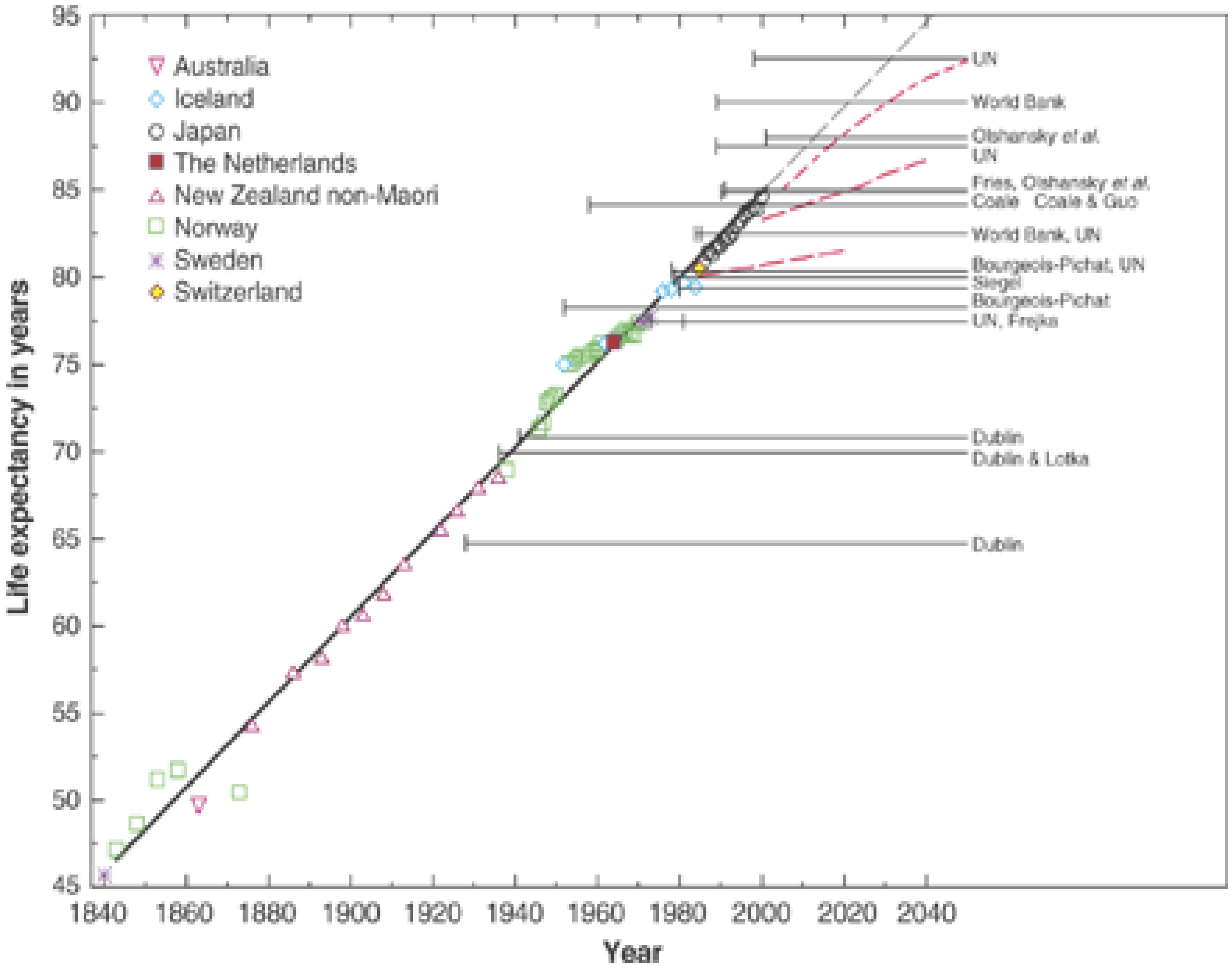
Two levels of mortality forecasting/research

Individual level

- Predicting age at death
- Risk factor approach
- Heterogeneus treatment effects
- ML/AI as emerging tool
- New knowledge gained?

Population level

- Mortality for groups, countries, women/men, etc
- Statistical extrapolation
- ML/AI as emerging tool
- Fundamentals based process models – little ML



Broken Limits to Life Expectancy, Volume: 296, Issue: 5570, Pages: 1029-1031, DOI: (10.1126/science.1069675)

Individual-level prediction of lifetime

MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH

MPIDR Working Paper WP 2023-008 | February 2023
Revised January 2024
https://doi.org/10.4054/MPIDR-WP-2023-008

The Limits of Predicting Individual-Level Longevity

Luca Badolato
Ari Decter-Frain
Nicholas J. Irons
Maria Miranda
Erin Walk
Elnura Zhalieva
Monica Alexander
Ugofilippo Basellini | basellini@demogr.mpg.de
Emilio Zagheni | office-zagheni@demogr.mpg.de

The limits of predicting longevity

- Predict individual-level mortality from survey data
- 150+ predictors
- broad range of classic statistical and machine learning survival analysis models
  - Cox models
  - random forests
  - deep neural networks
- U.S. Health and Retirement Study, 40k individuals, 1992-2018
- Predictors: Childhood, cognitive, demographic, behavioral, labor market, mental health, physical health, social, support, wealth

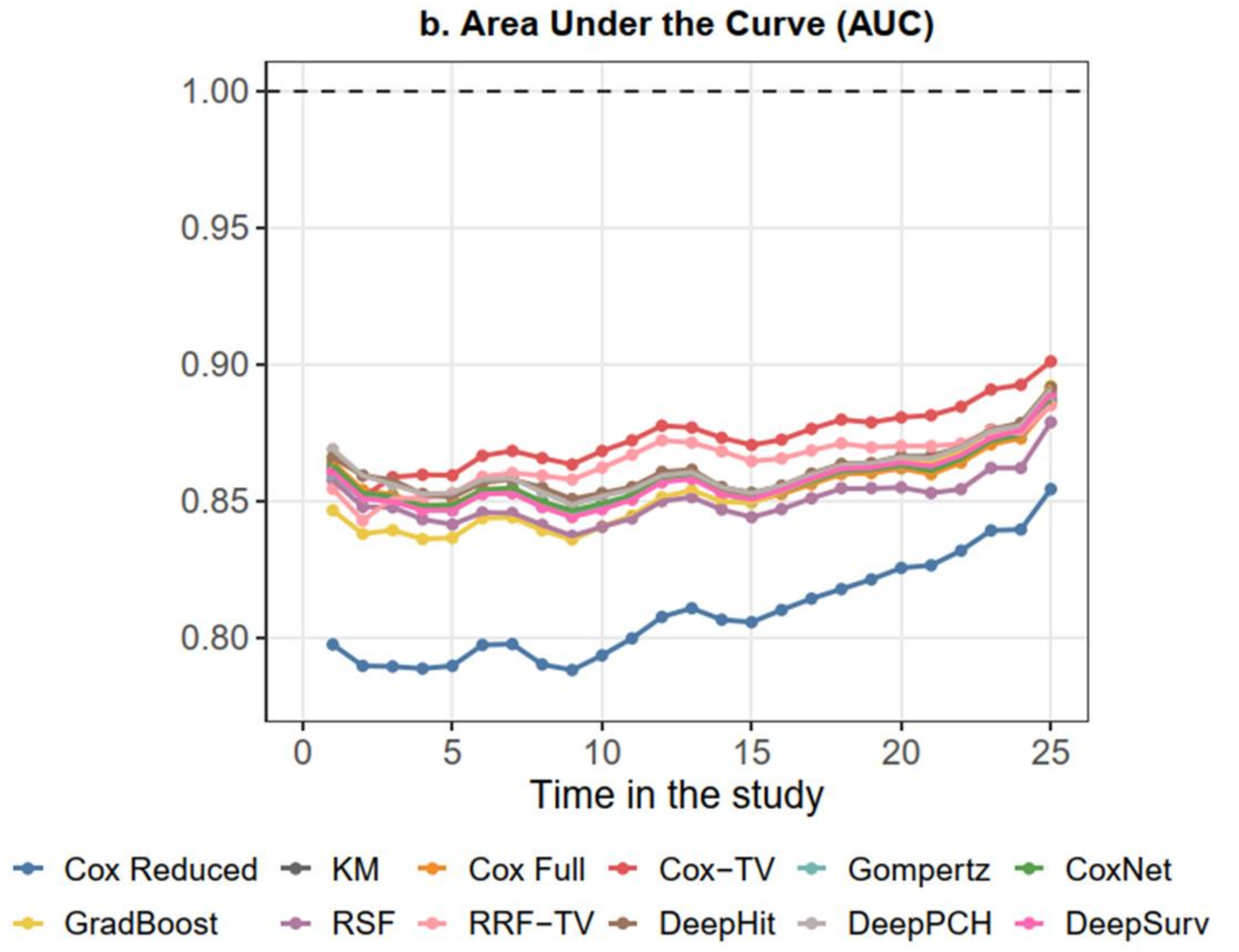
Methods to be compared

- Traditional
  - Cox proportional hazards model, with and without time-varying predictors

- Machine learning: forest based
  - Random survival forest (no prop hazard assumption)
  - Gradient-boosted trees with proportional hazards
  - Relative risk forest with time-varying predictors

- Machile learning: deep learning
  - Non-linear Cox model (DeepSurv)
  - Non-parametric predictor-based survival time model (DeepHit)

  - …and many more


b. Area Under the Curve (AUC)
1.00
0.95
0.90
0.85
0.80
0
5
10
15
20
25
Time in the study
Cox Reduced
KM
Cox Full
Cox−TV
Gompertz
CoxNet
GradBoost
RSF
RRF−TV
DeepHit
DeepPCH
DeepSurv

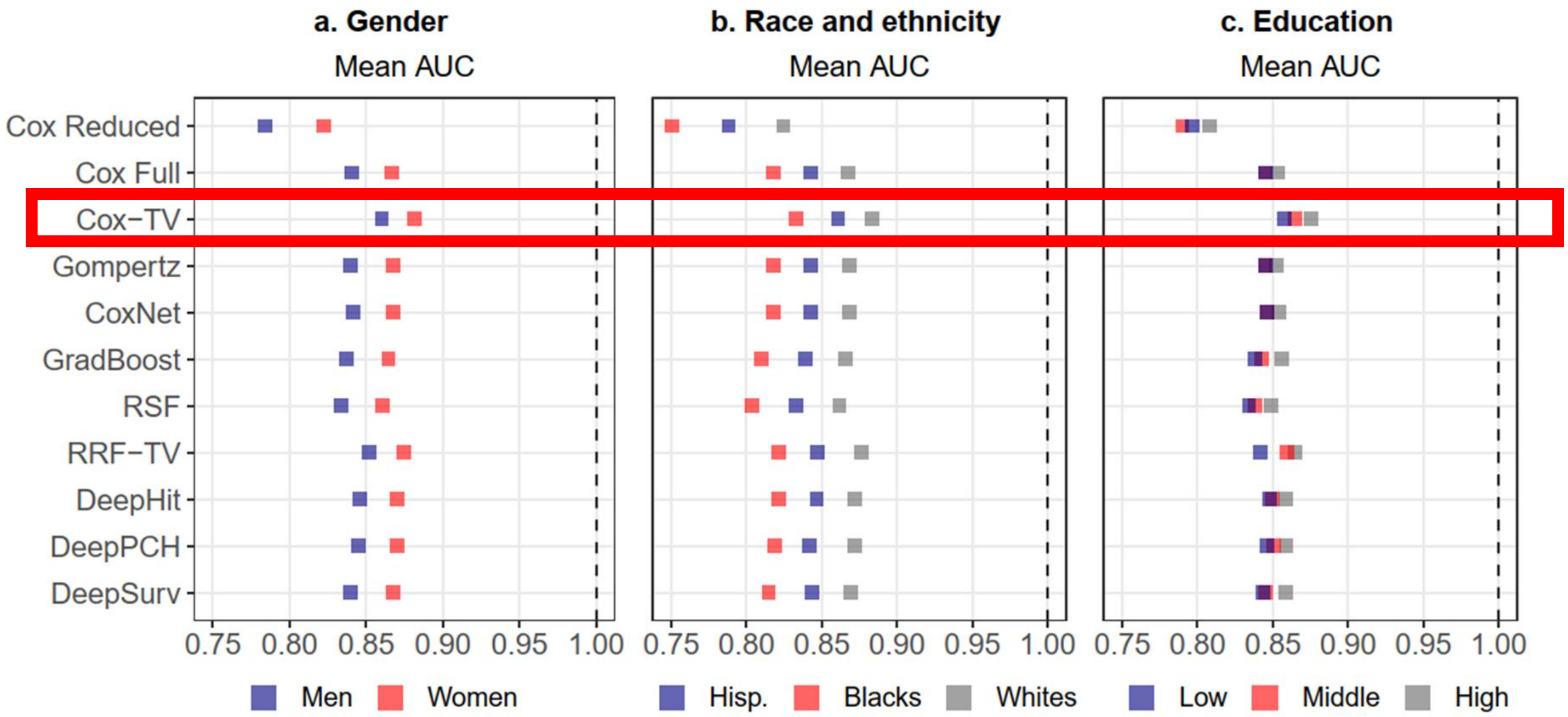


Figure 2: Integrated Brier Score and Mean Area Under the Curve by gender (panel a.), race and ethnicity (panel b.), and education (panel c.). The dashed vertical lines indicate the optimal Brier Score (0.00) and the Area Under the Curve (1.00). The Mean AUC for Kaplan-Meier, 0.5 by construction, has not been reported.

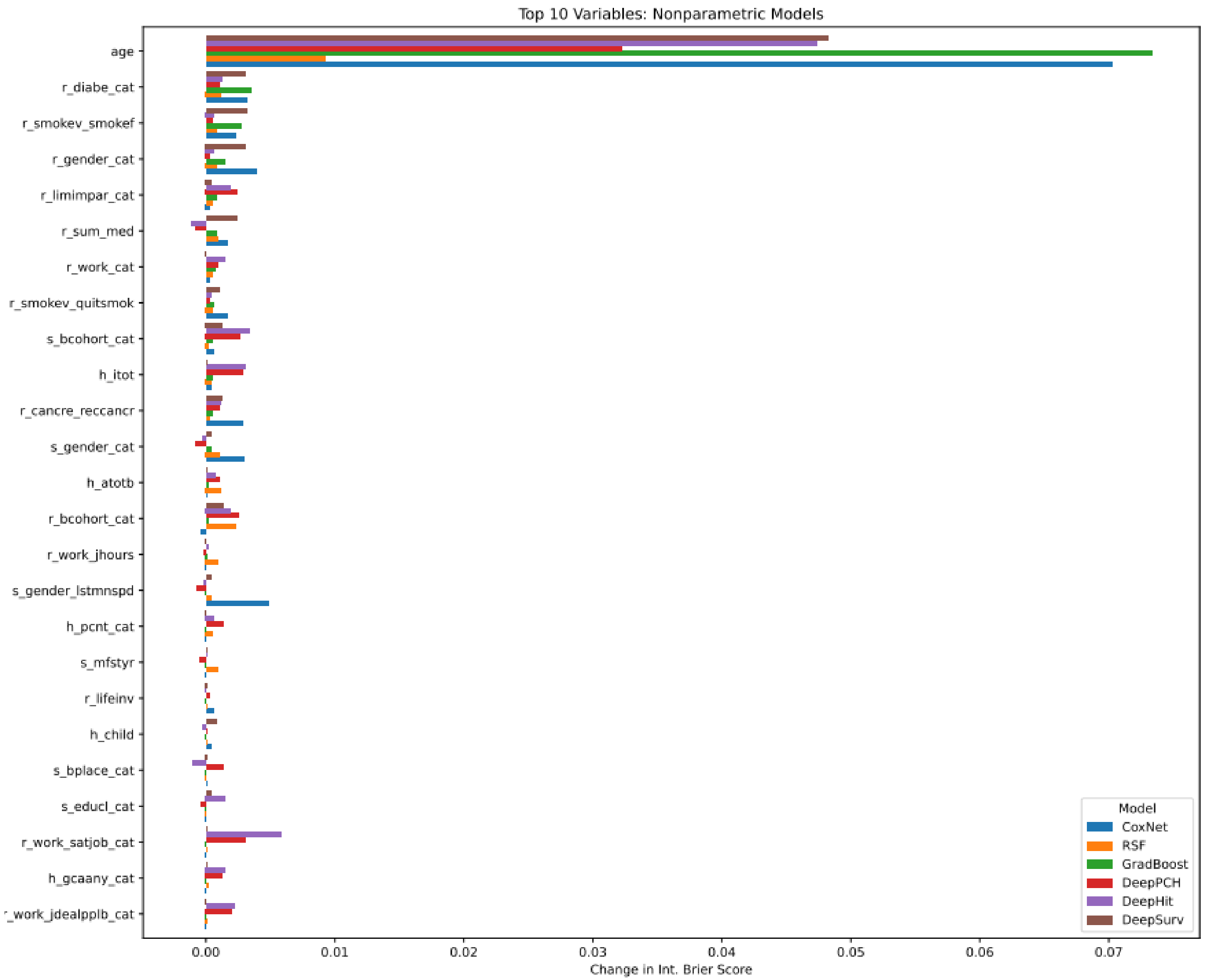


Figure 4: Comparative importance of the set of top 10 variables for each model.

- Age, diabetes, smoking, gender, medications, working status, ….

Summary 1/3

- Cox, David R (1972). "Regression Models and Life-Tables". *Journal of the Royal Statistical Society, Series B*. **34** (2): 187–220.

- Key predictors of mortality remain the same, no major new discoveries

- Perhaps the insights are hiding at deeper levels?

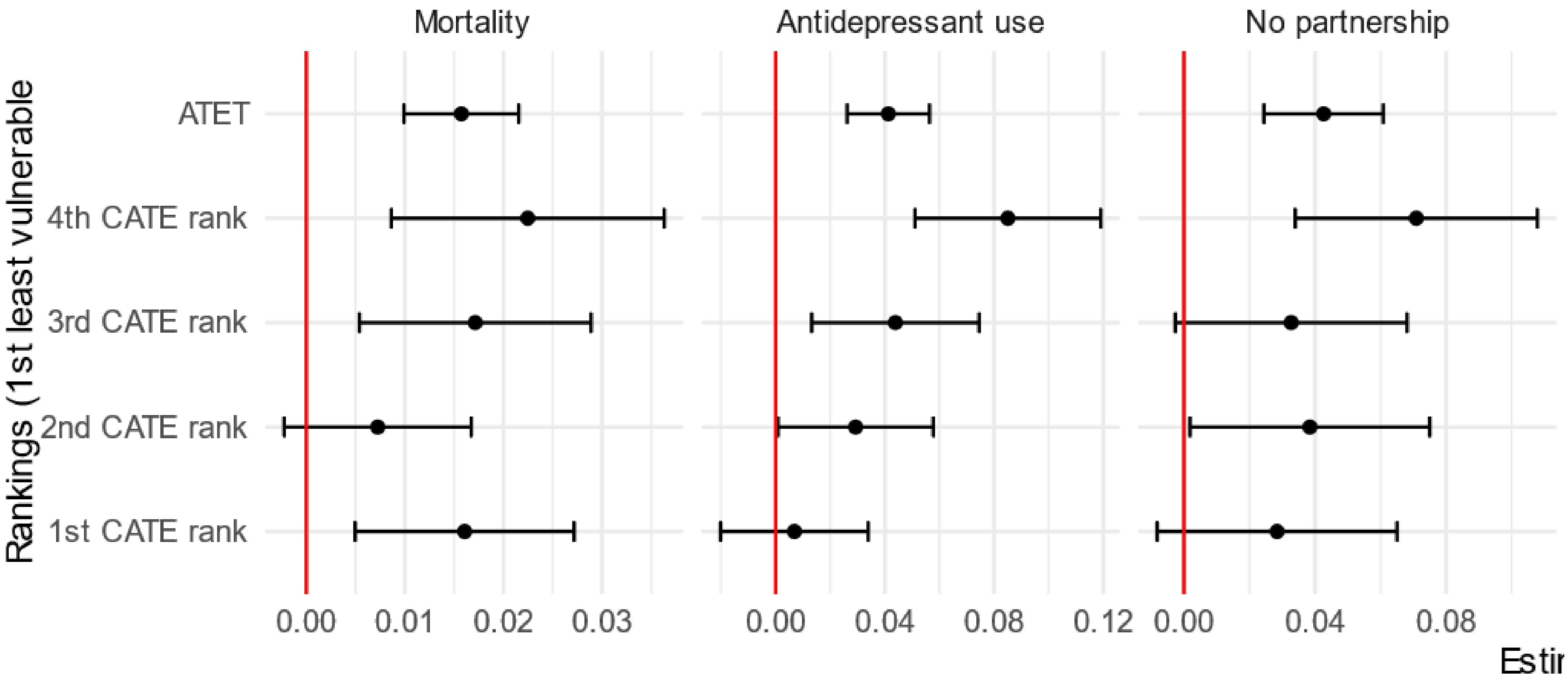
Heterogenous ”treatment” effects

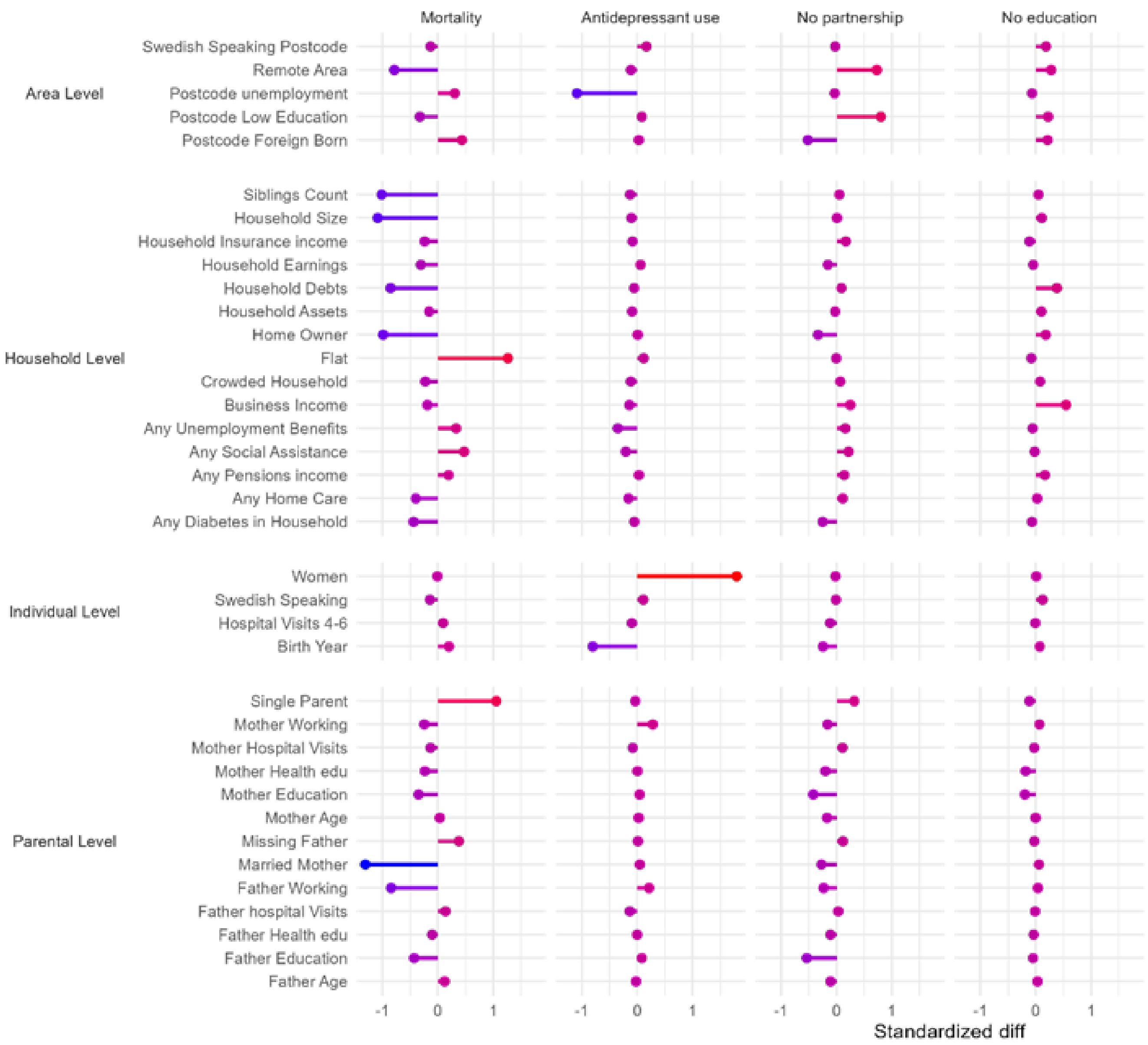
- The effect of a risk factor on outcome (mortality) may vary by sociodemographic groups
- Traditional interaction-based regression approaches limiting
- Machine learning methods (causal forests and the like) helpful

- Application: Type 1 diabetes and mortality

- Hiilamo, Myrskylä et al. 2025 American Journal of Epidemiology

Type 1 Diabetes and outcomes at age 30





- T1D predicts higher risk of death among those with low socioconomic status family

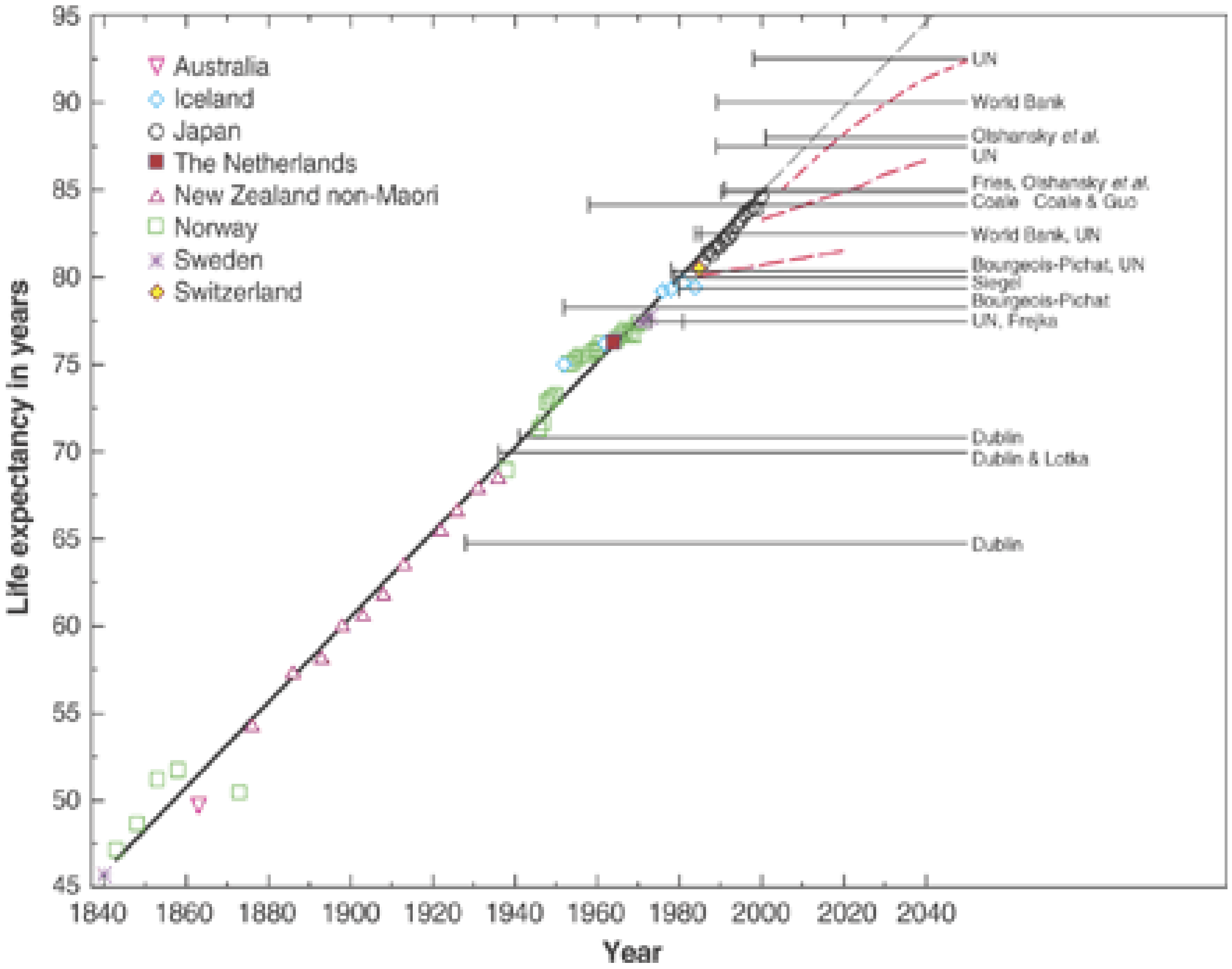
Summary 2/3

- Cox, David R (1972). "Regression Models and Life-Tables". *Journal of the Royal Statistical Society, Series B*. **34** (2): 187–220.

- Key predictors of mortality remain the same, no major new discoveries

- Perhaps the insights are hiding at deeper levels?
    - ML tools help to discover heterogeneity in “treatment effects” but the heterogeneity clusters around known risk factors

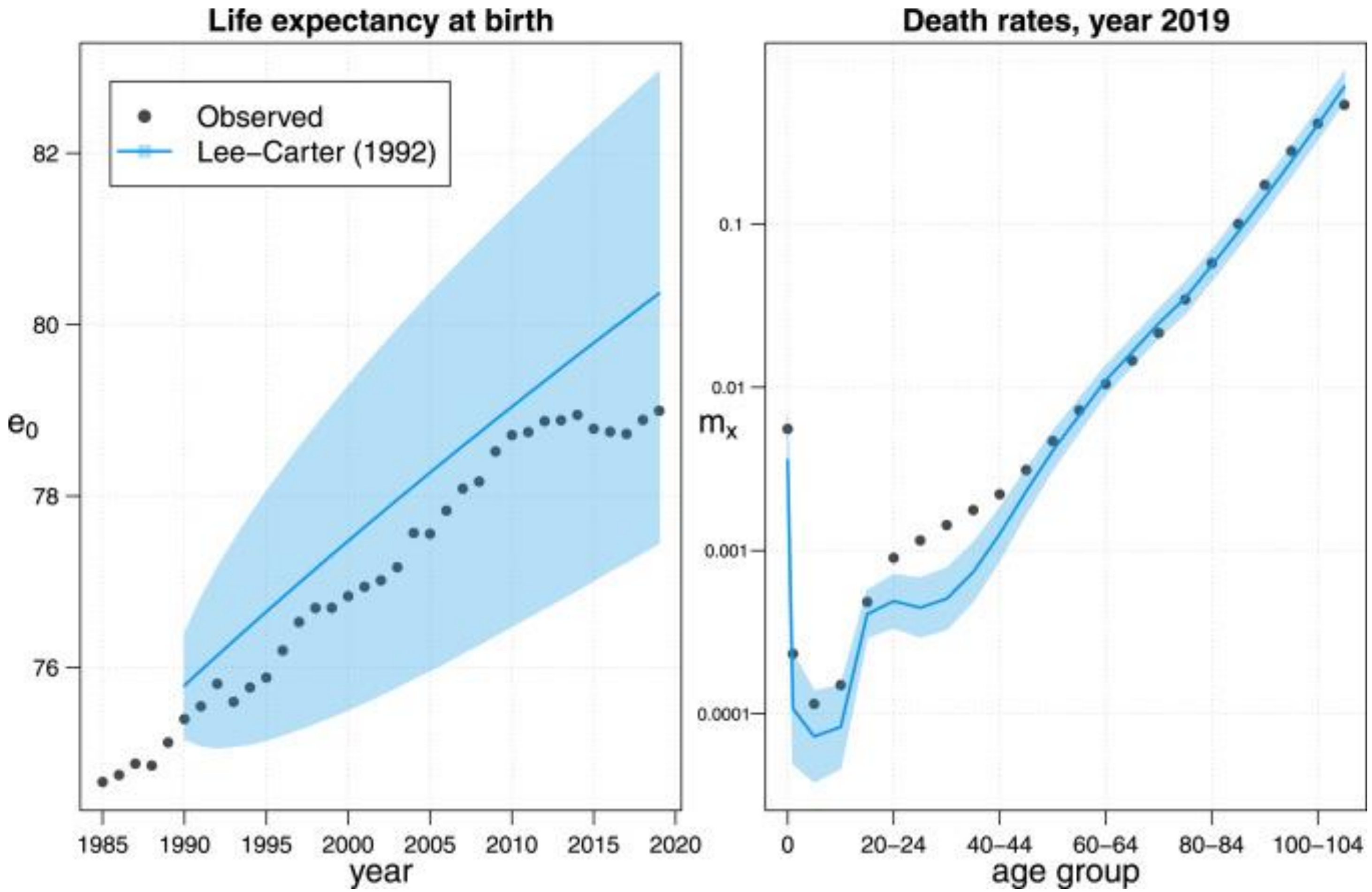
Science | AAAS



Broken Limits to Life Expectancy, Volume: 296, Issue: 5570, Pages: 1029-1031, DOI: (10.1126/science.1069675)

Progress!

- Lee-Carter 1992



ML Progress! Basellini & Camarda 2023

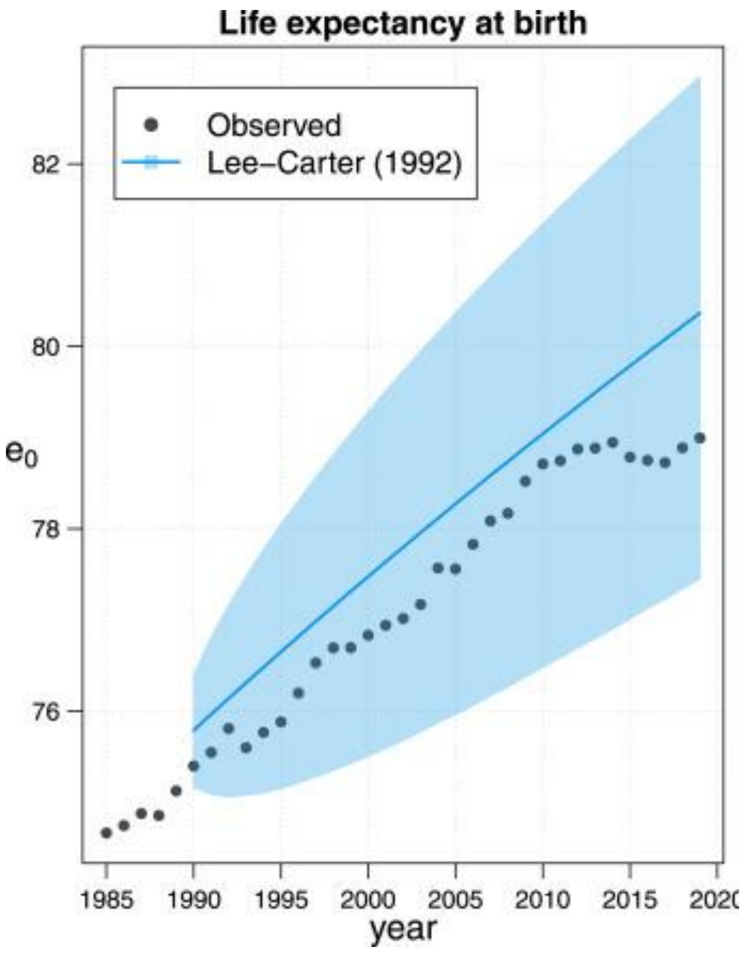
- Machine learning techniques have been implemented within the LC framework. Deprez et al. (2017) applied a regression tree boosting machine approach to improve the goodness-of-fit of the LC model and of the cohort-based extension of Renshaw and Haberman (2006). Levantesi and Pizzorusso (2019) extended this work by considering three alternative tree-based machine learning techniques. Moreover, these authors introduce an LC model enhanced by machine learning, whereby an additional set of LC parameters, derived from machine learning, is included in the LC model. Nigri et al. (2019) introduced a recurrent neural network (RNN) approach to model and forecast ; and Marino et al. (2022) extended this approach to derive prediction intervals. Richman and Wüthrich (2021) employed neural networks to estimate the parameters of the multipopulation coherent (Li & Lee, 2005) method (see 'Coherent mortality forecasts').

ML progress!

$$\ln(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t}$$

$$\sum_x b_x = 1 \qquad \text{and} \qquad \sum_t k_t = 0\,.$$

$$k_t = k_{t-1} + \delta + \beta D_f + e_t$$



Apicella et al. 2025

- Neural network Lee–Carter model and the actuarial relevance of longevity risk assessment

$$k_t = k_{t-1} + \delta + \beta D_f + e_t$$

- Neural network model for the “linear drift” component

Improved accuracy in mortality forecasting

Table 5. Forecast accuracy measures for the Danish population.

Model	RMSE	MAE	MAPE	PBS	WS	PS	CRPS
RW	13.512	11.076	13.137	3.469	59.098	7.298	7.235
ARX-NN(2)	9.658	7.538	8.835	2.106	34.754	5.392	5.344
ARX-NN(3)	9.646	7.446					
ARX-NN(4)	9.721	7.548					
ARX-NN(5)	9.762	7.565					
ARX-NN(6)	9.489	7.388					
ARX-NN(7)	9.258	7.204					
ARX-NN(8)	9.505	7.329					

Notes: Testing set = TE_1. ARX-NN = Autor
and lagged variables determined using t
size in parenthesis. The best ARIMA mod

Table 11. Forecast accuracy measures for the Norwegian population.

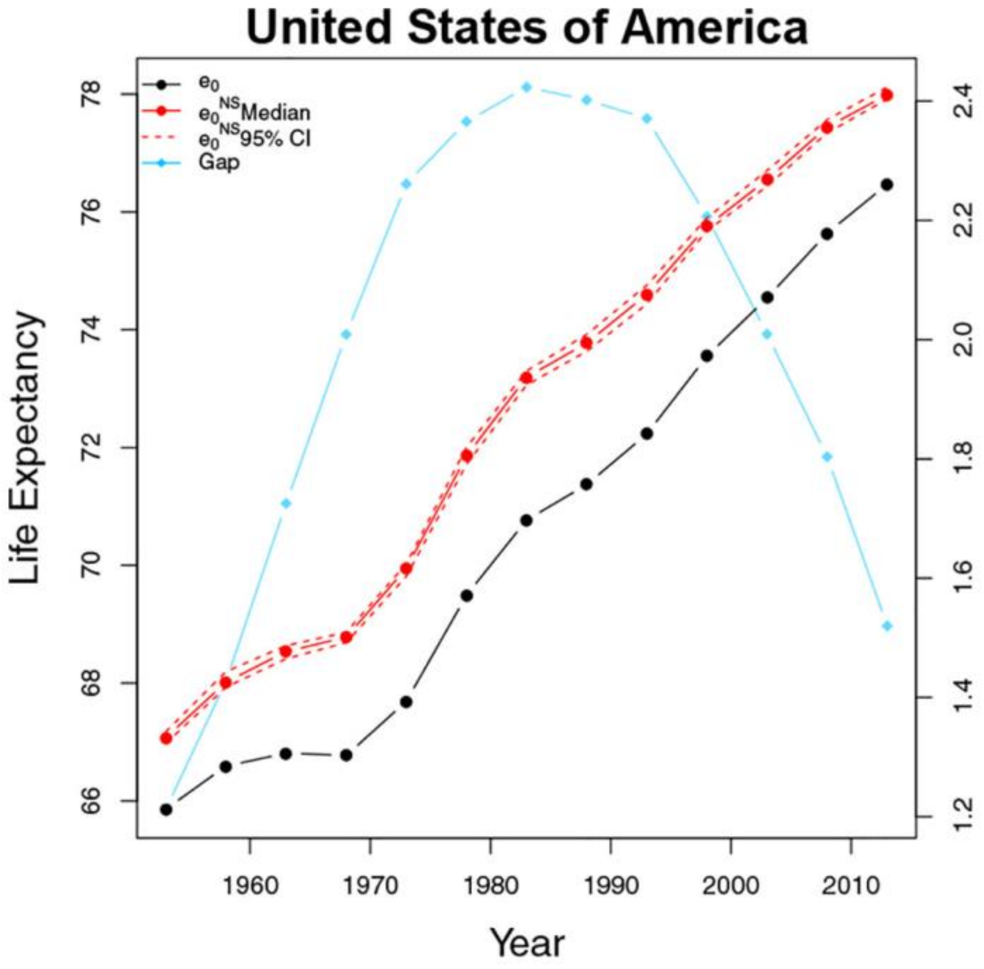
Model	RMSE	MAE	MAPE	PBS	WS	PS	CRPS
RW	6.084	4.935	6.060	3.473	75.858	4.637	4.598
ARX-NN(2)	7.154	6.101	8.593	1.245	43.553	4.615	4.575
ARX-NN(3)	7.272	6.241	8.738	1.243	44.518	4.664	4.623
ARX-NN(4)	5.482	4.444	6.447	1.377	40.673	3.606	3.574
ARX-NN(5)	5.175	4.264	6.123	1.506	41.541	3.438	3.408
ARX-NN(6)	5.034	4.209	5.798	1.543	39.128	3.188	3.160
ARX-NN(7)	5.060	4.232	5.968	1.477	39.563	3.283	3.255
ARX-NN(9)	5.355	4.504	6.420	1.490	41.062	3.484	3.454

Notes: Testing set = TE_1. ARX-NN = Autoregressive neural network model with 'Trend' as exogeneous variable and lagged variables determined using the Wang and Van Keilegom test and the Teräsvirta test. Hidden layer size in parenthesis. The best ARIMA model coincides with the RW model.

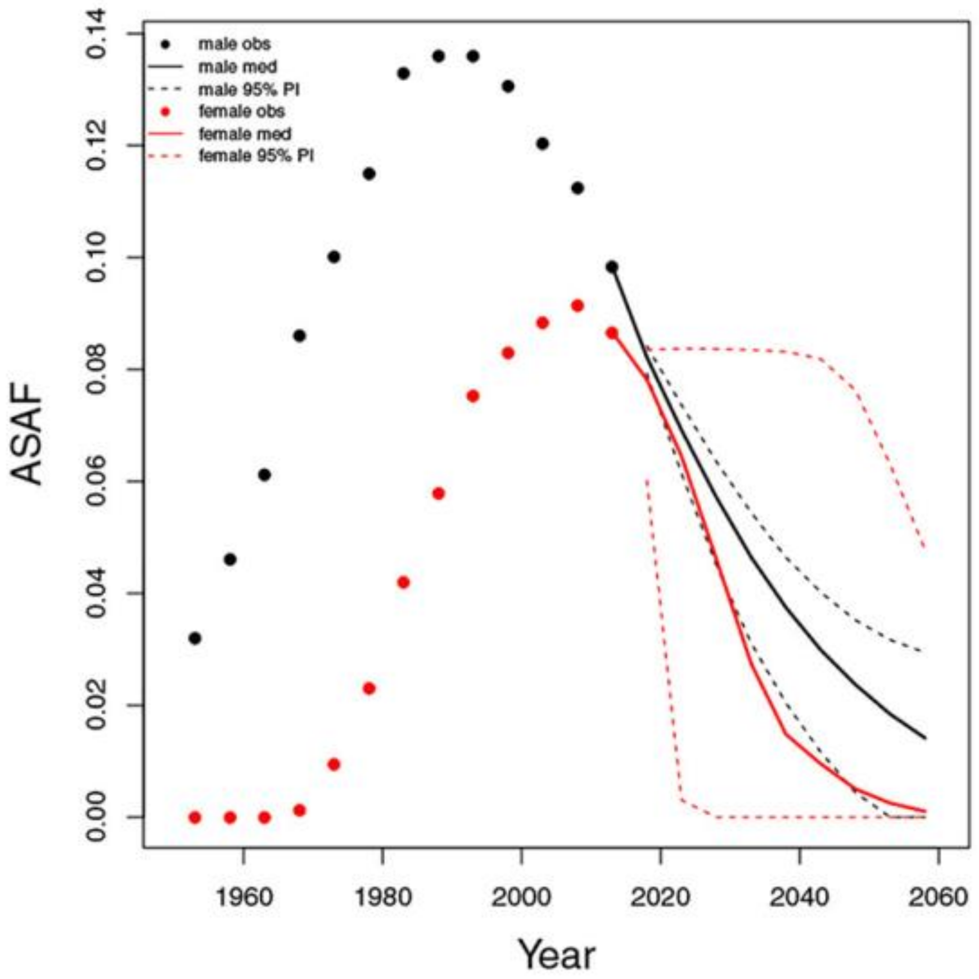
Fundamentals based approach to mortality

- Behaviors matter: Smoking, obesity, alcohol consumption – with these three, something can be forecasted.
- The most widely used determinant is smoking
  - Lots of data
  - Long lag to health effects

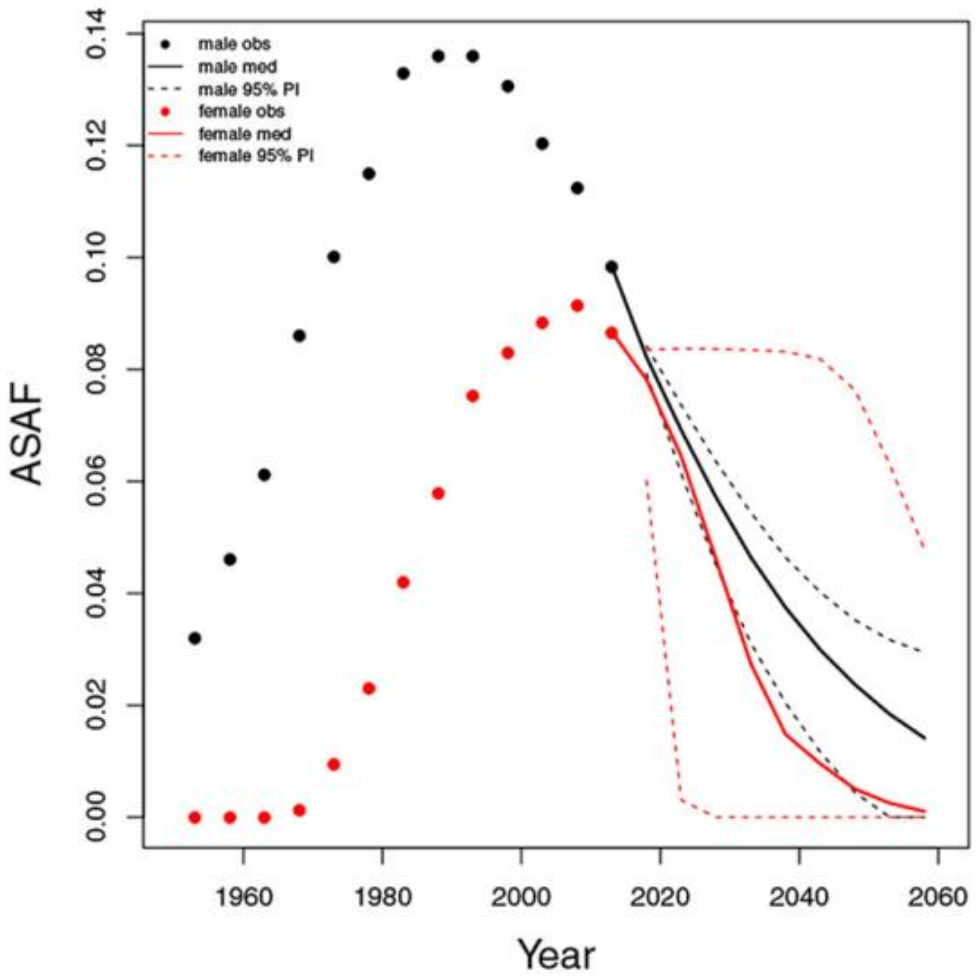
Smoking attributable mortality, example: US, black: true life expectancy, red: smoking-eliminated life expectancy

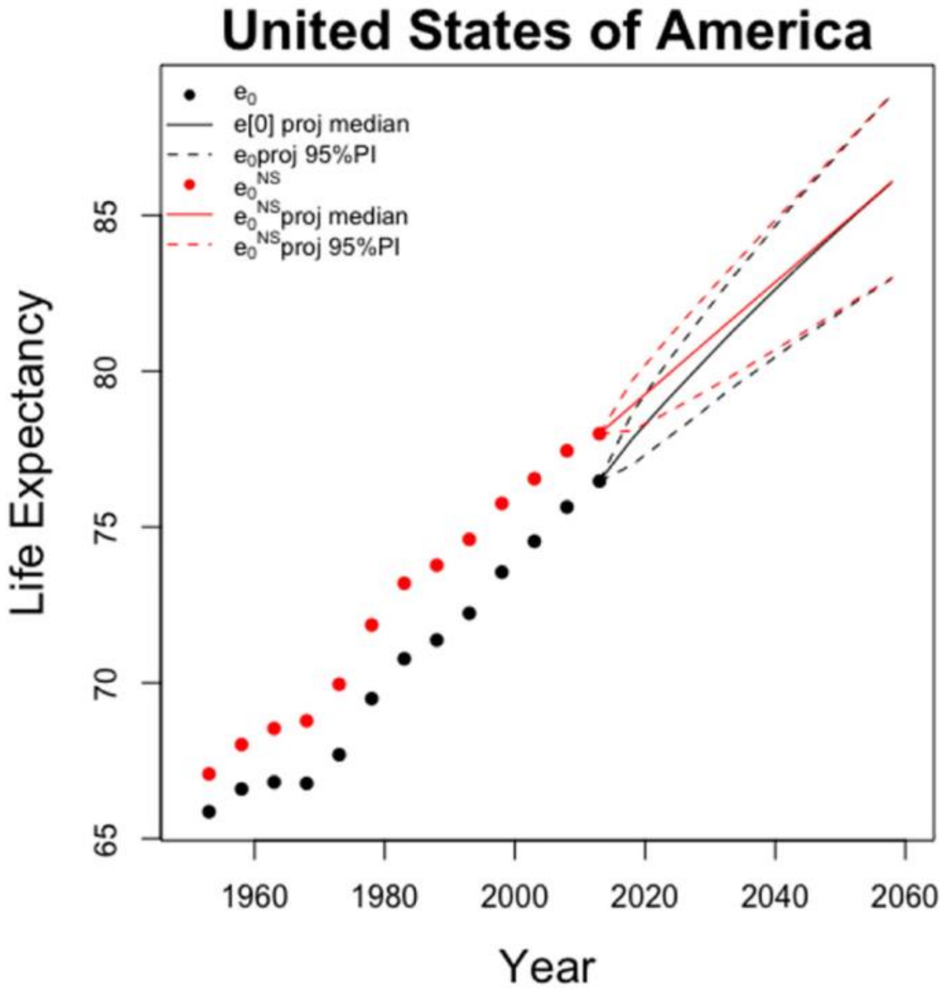


Forecasts of smoking attributable mortality, example: US, black: men, red: women. Li & Raftery 2021



Forecasts of smoking attributable mortality, example: US, black: men, red: women. Li & Raftery 2021





Period	Num	Sex	Method	MAE
Train:1950–2000	67	M	Lee-Carter	2.043
			Lee-Miller	1.536
			H-U FDA	2.206
			bayesLife	1.273
			smokeLife	**0.962**
Test: 2000–2015		F	Lee-Carter	1.210
			Lee-Miller	0.748
			H-U FDA	1.430
			bayesLife	0.876
			smokeLife	**0.728**

Period	Num	Sex	Method	MAE
Train:1950–2010	68	M	Lee-Carter	1.741
			Lee-Miller	0.853
			H-U FDA	1.364
			bayesLife	0.688
			smokeLife	**0.523**
Test: 2010–2015		F	Lee-Carter	1.025
			Lee-Miller	0.486
			H-U FDA	0.895
			bayesLife	0.464
			smokeLife	**0.319**

Summary 3/3

- Cox, David R (1972). "Regression Models and Life-Tables". *Journal of the Royal Statistical Society, Series B*. **34** (2): 187–220.

- Key predictors of mortality remain the same, no major new discoveries

- Population Level Forecasting
  - Lots of ML/AI/NN activity – much of that without ”real” determinants
  - Old-faschoned determinants still matter
  - Are fundamentals underrated?
  - ML-boosted model that accounts for smoking?

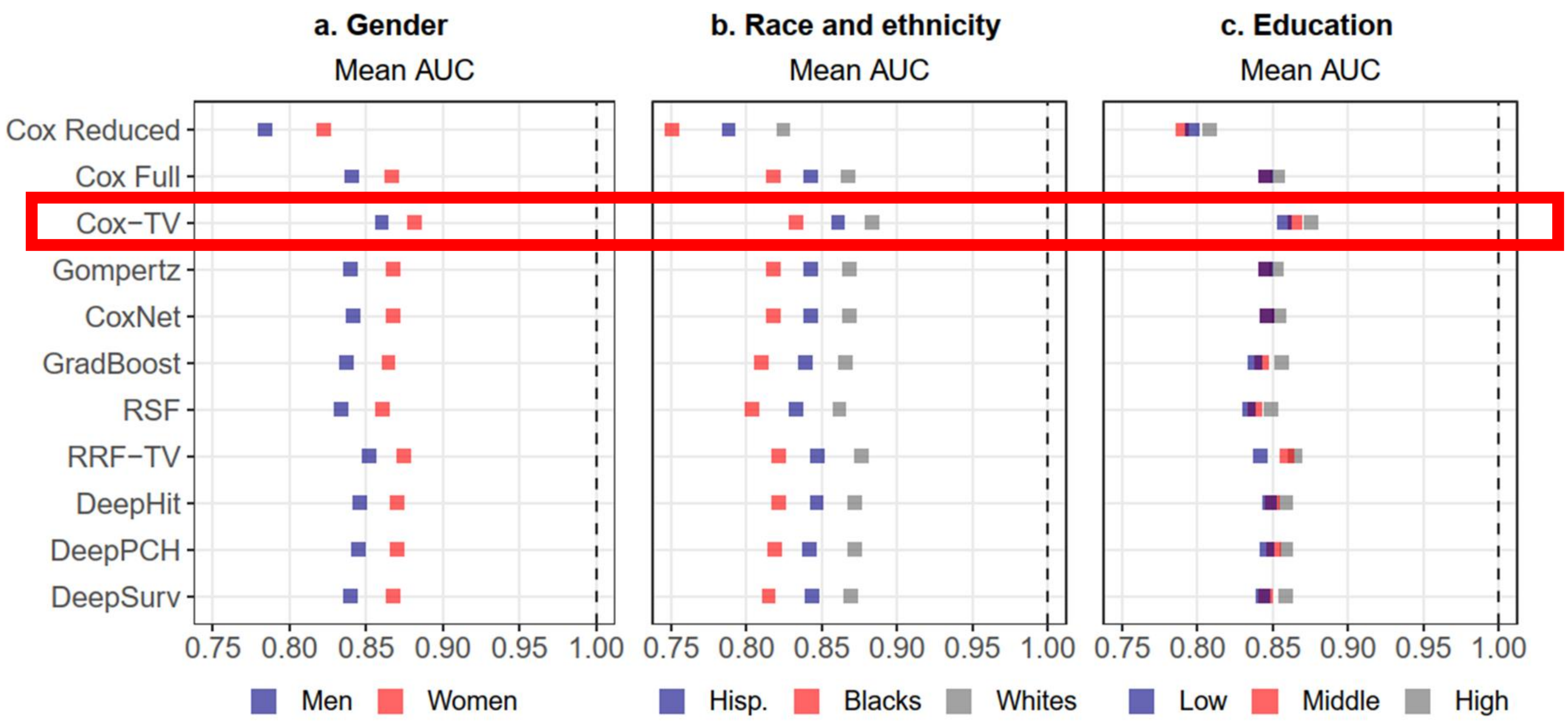


Figure 2: Integrated Brier Score and Mean Area Under the Curve by gender (panel a.), race and ethnicity (panel b.), and education (panel c.). The dashed vertical lines indicate the optimal Brier Score (0.00) and the Area Under the Curve (1.00). The Mean AUC for Kaplan-Meier, 0.5 by construction, has not been reported.