

# Modèle prédictif évolutif de survenue de décès dans les 7 jours chez les patients atteints de sepsis sévère

Adrien FRANCAIS

MASTER 2 MASSS (Modélisation et Apprentissage de la Statistique en Sciences Sociales)

Département de statistique de L'UPMF

Stage de fin d'études en Statistique



# Plan

- 1 Introduction
  - **Structure d'accueil**
  - Description du sujet
- 2 Population et statistiques descriptives
- 3 Analyse des données corrélées
  - Introduction
  - Détail des différentes approches
  - Bilan
- 4 Validation du modèle
  - Méthodes de validation
  - Résultats obtenus
  - Utilisation pratique du modèle sous un tableur
- 5 Conclusion

# Structure d'accueil

## L'Institut Albert Bonniot

Institut Fédératif de Recherche sur le cancer du poumon  
Unité Inserm U823 créée en 2007

## L'Equipe 11

"Epidémiologie des cancers et des affections graves"  
Axes de recherche : réanimation, asthme et cancer

## L'association OUTCOMEREA

Association loi 1901 créée en 1999  
Promouvoir et développer la recherche et l'enseignement de la réanimation  
Objectif : amélioration du pronostic en réanimation

# Plan

- 1 Introduction
  - Structure d'accueil
  - **Description du sujet**
- 2 Population et statistiques descriptives
- 3 Analyse des données corrélées
  - Introduction
  - Détail des différentes approches
  - Bilan
- 4 Validation du modèle
  - Méthodes de validation
  - Résultats obtenus
  - Utilisation pratique du modèle sous un tableur
- 5 Conclusion

# Description du sujet

- **Sepsis sévère** : maladie grave assez fréquente en réanimation
- Facteurs pronostiques connus du décès : gravité, maladie chronique, germes en cause, site infecté, traitement en cours et le moment où il a été mis en place
- **But ultime de la réanimation** : suppléer les organes défaillants
- Besoin de déterminer les probabilités de décès afin d'informer les familles et guider les traitements à mettre en place
  
- Trois types de sepsis sévères : à *l'admission*, *nosocomial précoce* et *nosocomial tardif*
- **Objectif** : créer un modèle simple utilisable par les médecins qui permet de prédire la mortalité dans les 7 jours suivant l'infection, à travers les variables à l'admission et de gravité journalières
- Problème des **données corrélées** : un patient peut avoir plusieurs épisodes de sepsis sévères différents

# Tableau de matière

- 1 Introduction
  - Structure d'accueil
  - Description du sujet
- 2 Population et statistiques descriptives
- 3 Analyse des données corrélées
  - Introduction
  - Détail des différentes approches
  - Bilan
- 4 Validation du modèle
  - Méthodes de validation
  - Résultats obtenus
  - Utilisation pratique du modèle sous un tableur
- 5 Conclusion

# Population étudiée

## Sélection de la population

- Base de construction : 1763 épisodes de sepsis pour 1482 patients
- Base de validation : 1030 épisodes de sepsis (Procédure SURVEYSELECT)

## Description de la population

- 63% d'hommes, 13% de pneumonie, un score de dysfonction d'organe élevé
- Variables testées : sexe, symptôme, type du malade, diabète, diagnostic principal, immunodépression, maladie chronique
- Modèle de régression logistique informatif de décès
- Gravité du patient, jour du sepsis, site d'infection

# Plan

- 1 Introduction
  - Structure d'accueil
  - Description du sujet
- 2 Population et statistiques descriptives
- 3 Analyse des données corrélées**
  - Introduction**
  - Détail des différentes approches
  - Bilan
- 4 Validation du modèle
  - Méthodes de validation
  - Résultats obtenus
  - Utilisation pratique du modèle sous un tableur
- 5 Conclusion

# Données corrélées

- Pas d'indépendance entre les observations d'un même patient
- Notion de *cluster* de taille variable

## Analyse de 5 méthodes

- 1 Régression logistique simple
- 2 Modèle linéaire généralisé avec la PROC GENMOD
- 3 Modèle logistique conditionnel avec PROC LOGISTIC
- 4 Modèle mixte avec PROC NL MIXED
- 5 Méthode hybride qui combine les effets fixes et les effets aléatoires

# Plan

- 1 Introduction
  - Structure d'accueil
  - Description du sujet
- 2 Population et statistiques descriptives
- 3 Analyse des données corrélées**
  - Introduction
  - Détail des différentes approches**
  - Bilan
- 4 Validation du modèle
  - Méthodes de validation
  - Résultats obtenus
  - Utilisation pratique du modèle sous un tableur
- 5 Conclusion

# Régression logistique simple

$$Y_i = \beta_0 + \beta_1 x_i' + \epsilon_i \quad (1)$$

avec  $i$  représentant l'épisode de sepsis.  
ce qui équivaut en terme de probabilité

$$p_i = Pr(Y_i = 1) = \frac{\exp(\beta x_i')}{1 + \exp(\beta x_i')} \quad (2)$$

avec  $p_i$  la probabilité du patient de décéder dans les 7 jours suivant l'épisode de sepsis,  $g$  la fonction de lien et  $\mu_i$  la moyenne des réponses au prédicteur linéaire  $\beta x_i'$ .

## Résultats

- Estimateurs consistants et non biaisés
- Tests de Wald surévalués
- Tenir compte de la corrélation intra-patient

# Modèle linéaire généralisé avec la PROC GENMOD

$$Y_{ij} = \beta_0 + \beta_1 \mathbf{x}'_{ij} + \epsilon_{ij} \quad (3)$$

avec  $i$  représentant le patient et  $j$  les différentes mesures intra-patient.

$$g(\mu_{ij}) = \log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \beta \mathbf{x}'_{ij} \quad (4)$$

ce qui équivaut en terme de probabilité

$$p_{ij} = Pr(Y_{ij} = 1) = \frac{\exp(\beta \mathbf{x}'_{ij})}{1 + \exp(\beta \mathbf{x}'_{ij})} \quad (5)$$

avec  $p_{ij}$  la probabilité du patient  $i$  de décéder dans les 7 jours suivant le  $j^{ieme}$  épisode de sepsis,  $g$  la fonction de lien et  $\mu_{ij}$  la moyenne des réponses au prédicteur linéaire  $\beta \mathbf{x}'_{ij}$ .

# Modèle linéaire généralisé avec la PROC GENMOD

- Méthode de maximum de vraisemblance
- Utilisation des matrices de covariance et de corrélation
- Résidus de Pearson :

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}} \quad (6)$$

avec  $y_{ij}$  l'événement que l'on modélise et  $\mu_{ij}$  le vecteur des moyennes correspondantes.

*Choix entre 2 structures de corrélation*

## 1 Exchangeable :

$$\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \alpha & j \neq k \end{cases}$$

## 2 Unstructured :

$$\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \alpha_{jk} & j \neq k \end{cases}$$

# Modèle logistique conditionnel avec PROC LOGISTIC

$$\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \alpha_i + \beta \mathbf{x}_{ij} \quad (7)$$

- Biais résultant de l'omission de variables explicatives
- $\alpha_i$  représente toutes les différences parmi les individus
- $\alpha_i$  constante fixe unique par patient

## Résultats

- 1 Algorithme n'utilisant que les strates informatives
- 2 Pas d'estimation pour les variables constantes au cours du temps
- 3 Pas de convergence

# Modèle mixte avec PROC NL MIXED

$$\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \alpha_i + \beta x_{ij} \quad (8)$$

- $\alpha_i$  : terme aléatoire alloué à chaque cluster
- $\alpha_i$  suit une loi normale de moyenne nulle et de variance  $S^2$
- Préciser les valeurs de départ pour la convergence de l'algorithme
- Chaque patient aura un risque de base de décès

## Résultats

- 1 Convergence lente
- 2 Estimation beaucoup plus grande des paramètres
- 3 Forte variance du terme  $\alpha_i$
- 4 Interprétation spécifique au patient

# Méthode hybride

- Vertus de la régression logistique conditionnelle et celles du MLG
- Décomposer les variables dépendantes du temps à travers leurs variations au sein du cluster
- Modèle avec les moyennes et la déviance de ces variables

Tester si les coefficients des variables de déviation sont les mêmes que les moyennes correspondantes. L'hypothèse nulle testée est :

$H_0$  : toutes les différences des paramètres dépendants du temps sont égales à 0, ce qui équivaut à un modèle à effets aléatoires

VS

$H_1$  : au moins une différence de paramètre est significative, ce qui équivaut à un modèle à effets fixes

C'est une approche de test entre effets fixes vs effets aléatoires.

- Approche réalisée avec modèle mixte et MLG
- Modèle à effets aléatoires rejeté au profit d'un modèle à effets fixes

<b>Contrastes</b>				
<b>Libellé</b>	<b>Degrés de lib. num.</b>	<b>Den DDL</b>	<b>Valeur F</b>	<b>Pr &gt; F</b>
<i>Test of fixed vs. random</i>	4	1457	12.78	<.0001

# Plan

- 1 Introduction
  - Structure d'accueil
  - Description du sujet
- 2 Population et statistiques descriptives
- 3 Analyse des données corrélées**
  - Introduction
  - Détail des différentes approches
  - Bilan**
- 4 Validation du modèle
  - Méthodes de validation
  - Résultats obtenus
  - Utilisation pratique du modèle sous un tableur
- 5 Conclusion

# Bilan

Paramètre	régression logistique simple		régression logistique généralisée à effets fixes		régression logistique à effets aléatoires	
	estimation	erreur standard	estimation	erreur standard	estimation	erreur standard
Intercept	-5.3957	0.3121	-5.3650	0.3140	-9.2429	1.0765
LOD le jour du sepsis	0.2590	0.0306	0.2409	0.0301	1.1527	0.3719
Jour du sepsis : 3-7ème	0.7601	0.2163	0.6549	0.2095	1.5108	0.3974
Jour du sepsis : 8-28ème	0.7141	0.1999	0.9145	0.1921	0.4207	0.06663
Etat de choc septique	0.4745	0.1729	0.3987	0.1655	0.6895	0.2821
Administration d'inotropes forte	0.7221	0.1663	0.6430	0.1655	1.1372	0.3091
SAPS à l'admission	0.0113	0.0052	0.0146	0.0053	0.02293	0.009688
Exactement une maladie chronique	0.5431	0.1710	0.5938	0.1773	1.0200	0.3208
Au moins deux maladies chroniques	0.7924	0.2458	0.7670	0.2573	1.3172	0.4630
Scale ou s2*	1.0000	0.0000	1.0000	.	6.3773 *	1.9799

- **Ecart-types robustes** proche du modèle linéaire généralisé (MLG) mais pas bonne dans la théorie
- **MLG** : corrélation intra-cluster, estimateurs fixes, tests significatifs, modèle reproductible
- **Régression linéaire mixte** : estimateurs différents à cause de la variabilité forte du terme aléatoire
- Validation effectuée grâce à la **discrimination et la calibration** pour chaque type de sepsis (base de construction et validation)

# Plan

- 1 Introduction
  - Structure d'accueil
  - Description du sujet
- 2 Population et statistiques descriptives
- 3 Analyse des données corrélées
  - Introduction
  - Détail des différentes approches
  - Bilan
- 4 Validation du modèle**
  - Méthodes de validation**
  - Résultats obtenus
  - Utilisation pratique du modèle sous un tableur
- 5 Conclusion

# Calibration

- Dans quelle mesure le risque prédit est-il proche du risque réel ?
- Test du Chi2 de Hosmer-Lemeshow

## Détail du calcul

- 1 Classer les probabilités de survenue dans l'ordre croissant
- 2 Créer 10 groupes d'effectifs de taille proche
- 3 Sommer les probabilités de décès dans chaque groupe donne le nombre de décès prédits
- 4 Test du Chi2 entre les décès prédits et observés
- 5 Comparer cette valeur au Chi2 de degré 8

# Discrimination

- Permet de classifier correctement les individus
- Sensibilité et Spécificité

## Sensibilité d'un test

- 1 Probabilité que le signe soit présent si le sujet est atteint de la maladie considérée
- 2 Capacité à prédire la mortalité pour les patients effectivement décédés

## Spécificité d'un test

- 1 Probabilité que le signe soit absent si le sujet n'est pas atteint de la maladie
- 2 Capacité à prédire la survie pour les patients ayant effectivement survécus

# Courbe ROC et Aire sous la Courbe (AUC)

## Objectif :

Modèle très sensible (ne pas laisser "passer" de décès) et spécifique (ne pas faire croire à un décès et provoquer des examens complémentaires inutiles).

- Obtenir les plus fortes valeurs pour ces 2 paramètres (variation en sens inverse)
- Courbe ROC (Receiver Operating Characteristics) et prendre le maximum
- Critère statistique : AUC (Area Under Curve)

$$AUC = \sum_i 0.5 * (1MSPEC_{i+1} - 1MSPEC_i) (SENSIT_{i+1} + SENSIT_i)$$

- Si AUC proche de 0.5, comme une pièce de monnaie !!!
- Si AUC supérieure à 0.8, le modèle discrimine bien les patients

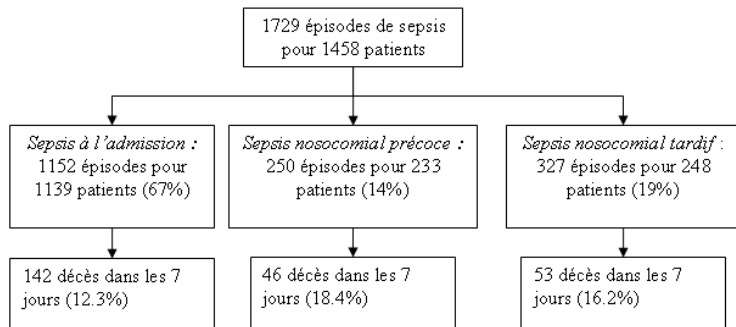
# Plan

- 1 Introduction
  - Structure d'accueil
  - Description du sujet
- 2 Population et statistiques descriptives
- 3 Analyse des données corrélées
  - Introduction
  - Détail des différentes approches
  - Bilan
- 4 Validation du modèle**
  - Méthodes de validation
  - Résultats obtenus**
  - Utilisation pratique du modèle sous un tableur
- 5 Conclusion

# Résultats obtenus

Calibration et discrimination selon chaque type de sepsis sévère :

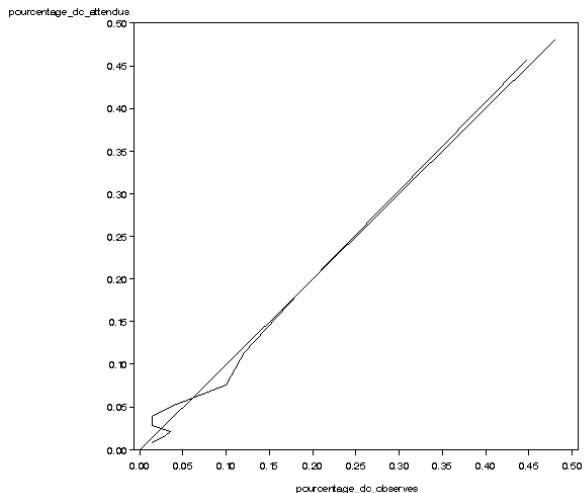
- sepsis à l'admission
- sepsis nosocomial précoce
- sepsis nosocomial tardif



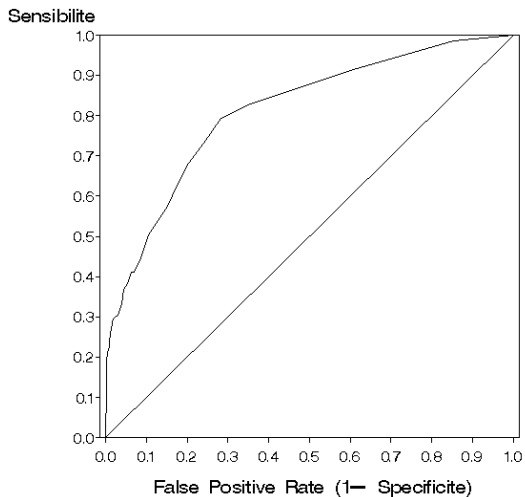
# Application sur les sepsis à l'admission

<i>groupe</i>	<i>effectif</i>	<i>décès observé</i>	<i>décès attendu</i>	<i>vivant observé</i>	<i>vivant attendu</i>	<i>total</i>
1	112	2	1,2	110	110,8	0,6
2	113	4	2,1	109	110,9	2,5
3	113	5	3,0	108	110,0	3,9
4	113	2	4,1	111	108,9	5,0
5	114	2	5,4	112	108,6	7,3
6	112	6	7,5	106	104,5	7,6
7	113	14	10,6	99	102,4	8,8
8	113	17	16,0	96	97,0	8,9
9	113	26	25,8	87	87,2	8,9
10	123	63	64,4	60	58,6	<b>8,9</b>

# Application sur les sepsis à l'admission



# Application sur les sepsis à l'admission



AUC : 0.825

# Bilan de la validation

Bonnes calibrations et discriminations pour les 3 types de sepsis

Application sur la base de validation

type de sepsis	effectif	effectif vivant	effectif mort	Chi2HL	AUC
<i>à l'admission</i>	709	620	89	16,5	0,783
<i>nosocomial précoce</i>	111	96	15	14,6	0,70
<i>nosocomial tardif</i>	136	118	18	4,9	0,78

Bons résultats

dans l'ensemble mais discrimination assez moyenne

Intéressant de valider sur une base externe ou avec le bootstrap

- Presque "trop" simple car absence de variables supposées prépondérantes (antibiothérapie précoce et efficace, type et site d'infection)
- Variables "logiques" dans le modèle connues dans la littérature du sepsis

# Plan

- 1 Introduction
  - Structure d'accueil
  - Description du sujet
- 2 Population et statistiques descriptives
- 3 Analyse des données corrélées
  - Introduction
  - Détail des différentes approches
  - Bilan
- 4 Validation du modèle**
  - Méthodes de validation
  - Résultats obtenus
  - **Utilisation pratique du modèle sous un tableur**
- 5 Conclusion

# Utilisation pratique du modèle sous un tableur

**Programme permettant de prédire le décès dans les 7 jours suivant l'infection au sepsis sévère**

**Prédiction établie à partir de variables à l'entrée en réanimation et le jour du sepsis**

<i>Score de gravité LOD le jour du sepsis</i>	20
<i>Jour d'acquisition du sepsis en 3 classes</i>	1
<i>Patient en choc septique (0 ou 1)</i>	1
<i>Type d'inotrope en 2 classes</i>	4
<i>SAFS à l'admission</i>	40
<i>Nombre de maladies chroniques du patient en 3 classes</i>	3
calcul du logit	4,01
<b>calcul de la probabilité associée</b>	<b>0,98</b>

Cette prédiction tient compte de la corrélation des observations quand on a plusieurs épisodes de sepsis par patient.

En fonction de la probabilité prédite, administrer les soins adéquats

Cette probabilité prédite est une aide pour le médecin mais il ne faut pas se tenir qu'à cette probabilité.

Il ne faut pas "délaisser" un patient si sa probabilité de mourir est faible. C'est juste un "plus" pour le médecin.

# Conclusion

- Modèle reproductible et sensé (gravité du patient, dysfonctions d'organe, maladies chroniques et le jour de la contraction)
- Modèle linéaire généralisé pour traiter les données corrélées
- Utilisation future par les médecins
- Ecriture future d'un article
  
- Nouvelles techniques et acquisition d'expérience
- Utilisation de SAS et R
- Bonne autonomie et bon encadrement
- Statisticien depuis septembre 2007

# C'est fini !

Merci de votre attention !