Contexte
Méthodes
Application
Conclusion
Références

Construction et estimation des capacités d'un score pronostique : intérêts de la pénalisation de LASSO et de l'estimateur bootstrap 0.632+ appliqués aux courbes ROC dépendantes du temps

Marie-Cécile Fournier[a,b], Florence Gillaizeau[a,b,c], Awena Le Fur[b,c], Jacques Dantal[b,c], Yohann Foucher[a,c]

marie-cecile.fournier@univ-nantes.fr

[a] *EA 4275 - SPHERE - Biostatistique, Pharmacoépidémiologie et Mesures Subjectives en Santé, Université de Nantes*
[b] *INSERM UMR1064, Institut Transplantation-Urologie-Néphrologie, Nantes*
[c] *Centre Hospitalier Universitaire de Nantes*

Contexte

Contexte
Méthodes
Application
Conclusion
Références

- Nombre de variables $\gg$ Nombre individus
  Ex : construction de signatures à partir de techniques des puces à ADN

- Données censurées (survie) : faible nombre d'évènements observés
  Ex : cohorte de faible taille, temps de suivi court ...

$\Rightarrow$ ⚠Risque de sur-ajustement (overfitting)

Objectifs :

1. Construire un score pronostique
2. Estimer correctement ses capacités pronostiques

Le modèle de Cox

Contexte
Méthodes
Application
Conclusion
Références

Modèle de Cox

$$h(t|X_1 = x_1, ..., X_p = x_p) = h_0(t) exp(\beta_1 x_1 + ... + \beta_p x_p)$$

- $h_0(t)$ risque de base
- $X = (X_1, ..., X_p)$ vecteur des P variables
- $\beta = (\beta_1, ..., \beta_p)$ coefficients de régression estimés par maximisation de la vraisemblance partielle VP :

$$VP(\beta) = \prod_{j=1}^{N} \left[\frac{exp(\beta x_j)}{\sum_{i \in R_j} exp(\beta x_i)} \right]^{\delta_j}$$

- R_j individus à risque au temps t_j
- δ_j=1 si l'évènement a eu lieu, 0 sinon.

La pénalisation de LASSO

Contexte
Méthodes
Application
Conclusion
Références

- Sélection des variables sur autre critère que p-value *

Modèle de Cox avec pénalisation de Lasso

$$\widehat{\beta} = argmax_{\beta} \left\{ VP(\beta) - \lambda \sum_{p=1}^{P} |\beta_p| \right\}$$

$\lambda \geq 0$, le paramètre de pénalisation

- Nous obtenons le score pronostique $\widehat{\eta} = \widehat{\beta} x$
- Si $\widehat{\eta}_j > c$ $(\forall c \in \mathbb{R})$, alors le patient j est dit à haut risque de subir l'évènement avant τ *(temps de pronostic)*.

* Tibshirani, 1996

Courbes ROC dépendantes du temps †

Contexte
Méthodes
Application
Conclusion
Références

2 catégories d'erreurs possibles :

- Le taux de faux négatifs : $TFN_\tau(c) = P(\eta \leq c | T \leq \tau)$
- Le taux de faux positifs : $TFP_\tau(c) = P(\eta > c | T > \tau)$

T : temps d'évènement

Taux d'erreurs estimés non paramétriquement ‡

◇ Courbe ROC dépendante du temps (ROCt) :

1-$TFN_\tau(c)$ en fonction de $TFP_\tau(c)$ pour toutes les valeurs de c.

◇ Aire sous la courbe ROCt (AUCt)

†Haegerty et al., 2000
‡Akritas, 1994

Evaluation des capacités pronostiques du score (1)

Contexte
Méthodes
Application
Conclusion
Références

- B échantillons de taille N avec remise (Bootstrap)
- Pour les B échantillons de patients inclus, estimation des paramètres β par maximisation de la vraisemblance pénalisée

Estimateur apparent

$$\overline{TFN}_\tau(c) = B^{-1} \sum_{b=1}^{B} \widehat{TFN}_\tau^{b+}(c) \quad \text{et} \quad \overline{TFP}_\tau(c) = B^{-1} \sum_{b=1}^{B} \widehat{TFP}_\tau^{b+}(c)$$

$\Longrightarrow$ Sur-estimation des capacités pronostiques

Bootstrap Cross Validation

$$\widehat{TFN}_\tau^{BCV}(c) = B^{-1} \sum_{b=1}^{B} \widehat{TFN}_\tau^{b-}(c)$$

$$\widehat{TFP}_\tau^{BCV}(c) = B^{-1} \sum_{b=1}^{B} \widehat{TFP}_\tau^{b-}(c)$$

$\Longrightarrow$ Sous-estimation des capacités pronostiques

Evaluation des capacités pronostiques du score (2) †

Contexte
Méthodes
Application
Conclusion
Références

Probabilité qu'un individu soit :

- inclus dans l'échantillon bootstrap : $\lim_{N\to+\infty} 1-\left[1-\frac{1}{N}\right]^N = 0,632$
- non inclus dans l'échantillon bootstrap : $\lim_{N\to+\infty} \left[1-\frac{1}{N}\right]^N = 0,368$

Bootstrap 0.632

$$\widehat{TFN}_\tau^{0.632}(c) = 0,368\,\overline{TFN}_\tau(c) + 0,632\,\widehat{TFN}_\tau^{BCV}(c)$$

$$\widehat{TFP}_\tau^{0.632}(c) = 0,368\,\overline{TFP}_\tau(c) + 0,632\,\widehat{TFP}_\tau^{BCV}(c)$$

$\Longrightarrow$ Sur-estimation possible des capacités pronostiques

† Efron, 1983 ; Foucher & Danger, 2012

Evaluation des capacités pronostiques du score (3) ‡

Contexte
Méthodes
Application
Conclusion
Références

Bootstrap 0.632+

$$\widehat{TFN}_{\tau}^{0.632+}(c) = \{1 - \hat{r}(c)\}\,\overline{TFN}_{\tau}(c) + \hat{r}(c)\,\widehat{TFN}_{\tau}^{BCV}(c)$$

$$\widehat{TFP}_{\tau}^{0.632+}(c) = \{1 - \hat{r}(c)\}\,\overline{TFP}_{\tau}(c) + \hat{r}(c)\,\widehat{TFP}_{\tau}^{BCV}(c)$$

Amélioration de l'estimateur 0.632 : pondération variable selon le taux de sur-ajustement $\hat{r}(c)$

‡ Efron & Tibshirani, 1997 ; Foucher & Danger, 2012

Estimation du paramètre de pénalisation

Contexte
Méthodes
Application
Conclusion
Références

- **Classiquement :**
  Par Validation croisée réalisée pour chaque échantillon de bootstrap : tout le modèle est réestimé à chaque itération

- **En pratique** :
  Réestimation coûteuse en temps de calcul

  → Estimation a priori de λ sur tout l'échantillon [§]

[§] Foucher & Danger, 2012 ; Schumacher et al., 2007

Application

Contexte
Méthodes
Application
Conclusion
Références

- Cohorte DIVAT Nantes
- Patients transplantés rénaux entre 2000 et 2010

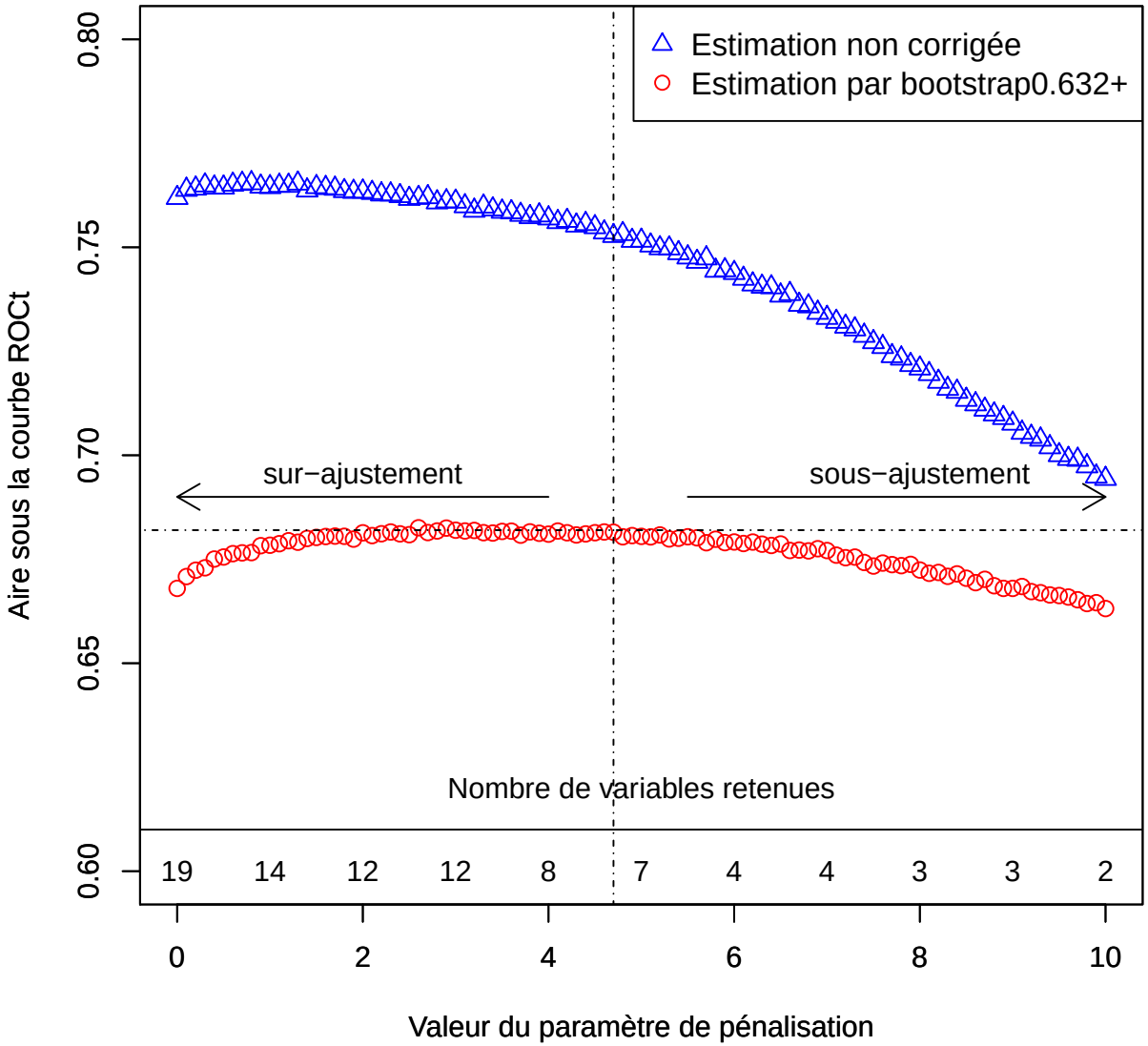
Objectif :

Proposer un score pronostique de la survenue d'un diabète post transplantation.
Score à calculer au moment de la greffe.

- n = 444 et seulement 58 évènements observés

Choix du paramètre de pénalisation
Détermination du score pronostique

Contexte
Méthodes
Application
Conclusion
Références

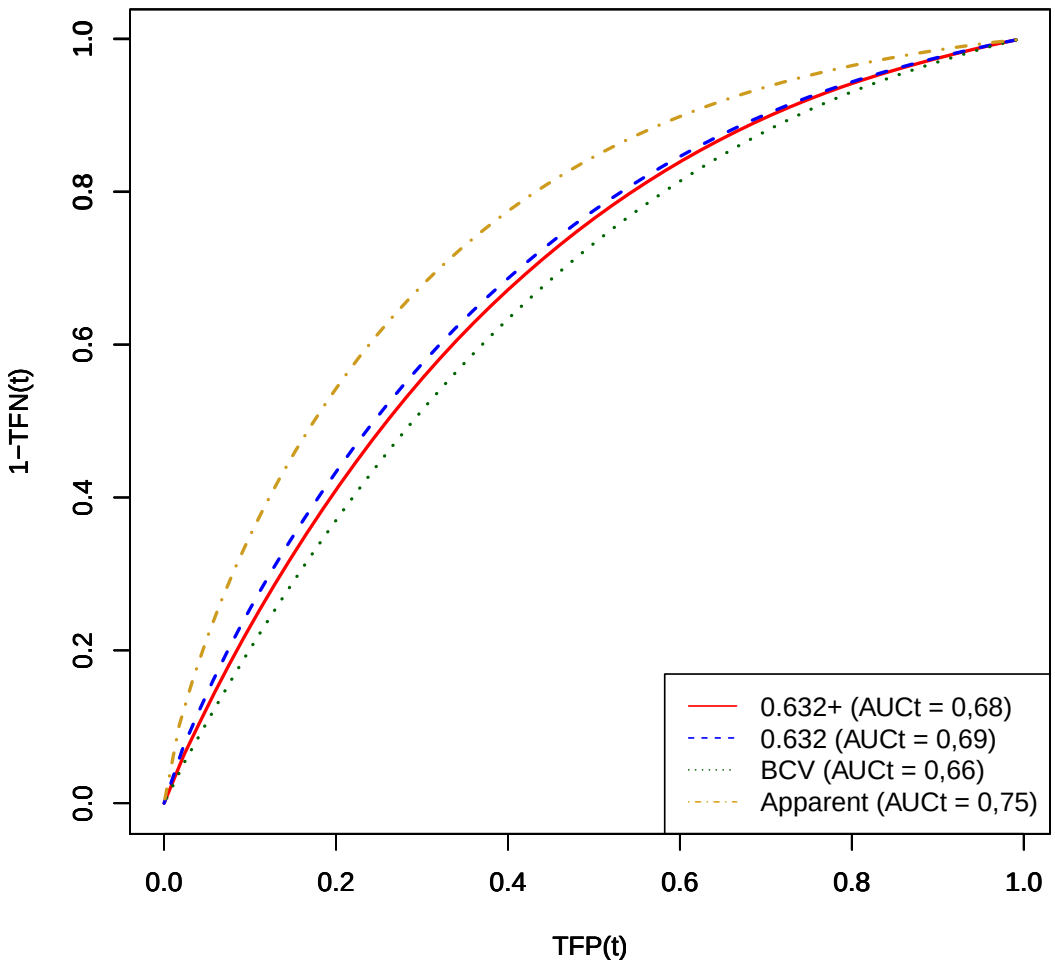


PACKAGE R : **ROC632** disponible sur www.divat.fr

Evaluation des capacités pronostiques
Apport de la correction

Contexte
Méthodes
Application
Conclusion
Références

Marqueurs pronostiques du diabète post-transplantation	coefficients
Taux de vitamine D à la greffe < 10 ng/mL	0,1592
Age du receveur ≥ 55 ans	0,3691
Patient traité par corticostéroïdes	0,2590
Patient traité par tacrolimus	0,6283
Indice de Masse Corporelle du receveur normalisé (centré-réduit)	0,3616
Patient ayant des antécédents cardiovasculaires	0,0647
Patient ayant des antécédents de dyslipidémie	-0,0761



Conclusion

Contexte
Méthodes
Application
Conclusion
Références

- Pas de découpage de l'échantillon initial, ce qui évite :
  - l'aggravation de la faible puissance
  - des intervalles de confiance plus larges
  - la répétition du découpage jusqu'à l'obtention de bons résultats
- Sélection sur un autre critère que la p-value

LIMITES

- Détermination du paramètre de pénalisation subjective
- Absence d'intervalle de confiance pour la courbe ROCt corrigée

Références

Contexte
Méthodes
Application
Conclusion
Références

- Akritas. Nearest neighbor estimation of a bivariate distribution under random censoring. The annals of statistics. 1994 Sep ;22(3) :1299-1327.
- Efron. Estimating the error rate of a prediction rule : improvement on cross validation. Journal of the american statistical association. 1983 Jun ;78(382)n :316-331.
- Efron, Tibshirani. Improvements on cross validation : the 632+ bootstrap method. Journal of the american statistical association. 1997 ;92(438) :548-560.
- Foucher, Danger. Time dependant ROC curves for the estimation of true prognostic capacity of microarray. Statistical applications in genetics and molecular biology. 2012 ;11(6).
- Haegerty, Lumley, Pepe. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics, 2000 Jun ;56(2) :337-344.
- Schumacher, Binder, Gerds. Assessment of survival prediction models based on microarray data. Bioinformatics, 23 :1768-1774,2007.
- Tibshirani. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society series B, 1996

Merci de votre attention