Case-Base Neural Networks: survival analysis with time-varying, higher-order interactions

Jesse Islam

Co-Authors:

Maxime Turgeon

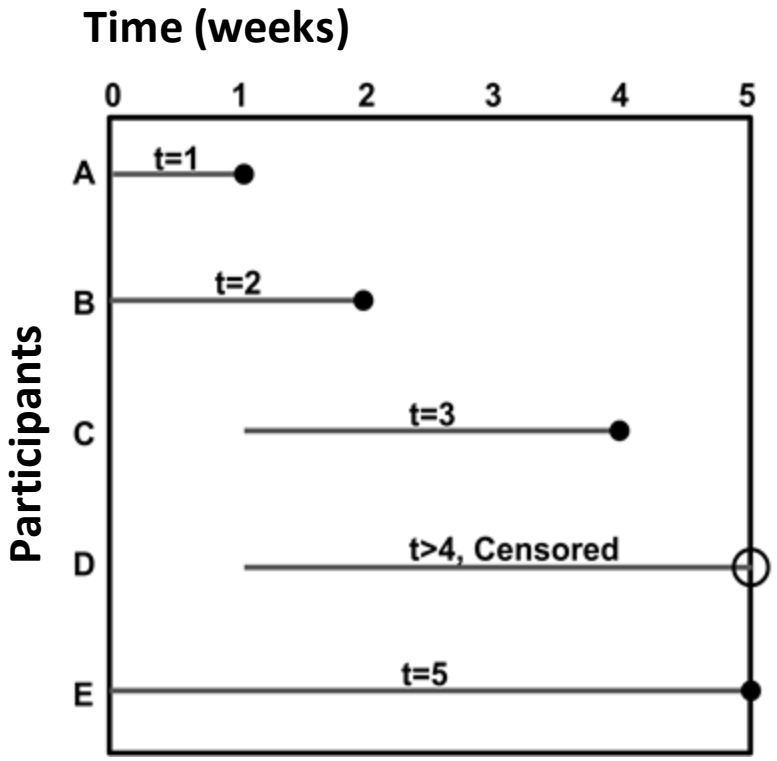
Rob Sladek

Sahir Bhatnagar

What is survival analysis?

- Any dataset concerning time to an event.
  - Time to death.
  - Time to graduation.
  - Time to getting a disease.
- Dataset consists of individuals who were followed over time.
  - Study may have a fixed duration or be open ended.
  - The event is not necessarily experienced until the study is over (“Censored”).
  - Participants may drop out early for any unrelated reason (“Censored”).

Survival time till assignment completion



ID	Survival Time (T)	Event?
A	1	1
B	2	1
C	3	1
D	4	0
E	5	1

Censored: Individual may experience the event of interest after follow-up has ended.

Hazard function:

$$h(X,t) = h_0(t)e^{\beta X}$$

Instantaneous potential of experiencing an event at time t, given you survived up to time t.

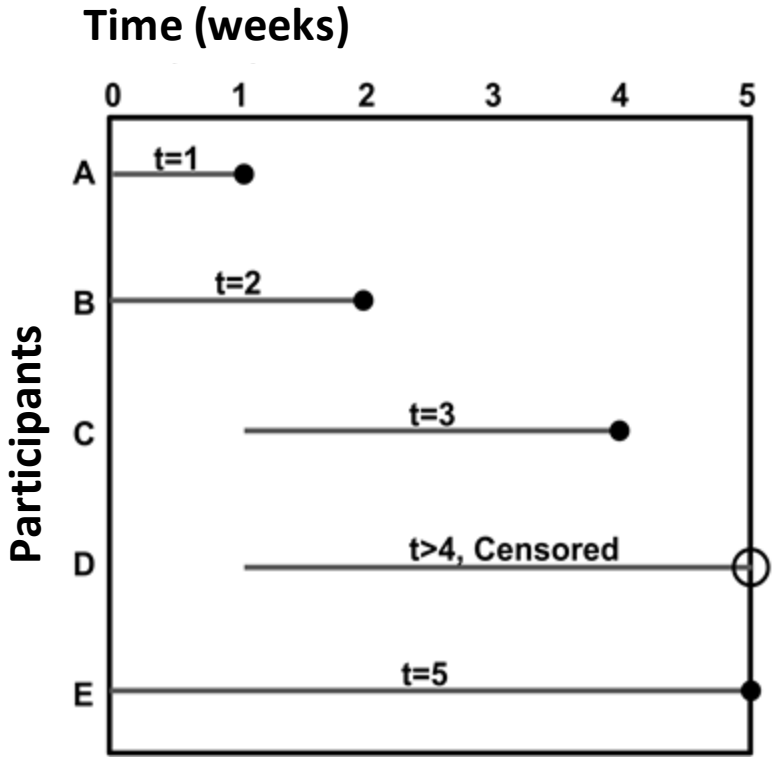
Hazard ratio:

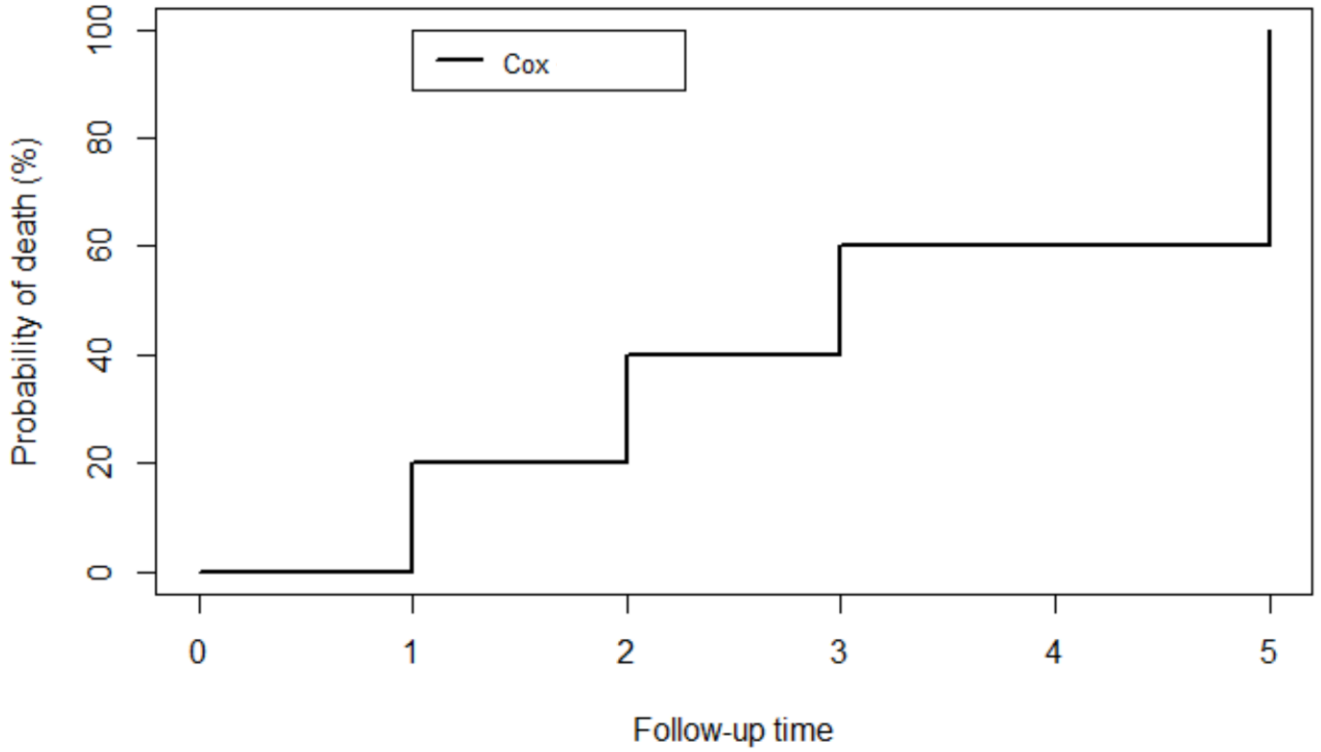
$$\frac{h(X,t)}{h(0,t)} = \frac{h_0(t)e^{\beta X}}{h_0(t)e^{0}} = e^{\beta X}$$

Cox regression: Assumes proportional hazards…
Effect of covariates do not vary with time.

- $h(X,t)$: hazard function.
- $h_0(t)$: baseline hazard.
- βX : linear predictor

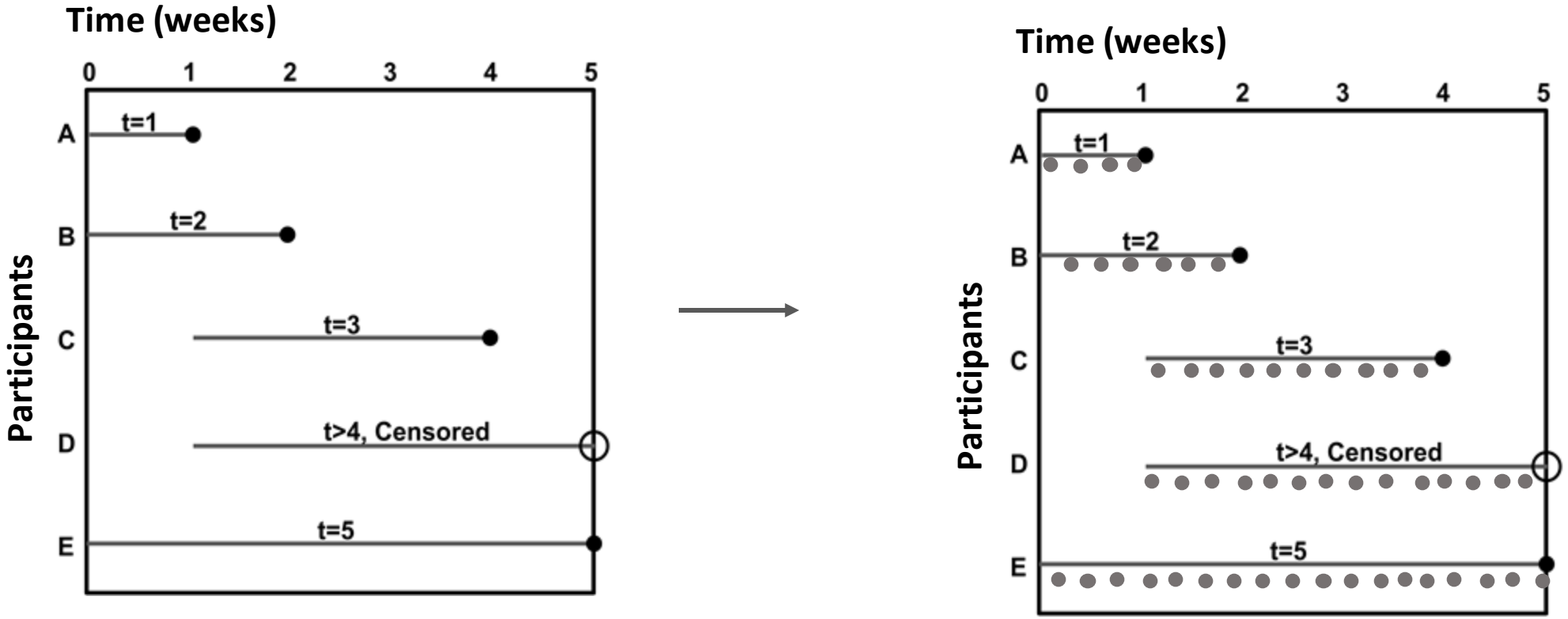
3-week risk?


Time (weeks)
0
1
2
3
4
5
Participants
A
t=1
B
t=2
C
t=3
D
t>4, Censored
E
t=5


Cox
Probability of death (%)
0
20
40
60
80
100
0
1
2
3
4
5
Follow-up time

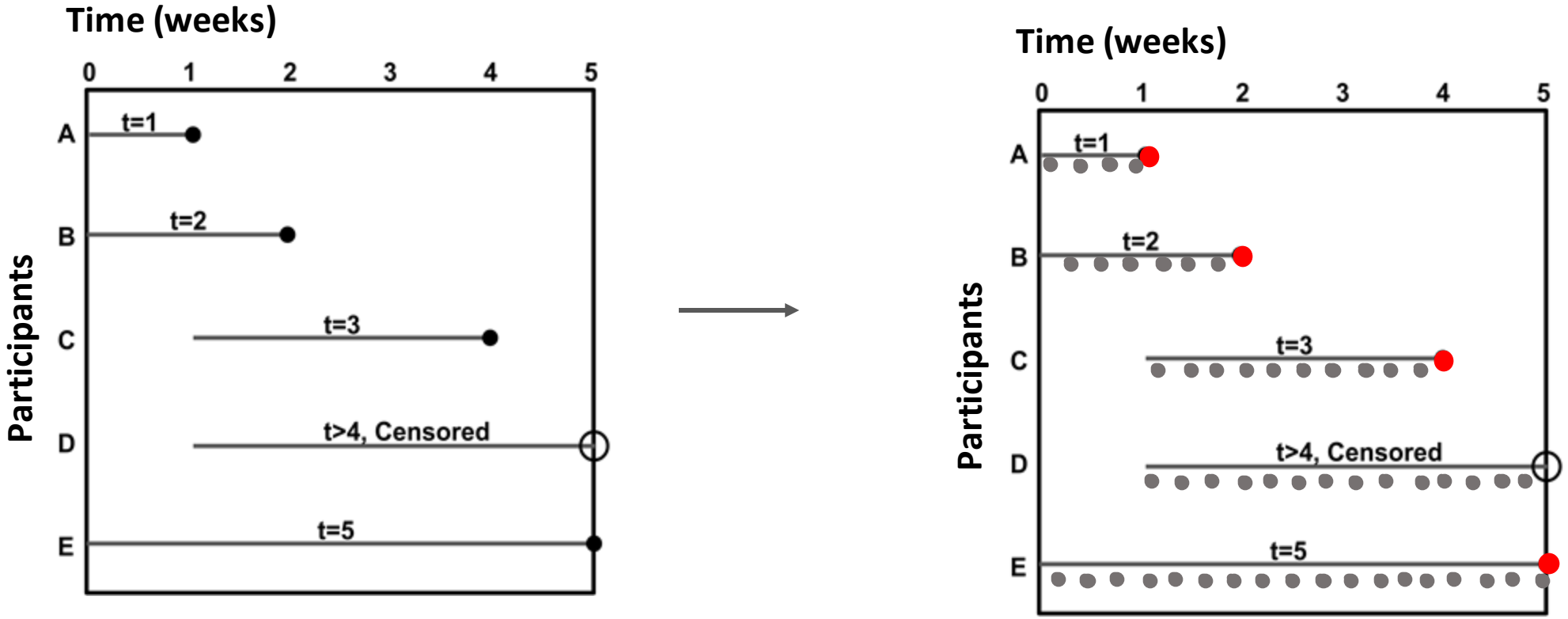
Case-base sampling with logistic regression

Case-base sampling



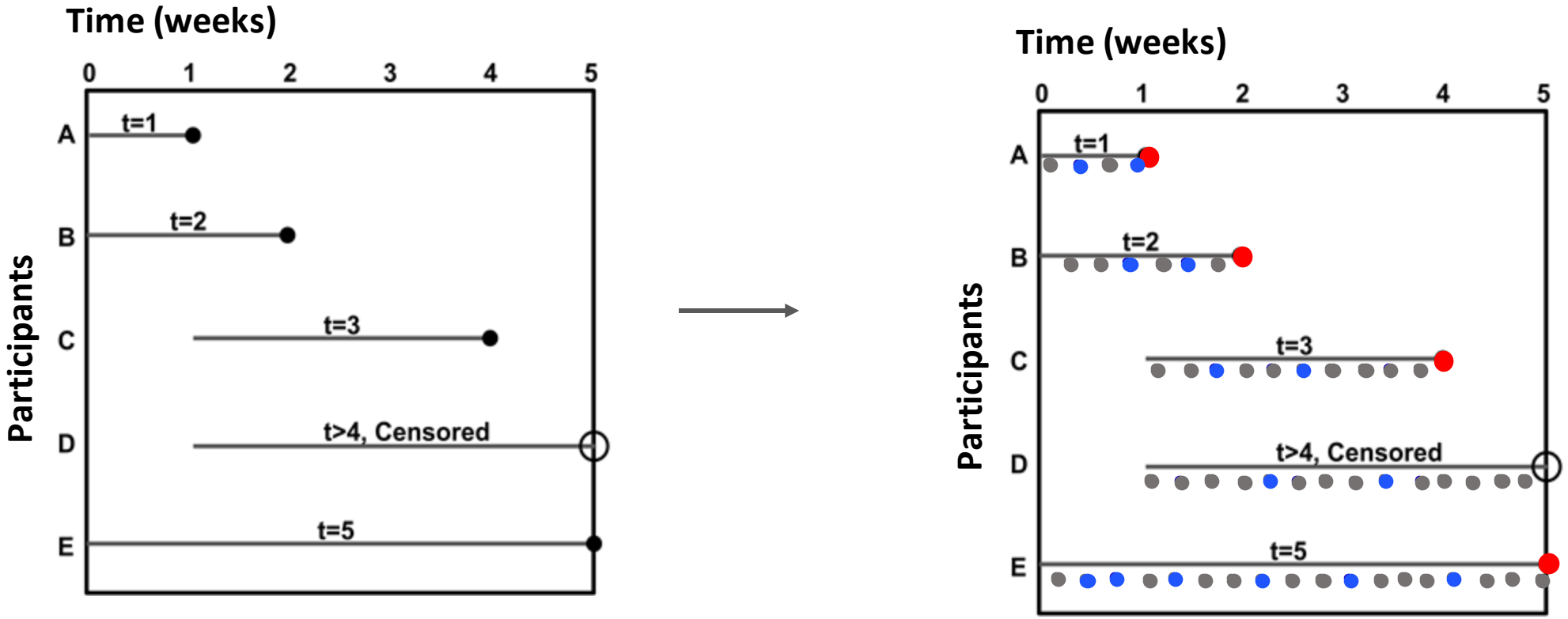
- Base: All the person-moments experienced in the study.

Case-base sampling



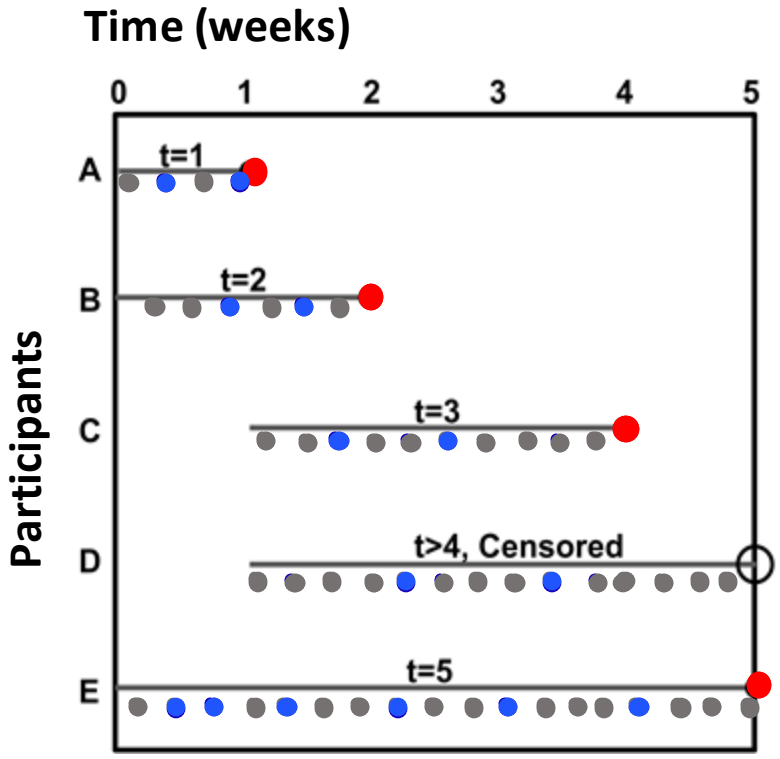
- Base: All the person-moments experienced in the study.
- Case series: all the person-moments where an event occurred.

Case-base sampling



- Base: All the person-moments experienced in the study.
- Case series: all the person-moments where an event occurred.
- Base series: sample of the base.

Case-base sampling and logistic regression



$$e^{\beta(x,t)} = \frac{Pr(Y=1|x,t)}{Pr(Y=0|x,t)}$$

$$\frac{Pr(Y=1|x,t)}{Pr(Y=0|x,t)} = \frac{h(x,t)*B(x,t)}{b[B(x,t)/B]}$$

$$\frac{h(x,t)*B(x,t)}{b[B(x,t)/B]} = \frac{h(x,t)*B}{b}$$

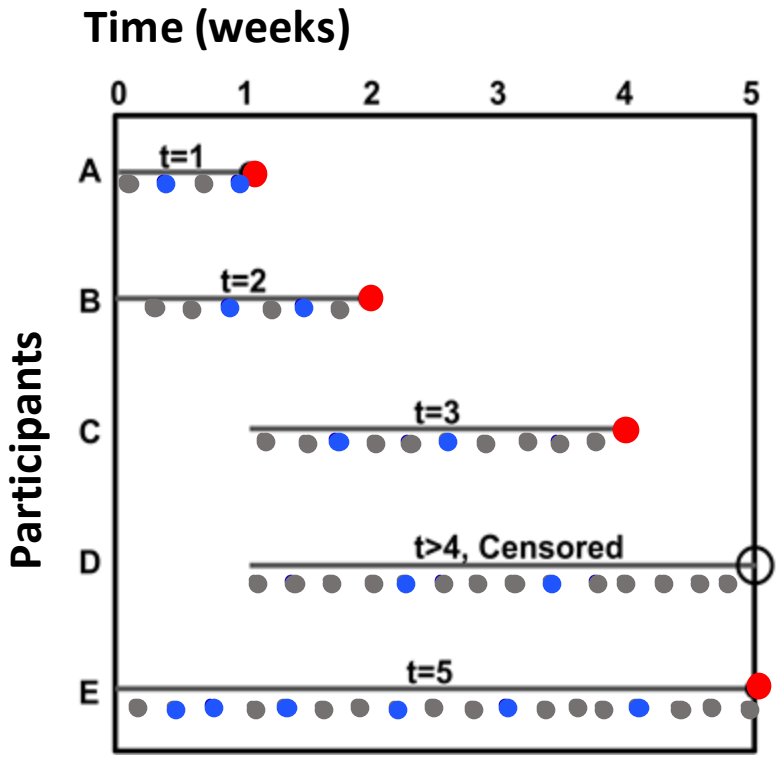
$$h(x,t) = e^{\beta(x,t)}\frac{b}{B}$$

$$ln(h(x,t)) = \beta(x,t) + ln\left(\frac{b}{B}\right)$$

b = # Blue

B = # Moments

Case-base sampling and logistic regression



$$e^{\beta(x,t)} = \frac{Pr(Y=1|x,t)}{Pr(Y=0|x,t)}$$

...

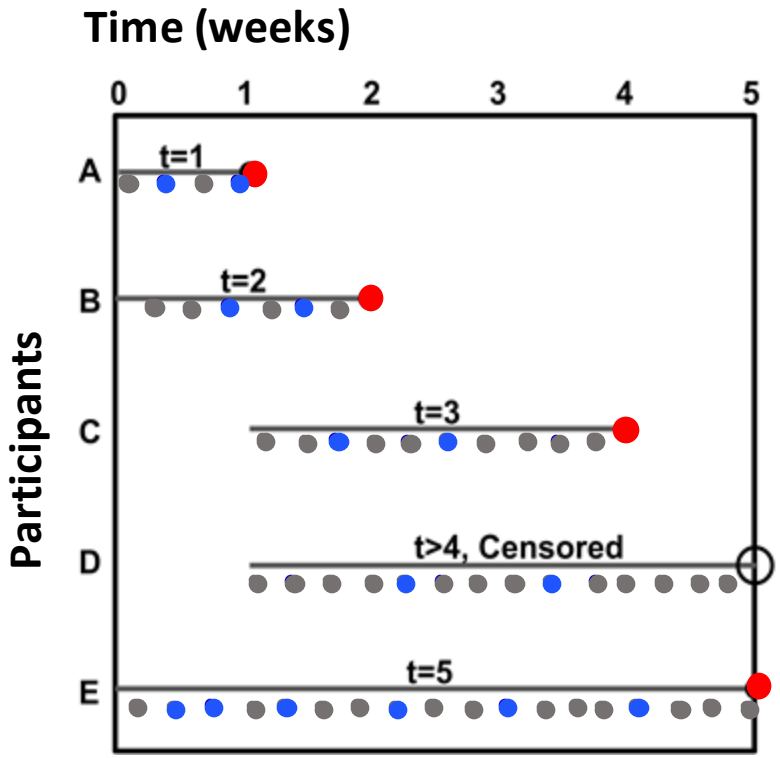
$$ln(h(\hat{x},t)) = \hat{\beta}(x,t) + ln\left(\frac{b}{B}\right)$$

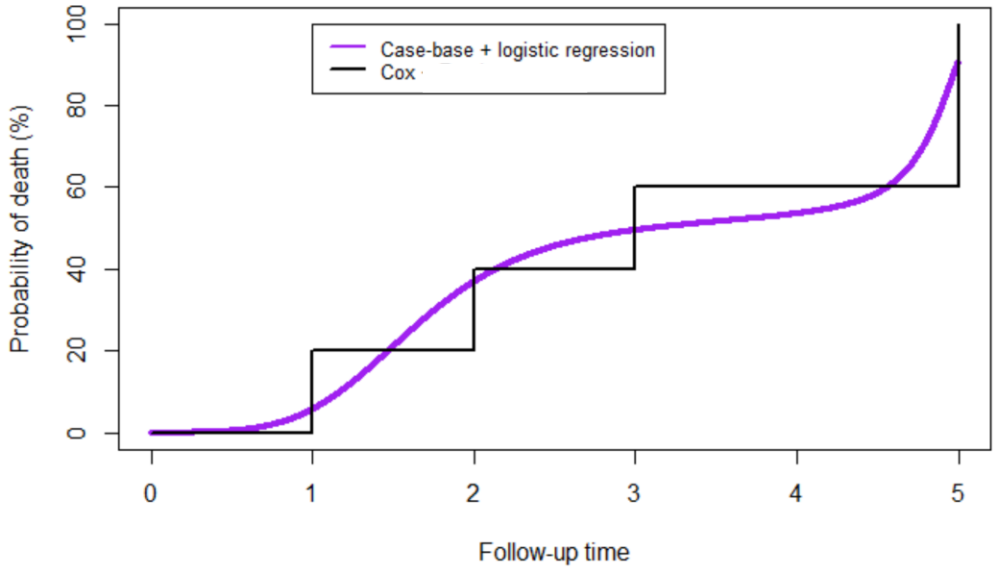
b = # (sample of moments)
B = # All moments

To have a flexible baseline hazard:

$$ln(\hat{h}(x,t)) = \hat{\beta}_{t_1}t + \hat{\beta}_{t_2}t^2 + \hat{\beta}_{t_3}t^3 + \hat{\beta}x + ln\left(\frac{b}{B}\right)$$

Case-base sampling and logistic regression





Case-base sampling permits flexible baseline hazard.

$$h(X,t) = h_0(t)e^{\beta X}$$

- $h(X,t)$: hazard function.
- $h_0(t)$: baseline hazard.
- βX : linear predictor

What about flexibility in covariates?

Case-base sampling permits flexible baseline hazard.

$$h(X,t) = h_0(t)e^{\beta X}$$

- $h(X,t)$: hazard function.
- $h_0(t)$: baseline hazard.
- βX : linear predictor

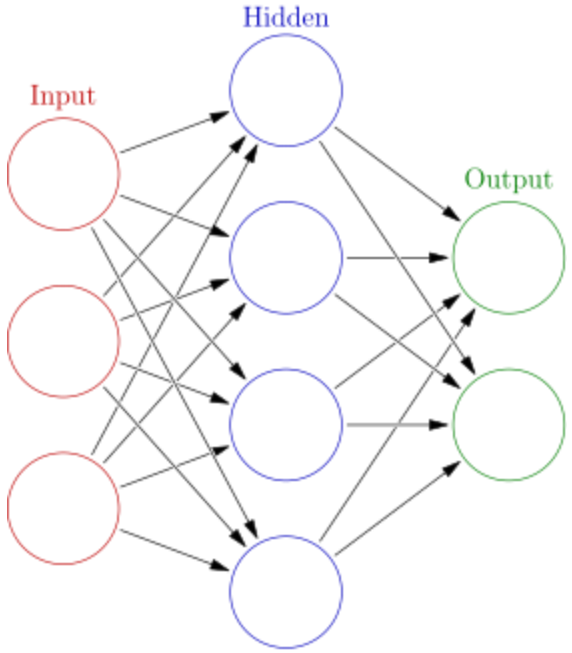
Exhaustive search with regression is hard

Many covariates.

- How many contribute?
- Interactions?
- Non-linearity?
- Genotypes, CT scans, etc.

Ideally, the model learns from the data.

- Neural networks can be used.
- *Case-base + NN = CBNN*



https://en.wikipedia.org/wiki/Artificial_neural_network

State of neural network survival analysis

DeepSurv – Cox neural networks.

- Cox regression extended using neural networks.
- Only uses proportional hazards (PH).

DeepHit – First Hitting Time neural networks.

- Inverse Gaussian distribution used as baseline hazard.
- Does not let model determine baseline hazard.

DeepSurvivalMachines (DSM) – Mixture model used for baseline hazard.

- User specifies a set of distributions to be used as the baseline hazard.
- Does not permit time-varying interactions.

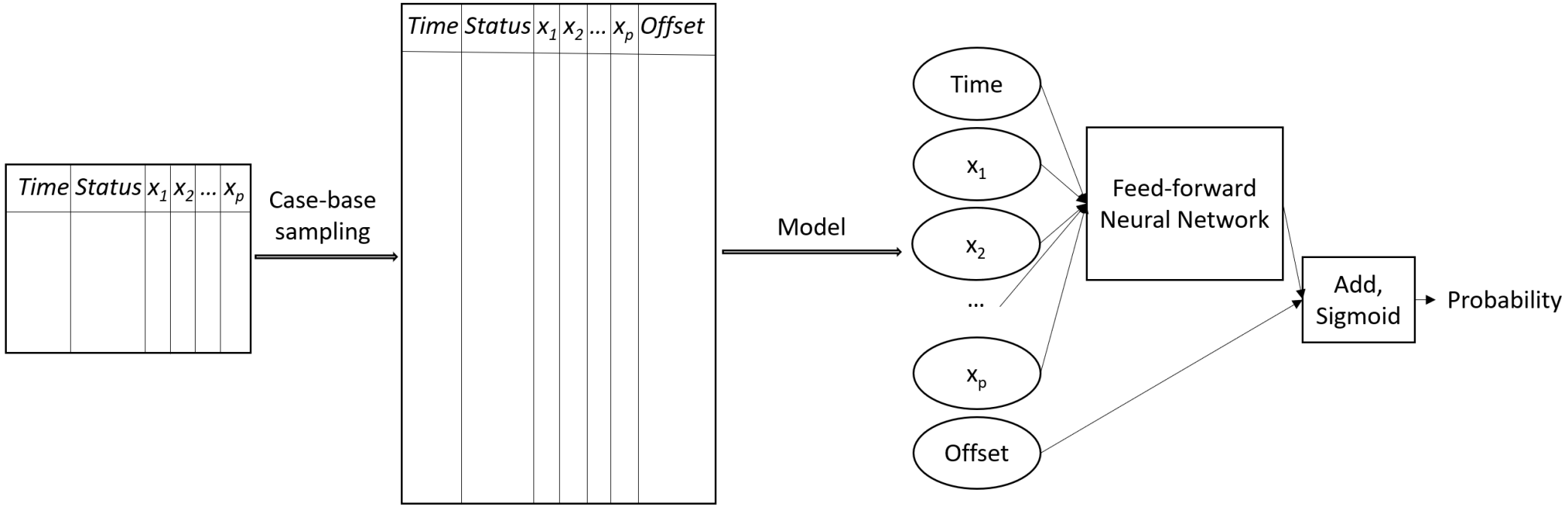
Need a parametric method that permits non-PH and flexible baseline hazard.

Proposal

Case-Base Neural Networks (CBNN)

- Provides a flexible baseline hazard.
- Permits time-varying interactions among covariates.

CBNN steps



1. Case-base sampling.
2. Neural network model.
3. Set offset to 0 when predicting on new data.

CBNN to hazard

Sigmoid to hazard

$$\log\left(h(t \mid X_i)\right) = \log\left(\frac{\text{sigmoid}\left(f_\theta(X,T) + \log\left(\frac{B}{b}\right)\right)}{1 - \text{sigmoid}\left(f_\theta(X,T) + \log\left(\frac{B}{b}\right)\right)}\right) + \log\left(\frac{b}{B}\right)$$

Sigmoid to hazard

$$\log\left(h(t \mid X_i)\right) = \log\left(\frac{\text{sigmoid}\left(f_\theta(X,T) + \log\left(\frac{B}{b}\right)\right)}{1 - \text{sigmoid}\left(f_\theta(X,T) + \log\left(\frac{B}{b}\right)\right)}\right) + \log\left(\frac{b}{B}\right)$$

$$= \log\left(\frac{\frac{\exp\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)}{\exp\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)+1}}{1 - \frac{\exp\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)}{\exp\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)+1}}\right) + \log\left(\frac{b}{B}\right)$$

Sigmoid to hazard

$$\log\left(h(t \mid X_i)\right) = \log\left(\frac{\text{sigmoid}\left(f_\theta(X,T) + \log\left(\frac{B}{b}\right)\right)}{1 - \text{sigmoid}\left(f_\theta(X,T) + \log\left(\frac{B}{b}\right)\right)}\right) + \log\left(\frac{b}{B}\right)$$

$$= \log\left(\frac{\frac{\exp\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)}{\exp\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)+1}}{1 - \frac{\exp\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)}{\exp\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)+1}}\right) + \log\left(\frac{b}{B}\right)$$

$$= \log\left(\exp\left(f_\theta(X,T) + \log\left(\frac{B}{b}\right)\right)\right) + \log\left(\frac{b}{B}\right)$$

Sigmoid to hazard

$$
\begin{aligned}
\log\left(h(t \mid X_i)\right) &= \log\left(\frac{\text{sigmoid}\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)}{1-\text{sigmoid}\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)}\right)+\log\left(\frac{b}{B}\right)\\
&= \log\left(\frac{\frac{\exp\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)}{\exp\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)+1}}{1-\frac{\exp\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)}{\exp\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)+1}}\right)+\log\left(\frac{b}{B}\right)\\
&= \log\left(\exp\left(f_\theta(X,T)+\log\left(\frac{B}{b}\right)\right)\right)+\log\left(\frac{b}{B}\right)\\
&= f_\theta(X,T)+\log\left(\frac{B}{b}\right)+\log\left(\frac{b}{B}\right)\\
&= f_\theta(X,T)
\end{aligned}
$$

Metrics and hyperparameters

Right-censored Brier score

$$\mathrm{BS}(t) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\left(\widehat{CI}(X,t)-1\right)^2 \cdot 1_{S_i \le t, \delta_i = 1}}{\hat{G}(S_i)} + \frac{\left(\widehat{CI}(X,t)\right)^2 \cdot 1_{S_i > t}}{\hat{G}(t)}\right)$$

- S_i <- Survival time of i-th individual.
- t <- Survival time of interest.
- $\widehat{CI}(X,t)$ <- Cumulative incidence.
- $\hat{G}(m)$<- Inverse probability censoring weighting (IPCW) at time *m*.
- δ_i <- Indicator: 1 = event , 0 = censored.

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, *18*(17-18), 2529-2545.

Index of Prediction Accuracy

$$\mathrm{IPA}(t) = 1 - \frac{BS_{model}(t)}{BS_{null}(t)}$$

- IPA > 0: model performs better than null.
- IPA < 0: model performs worse than null.

Kattan, M. W., & Gerds, T. A. (2018). The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagnostic and prognostic research*, *2*(1), 1-7.

Hyperparameters

- Epochs = 2000
- Batch size = 512
- Learning rate = 10e-3
- Decay = 10e-7
- Hidden layers = {50,50,25,25}
  - 50% dropout after each hidden layer.
- 60/20/20% train/validation/test.
- Stopping condition: minimum change in loss = 10e-7.

Simulation studies

Simulated covariates

$$z_1 \sim \text{Bernoulli}(0.5)$$

$$z_2 \sim \begin{cases} N(0, 0.5) & \text{if } z_1 = 0 \\ N(10, 0.5) & \text{if } z_1 = 1 \end{cases}$$

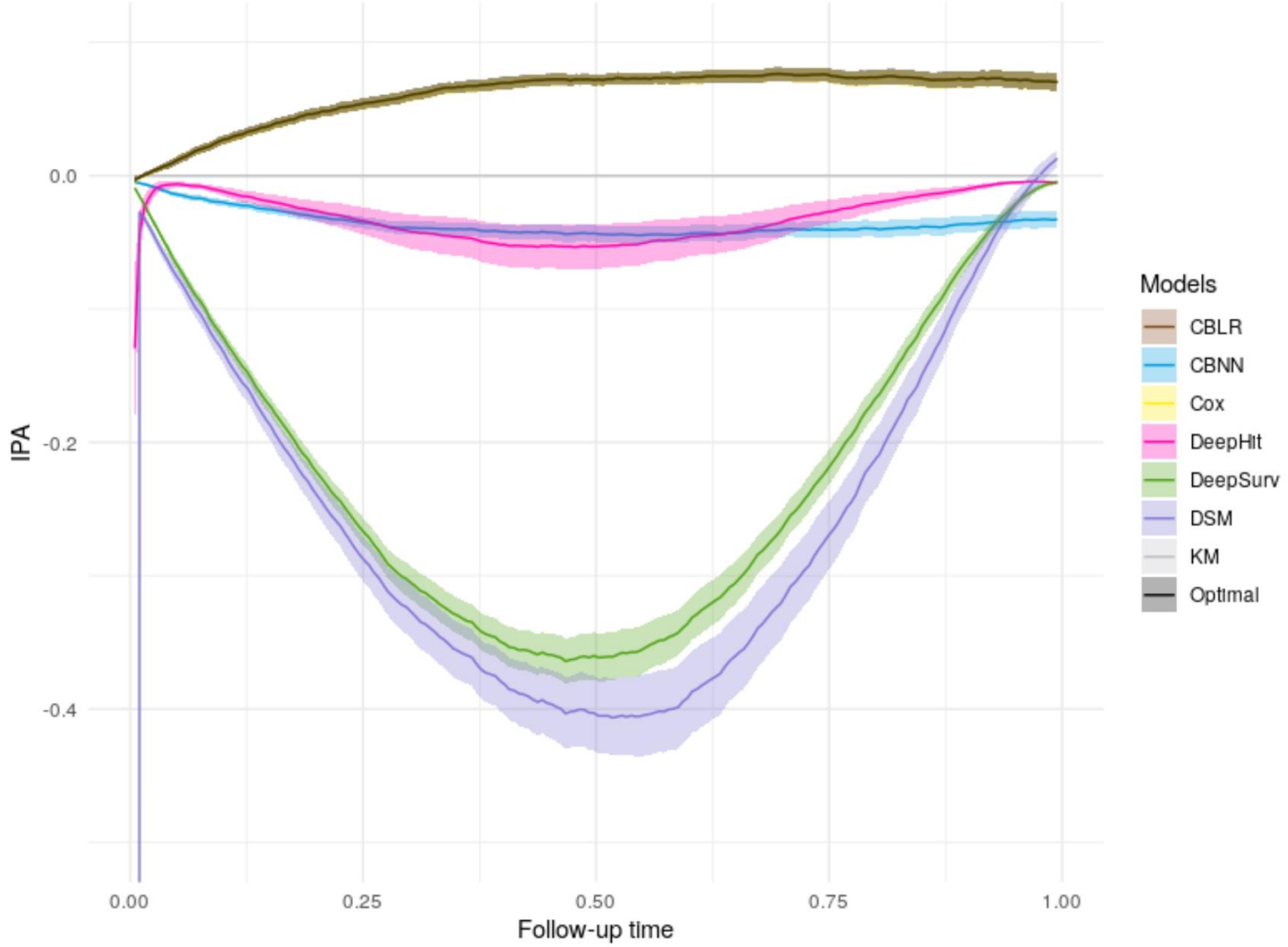
$$z_3 \sim \begin{cases} N(8, 0.5) & \text{if } z_1 = 0 \\ N(-3, 0.5) & \text{if } z_1 = 1 \end{cases}$$

Simple simulation

$$h(t \mid X_i) = \lambda \cdot e^{\beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3}$$

- $\beta 1 = \beta 2 = \beta 3 = 0.1$
- $\lambda = 1.0$

Simple simulation result



Complex simulation

$$h(t \mid X_i) = \sum_{i=1}^{5}(\gamma_i \cdot basis_i) + \beta_1 \cdot z_1 + \beta_2 \cdot z_2 + \beta_3 \cdot z_3 + \boxed{\tau_1 \cdot z_1 \cdot z_2 \cdot time} + \boxed{\tau_2 \cdot z_1 \cdot z_3 + \tau_3 \cdot z_2 \cdot z_3}$$

Time-varying interaction

Interactions

$\beta 1 = \beta 2 = \beta 3 = 1$

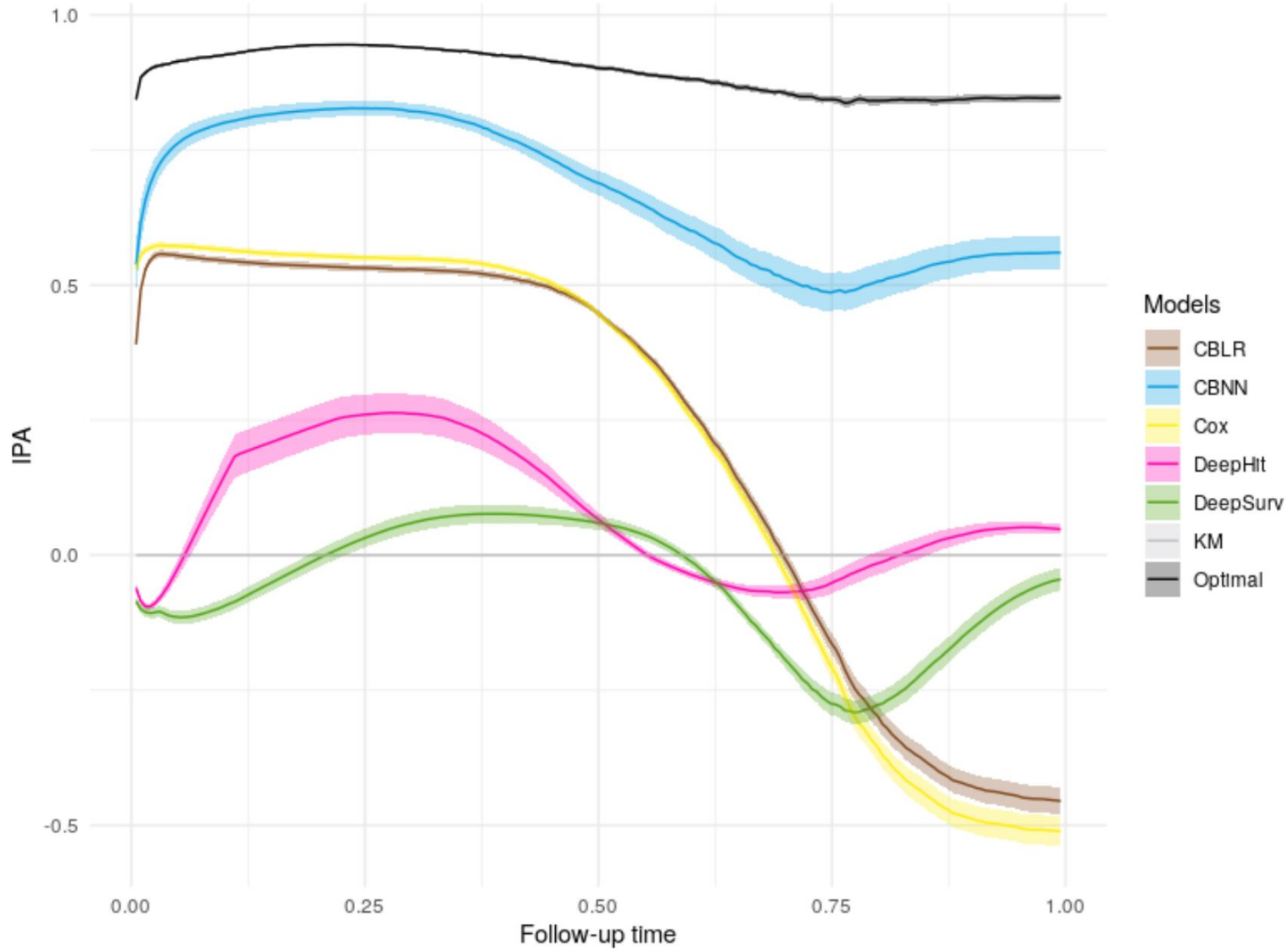
$\tau 1 = 10, \tau 2 = 2, \tau 3 = 2$

$\gamma 1 = 3.9, \gamma 2 = 3, \gamma 3 = -0.43, \gamma 4 = 1.33, \gamma 5 = -0.86$

- Breast cancer dataset from the *Flexsurv* package.
- 686 patients with primary node positive breast cancer.
  - 43% die over 7.28 years.
  - Breast cancer dataset originally used to demonstrate the benefit of flexible baseline hazards.

Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. *Stat Med* 2013;**32**(23):4118-4134.

Complex simulation result



Real data studies

SUPPORT study

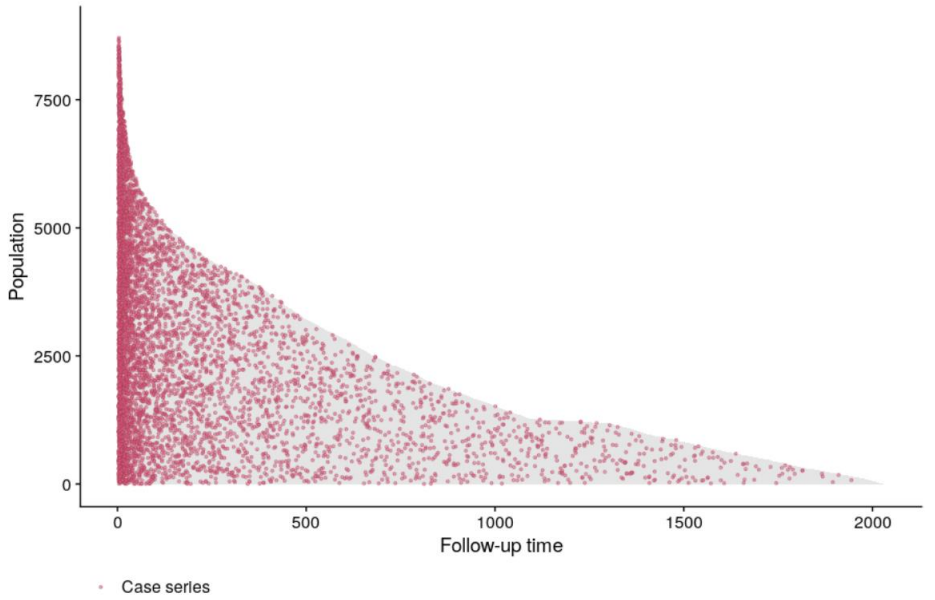
Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT) Phase I

8873 hospitalized adults.

- Followed up to 5.56 years.
- 68% incidence (death).
- 14 covariates (after imputation).

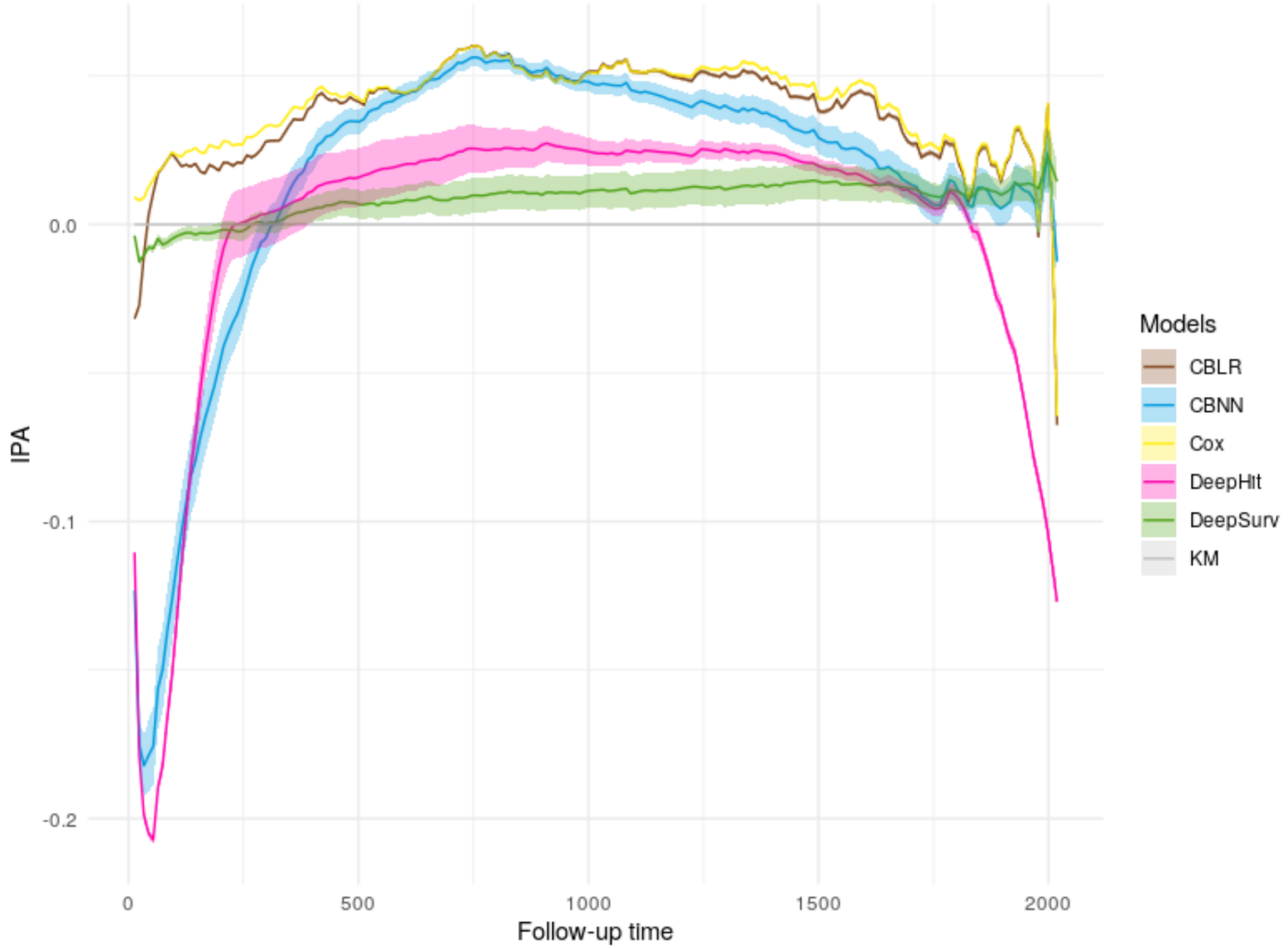
Requires imputation. For comparison with competitors a preprocessed version from DeepSurv is used.

- Age, sex, race, number of comorbidities, presence of diabetes/dementia/cancer, blood pressure, heart/respiration rate, temperature, white blood cell count, sodium and creatinine.



Knaus, 1995

SUPPORT result



METABRIC study

Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)

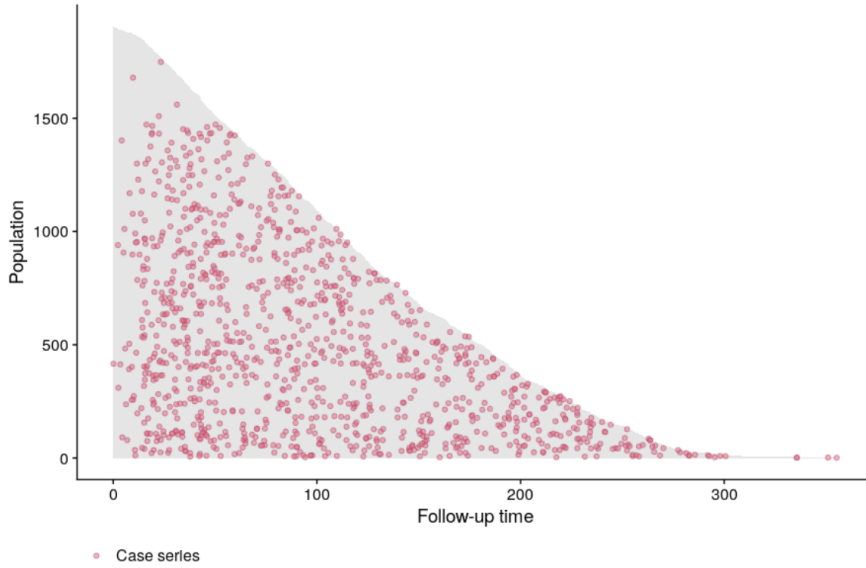
1980 individuals:

- 57.72% die due to breast cancer.
- with 30 years of follow-up.

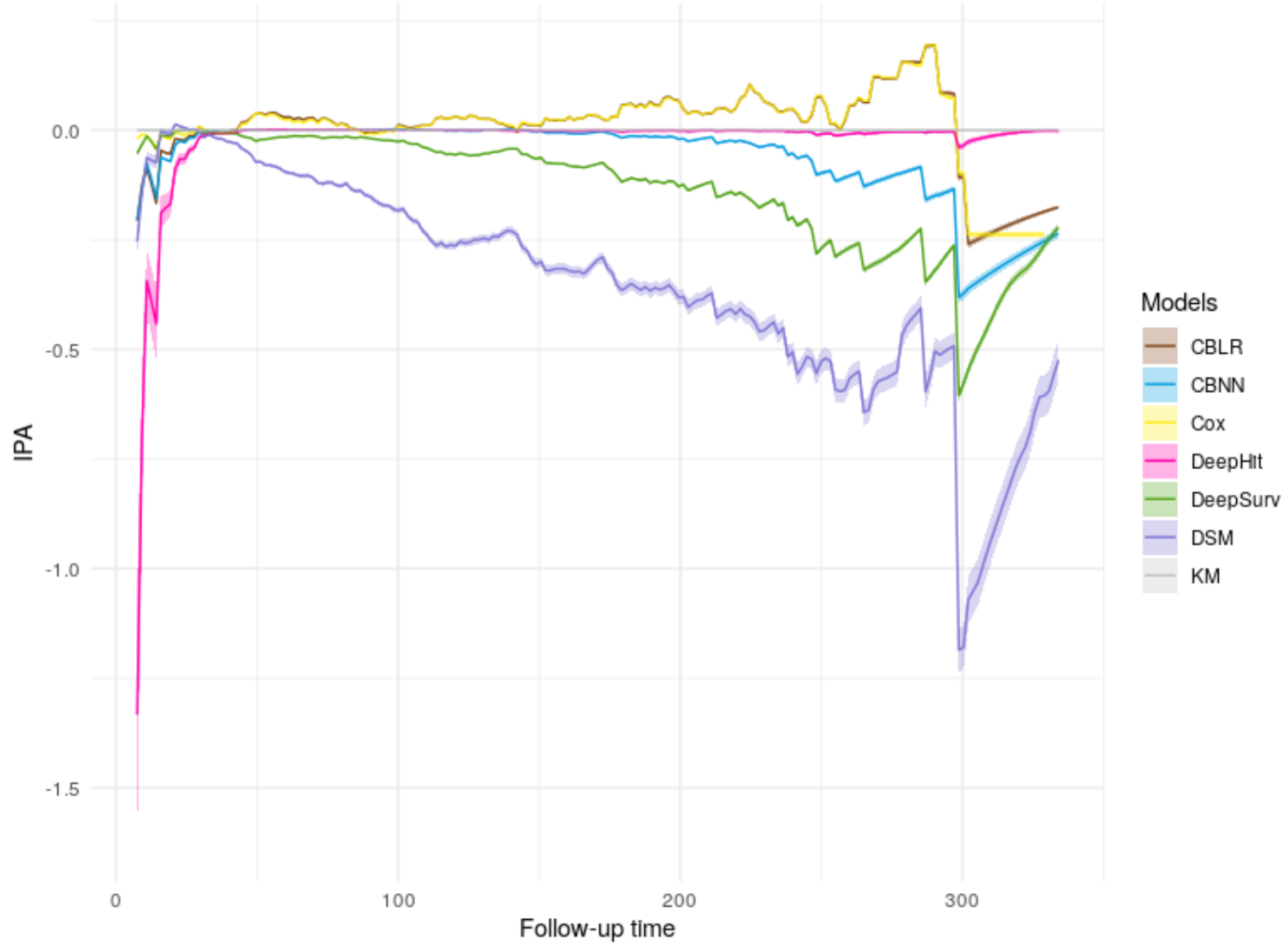
There are 9 covariates in total for this study (from Deepsurv):

- 4 genes (MKI67, EGFR, PGR, and ERBB2).
- 5 clinical features:
  - (hormone treatment/radiotherapy/chemotherapy/ER-positive indicator and age at diagnosis).

Pre-processed version from DeepSurv is used.



METABRIC result



Conclusion

If time varying interactions and a flexible baseline hazard without user specification are of interest, CBNN Should be strongly considered.

- Provides a parametric, flexible baseline hazard.
- Permits time-varying effects of covariates.
- Applicable to high-dimensional datasets.

https://github.com/Jesse-Islam/cbnn

https://github.com/Jesse-Islam/cbnnManuscript