R^2-type curve to evaluate dynamic predictions

MC. Fournier[1,2,*], E. Dantan[1], P. Blanche[3]

[1] EA 4275 - SPHERE, methodS in Patient-centered outcomes & HEalth ResEarch, Nantes University,
[2] INSERM UMR1064, Transplantation, Urology and Nephrology Institute (ITUN),
[3] LMBA Université de Bretagne Sud, Campus de Tohannic, Vannes
*marie-cecile.fournier@univ-nantes.fr

SPHERE methodS in Patient-centered outcomes & HEalth ResEarch

DIVAT www.divat.fr Données Informatiques VAlidées en Transplantation

Nantes Métropole

UNIVERSITÉ DE NANTES

INTRODUCTION

- Era of **personalized medicine**
- Prediction scores of major clinical events can help physicians in the patient taking care
- **Chronic disease context** :
  - ▷ longitudinal markers: routinely measured to assess the patient's health evolution
    $\Rightarrow$ may bring information to update predictions all along the patient follow-up

Dynamic predictions

- Computed from:
  - ▷ landmarking *(Nicolaie et al. 2013; Van Houwelingen and Putter 2012)*
  - ▷ joint modeling *(Rizopoulos 2011; Proust-Lima and Taylor 2009)*

Accuracy should be assessed accounting for dynamic setting and censoring issue

- **Discrimination**: *subjects with high/low predicted risk are more/less likely to experience the event*
- **calibration**: *if x subjects out of 100 experienced the event, we expect a mean predicted values at x for these subjects* (Steyerberg et al. 2010)
  - ▷ Dynamic ROC curve, easily interpretable, evaluates discrimination but not calibration.
  - ▷ Brier Score (a mean squared error of prediction) assesses both discrimination and calibration, but the trend according to landmark times is not straightforward
  - ⚠ **the curve of Brier Scores according to s can be misleading**

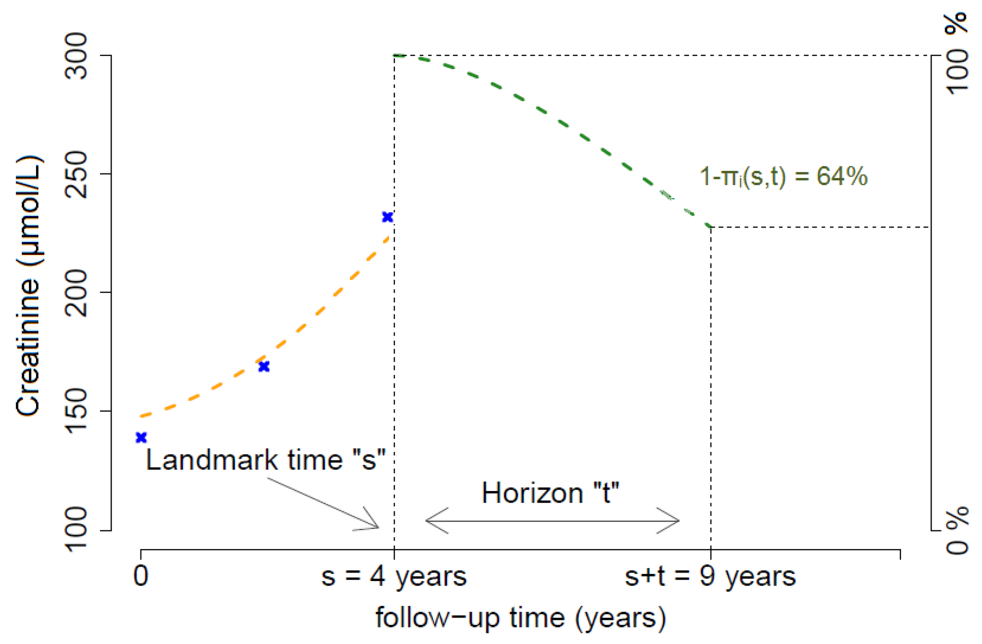
OBJECTIVE

To provide an **R^2-type criterion** to evaluate dynamic prediction

Principle: to introduce a benchmark value to standardize the Brier Score

MATERIALS AND METHODS

- Notations :
  - ▷ i: the subject ; s: the landmark time; t: the horizon window.
  - ▷ T: the time-to-event; C: the censoring time
  - ▷ $\widetilde{T} = \min(T, C)$: the observed time of follow-up and $\Delta = \mathbb{1}\{T \le C\}$, with $\mathbb{1}\{\cdot\}$ the indicator function.
  - ▷ $D(s,t) = \mathbb{1}\{s < T \le s+t\}$: the indicator of event in $(s, s+t)$
  - ▷ $\pi(s,t)\ \mathbb{P}(D(s,t) = 1 | \mathscr{H}^{\pi}(s), T > s)$: the subject-specific dynamic prediction with $\mathscr{H}^{\pi}(s)$: the observed subject-specific characteristics at landmark time s.



- **Brier Score** (the lower the better) : $\mathrm{BS}_{\pi}(s,t) = \mathbb{E}\left[\left(D(s,t) - \pi(s,t)\right)^2 \middle| T > s\right]$,
  - ▷ BS $\approx$ Bias ^{2} + Variance
  - ▷ Evaluates both discrimination and calibration:

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

  - ▷ Depends on the proportion of events in $(s, s+t)$ through the calibration term: an increasing or decreasing trend can be due to changes in:
    - the quality of the predictions
      AND/OR
    - in the at-risk population

- **R^2 criterion** (the higher the better)
  - ▷ Benchmark value : the best "null" model (or marginal) gives the same predicted risk for all subject: $\pi_0(s,t) = P(s < T < s+t | T > s) = 1 - S(s+t|s)$ with $S(\cdot)$ is the survival function. It can be estimated from the Kaplan-Meier estimator. It is free of any choice of modelisation.
    $\mathrm{BS}_0(s,t) = \mathrm{Var}\{D(s,t)|T > s\} = S(s+t|s)\{1 - S(s+t|s)\}$

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

  - ▷ the scale can be easily understanding compared to those of the Brier Score.
    INTERPRETATION:
    - $R^2(s,t) = 1 \Leftrightarrow \pi(s,t) = D(s,t)$: the prediction tool perfectly distinguish patients that will experience an event in (s, s+t] from those who will not.
    - $R^2(s,t) \approx 0 \Leftrightarrow \pi(s,t) \approx \pi_0(s,t)$
    - $R^2(s,t) < 0$ when the subject-specific information is wrongly used ($\Rightarrow$ extreme cases where the predictions performed worst than the marginal ones, with over fitted predictions for example).

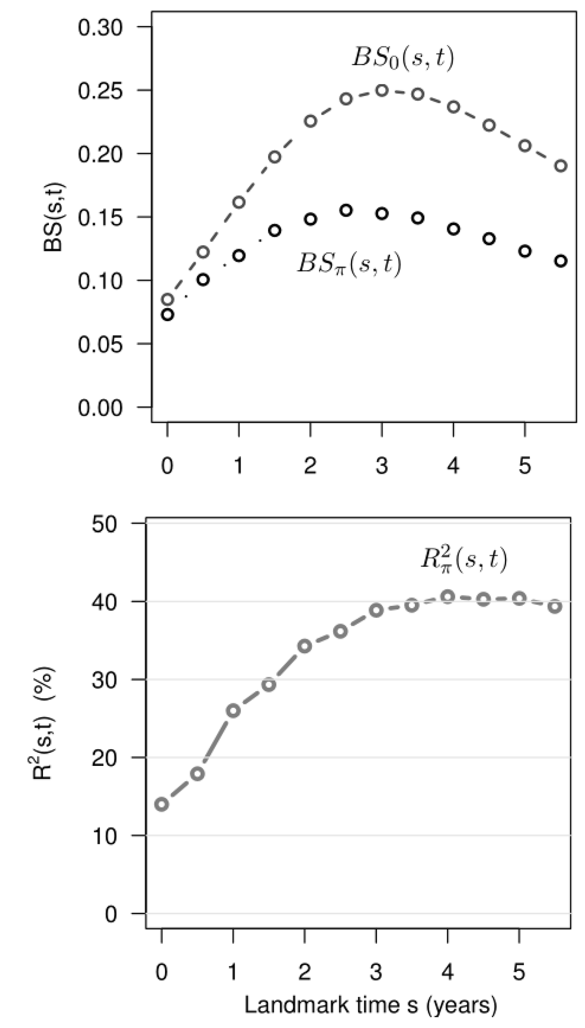
- Use of the Inverse Probability of Censoring Weighting (IPCW) to make inference (like in *Blanche et al. 2015*)
- Pointwise confidence intervals are constructed using a Wald-type confidence intervals
- Confidence bands over the landmark times are computed using a resampling method

SIMULATION STUDY

- Simulations studies have been carried out to
  - ▷ **show the usefulness of R^2 curve in contrast to the Brier Score or the AUC curves** ;
  - ▷ **study the behaviour of the inference** of R^2 curve.
- Data were simulated from a shared random effect joint models for longitudinal and time to event data. 500 simulations were done with a sample size of 1,000 and 3,000.

RESULTS OF THE SIMULATION STUDY

- In the scenario presented here, the proportion of events increase considerably according to the landmark time (10% at $s = 0$ to 55% at $s = 5.5$).



Estimates of $R^2_\pi(s,t)$ for landmark time $s \in \{0, 0.5, \ldots, 5.5\}$ with $n = 1,000$ and $n = 3,000$. Coverage of simultaneous confidence band 93.0% (for $n = 3,000$) and 93.8% (for $n = 1,000$).

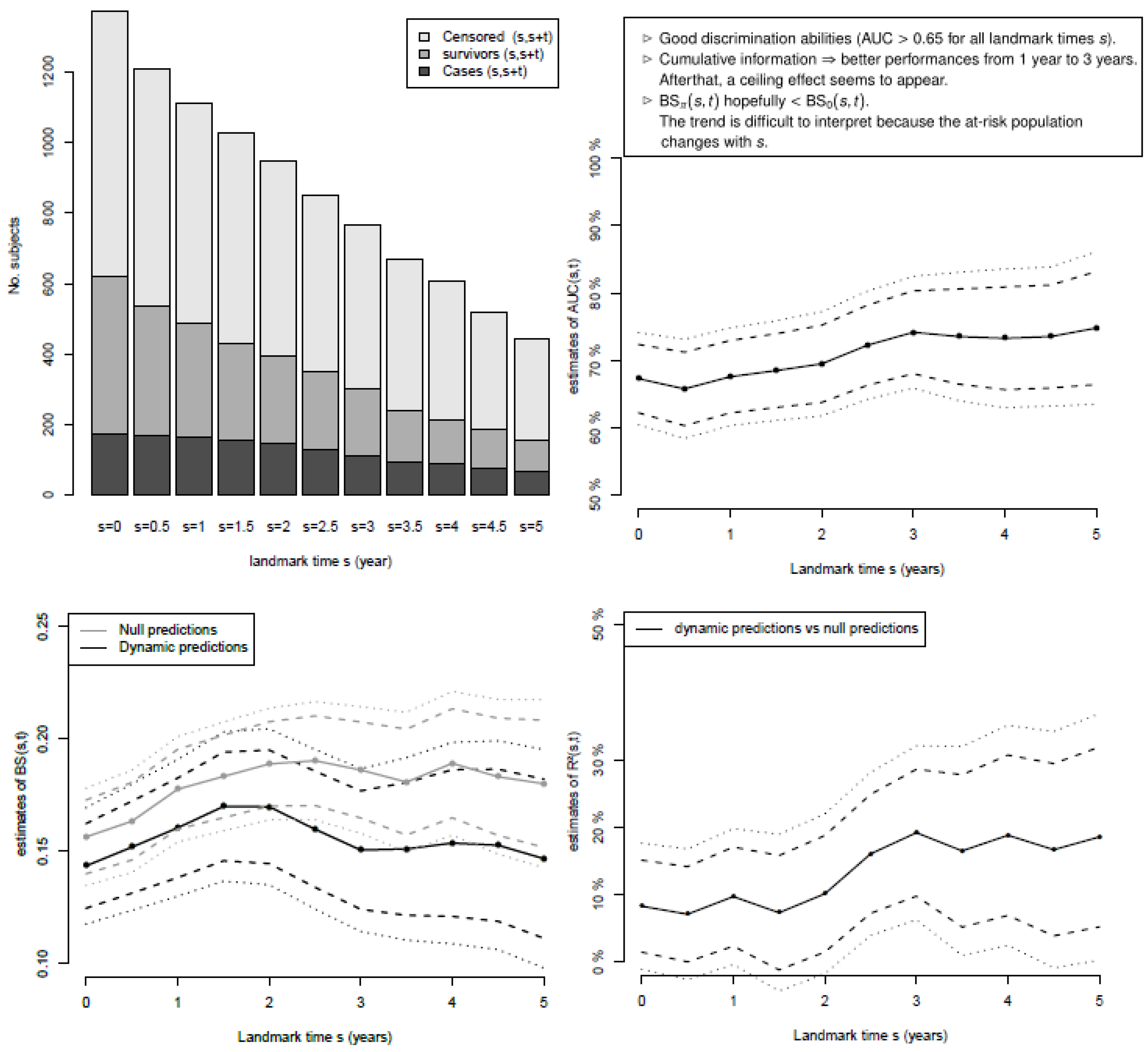
landmark time s	Status in (s, s+t) Dead	Alive	Censored	Bias ($\times$ 100)	CP(%)	a.s.e. ($\times$ 100)	a.s.e/s.e
$n = 3,000$							
0	269	2531	200	-0.1	93.4	1.78	0.97
0.5	404	2330	265	-0.0	95.2	1.78	0.98
1	564	2097	338	0.0	95.2	1.84	0.99
1.5	739	1842	414	-0.0	93.8	1.79	0.96
2	919	1578	491	-0.0	93.6	1.74	0.99
2.5	1091	1320	562	0.1	94.0	1.74	1.02
3	1240	1078	624	0.1	94.2	1.77	0.97
3.5	1357	861	673	0.1	95.2	1.85	1.00
4	1432	673	704	-0.1	94.2	1.97	0.97
4.5	1459	514	719	-0.0	94.2	2.16	0.96
5	1434	383	714	-0.1	93.4	2.39	0.97
5.5	1359	280	691	-0.0	93.8	2.71	0.95
$n = 1,000$							
0	90	844	67	-0.1	94.8	3.09	0.98
0.5	135	777	89	0.0	94.0	3.09	0.97
1	188	699	112	-0.0	94.6	3.20	0.96
1.5	246	614	138	0.0	94.2	3.10	0.97
2	306	526	164	-0.1	95.0	3.03	0.98
2.5	363	440	187	-0.0	94.0	3.03	0.98
3	413	360	208	-0.1	95.2	3.07	0.97
3.5	453	287	224	-0.0	96.0	3.21	1.05
4	478	224	235	-0.3	94.2	3.42	1.01
4.5	486	171	239	-0.3	95.4	3.75	0.99
5	478	128	238	-0.5	93.6	4.16	0.95
5.5	453	93	230	-0.5	94.2	4.73	0.94

acronyms: CP: Coverage Probability; a.s.e.: asymptotic standard error; s.e.: standard error (empirical)

- BS curve ↗ (at least at the beginning) = accuracy of predictions ↘ ... But surprinsingly, NOT: R^2-curve ↗. This is due to the fact that the BS curve of the marginal predictions follows a parallel trend.
- Satisfied results concerning the behaviour of the estimations

APPLICATION IN RENAL TRANSPLANTATION

- Context:
  - ▷ 4,121 kidney recipients from the French prospective DIVAT cohort (www.divat.fr)
  - ▷ Divided into training (2/3: n=2,749) and validation (1/3: n=1,372)
  - ▷ Longitudinal marker: **Serum creatinine**, yearly measured
  - ▷ Event: **Kidney graft failure** (return to dialysis or death with a functioning graft).
  - ▷ landmark times $s \in \{0, 0.5, \ldots, 5\}$ and time horizon $t = 5$ years for a medium-term prognosis.
  - ▷ Some scores already exist in kidney transplantation (Foucher et al. 2010; Lorent et al. 2016) but they did not integrate repeated measurements.
- Dynamic predictions calculated on validation sample from a shared random effect joint model estimated on the learning data set (corresponding to a simplified version of a previous work (Fournier et al. 2016)).



- ▷ Intervals confidence and confidence bands are rather large because of the important censoring process.

CONCLUSION

- R^2 criterion is closely related to the popular concept of **"explained variation"**
  - ▷ summarizes calibration AND discrimination simultaneously
  - ▷ has an understandable trend
- Others simulations are in process to show difference of interpretations between AUC curve and R^2 curve.

REFERENCES

Blanche, P. et al. (2015). Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. Biometrics

Foucher, Y. et al. (2010). A clinical scoring system highly predictive of long-term kidney graft survival. Kidney Int

Fournier, MC. et al. (2016). A joint model for longitudinal and time-to-event data to better assess the specific role of donor and recipient factors on long-term kidney transplantation outcomes. Eur J Epidem

Lorent, M. et al. (2016). Mortality Prediction after the First Year of Kidney Transplantation: An Observational Study on Two European Cohorts. PloS One

Nicolaie, M. et al. (2013). Dynamic prediction by landmarking in competing risks. SiM

Proust-Lima, C. and Taylor, J. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. Biostatistics

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics

Steyerberg E. Clinical Prediction Models. Springer, 2009.

Van Houwelingen, H. and Putter, H. (2012). Dynamic Prediction in Clinical Survival Analysis. CRC Press.