Courbes ROC dépendantes du temps et transplantation rénale

Yohann Foucher

ITERT & INSERM U643
Yohann.Foucher@univ-nantes.fr

19 Janvier 2010

Table of contents

Plan

Introduction (1)

Evolution des critères de jugements

- Amélioration du pronostic (pas de rejet aigu, chronique, etc.)
- Nécessité de marqueurs précoces du pronostic
- Aujourd'hui, la référence reste la biopsie précoce du greffon
  - Problème : Acte couteux et invasif.
- Concentration de la communauté médicale sur la clairance à la créatinine (CrCl)
  - Très corrélée à la survie : Kaplan-Meier (Hariharan, KI, 2002)
  - En pratique, de nombreuses études se basent sur la CrCl à 6-12 mois

Problème

- CrCl semble un mauvais marqueur pronostique : Courbes ROC (Kaplan et al., AJT, 2003)

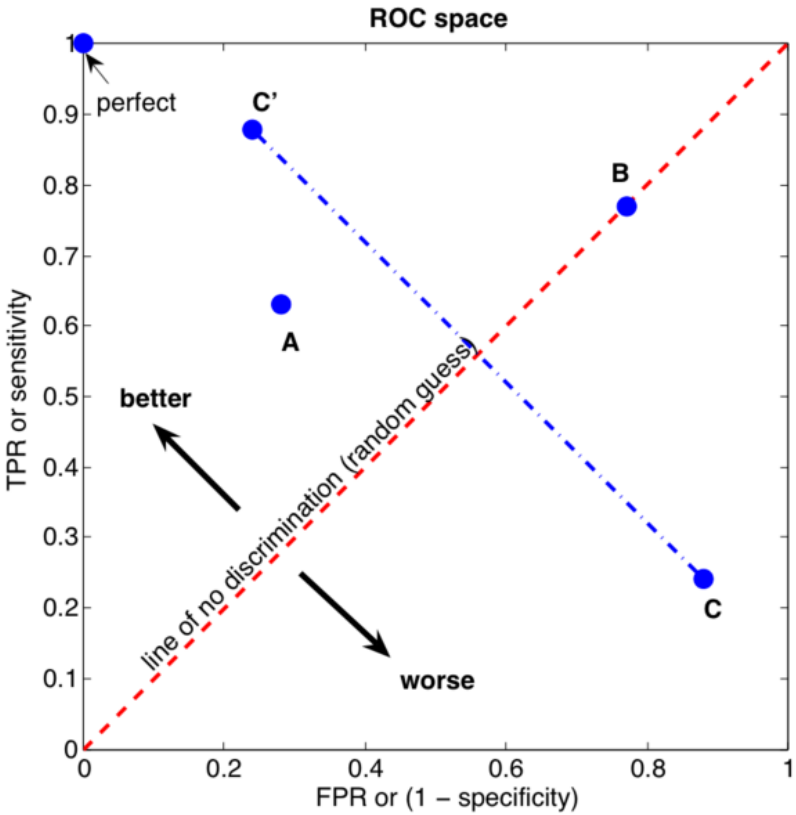
Introduction (2)

Rappels relatifs aux courbes ROC (1)

- Soit X la valeur du marqueur diagnostique et D l'événement à prédire ($D = 1$ si malade et $D = 0$ sinon).
- Les patients sont à risque si X est supérieur à un seuil c.
  - Sensibilité : $SE = P(X > c|D = 1)$.
  - Proportion de tests positifs chez les patients malades.
- Les patients ne sont pas à risque si X est inférieur à c.
  - Spécificité : $SP = P(X \leq c|D = 0)$.
  - Proportion de tests négatifs chez les patients non-malades.
- Un test diagnostique parfait : $SE = SP = 1$.
- Courbes ROC : Pour tous les seuils c possibles, on calcul $1 - SP$ et SP.

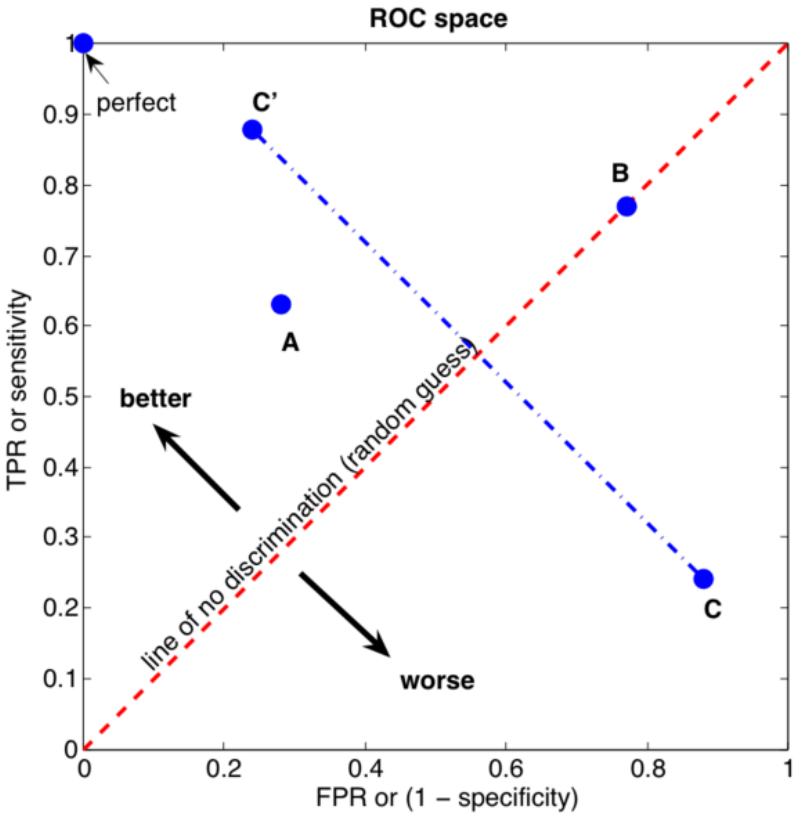
Introduction (3)

Rappels relatifs aux courbes ROC (2)



Introduction (3)

Rappels relatifs aux courbes ROC (2)



Grille d'évaluation

- Si AUC $= 0{,}5$: Aucune discrimination
- Si $0{,}7 \leq$ AUC $< 0{,}8$: Discrimination acceptable
- Si $0{,}8 \leq$ AUC $< 0{,}9$: Discrimination excellente
- Si AUC $\geq 0{,}9$: Discrimination hors du commun

Introduction (4)

Rappels relatifs aux courbes ROC (3)

- Courbes ROC très bien adaptées aux analyses diagnostiques
- Problèmes pour les analyses pronostiques basées sur des données incomplètes. Ex : Si on étudie les capacités de X à prédire le décès dans les 5 ans qui suivent sa mesure.
  - Sensibilité : $SE = P(X > c|D = 1)$.
  - Proportion de tests positifs chez les patients décédés dans les 5 ans.
    - Pas de problème d'estimation.
  - Spécificité : $SP = P(X \leq c|D = 0)$.
  - Proportion de tests négatifs chez les patients vivant à 5 ans.
    - Comment traité un patient vivant à 4 ans dont on a pas de nouvelles ensuite ?

Introduction (5)

Zoom sur les courbes ROC de Kaplan et al. (AJT, 2003)

" Observed patients were included that possessed follow-up information for a given time period (either 2 years or 7 years), and baseline levels at 6 months and 1 year were both tested as explanatory variables. "

- Seuls les patients ayant un suivi assez important sont inclus donc :
  - Les patients n'ayant pas d'événement ont moins de chance d'être inclus
  - Ce déséquilibre est d'autant plus fort que le suivi est long
  - La même analyse à partir de la même cohorte n'utilise pas les mêmes patients selon le temps de pronostic.
  - Etc.... Tous les biais liés à la non-prise en compte de la censure à droite.

Introduction (6)

Zoom sur les courbes ROC de Kaplan et al. (AJT, 2003)

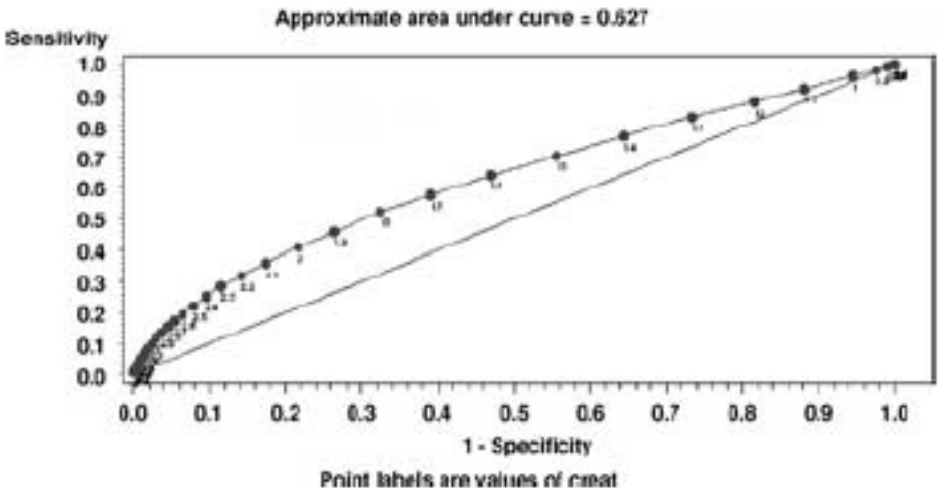


Figure 1: ROC curve of creatinine for overall graft failure (2-year graft survival 1-year creatinine as a predictive quantity).

Introduction (7)

Zoom sur les courbes ROC de Kaplan et al. (AJT, 2003)

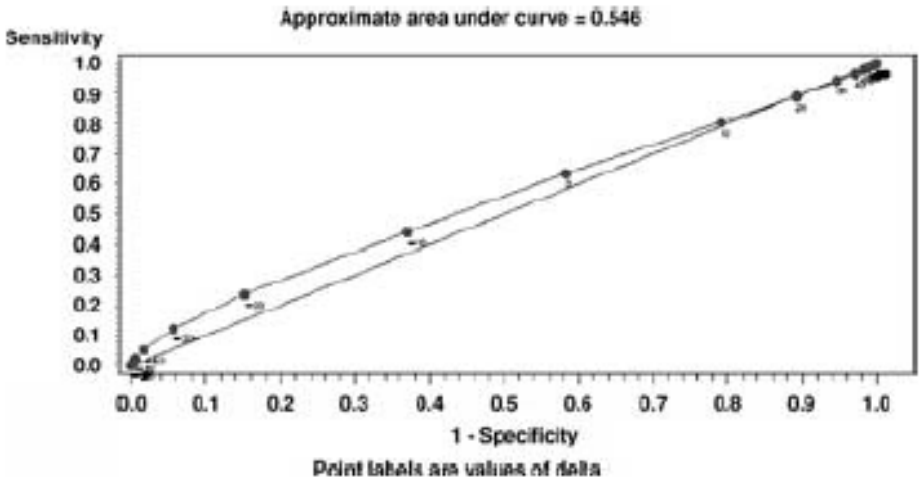


Figure 2: ROC plot for 7-year graft outcome with relative percent change in 1/creatinine levels from 6 months to 1 year as a predictor.

Introduction (8)

Intérêt des courbes ROC dépendantes du temps

- Développées initialement par Heagerty et al. (Biometrics, 2000)
- Permettent l'évaluation des capacités pronostiques d'un marqueur avec données censurées à droite

Applications spécifiques à la transplantation

1. Capacité de la CrCl à pronostiquer le retour en dialyse.
2. Capacité d'un score à pronostiquer le retour en dialyse.
3. Capacité de la CrCl à pronostiquer le retour en dialyse et/ou le décès du patient (généralisation aux risques compétitifs).
4. Sélection des gènes pronostiques de la dégradation de la CrCl à partir de puces (prise en compte du problème de dimension et de la troncature).

Plan

Plan

Données (1)

Critères d'inclusions $\longrightarrow$ 3040 patients

- Centres : Nantes (depuis 1996), Nancy (depuis 1998), Paris Necker (depuis 1996), Toulouse (depuis 2003) et Montpellier (depuis 2003).
- Uniquement les patients non retournés en dialyse, non décédés ou non perdus de vue dans la première année de greffe.
- Greffon issu de cadavre.
- Receveur âgé d'au moins 18 ans au moment de la transplantation.
- Aucune greffe de pancréas associée.

Données (2)

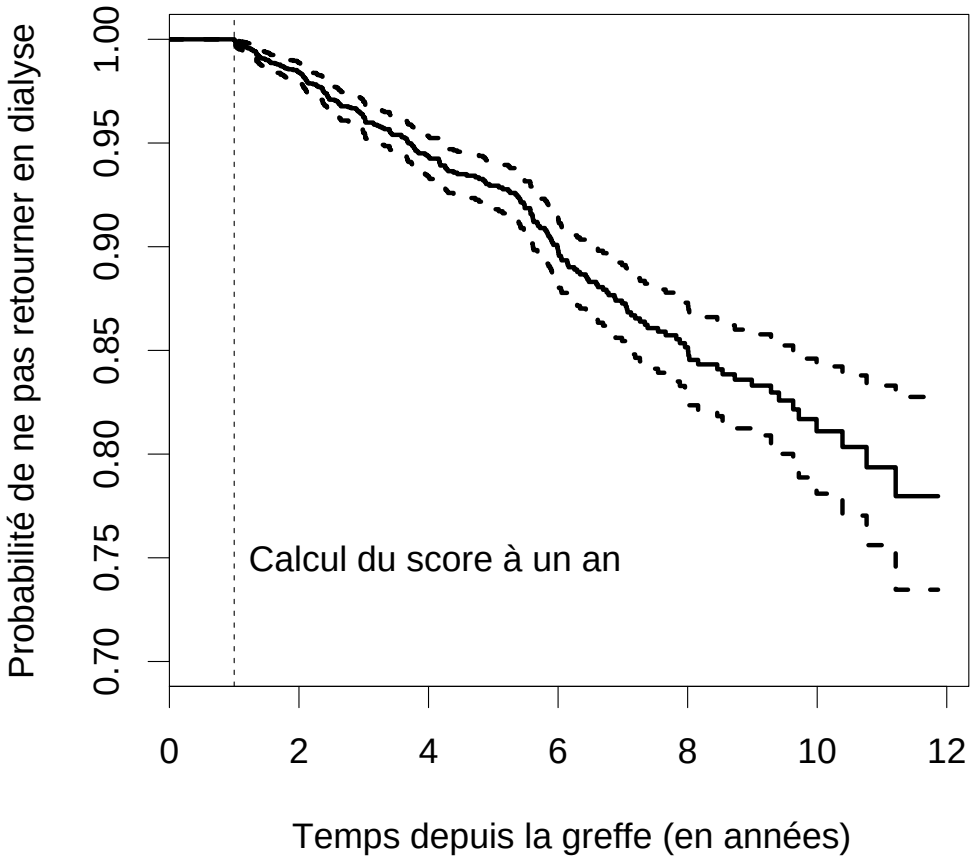
Evénement étudié

- Temps entre la greffe et le retour en dialyse.
- Décès = Censures à droite (perdus de vue).
- 230 retours en dialyse.

Données (2)

Evénement étudié

- Temps entre la greffe et le retour en dialyse.
- Décès = Censures à droite (perdus de vue).
- 230 retours en dialyse.



Plan

Courbes ROC temps-dépendantes (1)

- Théorie proposée par Heagerty et al. (Biometrics, 2001).
- Soit T le temps d'apparition de l'événement étudié.
- Soit X le marqueur de substitution composite.

	ROC	ROC(t)
Sensibilité	$se(c) = P(X > c \mid D = 1)$	$se(c,t) = P(X > c \mid D(t) = 1)$
Spécificité	$sp(c) = P(X \leq c \mid D = 0)$	$sp(c,t) = P(X \leq c \mid D(t) = 0)$

- ROC(t) $\longrightarrow$ AUC(t)

Courbes ROC temps-dépendantes (2)

- L'estimation la plus simple fait appel à deux estimateurs bien connus :
  - L'estimateur de Kaplan et Meier de la survie

$$\hat{S}_{KM}(t|X \leq c) = P(T > t|X \leq c) = \prod_{t_i \leq \tau; x_i \leq c} \{1 - d_i/r_i\}$$

    - d_i : nombre de décès au temps t_i
    - r_i : effectif à risque au temps t_i
  - L'estimateur empirique de la fonction de répartition

$$\hat{F}_X(c) = P(X \leq c) = N^{-1} \sum_i \Delta(x_i \leq c)$$

Courbes ROC temps-dépendantes (3)

Calcul de la sensibilité

$$se(c,t) = P(X > c | D(t) = 1)$$

$$\downarrow$$

$$se(c,t) = P(T \leq t | X > c) P(X > c) / P(T \leq t)$$

$$\downarrow$$

$$se(c,t) = \{1 - S(t|X > c)\}\{1 - P(X \leq c)\} / \{1 - S(t)\}$$

$$\downarrow$$

$$\hat{se}_{KM}(c,t) = \{1 - \hat{S}_{KM}(t|X > c)\}\{1 - \hat{F}_X(c)\} / \{1 - \hat{S}_{KM}(t)\}$$

Courbes ROC temps-dépendantes (4)

Calcul de la spécificité

$$sp(c,t) = P(X \leq c | D(t) = 0)$$

$$\downarrow$$

$$sp(c,t) = P(T > t | X \leq c) P(X \leq c) / P(T > t)$$

$$\downarrow$$

$$sp(c,t) = S(t\ X \leq c) P(X \leq c) / S(t)$$

$$\downarrow$$

$$\hat{sp}_{KM}(c,t) = \hat{S}_{KM}(t | X \leq c) \hat{F}_X(c) / \hat{S}_{KM}(t)$$

Courbes ROC temps-dépendantes (5)

Procédure d'estimation du score (DIVAT, $n \approx 3000$)

1. Analyse de tous les facteurs de risque à 1 an de greffe
2. Temps de pronostic = 8 ans.
3. Estimation d'un modèle de Cox par maximisation de la vraisemblance partielle.
4. Score initial = somme des coefficients de régression ($\log RR$) multipliés par les variables correspondantes.
5. Correction des coefficients par maximisation de l'aire sous la courbe ROC temps dépendante.
6. Estimation de l'intervalle de confiance par re-échantillonnage (bootstrap).
7. Calcul du seuil du score optimal par maximisation jointe de la sensibilité et de la spécificité.

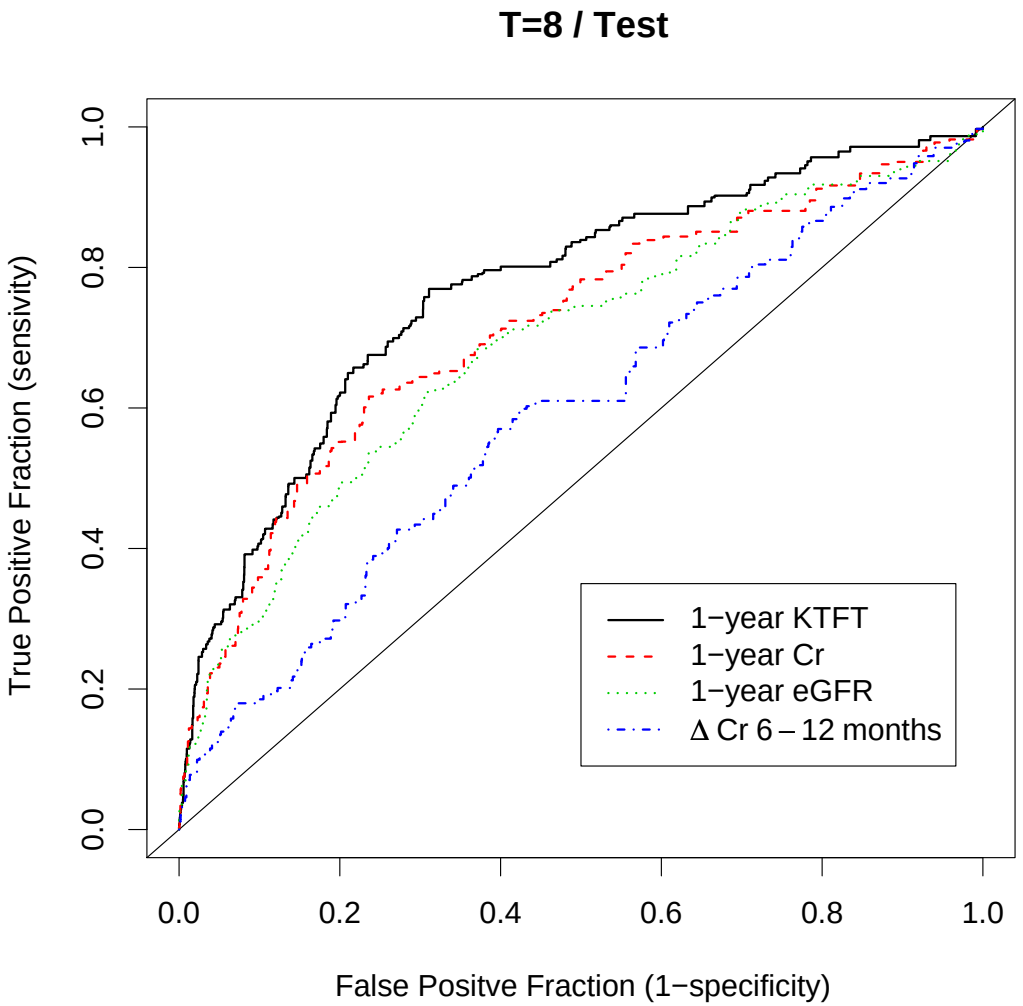
Courbes ROC temps-dépendantes (6)

Validation du score (Tours, Caen, Strasbourg, $n \approx 300$)

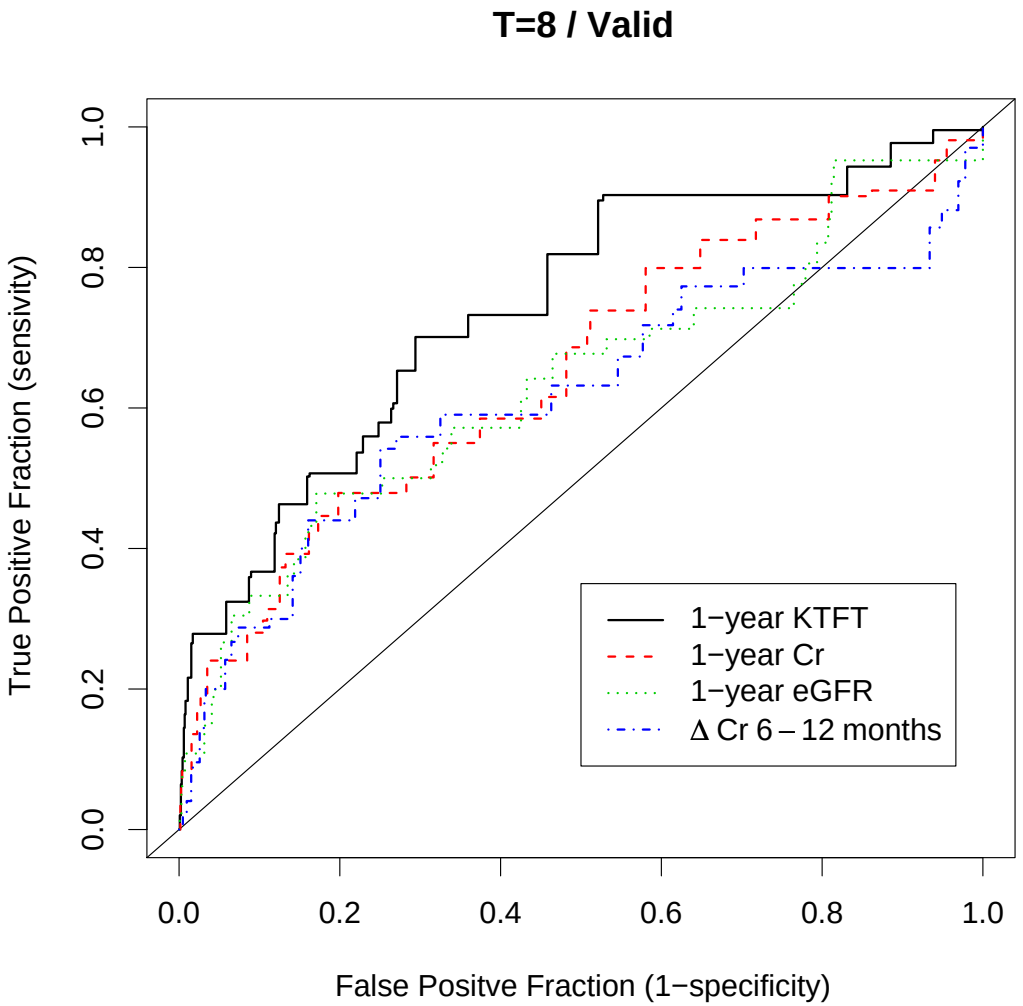
1. Calcul du score à partir des paramètres précédents.
2. Calcul de l'aire sous la courbe ROC.
3. Calcul de l'intervalle de confiance de l'aire sous la courbe.
4. Application du seuil précédent (estimateur de Kaplan et Meier dans chacun des groupes)

Plan

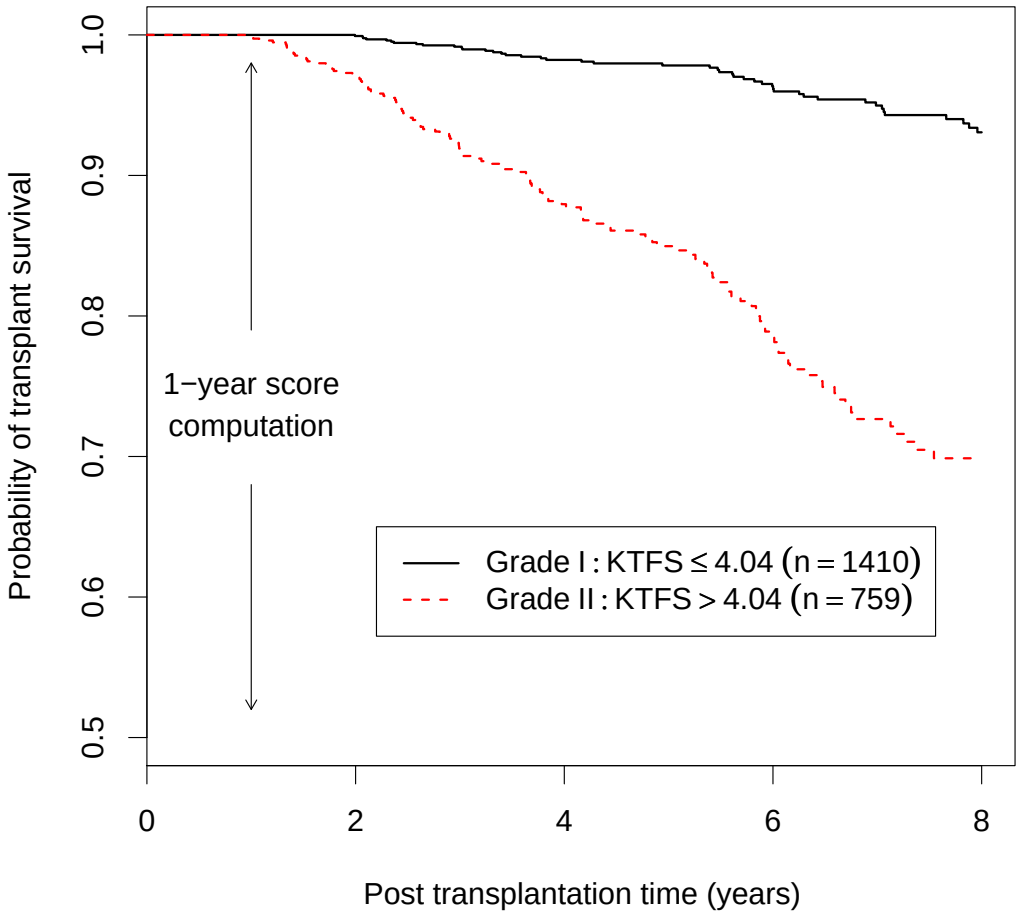
Capacités pronostiques du score à 8 ans - Test



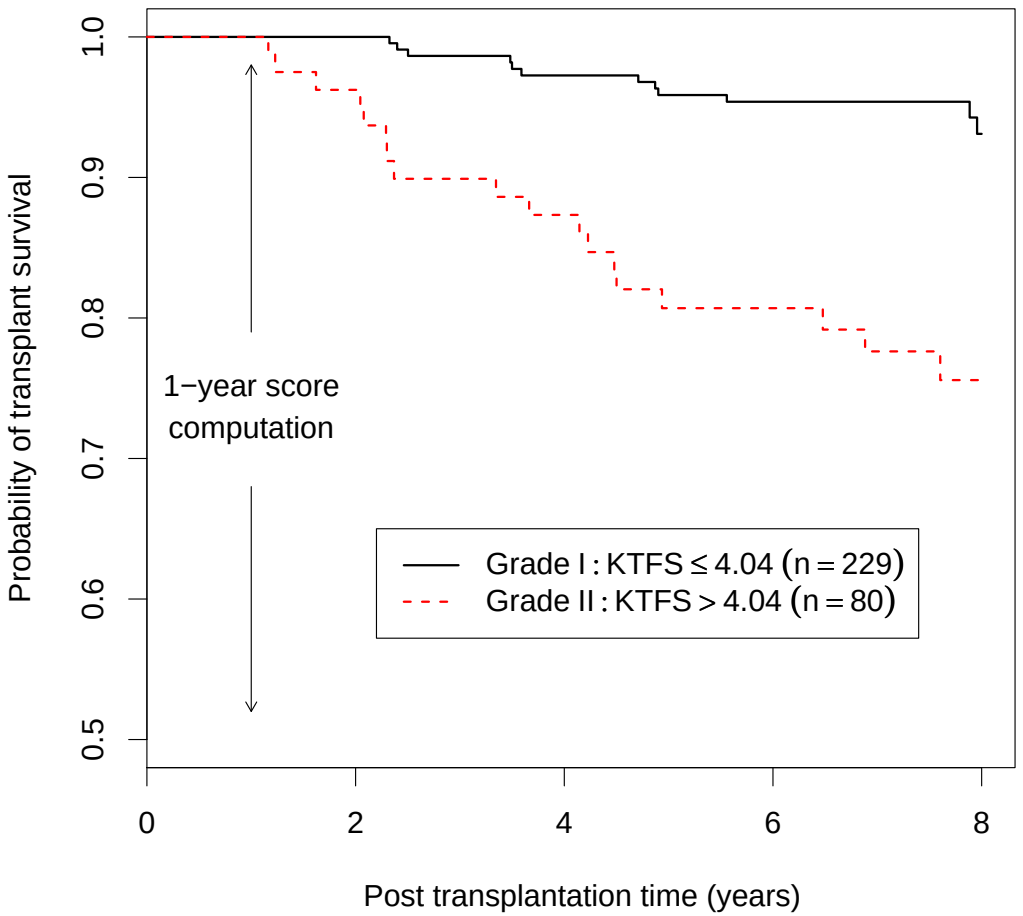
Capacités pronostiques du score à 8 ans - Validation



Discrimination de 2 populations - Test



Discrimination de 2 populations - Validation



Conclusions

Retombées

- On montre que la capacité pronostique de la CrCl n'est pas si mauvaise
- Proposition pour la première fois d'un marqueur composite et précoce du retour en dialyse avec une méthodologie adaptée et une validation externe

Problème

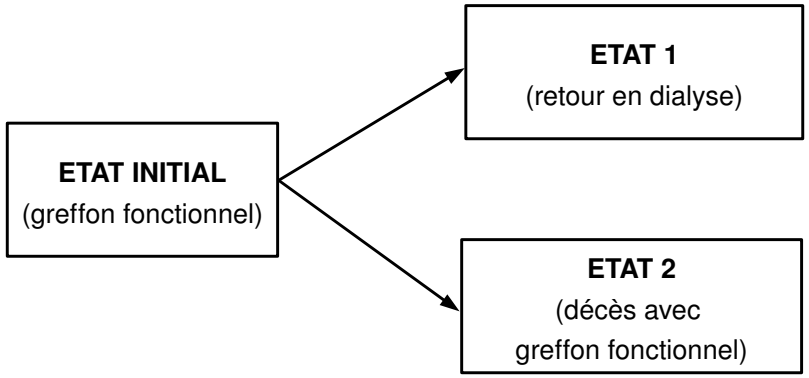
- Une des limites importantes est la censure des décès (certains peuvent être dus à la transplantation)
- Nécessité d'une méthode adaptée au pronostic à 3 classes : greffon fonctionnel, retour en dialyse et décès

Plan

Plan

Courbes ROC temps dépendantes à trois classes (1)

Définitions (1)



- T = le temps écoulé entre l'origine et le 1er échec observé X
- $X = 1$ pour le retour en dialyse et $X = 2$ pour le décès du patient avec un greffon fonctionnel
- P_i = probabilité que l'échec terminal soit i $(i = 1, 2)$
- $f_i(t) = \lim_{\Delta t \to 0^+} P(t < T < t + \Delta t | X = i)/\Delta t$
- y la valeur du marqueur pronostique étudié à l'origine

Courbes ROC temps dépendantes à trois classes (2)

Définitions (2)

- Principe semi-markovien :

$$\lambda_i(t|y) = \lambda_{0i}(t)exp(\beta_i y), \quad i = 1, 2$$

- Soit $z_i = \beta_i y$, le score associé à la transition vers l'échec i.
- Soit $g(z_i)$, la densité de z_i.
- Quand z_i augmente, le risque lié à l'échec i augmente.

Observations à $t = 0$	Pronostic
$z_1 > c_1(\tau)$ et $z_2 \leq c_2(\tau)$	Passage à l'état 1 avant τ
$z_1 > c_1(\tau)$ et $z_2 > c_2(\tau)$	Passage à l'état 1 ou à l'état 2 avant τ
$z_1 \leq c_1(\tau)$ et $z_2 \leq c_2(\tau)$	Aucune transition avant τ
$z_1 \leq c_1(\tau)$ et $z_2 > c_2(\tau)$	Passage à l'état 2 avant τ

Courbes ROC et risques compétitifs (3)

Indicateurs de la capacité de pronostic de y

- Sensibilité propre à l'événement i :

$$\begin{aligned} \mathcal{S}e_i(\tau) &= P(z_i > c_i(\tau) | T \leq \tau, X = i) \\ &= \int_{c_i(\tau)}^{\infty} F_i(\tau|z_i) g(z_i) dz_i \Big/ \int_{-\infty}^{\infty} F_i(\tau|z_i) g(z_i) dz_i \end{aligned}$$

- Sensibilité marginale :

$$\begin{aligned} \mathcal{S}e(\tau) &= P(\overline{z_1 > c_1(\tau), z_2 > c_2(\tau)} | T \leq \tau) \\ &= 1 - \left\{ \sum_{i=1}^{2} P_i \int_{-\infty}^{\omega_i} F_i(\tau|z_i) g(z_i) dz_i \right\} \Big/ \left\{ \sum_{i=1}^{2} P_i \int_{-\infty}^{\infty} F_i(\tau|z_i) g(z_i) dz_i \right\} \end{aligned}$$

où $\omega_1 = min(c_1(\tau), \gamma^{-1} c_2(\tau))$ et $\omega_2 = min(\gamma c_1(\tau), c_2(\tau))$ avec $\gamma = \beta_1/\beta_2 > 0$.

- Développements similaires pour $\mathcal{S}p_i$, $\mathcal{V}PP_i$ et $\mathcal{V}PN_i$
- On en déduit $ROC_i(\tau)$, $ROC(\tau)$, $AUC_i(\tau)$ et $AUC(\tau)$.

Courbes ROC et risques compétitifs (4)

Estimation des seuils de décision : $\hat{c}_1(\tau)$ et $\hat{c}_2(\tau)$

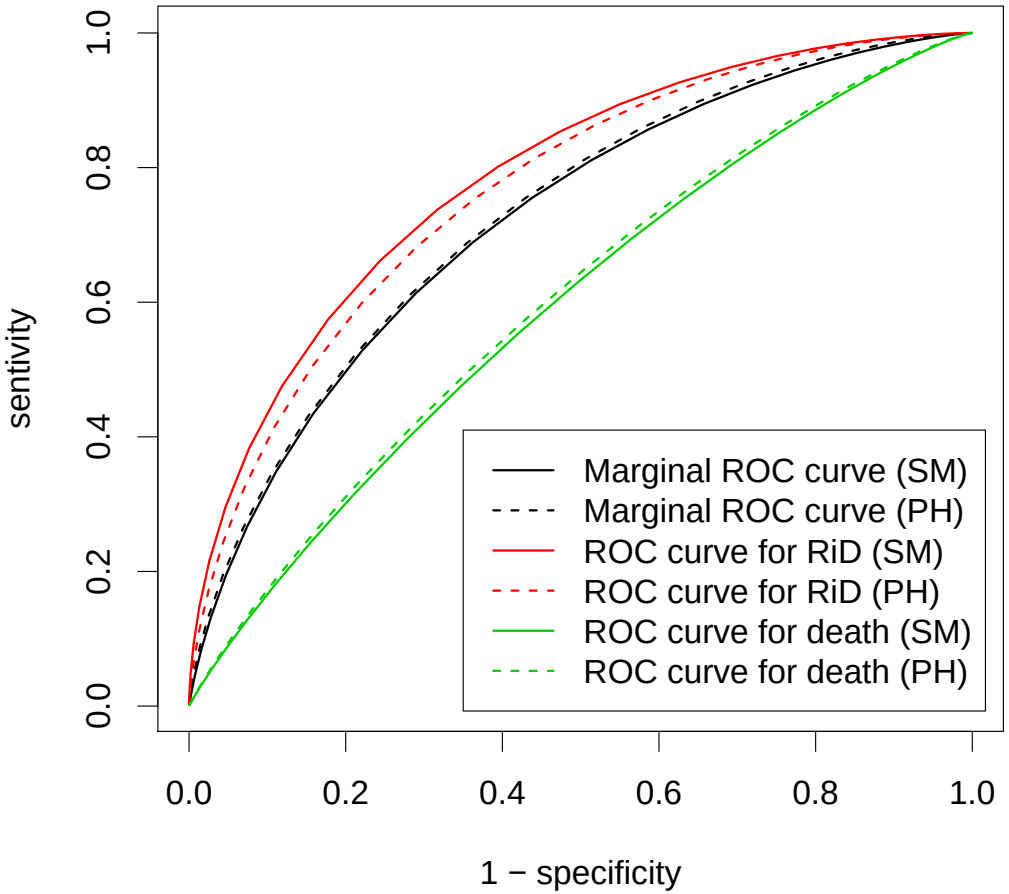
- Fonction de coût représente la somme des erreurs, issues du pronostic basé sur $c_1(\tau)$ et $c_2(\tau)$, pondérées selon leur importance en pratique :
  - ϕ_p = poids des faux positifs et ϕ_n = poids des faux négatifs
    - $\phi_n = 1 - \phi_p = P(T > \tau) = \sum_{i=1}^{2} P_i \int_{-\infty}^{\infty} S_i(\tau|z_i) g(z_i) dz_i$
  - ϕ_i poids des erreurs relatives à l'échec i $(i = 1, 2)$
    - $\phi_1 = AUC_1(\tau)$ et $\phi_2 = AUC_2(\tau)$

$$
\begin{aligned}
\mathcal{C}(\tau) \quad \propto \quad & \phi_p \{ \phi_1 \left(P_1 \int_{c_1(\tau)}^{\infty} S_1(\tau|z_1) g(z_1) dz_1 + P_2 \int_{\gamma c_1(\tau)}^{\infty} S_2(\tau|z_2) g(z_2) dz_2 \right) \\
+ \quad & \phi_2 \left(P_1 \int_{c_2(\tau)/\gamma}^{\infty} S_1(\tau|z_1) g(z_1) dz_1 + P_2 \int_{c_2(\tau)}^{\infty} S_2(\tau|z_2) g(z_2) dz_2 \right) \} \\
+ \quad & \phi_n \{ \sum_{i=1}^{2} \phi_i P_i \int_{-\infty}^{c_i(\tau)} F_i(\tau|z_i) g(z_i) dz_i \}
\end{aligned}
$$

Plan

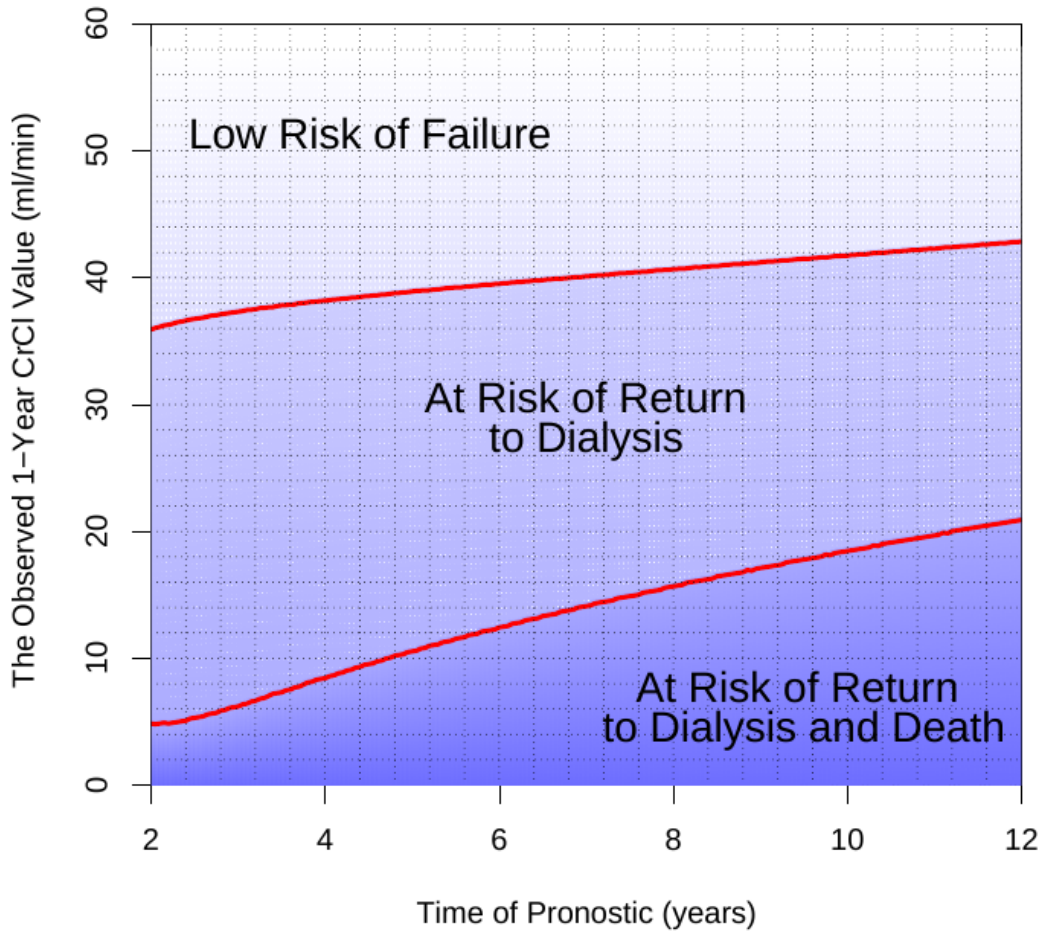
Résultats (1)

Courbes ROC à 10 ans



Résultats (2)

Estimation des seuils de décision selon le temps de pronostic



Plan

Plan

Problématique (1)

Objectif

- Identifier une signature basée sur quelques gènes pour pronostiquer la dégradation de la fonction rénale

Données disponibles

- 137 patients hyper-stables depuis au moins 5 ans.
- Prélèvements sanguins réalisés à l'inclusion.
- Données cliniques issues de DIVAT.
- Données génomiques : Puces ADN réalisées par TcLand.

Problèmes méthodologiques importants

- Overfitting
- Censures et troncature des trajectoires

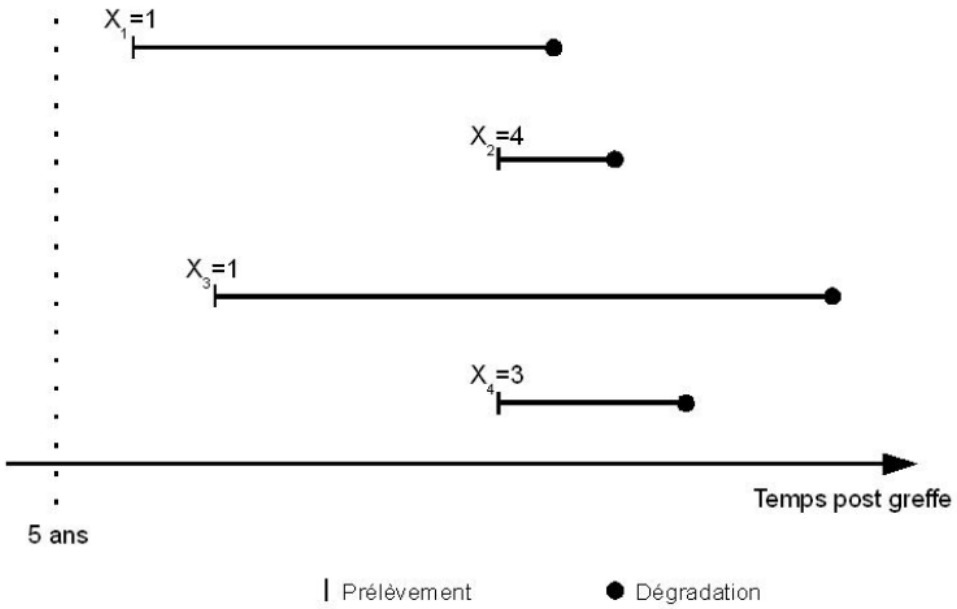
Problématique (2)

Overfitting

- Nombre de paramètres >>> Nombre d'observations
- Résultats toujours prometteurs, mais déception après validation...
- *Overfitted data* = *Overoptimistic results* (Ransohoff, Nature Rewiev Cancer, 2004)
- Solutions les plus courantes :
  - Division de l'échantillon en test/validation.
  - Ré-échantillonnage *leave and out*.
  - Ré-échantillonnage par bootstrap
  - Etc...

Problématique (4)

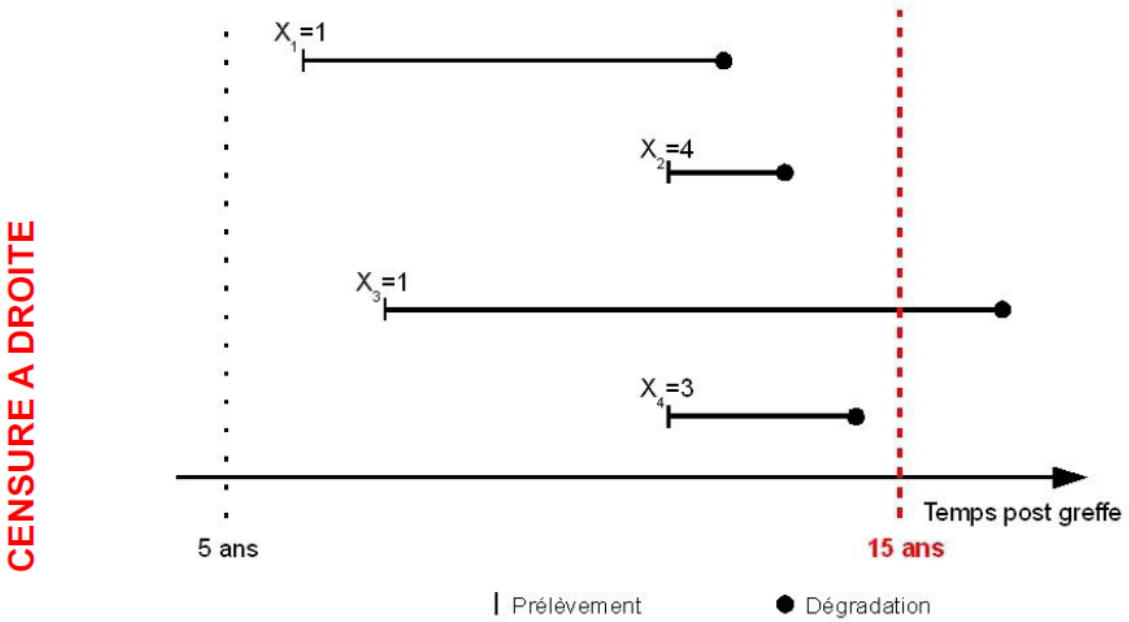
L'importance de prendre en compte la censure et la troncature



Hypothèse de travail : Expressions fortes sont associées à une dégradation plus rapide

Problématique (4)

L'importance de prendre en compte la censure et la troncature

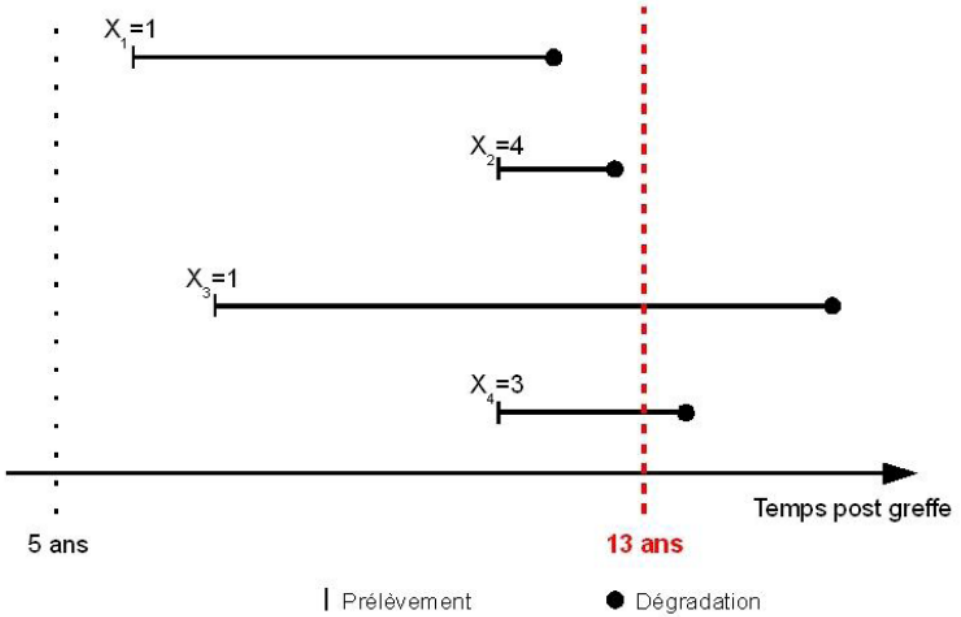


STABLES	DEGRADES
$X_3=1$	$X_1=1$
	$X_2=4$
	$X_4=3$
Moy = 1,0	Moy = 2,7

Problématique (5)

L'importance de prendre en compte la censure et la troncature

CENSURE A DROITE

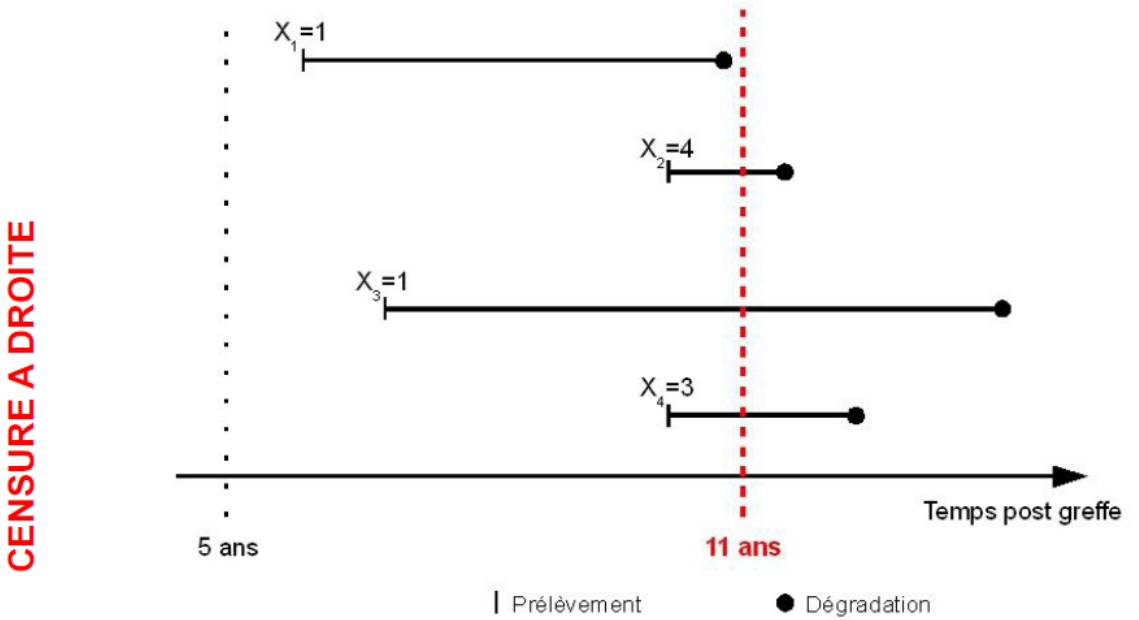


STABLES	DEGRADES
X_3=1	X_1=1
	X_2=4
	X_4=3
Moy = 1,0	Moy = 2,7

STABLES	DEGRADES
X_3=1	X_1=1
X_4=3	X_2=4
Moy = 2,0	Moy = 2,5

Problématique (6)

L'importance de prendre en compte la censure et la troncature

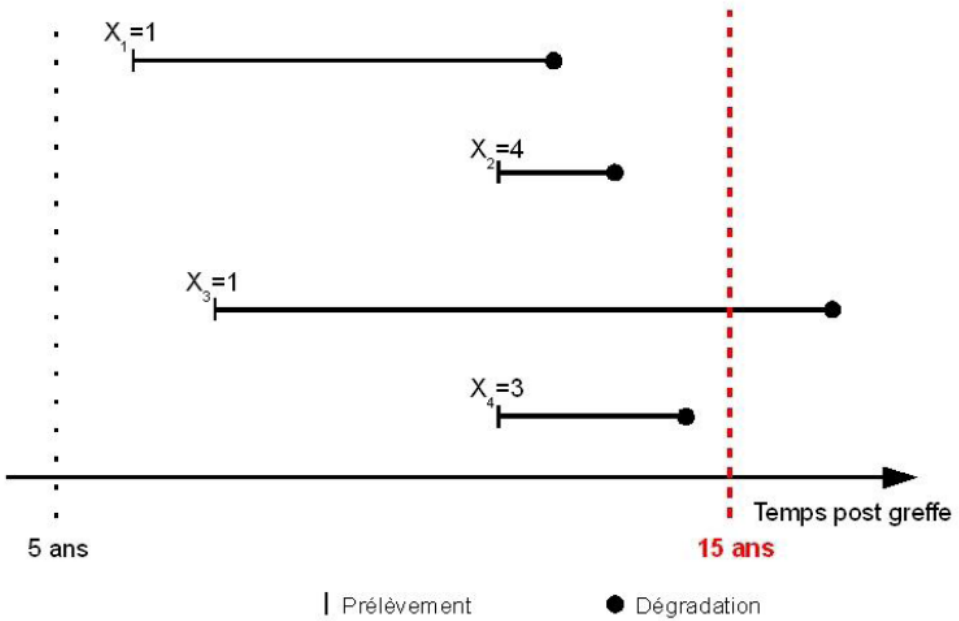


STABLES	DEGRADES
$X_3=1$	$X_1=1$
	$X_2=4$
	$X_4=3$
Moy = 1,0	Moy = 2,7

STABLES	DEGRADES
$X_2=4$	$X_1=1$
$X_3=1$	
$X_4=3$	
Moy = 2,7	Moy = 1,0

Problématique (7)

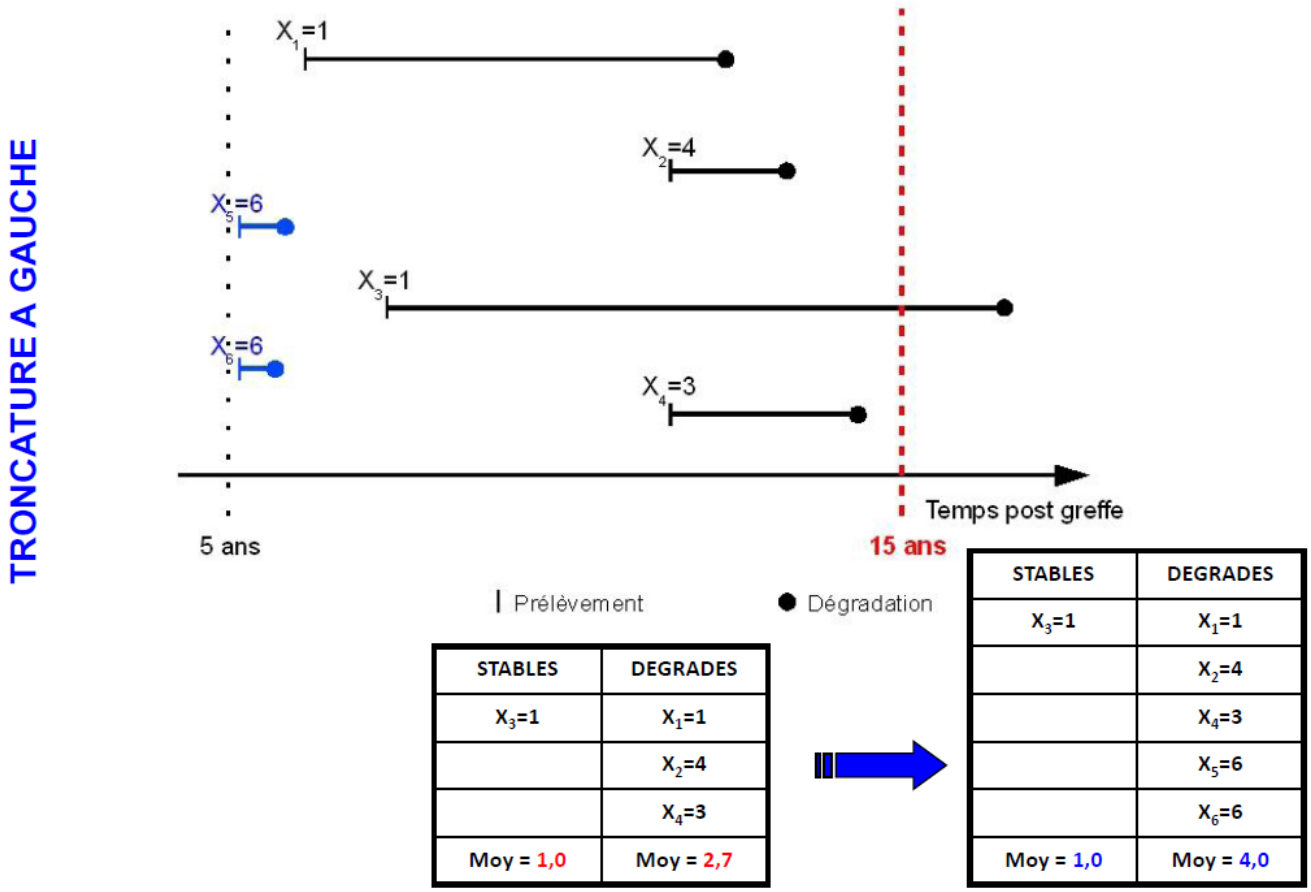
L'importance de prendre en compte la censure et la troncature



STABLES	DEGRADES
X_3=1	X_1=1
	X_2=4
	X_4=3
Moy = 1,0	Moy = 2,7

Problématique (8)

L'importance de prendre en compte la censure et la troncature



STABLES	DEGRADES
$X_3=1$	$X_1=1$
	$X_2=4$
	$X_4=3$
Moy = 1,0	Moy = 2,7

STABLES	DEGRADES
$X_3=1$	$X_1=1$
	$X_2=4$
	$X_4=3$
	$X_5=6$
	$X_6=6$
Moy = 1,0	Moy = 4,0

Problématique (9)

Données incomplètes (censures et troncatures)

- Méthodes basées sur Kaplan-Meier et LogRank
  - Une seule variable
  - Variables à catégoriser
    - Combien de classes ?
    - Nécessité de répéter les tests pour trouver le cut-off optimal.
- Modèles de régressions (Cox, AFT) : plusieurs variables quantitatives.
  - Hypothèses : loglinéarité et proportionnalité des risques
  - Dépend de l'incidence de l'événement (échantillon représentatif ?)
- Confusion entre corrélation et prédiction.

Problématique (10)

L'intérêt des courbes ROC dépendantes du temps

1. Proposer une méthode non-paramétrique
2. L'objectif est le pronostic et pas la corrélation
3. Obtenir une signature indépendante de l'incidence de l'événement
4. Adapter l'estimateur de Bootstrap .632+ pour contrôler l'*overfitting*
5. Prendre en compte plusieurs gènes et leurs valeurs quantitatives

Plan

Méthodes (1)

Définitions (1)

- N représente le nombre d'individus ($j = 1, ..., N$).
- On pose $E_j = \beta X_j$ la variable pronostique pour le sujet j.
  - β sont les coefficients de régressions associés aux gènes X.
  - e_j et x_j sont les observations de E et X pour l'individu j, avec $e_j = \beta x_j$.
- C_j est le temps de dernier suivi (censure à droite) du sujet j.
  - Si $C_j < T_j$ l'événement n'est pas observé, on sait juste qu'il arrive après C_j : $\Delta_j = 0$.
  - Si $C_j \geq T_j$, l'événement est exactement observé : $\Delta_j = 1$.
- Le temps de la dernière observation du sujet j est noté Y_{1j}, avec $Y_{1j} = min(T_j, C_j)$.
- On considère aussi le cas où T_j peut être tronqué à gauche.
- Y_{1j} est observé si $T_j \geq Y_{0j}$, où Y_{0j} est le temps de trocature à gauche.

Méthodes (2)

Définitions (2)

- τ est le temps de pronostic
  - $se_\tau(c) = P(E > c | T \geq \tau)$
  - $sp_\tau(c) = P(E \leq c | T > \tau)$
- $roc_\tau = \{1 - sp_\tau(c), se_\tau(c), c \in \Re\}$
- $se_\tau(c)$ et $sp_\tau(c)$ sont estimées non paramétriquement.

Méthodes (2)

Estimation de $se_\tau(c)$

$$se_\tau(c) = P(E > c | T \geq \tau)$$

$$\downarrow$$

$$se_\tau(c) = P(T \leq \tau, E > c)/P(T \leq \tau)$$

$$\downarrow$$

$$se_\tau(c) = \{P(E > c) - P(T > \tau, E > c)\}/P(T \leq \tau)$$

$$\downarrow$$

$$\hat{se}_\tau(c) = \{(1 - \hat{G}(c)) - \hat{S}(c, \tau)\}/\{1 - \hat{S}(-\infty, \tau)\}$$

- $\hat{S}(c, \tau)$ est obtenue par la méthode d'Akritas (1994, Annals of Stat.)

Méthodes (3)

Estimation de $P(T > \tau, E > c)$

$$\hat{S}(c, \tau) = N^{-1} \sum_{j=1}^{N} \hat{S}(\tau | E = e_j) \delta(e_j > c)$$

- $\hat{S}(\tau | E = c)$ est estimée par la méthode des plus proches voisins
- Posons K_{jl} l'indicateur que le patient l est elligible comme voisin du patient j
- $K_{jl} = \delta(|\hat{F}(e_j) - \hat{F}(e_l)| < \lambda)$, λ est la proportion de voisins

$$\hat{S}(\tau | E = c) = \prod_{y_{1j} \leq \tau} \{1 - d(y_{1j})/R(y_{1j})\}$$

- $d(y_{1j}) = \sum_{l=1}^{N} K_{jl} \delta(y_{1j} = y_{1l}) \Delta_l \rightarrow$ nb. de décès
- $R(y_{1j}) = \sum_{l=1}^{N} K_{jl} \delta(y_{1j} \leq y_{1l}) \rightarrow$ nb. à risque

Méthodes (4)

Estimation de $P(T > \tau, E > c)$

- La méthode peut être adaptée pour prendre en compte la troncature à gauche des données
- Le nombre d'individus à risque devient

$$R(y_{1j}) = \sum_{l=1}^{N} K_{jl} \delta(y_{0j} \leq y_{1l} \leq y_{1l})$$

Méthodes (5)

Estimation de $sp_\tau(c)$

$$sp_\tau(c) = P(E \leq c | T > \tau)$$

$$\downarrow$$

$$sp_\tau(c) = P(E \leq c, T > \tau)/P(T > \tau)$$

$$\downarrow$$

$$sp_\tau(c) = \{P(T > \tau) - P(E > c, T > \tau)\}/P(T > \tau)$$

$$\downarrow$$

$$sp_\tau(c) = 1 - P(E > c, T > \tau)/P(T > \tau)$$

$$\downarrow$$

$$\hat{sp}_\tau(c) = 1 - \hat{S}(c, \tau)/\hat{S}(-\infty, \tau)$$

Méthodes (6)

L'estimation apparente de la capacité pronostique, $\widehat{auc}_\tau$

- Il s'agit de l'aire sous la courbe ROC obtenue à partir de
  1. $\hat{se}_\tau(c)$ et $\hat{sp}_\tau(c)$ estimées sur l'échantillon initial
  2. Les coefficients de régression β qui maximisent cette aire

$$\hat{\beta} = \text{argmax}_\beta \widehat{auc}_\tau$$

Problème

- $\widehat{auc}_\tau$ surestime les capacités pronostiques de E (surtout quand le nombre de gènes inclus est grand)

Méthodes (7)

Solutions pour corriger l'overfitting (Molinaro et al., Bioinformatics, 2005)

- Division de l'échantillon (apprentissage + validation)
  - Perte d'information et de puissance sur des tailles d'échantillon souvent faibles
- Ré-échantillonnage par Bootstrap
  - Sous-estimation de l'overfitting
- Division + Bootstrap
  - Sur-estimation de l'overfitting
- 0.632 Bootstrap
  - Sous-estimation de l'overfitting
- 0.632+ Bootstrap (Efron et Tibshirani, JASA, 1997)

Méthodes (8)

L'alogrithme bootstrap 0.632+ (1)

B échantillons de bootstrap ($b = 1, ..., B$) de taille N avec remplacement.

- Les données inclues dans les échantillons de bootstrap permettent d'estimer les coefficients optimaux $\hat{\beta}^b$.
- Ce seuil est appliqué sur les individus non-inclus pour calculer les sensibilités et spécificités :
  - $\hat{se}_\tau^b(c)$ et $\hat{sp}_\tau^b(c)$
- Les estimations des sensibilités et spécificités sont alors obtenues par les moyennes :
  - $\overline{se}_\tau^b(c) = B^{-1} \sum_{b=1}^{B} \hat{se}_\tau^b(c)$
  - $\overline{sp}_\tau^b(c) = B^{-1} \sum_{b=1}^{B} \hat{sp}_\tau^b(c)$
- La courbe ROC correspondante sous-estime la capacité pronostique.

Méthodes (9)

L'algorithme bootstrap 0.632+ (2)

- Correction par la méthode de bootstrap .632 :

$$\begin{aligned} se_\tau^{.632}(c) &= .368\ \overline{se}_\tau^*(c) + .632\ \overline{se}_\tau^b(c) \\ sp_\tau^{.632}(c) &= .368\ \overline{sp}_\tau^*(c) + .632\ \overline{sp}_\tau^b(c) \end{aligned}$$

  - 0.632 = probabilité qu'un individu soit dans l'échantillon d'apprentissage
  - $\overline{se}_\tau^*(c)$ et $\overline{se}_\tau^*(c)$ sont les moyennes des sensibilités et des spécificités obtenus à partir des mêmes paramètres $\hat{\beta}^{\boldsymbol{b}}$, mais tous les individus sont inclus dans le calcul des sensibilités et spécificités
- La courbe ROC correspondante sur-estime la capacité pronostique.

Méthodes (10)

L'algorithme bootstrap 0.632+ (3)

- Correction par la méthode de bootstrap .632+.
- Les sensibilités et spécificités si aucune information apportée par E :
  - $\hat{\gamma}_{sp,\tau}(c) = 1 - \hat{\gamma}_{se,\tau}(c) = \hat{G}(c)$.
- Les taux correspondants de non-information :

$$\begin{aligned} \hat{r}_{se,\tau}(c) &= \{\overline{se}^b_\tau(c) - \overline{se}^*_\tau(c)\}/\{\hat{\gamma}_{se,\tau}(c) - \overline{se}^*_\tau(c)\} \\ \hat{r}_{sp,\tau}(c) &= \{\overline{sp}^b_\tau(c) - \overline{sp}^*_\tau(c)\}/\{\hat{\gamma}_{sp,\tau}(c) - \overline{sp}^*_\tau(c)\} \end{aligned}$$

- Valeurs corrigées à 0 si $\hat{r}_{.,\tau}(c) < 0$ et à 1 si $\hat{r}_{.,\tau}(c) > 1$.

$$\begin{aligned} se^{.632+}_\tau(c) &= (1 - \psi(\hat{r}_{se,\tau}(c)))\ \overline{se}^*_\tau(c) + \psi(\hat{r}_{se,\tau}(c))\ \overline{se}^b_\tau(c) \\ sp^{.632+}_\tau(c) &= (1 - \psi(\hat{r}_{sp,\tau}(c)))\ \overline{sp}^*_\tau(c) + \psi(\hat{r}_{sp,\tau}(c))\ \overline{sp}^b_\tau(c) \end{aligned}$$

- $\psi(x) = .632/(1 - .368x)$

Plan

Application à des données déjà publiées (1)

L'étude DLBCL (Rosenwald et al., New Engl. J. Med., 2002)

- Etude du lymphome
- 240 patients avec un prélèvement de tissu au moment du diagnostic
- Résultats des auteurs sur la survie à 5 ans (nb. gènes = 16)
  - 60% pour les patients GCB
  - 35% for ABC-like
  - 39% pour un troisième groupe.

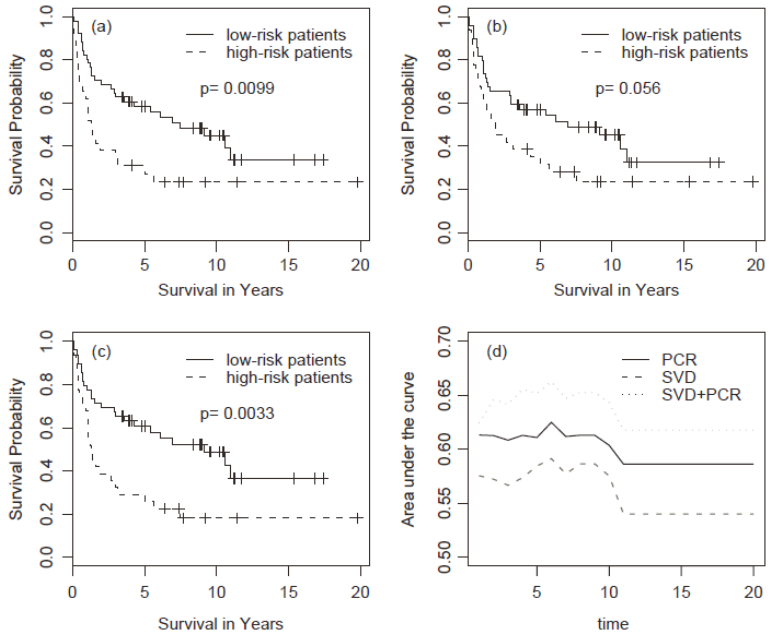
Problème

- Première étape de clustering n'est pas forcément adaptée aux données de survie
- Conclusions : marqueur pronostique ?
- Pas de prise en compte de l'overfitting dans les modèles de survie.
- Découpage Validation/Test

Application à des données déjà publiées (2)

Table 1. The results (P-values) for univariate Cox regression analysis for the first 10 PCR components, first 10 PC–PCR components and the top 10 PCs with the largest variances built using 7399, 1836 and 506 genes

	Number of genes used								
	7399			1836			506		
	PCR	PC	PC–PCR	PCR	PC	PC–PCR	PCR	PC	PC–PCR
1	8E−13	0.373	0	7E−13	2E−08	0	3E−13	2E−11	0
2	4E−09	0.505	3E−11	2E−06	2E−03	3E−13	1E−04	0.189	2E−12
3	9E−10	1E−05	2E−03	3E−11	0.565	2E−05	6E−09	0.481	2E−07
4	2E−05	0.907	0.233	4E−06	0.429	0.047	1E−05	0.014	4E−03
5	2E−03	0.942	0.223	8E−04	0.686	0.280	3E−04	0.846	0.086
6	4E−03	0.873	0.280	0.024	0.454	0.432	2E−03	0.784	0.234
7	0.031	0.784	0.829	0.013	0.358	0.551	0.027	0.486	0.190
8	0.078	0.553	0.824	0.121	0.064	0.716	0.056	0.951	0.546
9	0.298	0.421	0.652	0.499	0.376	0.839	0.241	0.124	0.516
10	0.251	0.029	0.899	0.356	0.365	0.518	0.143	0.112	0.946



Application à des données déjà publiées (3)

Problème (Li and Gui, Bioinformatics, 2004)

- Nombre de gènes : utilité limitée de la signature en pratique
- Capacités pronostiques tout juste acceptables
- Pas de prise en compte de l'overfitting
- Modèle de régression

Papier récent (Schumacher, Bioinformtics, 2007)

- Propose un algorithme de bootstrap .632+ pour le modèle de Li and Gui (2004)

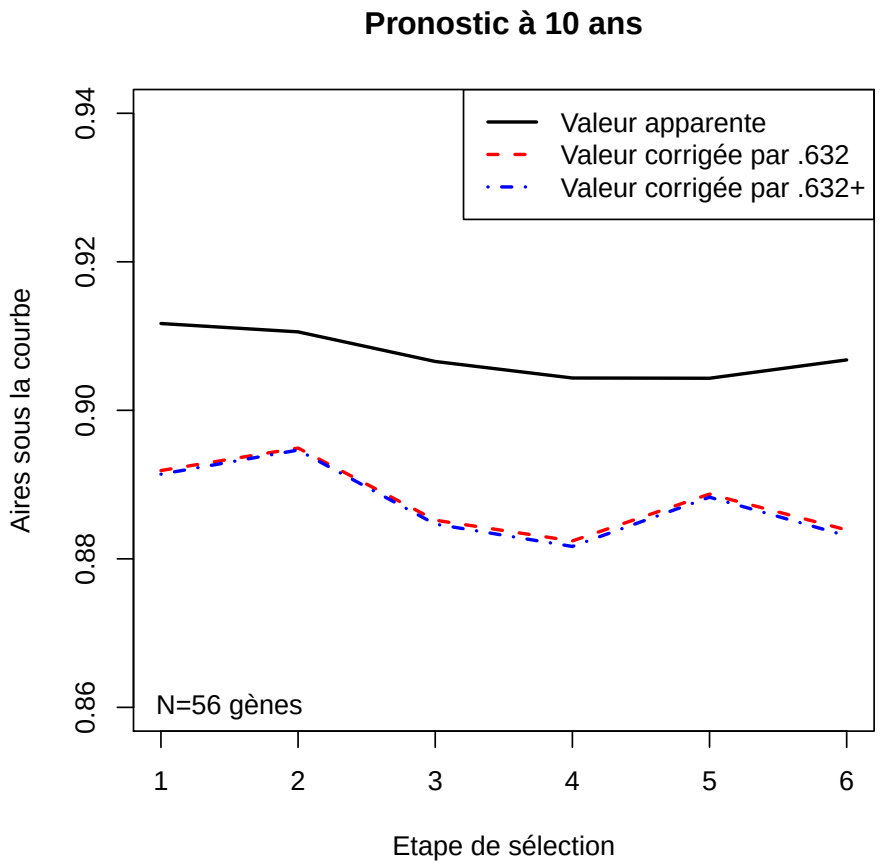
Application à des données déjà publiées (3)

Problème

- Nombre de gènes : utilité limitée de la signature en pratique
- Capacités pronostiques tout juste acceptable
- Pas de prise en compte de l'overfitting
- Découpage Validation/Test

Application à des données déjà publiées (4)

Résultats **péliminaires**



Plan

Conclusions (1)

Synthèse

- Méthode adaptée pour évaluer les performances d'un marqueur à pronostiquer un événement.
  - Un biomarqueur seul.
  - Un score composite.
- Elle peut être généralisée au pronostic de plusieurs événements.
- Elle peut être développée pour sélectionner des gènes ou protéines à partir de criblage à haut débit.

Conclusions (2)

Notions méthodologiques importantes

- La corrélation forte d'une variable avec la survie ne veut pas dire que cette dernière est un bon marqueur pronostique.
- Ne pas utiliser les courbes ROC classiques pour des données dépendantes du temps (données incomplètes : censure/troncature).
- Lorsqu'il y a plus de deux classes (que ce soit en pronostic ou en diagnostic), il peut exister des méthodes plus adaptées que le calcul de 3 courbes ROC différentes.
- Les données de puces sont souvent présentées comme des échantillons issus de plusieurs groupes. Attention au protocole lorsqu'il s'agit de pathologies chroniques.
  - A quel moment le prélèvement a été fait (troncature, censure) ?
  - Diagnostic / Pronostic