Objective: To provide an R²-type criterion to evaluate dynamic prediction

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

Materials and Methods

- Notations:
  - $s$: the subject; $a$: the landmark time; $t$: the horizon window.
  - $T$: the length of the time-to-event; $C$: the censoring time.
  - $D(s,t) = |t < s < t|$; the observed time of follow-up and $\Delta = |t \leq C|$, with $\Delta$ the indicator function.
  - $s(\pi): D(s,t_1) = 1$, $D(s,t_2) = 0$. $T = s$: the subject-specific dynamic prediction with $\pi(s)$: the observed subject-specific characteristics at landmark time $s$.

- $\text{Brier Score (the lower the better)}$: $\text{BS}_1(s,t) = \left[\left(\pi(s,t) - s(t)\right)^2\right] / T > s$.

- $\text{BS} = \text{Bias}^2 + \text{Variance}$
  - Evaluates both discrimination and calibration:
    \[
    \text{BS}_2(s,t) = \frac{\text{Var}(s(t) - s(t))}{T > s} + \left[\text{E}[(s(t) - s(t))^2] - \text{Var}(s(t))\right]/ T > s
    \]

- $R^2$ criterion (the higher the better)
  - Benchmark value: the best "null" model (or marginal) gives the same predicted risk for all subjects
  - $\text{RP}(s,t) = 1 - \text{BS}_2(s,t)$

- $R^2$ can be easily understood comparing to those of the Brier Score.

Interpretation:

- $\text{RP}(s,t) = 1 - \text{BS}_2(s,t)$: the prediction perfectly distinguishes patients who will experience an event in $(s, s + t)$ from those who will not.
- $\text{RP}(s,t) = 0$ when the subject-specific information is wrongly used (ie extreme cases where the predictions performed worse than the marginal ones, with over fitted predictions for example).
- Use of the Inverse Probability of Censoring Weighing (IPCW) to make inference (like in Steyerberg et al. 2010)
- 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 Influence 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 $s = 5.5$)
  - $\text{BS}_1(s,t) = \left[\left(\pi(s,t) - s(t)\right)^2\right] / T > s$.
  - $\text{BS}_2(s,t) = \frac{\text{Var}(s(t) - s(t))}{T > s} + \left[\text{E}[(s(t) - s(t))^2] - \text{Var}(s(t))\right]/ T > s$

- $\text{BS}_2$ curve (at least at the beginning) = accuracy of predictions $\pi(\ldots)$ ... But surprising: 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:
  - 1,421 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 $s = (0,0.5,...,5)$ and time horizon $t = 5$ years for a medium-term prognosis.
- Some scores already exist in kidney transplantation (Foucher et al. 2010; Lorenet 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 date set (corresponding to a simplified version of a previous work (Fournier et al. 2016)).

Conclusion

- $\text{RP}$ 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.

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

References


Nicolai, M et al. (2015). Dynamic prediction by landmarking in competing risks. SM


Steyerberg, E. C. Clinical Predictive Model. Springer. 2010