Faculty of Health Sciences

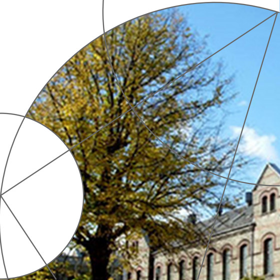
R^2-type Curves for Dynamic Predictions from Joint Longitudinal-Survival Models

Inference & application to prediction of kidney graft failure

Paul Blanche

joint work with

M-C. Fournier & E. Dantan (Nantes, France)



Context & Motivation

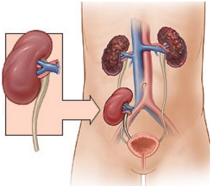
- Medical researchers hope to improve patient management using **earlier diagnoses**
- Statisticians can help by fitting **prediction models**
- The making of so-called **"dynamic"** predictions has recently received a lot of attention
- In order to be useful for medical practice, predictions should be "accurate"

How should we **evaluate dynamic prediction** *accuracy?*

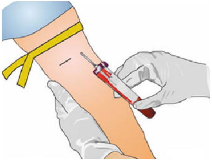
Data & Clinical goal

- **Data:**

  DIVAT cohort data
  of kidney transplant recipients
  (subsample, $n = 4,119$)

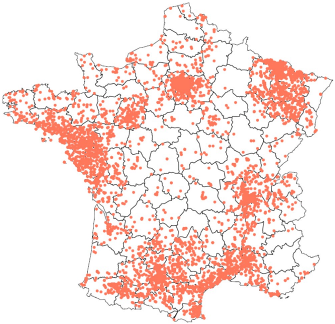


- **Clinical goal:**
  - Dynamic prediction of risk of kidney graft failure (death or return to dialysis)
  - Using repeated measurements of serum creatinine



DIVAT data sample (n=4,119)

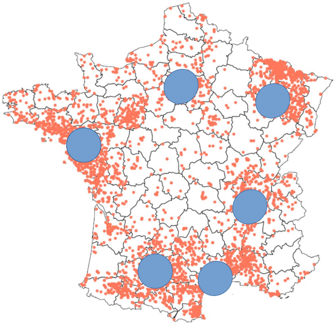
- French cohort
- Adult recipients
- Transplanted after 2000
- Creatinine measured every year



(www.divat.fr)

DIVAT data sample (n=4,119)

- French cohort
- Adult recipients
- Transplanted after 2000
- Creatinine measured every year
- 6 centers



(www.divat.fr)

Statistical challenges discussed

How to **evaluate** *and/or compare dynamic predictions?*

- **Using concepts** of:
  - Discrimination
  - Calibration

- **Accounting** for:
  - Dynamic setting
  - Censoring issue

Basic idea & concepts for evaluating predictions

Basic idea: comparing predictions and observations

Basic idea & concepts for evaluating predictions

Basic idea: comparing predictions and observations (*simple*!)

Basic idea & concepts for evaluating predictions

Basic idea: comparing predictions and observations (*simple*!)

Concepts:

- **Discrimination**:

- **Calibration**:

Basic idea & concepts for evaluating predictions

Basic idea: comparing predictions and observations (*simple*!)

Concepts:

- **Discrimination**:
A model has high discriminative power if the range of predicted risks is wide and subjects with low (high) predicted risk are more (less) likely to experience the event.

- **Calibration**:

Basic idea & concepts for evaluating predictions

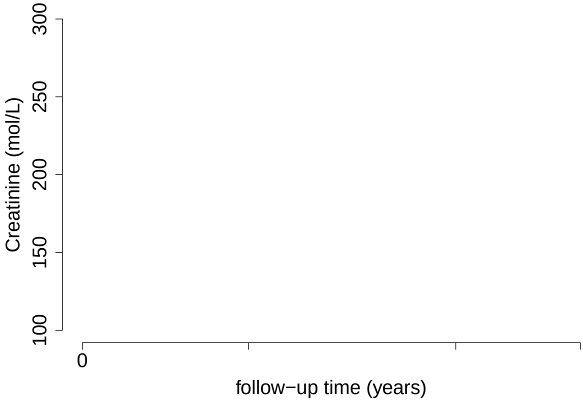
Basic idea: comparing predictions and observations (*simple*!)

Concepts:

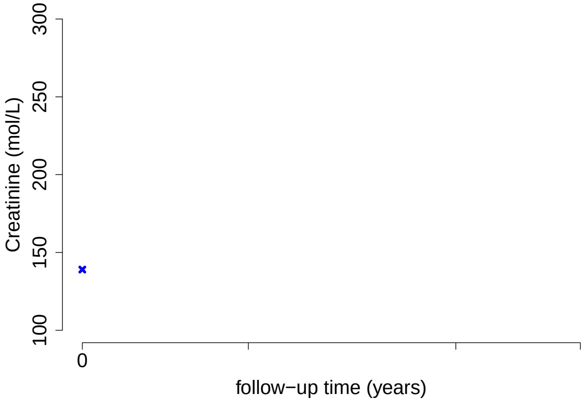
- **Discrimination**:
A model has high discriminative power if the range of predicted risks is wide and subjects with low (high) predicted risk are more (less) likely to experience the event.

- **Calibration**:
A model is calibrated if we can expect that x subjects out of 100 experience the event among all subjects that receive a predicted risk of x% ("**weak**" definition).

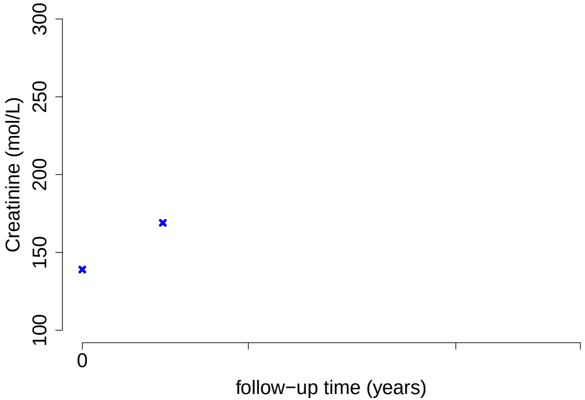
Dynamic prediction



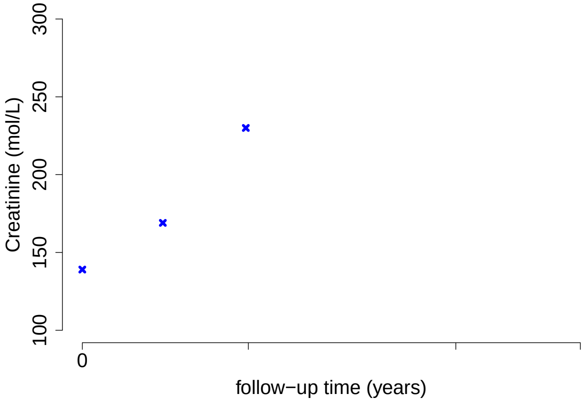
Dynamic prediction



Dynamic prediction

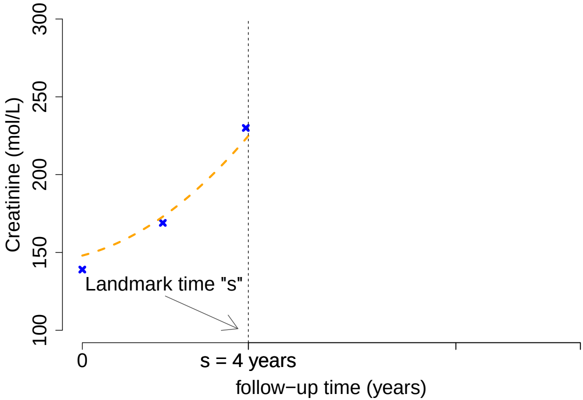


Dynamic prediction



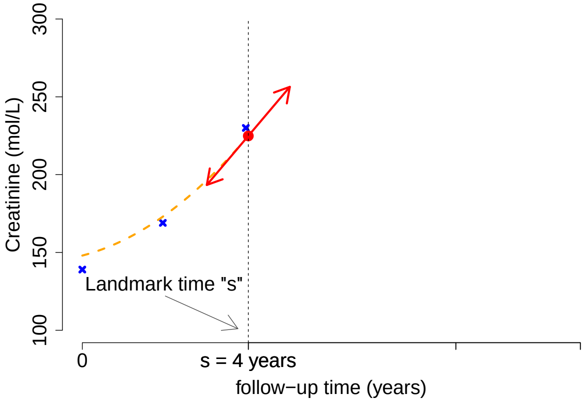
Dynamic prediction

- s: Landmark time at which predictions are made (varies)



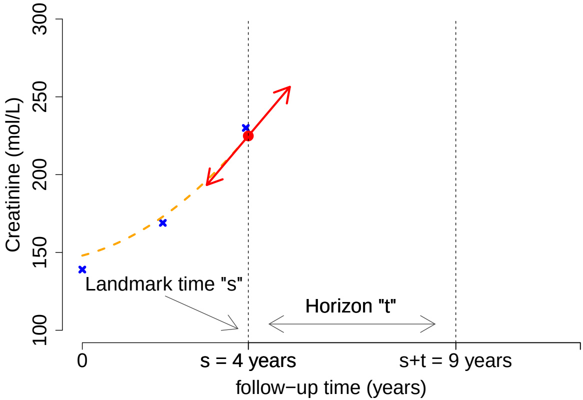
Dynamic prediction

- *s*: Landmark time at which predictions are made (varies)



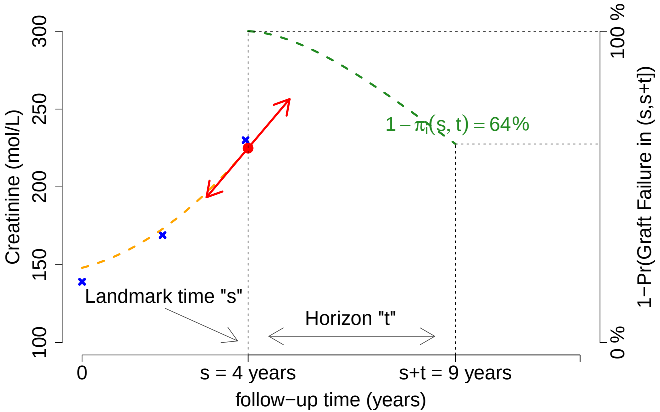
Dynamic prediction

- s: Landmark time at which predictions are made (varies)
- t: prediction horizon (fixed)

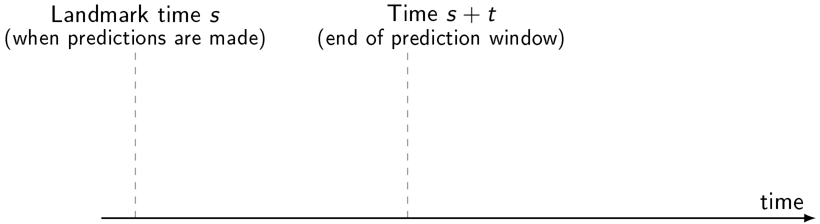


Dynamic prediction

- s: Landmark time at which predictions are made (varies)
- t: prediction horizon (fixed)

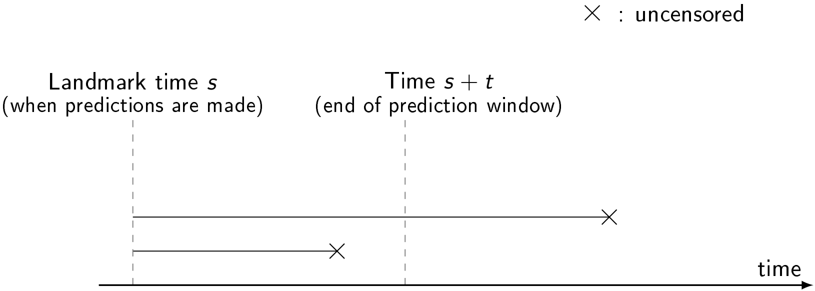


Right censoring issue



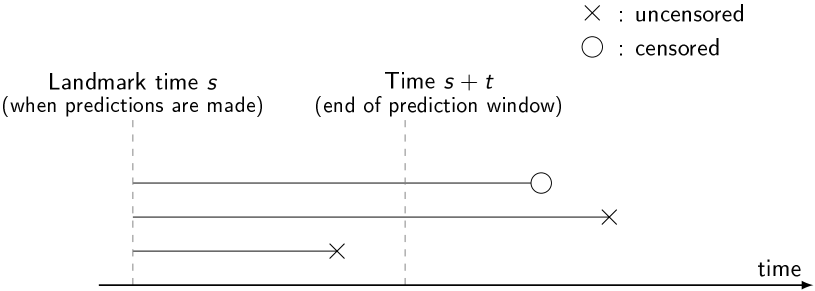
$$D_i(s,t) = \mathbb{1}\{\text{event occurs in } (s, s+t]\}$$

Right censoring issue



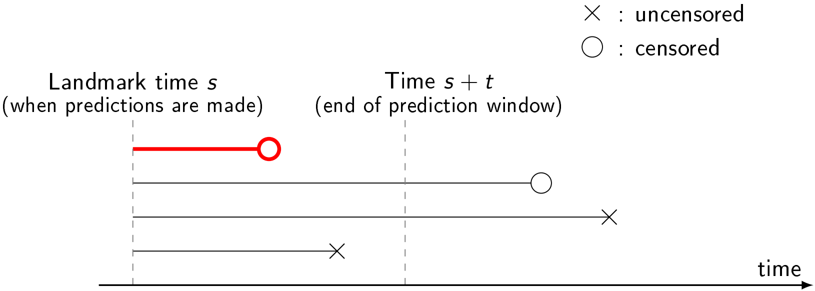
$$D_i(s,t) = \mathbb{1}\{\text{event occurs in } (s, s+t]\}$$

Right censoring issue



$$D_i(s,t) = \mathbb{1}\{\text{event occurs in } (s, s+t]\}$$

Right censoring issue



For subject i censored within $(s, s+t]$ the status

$$D_i(s,t) = \mathbb{1}\{\text{event occurs in } (s, s+t]\}$$

is unknown.

Notations for population parameters

- **Indicator of event** in $(s, s+t]$:

$$D_i(s,t) = \mathbb{1}\{s < T_i \leq s+t\}$$

where T_i is the time-to-event.

Notations for population parameters

- **Indicator of event** in $(s, s+t]$:

$$D_i(s,t) = \mathbb{1}\{s < T_i \leq s+t\}$$

where T_i is the time-to-event.

- **Dynamic predictions:**

$$\pi_i(s,t)$$

Notations for population parameters

- **Indicator of event** in $(s, s+t]$:

$$D_i(s,t) = \mathbb{1}\{s < T_i \le s+t\}$$

where T_i is the time-to-event.

- **Dynamic predictions:**

$$\pi_i(s,t) = \widehat{\mathbb{P}}_{\widehat{\boldsymbol{\xi}}}\Big(\qquad\qquad\qquad\qquad \Big)$$

  - $\widehat{\boldsymbol{\xi}}$: previously estimated parameters (from independent training data)

Notations for population parameters

- **Indicator of event** in $(s, s+t]$:

$$D_i(s,t) = \mathbb{1}\{s < T_i \leq s+t\}$$

where T_i is the time-to-event.

- **Dynamic predictions:**

$$\pi_i(s,t) = \widehat{\mathbb{P}}_{\widehat{\boldsymbol{\xi}}}\Big(D_i(s,t) = 1 \Big| \qquad\qquad\qquad \Big)$$

  - $\widehat{\boldsymbol{\xi}}$: previously estimated parameters (from independent training data)

Notations for population parameters

- **Indicator of event** in $(s, s+t]$:

$$D_i(s,t) = \mathbb{1}\{s < T_i \leq s+t\}$$

where T_i is the time-to-event.

- **Dynamic predictions:**

$$\pi_i(s,t) = \widehat{\mathbb{P}}_{\widehat{\boldsymbol{\xi}}}\Big(D_i(s,t) = 1 \Big| T_i > s, \mathcal{Y}_i(s), \quad \Big)$$

  - $\widehat{\boldsymbol{\xi}}$: previously estimated parameters (from independent training data)
  - $\mathcal{Y}_i(s)$: marker measurements observed before time s

Notations for population parameters

- **Indicator of event** in $(s, s+t]$:

$$D_i(s,t) = \mathbb{1}\{s < T_i \leq s+t\}$$

where T_i is the time-to-event.

- **Dynamic predictions:**

$$\pi_i(s,t) = \widehat{\mathbb{P}}_{\widehat{\boldsymbol{\xi}}}\Big(D_i(s,t) = 1 \Big| T_i > s, \mathcal{Y}_i(s), \mathbf{X}_i\Big)$$

  - $\widehat{\boldsymbol{\xi}}$: previously estimated parameters (from independent training data)
  - $\mathcal{Y}_i(s)$: marker measurements observed before time s
  - $\mathbf{X}_i$: baseline covariates

Predictive accuracy

How close are the predicted risks $\pi_i(s,t)$ to the "true underlying" risk $\mathbb{P}\Big($event occurs in (s,s+t]$\Big|$information at s$\Big)$?

Predictive accuracy

How close are the predicted risks $\pi_i(s,t)$ to the "true underlying" risk $\mathbb{P}\big(\text{event occurs in (s,s+t]}\big|\text{information at s}\big)$?

- **Prediction Error**:

$$\boxed{\mathrm{PE}_\pi(s,t) = \mathbb{E}\left[\Big\{D(s,t) - \pi(s,t)\Big\}^2 \Big| T > s\right]}$$

Predictive accuracy

How close are the predicted risks $\pi_i(s,t)$ to the "true underlying" risk $\mathbb{P}\Big(\text{event occurs in (s,s+t]}\Big|\text{information at s}\Big)$?

▶ **Prediction Error**:

$$\boxed{\text{PE}_\pi(s,t) = \mathbb{E}\left[\Big\{D(s,t) - \pi(s,t)\Big\}^2 \Big| T > s\right]}$$

- the lower the better
- PE $\approx$ Bias^2 + Variance
- evaluates both **Calibration** and **Discrimination**
- depends on $\mathbb{P}\Big(\text{event occurs in (s,s+t]}\Big|\text{at risk at s}\Big)$
- often called "Expected **Brier Score**"

How does the PE relate to calibration and discrimination?

$$\mathrm{PE}_\pi(s,t) = \mathbb{E}\left[\left\{D(s,t) - \pi(s,t)\right\}^2 \middle| T > s\right]$$

How does the PE relate to calibration and discrimination?

$$
\begin{aligned}
\mathrm{PE}_\pi(s,t) =& \mathbb{E}\Big[\Big\{D(s,t) - \mathbb{E}\big[D(s,t)\big|\mathcal{H}^\pi(s)\big] \\
&\qquad + \underbrace{\mathbb{E}\big[D(s,t)\big|\mathcal{H}^\pi(s)\big]}_{\text{“true underlying” risk}} - \pi(s,t)\Big\}^2 \Big| T > s\Big]
\end{aligned}
$$

How does the PE relate to calibration and discrimination?

$$
\begin{aligned}
\mathrm{PE}_\pi(s,t) =& \mathbb{E}\Big[\Big\{D(s,t) - \mathbb{E}\big[D(s,t)\big|\mathcal{H}^\pi(s)\big]\Big\}^2\Big|T>s\Big] \\
&+ \mathbb{E}\Big[\Big\{\underbrace{\mathbb{E}\big[D(s,t)\big|\mathcal{H}^\pi(s)\big]}_{\text{“true underlying” risk}} - \pi(s,t)\Big\}^2\Big|T>s\Big]
\end{aligned}
$$

$\mathcal{H}^\pi(s) = \{\mathcal{X}^\pi(s), T > s\}$ denotes the subject-specific history at time s.

How does the PE relate to calibration and discrimination?

$$\mathrm{PE}_\pi(s,t) = \underbrace{\mathbb{E}\left[\left\{D(s,t) - \mathbb{E}\left[D(s,t)\middle|\mathcal{H}^\pi(s)\right]\right\}^2\middle|T>s\right]}_{\text{Inseparability}}$$
$$+\underbrace{\mathbb{E}\left[\left\{\mathbb{E}\left[D(s,t)\middle|\mathcal{H}^\pi(s)\right] - \pi(s,t)\right\}^2\middle|T>s\right]}_{\text{Bias/Calibration}}$$

$\mathcal{H}^\pi(s) = \{\mathcal{X}^\pi(s), T > s\}$ denotes the subject-specific history at time s.

How does the PE relate to calibration and discrimination?

$$\mathrm{PE}_\pi(s,t) = \underbrace{\mathbb{E}\left[\mathbf{Var}\left\{D(s,t)\middle|\mathcal{H}^\pi(s)\right\}\middle| T>s\right]}_{\textbf{Discrimination}} + \underbrace{\mathbb{E}\left[\left\{\mathbb{E}\left[D(s,t)\middle|\mathcal{H}^\pi(s)\right] - \pi(s,t)\right\}^2\middle| T>s\right]}_{\textbf{Calibration}}$$

$\mathcal{H}^\pi(s) = \{\mathcal{X}^\pi(s), T > s\}$ denotes the subject-specific history at time s.

How does the PE relate to calibration and discrimination?

$$\text{PE}_\pi(s,t) = \underbrace{\mathbb{E}\Big[\textbf{Var}\big\{D(s,t)\big|\mathcal{H}^\pi(s)\big\}\Big| T > s\Big]}_{\textbf{Discrimination}}$$

$$+\underbrace{\mathbb{E}\Big[\Big\{\mathbb{E}\big[D(s,t)\big|\mathcal{H}^\pi(s)\big] - \pi(s,t)\Big\}^2\Big| T > s\Big]}_{\textbf{Calibration}}$$

$\mathcal{H}^\pi(s) = \{\mathcal{X}^\pi(s), T > s\}$ denotes the subject-specific history at time s.

- the more discriminating $\mathcal{H}^\pi(s)$ the smaller $\textbf{Var}\big\{D(s,t)\big|\mathcal{H}^\pi(s)\big\}$

How does the PE relate to calibration and discrimination?

$$\mathrm{PE}_\pi(s,t) = \underbrace{\mathbb{E}\Big[\mathbf{Var}\big\{D(s,t)\big|\mathcal{H}^\pi(s)\big\}\Big|T>s\Big]}_{\textbf{Discrimination}}$$

$$+\underbrace{\mathbb{E}\Big[\Big\{\mathbb{E}\big[D(s,t)\big|\mathcal{H}^\pi(s)\big]-\pi(s,t)\Big\}^2\Big|T>s\Big]}_{\textbf{Calibration}}$$

$\mathcal{H}^\pi(s)=\{\mathcal{X}^\pi(s), T>s\}$ denotes the subject-specific history at time s.

- the more discriminating $\mathcal{H}^\pi(s)$ the smaller $\mathbf{Var}\big\{D(s,t)\big|\mathcal{H}^\pi(s)\big\}$
- $\mathbb{E}\big[D(s,t)\big|\mathcal{H}^\pi(s)\big]-\pi(s,t)\equiv 0$ defines **"strong" calibration**.

How does the PE relate to calibration and discrimination?

$$\mathrm{PE}_\pi(s,t) = \underbrace{\mathbb{E}\Big[\mathbf{Var}\big\{D(s,t)\big|\mathcal{H}^\pi(s)\big\}\Big|T>s\Big]}_{\textbf{Does NOT depend on } \pi(s,t)} + \underbrace{\mathbb{E}\Big[\Big\{\mathbb{E}\big[D(s,t)\big|\mathcal{H}^\pi(s)\big] - \pi(s,t)\Big\}^2\Big|T>s\Big]}_{\textbf{Depends on } \pi(s,t)}$$

$\mathcal{H}^\pi(s) = \{\mathcal{X}^\pi(s), T > s\}$ denotes the subject-specific history at time s.

- the more discriminating $\mathcal{H}^\pi(s)$ the smaller $\mathbf{Var}\big\{D(s,t)\big|\mathcal{H}^\pi(s)\big\}$
- $\mathbb{E}\big[D(s,t)\big|\mathcal{H}^\pi(s)\big] - \pi(s,t) \equiv 0$ defines **"strong" calibration**.

R^2_π-type criterion

- **Benchmark** PE_0

The **best "null" prediction tool**, which gives the same (marginal) predicted risk

$$S(s+t|s) = \mathbb{E}\big[D(s,t)\big|\mathcal{H}^0(s)\big], \qquad \mathcal{H}^0(s) = \{T > s\}$$

to all subjects leads to

$$PE_0(s,t) = \mathrm{Var}\{D(s,t)|\, T > s\} = S(s+t|s)\Big\{1 - S(s+t|s)\Big\}.$$

R^2_π-type criterion

- **Benchmark** PE_0

The **best "null" prediction tool**, which gives the same (marginal) predicted risk

$$S(s+t|s) = \mathbb{E}\big[D(s,t)\big|\mathcal{H}^0(s)\big], \qquad \mathcal{H}^0(s) = \{T > s\}$$

to all subjects leads to

$$PE_0(s,t) = \text{Var}\{D(s,t)|\,T > s\} = S(s+t|s)\Big\{1 - S(s+t|s)\Big\}.$$

- **Simple idea**

$$\boxed{R^2_\pi(s,t) = 1 - \frac{\text{PE}_\pi(s,t)}{\text{PE}_0(s,t)}}$$

Why bother?

- $R^2_\pi(s,t)$ aims to circumvent the difficult interpretation of:
  - the scale on which $PE(s,t)$ is measured
  - interpretation for trend of $PE(s,t)$ vs s

Why bother?

- $R^2_\pi(s,t)$ aims to circumvent the difficult interpretation of:
  - the scale on which $PE(s,t)$ is measured
  - interpretation for trend of $PE(s,t)$ vs s

- Because the meaning of **the scale on which PE(s,t) is measured changes with s**, an increasing/decreasing trend can be due to changes in:
  - the quality of the predictions

and/or
  - the at risk population

Changes in the quality of the predictions

"Essentially, all models are wrong, but some are useful.", G. Box

Changes in the quality of the predictions

"Essentially, all models are wrong, but some are useful.", G. Box

▶ The prediction model from which we have obtained the predictions can be "more wrong" for some s than for some others.

- Calibration term of $PE(s,t)$ changes with s
- We can work on it!

Changes in the at risk population

An example:

- *Patients with cardiovascular history (CV) all die early.*
- *Only those without CV remain at risk for late s.*
- *Then:*
  - *the earlier s the more homogeneous the at risk population*
  - *CV is useful for prediction for early s but useless for late s.*

Changes in the at risk population

An example:

- *Patients with cardiovascular history (CV) all die early.*
- *Only those without CV remain at risk for late s.*
- *Then:*
  - *the earlier s the more homogeneous the at risk population*
  - *CV is useful for prediction for early s but useless for late s.*

▶ The available information can be more informative for some s than for some others.

- Discriminating term of $PE(s,t)$ changes with s
- This is just how it is, there is nothing we can do!

(we can only work with the data we have)

$R^2_\pi(s,t)$ interpretation

- **Always true:**

Measure of how the prediction tool $\pi(s,t)$ performs compared to the benchmark null prediction tool, which gives the same predicted risk to all subjects (marginal risk).

$R^2_\pi(s,t)$ interpretation

- **Always true:**

Measure of how the prediction tool $\pi(s,t)$ performs compared to the benchmark null prediction tool, which gives the same predicted risk to all subjects (marginal risk).

- **When** predictions are **calibrated** (strongly):

$$R^2_\pi(s,t) = \frac{\mathrm{Var}\{\pi(s,t)|T>s\}}{\mathrm{Var}\{D(s,t)|T>s\}}$$

explained variation

(after little algebra)

$R^2_\pi(s,t)$ interpretation

- **Always true:**
Measure of how the prediction tool $\pi(s,t)$ performs compared to the benchmark null prediction tool, which gives the same predicted risk to all subjects (marginal risk).

- **When** predictions are **calibrated** (strongly):

$$R^2_\pi(s,t) = \frac{\mathrm{Var}\{\pi(s,t)|T > s\}}{\mathrm{Var}\{D(s,t)|T > s\}}$$ **explained variation**

$$= \mathrm{Corr}^2\big\{D(s,t), \pi(s,t)\big| T > s\big\}$$ **correlation**

(after little algebra)

$R^2_\pi(s,t)$ interpretation

- **Always true:**
Measure of how the prediction tool $\pi(s,t)$ performs compared to the benchmark null prediction tool, which gives the same predicted risk to all subjects (marginal risk).

- **When** predictions are **calibrated** (strongly):

$$R^2_\pi(s,t) = \frac{\mathrm{Var}\{\pi(s,t)|T>s\}}{\mathrm{Var}\{D(s,t)|T>s\}}$$ **explained variation**

$$= \mathrm{Corr}^2\Big\{D(s,t),\pi(s,t)\Big|T>s\Big\}$$ **correlation**

$$= \mathbb{E}\Big\{\pi(s,t)\Big|D(s,t)=1, T>s\Big\} - \mathbb{E}\Big\{\pi(s,t)\Big|D(s,t)=0, T>s\Big\}.$$ **mean risk difference**

(after little algebra)

Observations & IPCW PE estimator

- **Observations** (i.i.d.)

$$\Big\{\big(\widetilde{T}_i, \Delta_i, \pi_i(\cdot,\cdot)\big), i = 1, \ldots, n\Big\} \quad \text{where } \widetilde{T}_i = T_i \wedge C_i,\ \Delta_i = \mathbb{1}\{T_i \leq C_i\}$$

Observations & IPCW PE estimator

- **Observations** (i.i.d.)

$$\Big\{\big(\widetilde{T}_i, \Delta_i, \pi_i(\cdot,\cdot)\big), i = 1, \dots, n\Big\} \quad \text{where } \widetilde{T}_i = T_i \wedge C_i,\ \Delta_i = \mathbb{1}\{T_i \le C_i\}$$

- Indicator of **"observed event occurrence"** in $(s, s+t]$:

$$\widetilde{D}_i(s,t) = \mathbb{1}\{s < \widetilde{T}_i \le s+t, \Delta_i = 1\} = \begin{cases} 1 & : \text{ event occurred} \\ 0 & : \text{ event did not occur} \\ & \text{or } \textbf{censored} \text{ obs.} \end{cases}$$

Observations & IPCW PE estimator

- **Observations** (i.i.d.)

$$\Big\{\big(\widetilde{T}_i, \Delta_i, \pi_i(\cdot,\cdot)\big), i = 1, \dots, n\Big\} \quad \text{where } \widetilde{T}_i = T_i \wedge C_i,\ \Delta_i = \mathbb{1}\{T_i \leq C_i\}$$

- Indicator of **"observed event occurrence"** in $(s, s+t]$:

$$\widetilde{D}_i(s,t) = \mathbb{1}\{s < \widetilde{T}_i \leq s+t, \Delta_i = 1\} = \begin{cases} 1 & : \text{event occurred} \\ 0 & : \text{event did not occur} \\ & \text{or } \textbf{censored} \text{ obs.} \end{cases}$$

- Inverse Probability of Censoring Weighting (IPCW) estimator:

$$\widehat{PE}_\pi(s,t) = \frac{1}{n}\sum_{i=1}^{n} \qquad \Big\{\widetilde{D}_i(s,t) - \pi_i(s,t)\Big\}^2$$

Observations & IPCW PE estimator

- **Observations** (i.i.d.)

$$\left\{\left(\widetilde{T}_i, \Delta_i, \pi_i(\cdot,\cdot)\right), i = 1, \ldots, n\right\} \quad \text{where } \widetilde{T}_i = T_i \wedge C_i,\ \Delta_i = \mathbb{1}\{T_i \le C_i\}$$

- Indicator of **"observed event occurrence"** in $(s, s+t]$:

$$\widetilde{D}_i(s,t) = \mathbb{1}\{s < \widetilde{T}_i \le s+t, \Delta_i = 1\} = \begin{cases} 1 & : \text{ event occurred} \\ 0 & : \text{ event did not occur} \\ & \text{ or } \textbf{censored} \text{ obs.} \end{cases}$$

- Inverse Probability of Censoring Weighting (IPCW) estimator:

$$\widehat{PE}_\pi(s,t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{W}_i(s,t) \left\{\widetilde{D}_i(s,t) - \pi_i(s,t)\right\}^2$$

Observations & IPCW PE estimator

- **Observations** (i.i.d.)

$$\left\{\left(\widetilde{T}_i, \Delta_i, \pi_i(\cdot,\cdot)\right), i = 1, \dots, n\right\} \quad \text{where } \widetilde{T}_i = T_i \wedge C_i,\ \Delta_i = \mathbb{1}\{T_i \leq C_i\}$$

- Indicator of **"observed event occurrence"** in $(s, s+t]$:

$$\widetilde{D}_i(s,t) = \mathbb{1}\{s < \widetilde{T}_i \leq s+t, \Delta_i = 1\} = \begin{cases} 1 & : \text{ event occurred} \\ 0 & : \text{ event did not occur} \\ & \text{or } \textbf{censored} \text{ obs.} \end{cases}$$

- Inverse Probability of Censoring Weighting (IPCW) estimator:

$$\widehat{PE}_\pi(s,t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{W}_i(s,t) \left\{\widetilde{D}_i(s,t) - \pi_i(s,t)\right\}^2$$

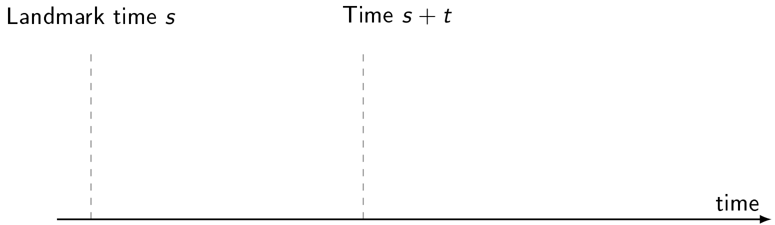
and

$$\widehat{R}^2_\pi(s,t) = 1 - \frac{\widehat{PE}_\pi(s,t)}{\widehat{PE}_0(s,t)}$$

Inverse Probability of Censoring Weights

$$\boxed{\widehat{W}_i(s,t) = \qquad\qquad + \qquad\qquad +}$$

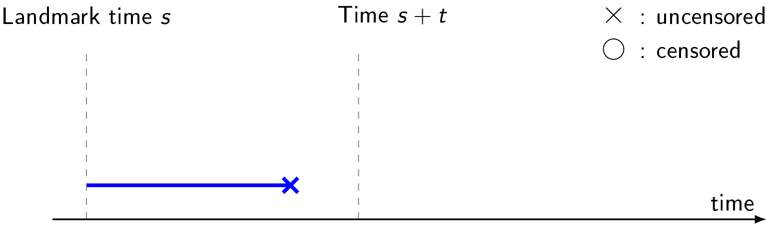
with $\widehat{G}(u|s)$ the Kaplan-Meier estimator of $\mathbb{P}(C > u | C > s)$.



Inverse Probability of Censoring Weights

$$\boxed{\widehat{W}_i(s,t) = \frac{\mathbb{1}\{s < \widetilde{T}_i \le s+t\}\Delta_i}{\widehat{G}(\widetilde{T}_i|s)} + \qquad\qquad +}$$

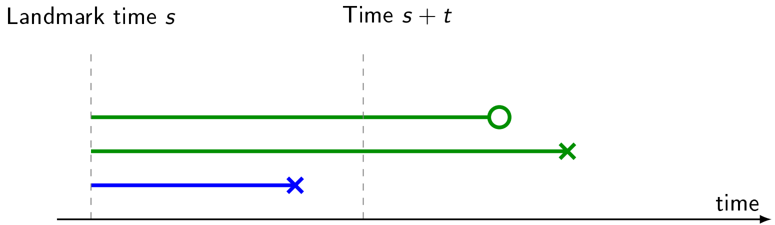
with $\widehat{G}(u|s)$ the Kaplan-Meier estimator of $\mathbb{P}(C > u | C > s)$.



Inverse Probability of Censoring Weights

$$\widehat{W}_i(s,t) = \frac{\mathbb{1}\{s < \widetilde{T}_i \leq s+t\}\Delta_i}{\widehat{G}(\widetilde{T}_i|s)} + \frac{\mathbb{1}\{\widetilde{T}_i > s+t\}}{\widehat{G}(s+t|s)} +$$

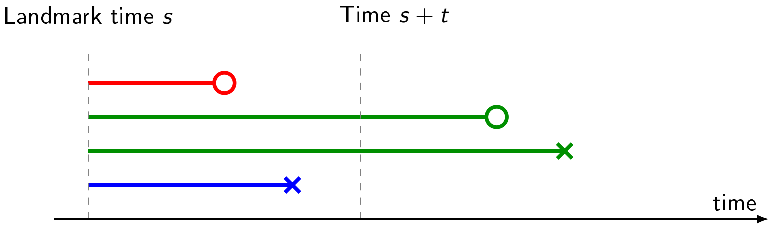
with $\widehat{G}(u|s)$ the Kaplan-Meier estimator of $\mathbb{P}(C > u | C > s)$.



Inverse Probability of Censoring Weights

$$\boxed{\widehat{W}_i(s,t) = \frac{\mathbb{1}\{s < \widetilde{T}_i \leq s+t\}\Delta_i}{\widehat{G}(\widetilde{T}_i|s)} + \frac{\mathbb{1}\{\widetilde{T}_i > s+t\}}{\widehat{G}(s+t|s)} + 0}$$

with $\widehat{G}(u|s)$ the Kaplan-Meier estimator of $\mathbb{P}(C > u | C > s)$.



Asymptotic i.i.d. representation

Lemma: Assume that the censoring time C is independent of $(T, \eta, \pi(\cdot,\cdot))$ and let θ denote either PE_π, R^2_π or a difference in PE or R^2_π, then

$$\boxed{\sqrt{n}\left(\widehat{\theta}(s,t) - \theta(s,t)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IF}_\theta(\widetilde{T}_i, \Delta_i, \pi_i(s,t), s, t) + o_p(1)}$$

where $\mathrm{IF}_\theta(\widetilde{T}_i, \Delta_i, \pi_i(s,t), s, t)$ being :

- zero-mean i.i.d. terms
- easy to estimate (using Nelson-Aalen & Kaplan-Meier)

Pointwise confidence interval (fixed s)

- Asymptotic normality:

$$\sqrt{n}\Big(\widehat{\theta}(s,t)-\theta(s,t)\Big) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0,\sigma^2_{s,t}\right)$$

- 95% confidence interval:

$$\left\{\widehat{\theta}(s,t) \pm z_{1-\alpha/2}\frac{\widehat{\sigma}_{s,t}}{\sqrt{n}}\right\}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of $\mathcal{N}(0,1)$.

- Variance estimator:

$$\widehat{\sigma}^2_{s,t} = \frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{\mathrm{IF}}_{\theta}(\widetilde{T}_i,\Delta_i,\pi_i(s,t),s,t)\right\}^2$$

Simultaneous confidence band over $s \in \mathcal{S}$

$$\left\{ \widehat{\theta}(s,t) \pm \widehat{q}_{1-\alpha}^{(\mathcal{S},t)} \frac{\widehat{\sigma}_{s,t}}{\sqrt{n}} \right\}, \quad s \in \mathcal{S}$$

Simultaneous confidence band over $s \in \mathcal{S}$

$$\left\{ \widehat{\theta}(s,t) \pm \widehat{q}_{1-\alpha}^{(\mathcal{S},t)} \frac{\widehat{\sigma}_{s,t}}{\sqrt{n}} \right\}, \quad s \in \mathcal{S}$$

Computation of $\widehat{q}_{1-\alpha}^{(\mathcal{S},t)}$ by the simulation algorithm

1. For $b = 1, \ldots, B$, say $B = 4000$, do:

Simultaneous confidence band over $s \in \mathcal{S}$

$$\left\{ \widehat{\theta}(s,t) \pm \widehat{q}_{1-\alpha}^{(\mathcal{S},t)} \frac{\widehat{\sigma}_{s,t}}{\sqrt{n}} \right\}, \quad s \in \mathcal{S}$$

Computation of $\widehat{q}_{1-\alpha}^{(\mathcal{S},t)}$ by the simulation algorithm ($\approx$ Wild Bootstrap):

1. For $b = 1, \ldots, B$, say $B = 4000$, do:

(*Conditional multiplier central limit theorem*)

Simultaneous confidence band over $s \in \mathcal{S}$

$$\left\{ \widehat{\theta}(s,t) \pm \widehat{q}_{1-\alpha}^{(\mathcal{S},t)} \frac{\widehat{\sigma}_{s,t}}{\sqrt{n}} \right\}, \quad s \in \mathcal{S}$$

Computation of $\widehat{q}_{1-\alpha}^{(\mathcal{S},t)}$ by the simulation algorithm ($\approx$ Wild Bootstrap):

1. For $b = 1, \ldots, B$, say $B = 4000$, do:
   1. Generate $\{\omega_1^b, \ldots, \omega_n^b\}$ from n i.i.d. $\mathcal{N}(0,1)$.

(*Conditional multiplier central limit theorem*)

Simultaneous confidence band over $s \in \mathcal{S}$

$$\left\{ \widehat{\theta}(s,t) \pm \widehat{q}_{1-\alpha}^{(\mathcal{S},t)} \frac{\widehat{\sigma}_{s,t}}{\sqrt{n}} \right\}, \quad s \in \mathcal{S}$$

Computation of $\widehat{q}_{1-\alpha}^{(\mathcal{S},t)}$ by the simulation algorithm ($\approx$ Wild Bootstrap):

1. For $b = 1, \dots, B$, say $B = 4000$, do:
   1. Generate $\{\omega_1^b, \dots, \omega_n^b\}$ from n i.i.d. $\mathcal{N}(0,1)$.
   2. Using the plug-in estimator $\widehat{\mathrm{IF}}_\theta(\cdot)$, compute:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \quad \frac{\widehat{\mathrm{IF}}_\theta(\widetilde{T}_i, \Delta_i, \pi_i(s,t), s, t)}{\widehat{\sigma}_{s,t}}$$

(*Conditional multiplier central limit theorem*)

Simultaneous confidence band over $s \in \mathcal{S}$

$$\left\{ \widehat{\theta}(s,t) \pm \widehat{q}_{1-\alpha}^{(\mathcal{S},t)} \frac{\widehat{\sigma}_{s,t}}{\sqrt{n}} \right\}, \quad s \in \mathcal{S}$$

Computation of $\widehat{q}_{1-\alpha}^{(\mathcal{S},t)}$ by the simulation algorithm ($\approx$ Wild Bootstrap):

1. For $b = 1, \ldots, B$, say $B = 4000$, do:
   1. Generate $\{\omega_1^b, \ldots, \omega_n^b\}$ from n i.i.d. $\mathcal{N}(0,1)$.
   2. Using the plug-in estimator $\widehat{\mathrm{IF}}_\theta(\cdot)$, compute:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega_i^b \frac{\widehat{\mathrm{IF}}_\theta(\widetilde{T}_i, \Delta_i, \pi_i(s,t), s, t)}{\widehat{\sigma}_{s,t}}$$

(*Conditional multiplier central limit theorem*)

Simultaneous confidence band over $s \in \mathcal{S}$

$$\left\{ \widehat{\theta}(s,t) \pm \widehat{q}_{1-\alpha}^{(\mathcal{S},t)} \frac{\widehat{\sigma}_{s,t}}{\sqrt{n}} \right\}, \quad s \in \mathcal{S}$$

Computation of $\widehat{q}_{1-\alpha}^{(\mathcal{S},t)}$ by the simulation algorithm ($\approx$ Wild Bootstrap):

1. For $b = 1, \ldots, B$, say $B = 4000$, do:
   1. Generate $\{\omega_1^b, \ldots, \omega_n^b\}$ from n i.i.d. $\mathcal{N}(0,1)$.
   2. Using the plug-in estimator $\widehat{\text{IF}}_\theta(\cdot)$, compute:

$$\Upsilon^b = \sup_{s \in \mathcal{S}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega_i^b \frac{\widehat{\text{IF}}_\theta(\widetilde{T}_i, \Delta_i, \pi_i(s,t), s, t)}{\widehat{\sigma}_{s,t}} \right|$$

(*Conditional multiplier central limit theorem*)

Simultaneous confidence band over $s \in \mathcal{S}$

$$\left\{ \widehat{\theta}(s,t) \pm \widehat{q}_{1-\alpha}^{(\mathcal{S},t)} \frac{\widehat{\sigma}_{s,t}}{\sqrt{n}} \right\}, \quad s \in \mathcal{S}$$

Computation of $\widehat{q}_{1-\alpha}^{(\mathcal{S},t)}$ by the simulation algorithm ($\approx$ Wild Bootstrap):

1. For $b = 1, \dots, B$, say $B = 4000$, do:
   1. Generate $\{\omega_1^b, \dots, \omega_n^b\}$ from n i.i.d. $\mathcal{N}(0,1)$.
   2. Using the plug-in estimator $\widehat{\mathrm{IF}}_\theta(\cdot)$, compute:

$$\Upsilon^b = \sup_{s \in \mathcal{S}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega_i^b \frac{\widehat{\mathrm{IF}}_\theta(\widetilde{T}_i, \Delta_i, \pi_i(s,t), s, t)}{\widehat{\sigma}_{s,t}} \right|$$

2. Compute $\widehat{q}_{1-\alpha}^{(\mathcal{S},t)}$ as the $100(1-\alpha)$th percentile of $\left\{\Upsilon^1, \dots, \Upsilon^B\right\}$

(*Conditional multiplier central limit theorem*)

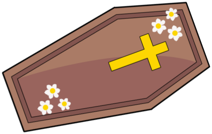
DIVAT sample

- Population based study of kidney recipients (n=4,119)
- Split the data into training (2/3) and validation (1/3) samples

DIVAT sample

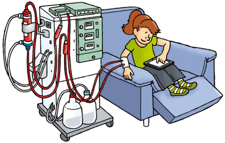
- Population based study of kidney recipients (n=4,119)
- Split the data into training (2/3) and validation (1/3) samples
- T: time from 1-year after transplantation to graft failure which is:

Death



OR

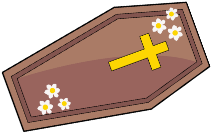
Return to dialysis



DIVAT sample

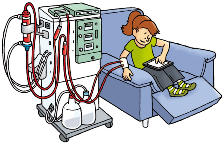
- Population based study of kidney recipients (n=4,119)
- Split the data into training (2/3) and validation (1/3) samples
- T: time from 1-year after transplantation to graft failure which is:

Death



OR

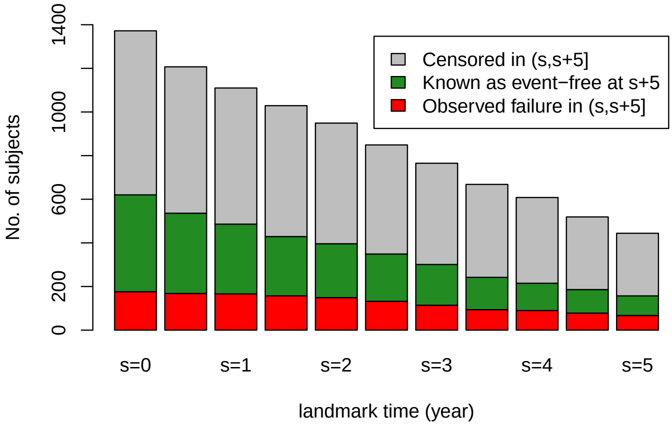
Return to dialysis



- Censoring due to:
  - delayed entries: 2000-2013
  - end of follow-up: 2014
- Baseline covariates: age, sex, cardiovascular history
- Longitudinal biomarker (yearly): serum creatinine

Descriptive statistics & censoring issue

- $s \in \mathcal{S} = \{0, 0.5, \ldots, 5\}$
- $t = 5$ years



Joint model

- **Longitudinal**

$$
\begin{aligned}
log\Big[Y_i(t_{ij})\Big] &= (\beta_0 + b_{0i}) + \beta_{0,\text{age}}\mathbf{AGE}_i + \beta_{0,\text{sex}}\mathbf{SEX}_i \\
&\quad + \Big(\beta_1 + b_{1i} + \beta_{1,\text{age}}\mathbf{AGE}_i\Big) \times t_{ij} + \epsilon_{ij} \\
&= \mathbf{m}_i(t) + \varepsilon_{ij}
\end{aligned}
$$

(fitted using R package **JM**)

Joint model

- **Longitudinal**

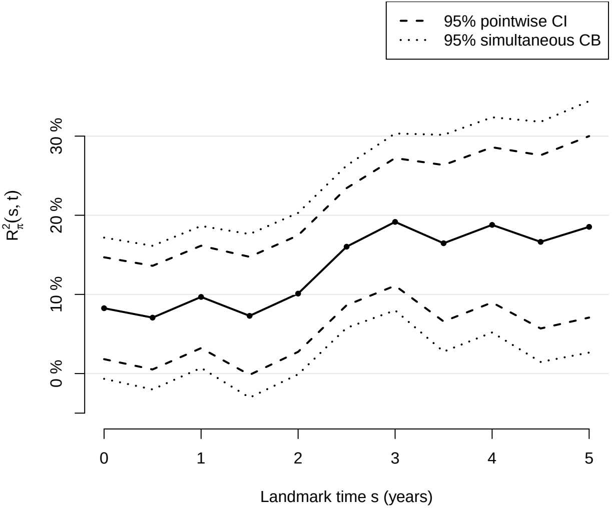
$$
\begin{aligned}
log\Big[Y_i(t_{ij})\Big] &= (\beta_0 + b_{0i}) + \beta_{0,\text{age}}\mathbf{AGE}_i + \beta_{0,\text{sex}}\mathbf{SEX}_i \\
&\quad + \big(\beta_1 + b_{1i} + \beta_{1,\text{age}}\mathbf{AGE}_i\big) \times t_{ij} + \epsilon_{ij} \\
&= \mathbf{m}_i(t) + \varepsilon_{ij}
\end{aligned}
$$

- **Survival (hazard)**

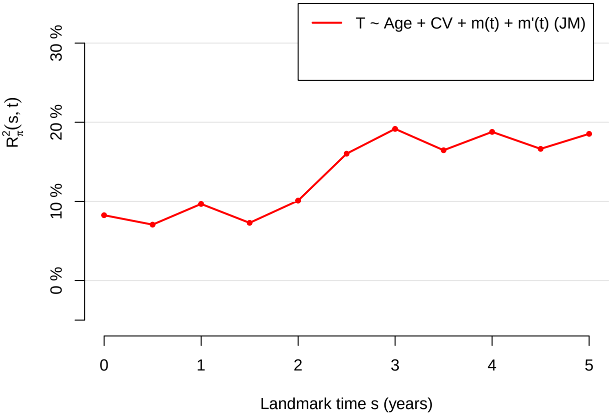
$$
\begin{aligned}
h_i(t) = h_0(t)\exp\Big\{&\gamma_{\text{age}}\mathbf{AGE}_i + \gamma_{\text{CV}}\mathbf{CV}_i \\
&+ \alpha_1\mathbf{m}_i(t) + \alpha_2\frac{d\mathbf{m}_i(t)}{dt}\Big\}
\end{aligned}
$$

(fitted using R package **JM**)

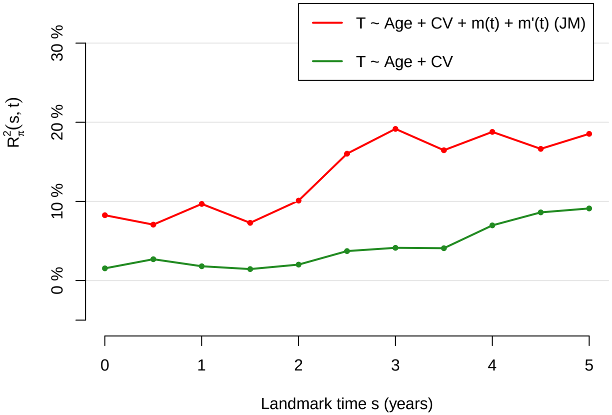
$R^2_\pi(s,t)$ vs s (t=5 years)



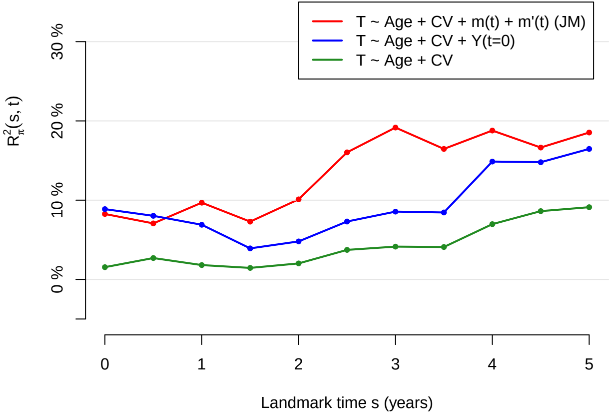
Comparing $R^2_\pi(s,t)$ vs s for different $\pi(s,t)$



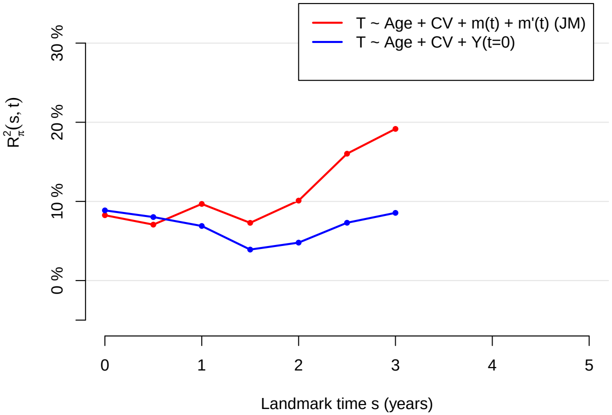
Comparing $R^2_\pi(s,t)$ vs s for different $\pi(s,t)$



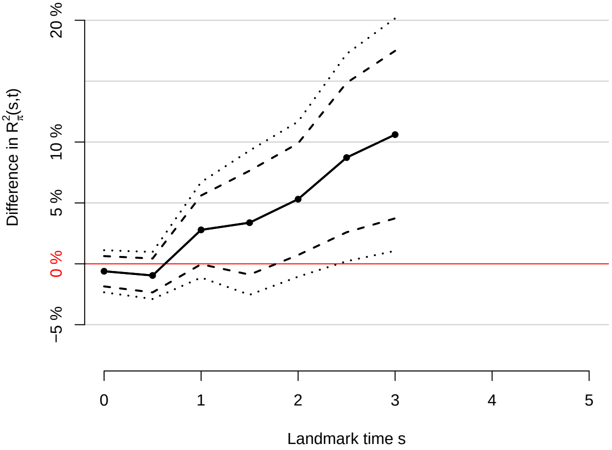
Comparing $R^2_\pi(s,t)$ vs s for different $\pi(s,t)$



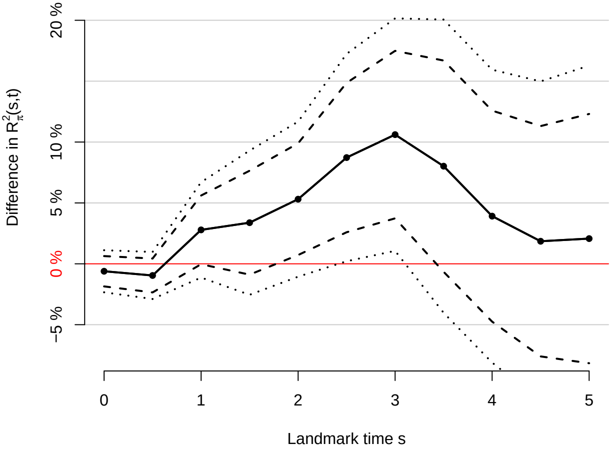
Comparing $R^2_\pi(s,t)$ vs s for different $\pi(s,t)$



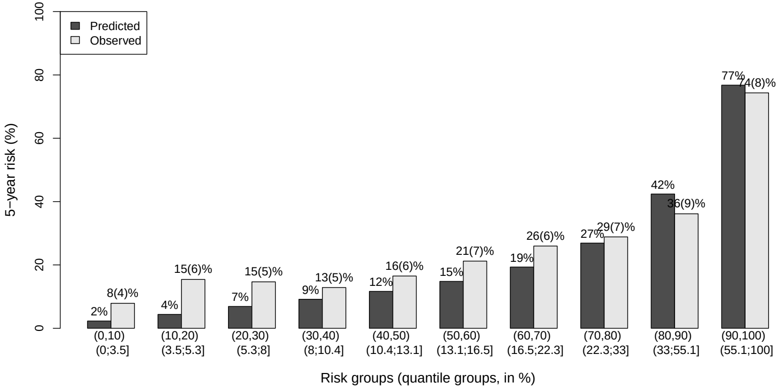
Comparing $R^2_\pi(s,t)$ vs s for different $\pi(s,t)$



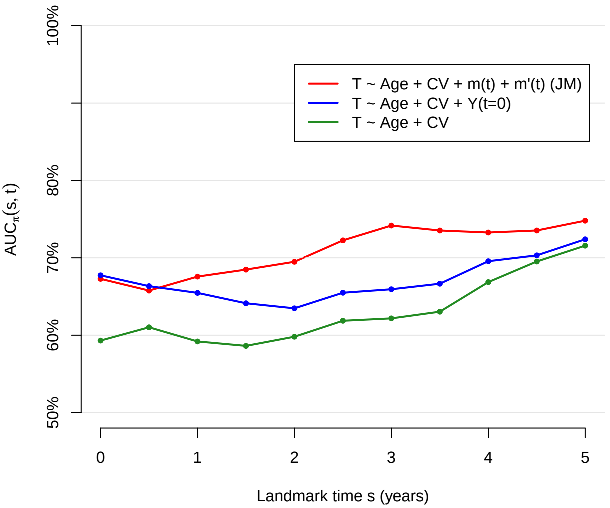
Comparing $R^2_\pi(s,t)$ vs s for different $\pi(s,t)$



Calibration plot (example for $s = 3$ years)



Area under the ROC(s, t) curve vs s



Summing up

- **R^2-type curve**
  - summarizes calibration and discriminating simultaneously
  - has an understandable trend
- **Simple model free inference**
  - predictions can be obtained from any model
  - we do not assume any model to hold
  - allows fair comparisons of different predictions
- **The method accounts** for:
  - Censoring
  - Dynamic setting (the at risk population changes)

Discussion

- The strong calibration assumption allows different interesting interpretations:
  - Explained variation
  - Correlation
  - Mean risk difference
- Unfortunately
  - the strong calibration cannot be checked (curse of dimensionality)
- However
  - weak and strong definitions are closely related:
    - strong calibration implies weak calibration
    - weak calibration can "often" be seen as a reasonable approximation of strong calibration in practice
  - weak calibration can be assessed (plots)

Discussion

- The strong calibration assumption allows different interesting interpretations:
  - Explained variation
  - Correlation
  - Mean risk difference
- Unfortunately
  - the strong calibration cannot be checked (curse of dimensionality)
- However
  - weak and strong definitions are closely related:
    - strong calibration implies weak calibration
    - weak calibration can "often" be seen as a reasonable approximation of strong calibration in practice
  - weak calibration can be assessed (plots)

Thank you for your attention!