Régression logistique

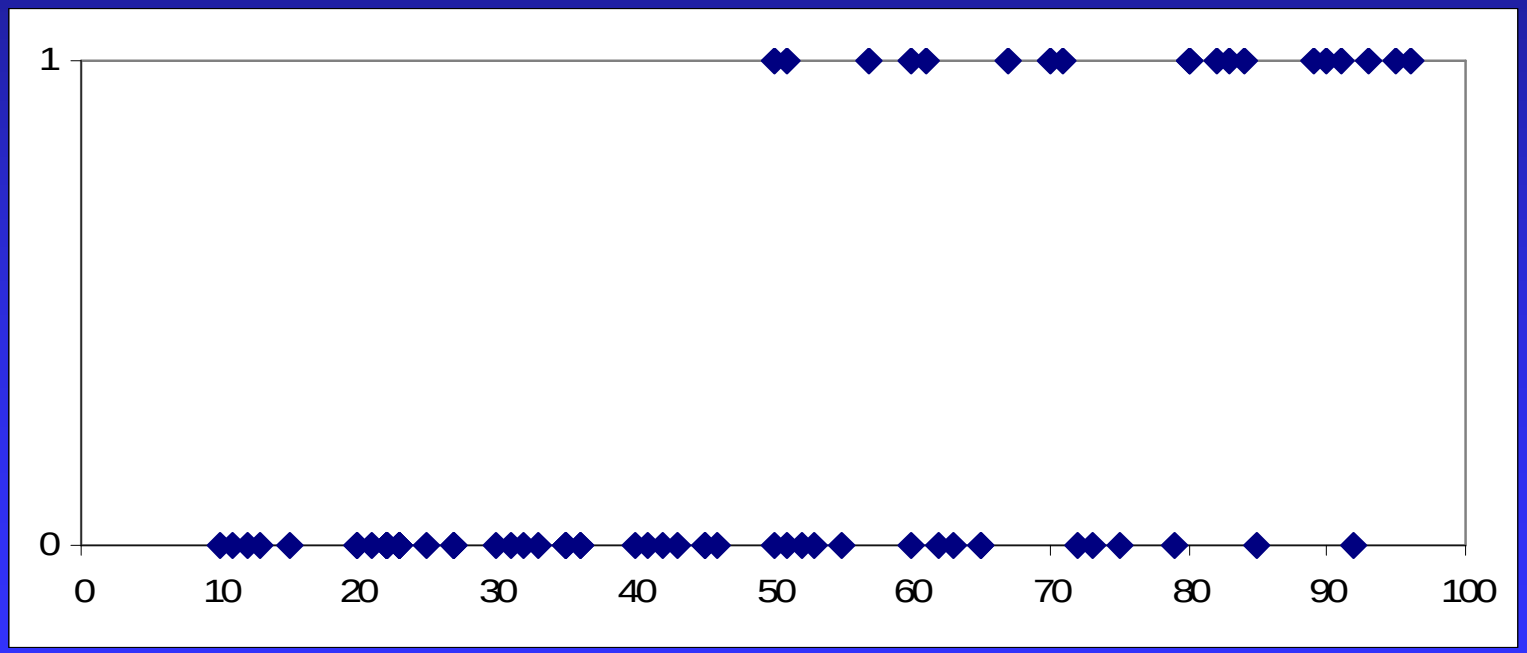
Dr Cécile Couchoud

But de la régression logistique

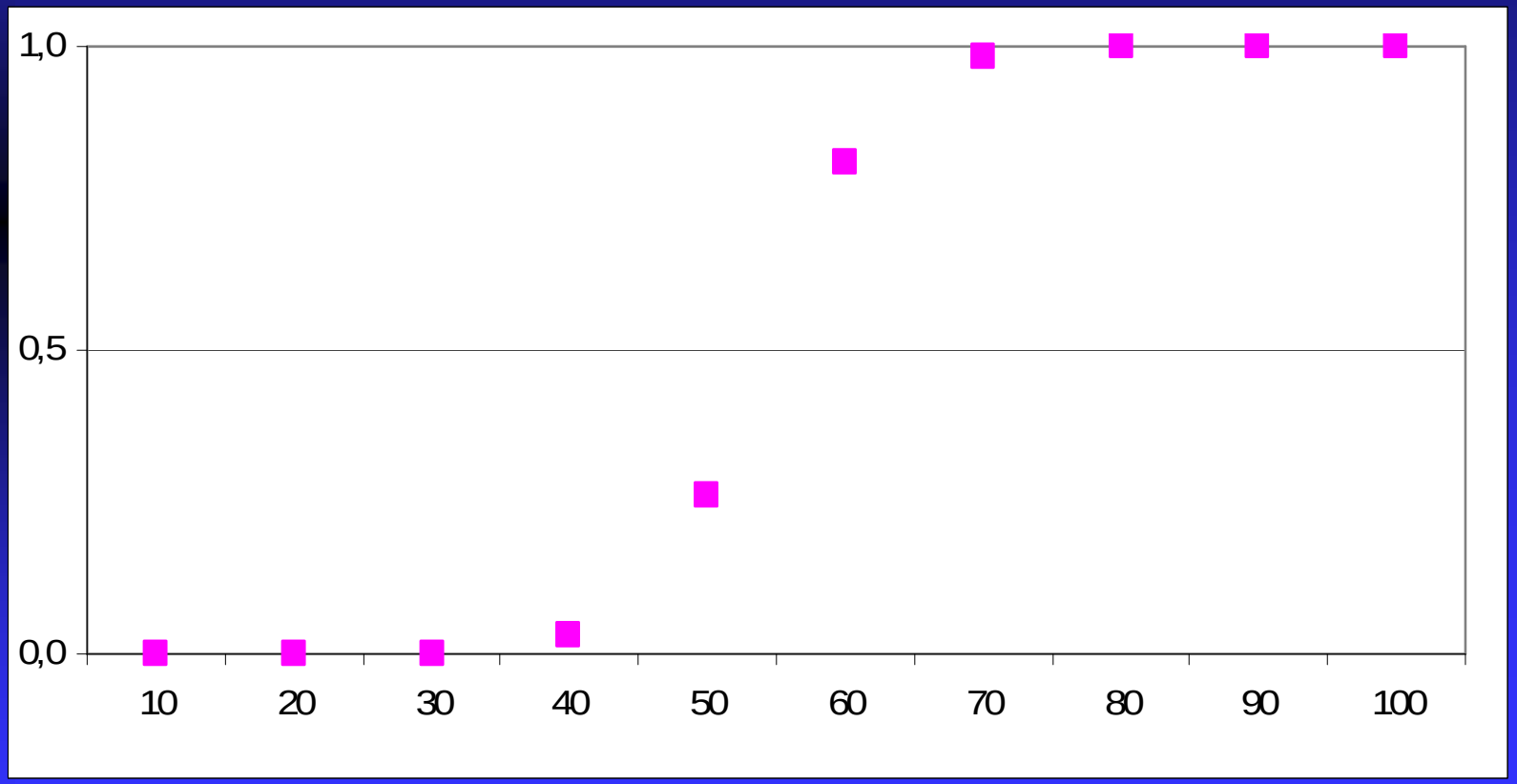
- Modélisation d'une variable dépendante (que l'on veut prédire ou expliquer) qualitative dichotomique : sains/malades, exposés/non exposés…
- Sa relation avec des variables explicatives : association ? prévisions ? causalité ?
- Si variables dépendantes à plusieurs modalités : régression polytomique

Exemple : cancer de la prostate selon l'âge

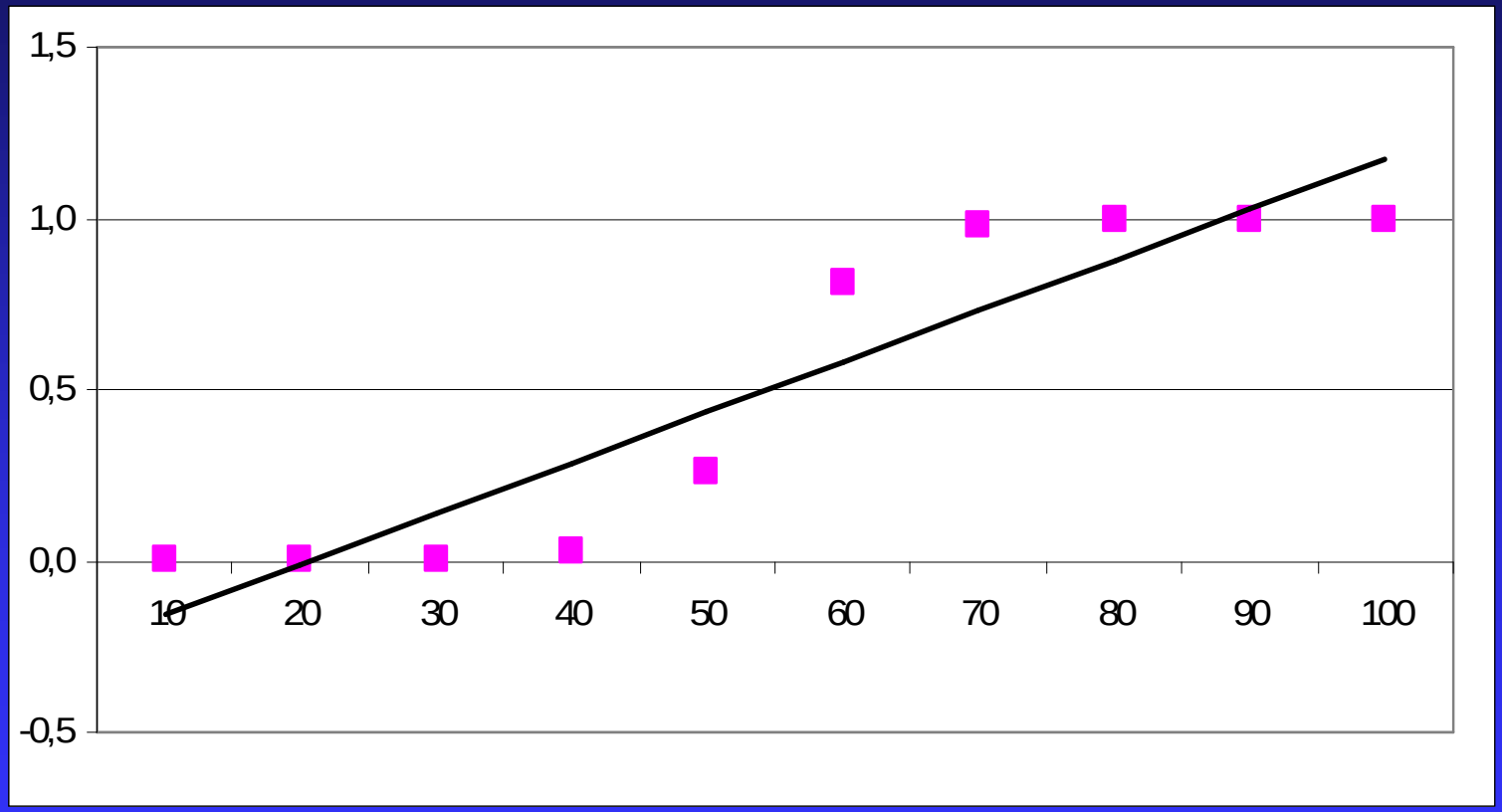
Représentation des patients



Exemple : probabilité cumulée d'avoir un cancer de la prostate selon l'âge

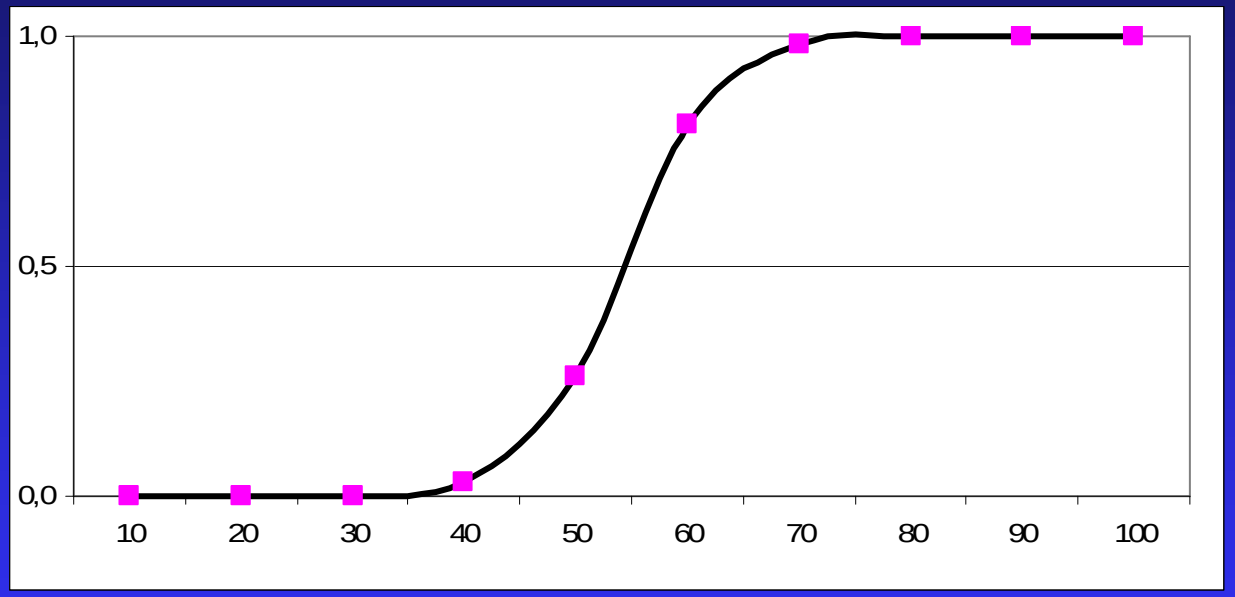


Modélisation linéaire ???



Attention ! Probabilité : par définition comprise entre 0 et 1

Relation sigmoïdale



Pour modéliser une probabilité, on utilise une fonction de répartition (ou fonction cumulative) d'une loi

Pourquoi « régression » ?

- Régression de « y » par rapport à « x »
  - Espérance mathématique de y conditionnelle à x : $E(y \mid x)$
  - Ex : fonction « doublement »
    - $E(y \mid x = 3) = 6$
- Lien entre espérance conditionnelle et probabilité

$$y = \begin{cases} 1 \text{ (ex : « malade »)} & \text{avec probabilité} = p \\ 0 \text{ (ex : « sain »)} & \text{avec probabilité} = 1 - p \end{cases}$$

  - $E(y \mid x) = 1 \times p + 0 \times (1-p) = p$

Pourquoi « logistique » ?

- Choix du modèle Logit = fonction de répartition de la loi logistique

$$P(M^{+}/X=x) = f(x) = \frac{1}{1+\exp^{(-\alpha+\beta x)}}$$

- *On aurait pu choisir le modèle Probit (fonction répartition de la loi Normale)*

- Permet le calcul de l'OR

Transformation logit

Par définition, $\text{Logit } P = \text{Ln} \dfrac{P}{1 - P}$

Pour $P = P(M^{+}/X=1)$, $\text{Logit } P = \alpha+\beta$

$$OR = \frac{P1 / (1 - P1)}{P0 / (1 - P0)}$$

$$\text{Ln OR} = \text{Logit } P1 - \text{Logit } P0 = \alpha+\beta - \alpha = \beta$$

Exemple : FR d'angor

- Logit [P(angor=1|diabète)] = $\beta 0+\beta 1$ diabète
  - Logit [P(angor =1| diabète =1)] = $\beta 0+\beta 1$
  - Logit [P(angor =1| diabète =0)] = $\beta 0$
- Odds = p/1-p $\Rightarrow$ OR = $p_1/(1-p_1) / p_0/(1-p_0)$

$$OR = \frac{e^{\text{Logit [P(diabète =1|sexe=1)]}}}{e^{\text{Logit [P(diabète =1|sexe=0)]}}} = \frac{e^{\beta 0+\beta 1}}{e^{\beta 0}}$$

- OR diabète = $e^{\beta 1}$

Voir cours sur mesures d'association

Likelihood Ratio	43,6700	1	0,0000

Previous Dataset Results Library

LOGISTIC angor = diab

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
diab	5,2226	3,2152	8,4833	1,6530	0,2475	6,6785	0,0000
CONSTANT	*	*	*	-3,0744	0,1781	-17,2664	0,0000

Convergence:	Converged
Iterations:	6
Final -2*Log-Likelihood:	483,2158
Cases included:	963

Test	Statistic	D.F.	P-Value
Score	52,6815	1	0,0000
Likelihood Ratio	43,6700	1	0,0000

Previous Dataset Results Library

LOGISTIC angor = diab

Exp(1.653) = 5,2226
Les diabétiques ont un risque 5 fois plus élevé d'avoir un angor

Au final

- Logit $[P(y=1|x_1,x_2 \ldots x_n)]$
$= \beta_0+\beta 1x_1 \ldots + \beta_n x_n$

- Estimation de $\beta 0,\beta_1, \beta_2 \ldots.,$
- Exponentielle de $\beta 0,\beta_1, \beta_2 \ldots., \beta_n$

  - Odds Ratio de chaque variable explicative $\beta 0,\beta_1, \beta_2 \ldots., \beta_n$

Score	21,0725	1	0,0000
Likelihood Ratio	22,5472	1	0,0000

Previous Dataset Results Library

LOGISTIC angor = diab SEXE

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
diab	5,1131	3,1253	8,3650	1,6318	0,2512	6,4970	0,0000
SEXE (2/1)	0,2918	0,1636	0,5207	-1,2316	0,2954	-4,1693	0,0000
CONSTANT	*	*	*	-2,6454	0,1915	-13,8112	0,0000

Convergence:	Converged
Iterations:	6
Final -2*Log-Likelihood:	462,8079
Cases included:	963

Interprétation ?

Test	Statistic	D.F.	P-Value
Score	70,6242	2	0,0000
Likelihood Ratio	64,0780	2	0,0000

Estimation du modèle

- Méthode du maximum de vraisemblance
- Algorithme numérique

- Hypothèses sous-jacentes :
  - Indépendance des observations
  - Si données répétées, structure hiérarchique..
    - autres modèles

Significativité des coefficients ?

- Test de Wald ou test du rapport de vraisemblance
  - H0 : $\beta_0=0$
- Intervalle de confiance
  - Contient la valeur 1 ?

Score	21,0725	1	0,0000
Likelihood Ratio	22,5472	1	0,0000

Previous Dataset Results Library

LOGISTIC angor = diab SEXE

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
diab	5,1131	3,1253	8,3650	1,6318	0,2512	6,4970	0,0000
SEXE (2/1)	0,2918	0,1636	0,5207	-1,2316	0,2954	-4,1693	0,0000
CONSTANT	*	*	*	-2,6454	0,1915	-13,8112	0,0000

Convergence:	Converged
Iterations:	6
Final -2*Log-Likelihood:	462,8079
Cases included:	963

Test	Statistic	D.F.	P-Value
Score	70,6242	2	0,0000
Likelihood Ratio	64,0780	2	0,0000

Intercept : constante du modèle

- β_0

= L' « effet » de la catégorie de référence

= probabilité de y lorsque toutes les co-variables sont nulles

- Logit [P(y=1|âge, sexe)] = β_0+β1âge+ $\beta_2$sexe
  - β_0= probabilité d'une femme d'âge 0
- Recodage de l'âge : centrer la variable autour de sa moyenne
  - β_0= probabilité d'une femme à l'âge moyen dans l'échantillon

Variable explicative dichotomique

- OR = $e^{\beta 1}$
- OR associé au passage de la catégorie de référence 0 à la catégorie 1
- OR ajusté sur les autres variables
- Exemple :
  - OR diab/non diab = 5.2
  - Les diabétiques ont une probabilité d'avoir un angor qui est multiplié par 5.2 par rapport aux non diabétiques

Variables explicatives à plusieurs modalités

- On choisit une catégorie de référence
- Variables indicatrices (dummy)
  - Ex : 1 variable :Tabac=1 si fumeur, =2 si ex fumeur = 3 si non fumeur
  - 2 variables : fum=1 si fumeur, exfum=1 si ex fumeur
- Logit $[P(y=1|x_1,x_2\ldots x_n)] = \beta_0+\beta 1 fum+ \beta_2 exfum$
  - OR fumeur/non = $e^{\beta 1}$
  - OR ex fumeur/non = $e^{\beta 2}$
  - OR fumeur/ex fumeur = $e^{\beta 1 - \beta 2}$

Si on n'avait pas créer de variables indicatrices

- Logit $[P(y=1|x_1,x_2\ldots x_n)] = \beta_0+\beta 1\text{tabac}$

  OR fumeur/exfumeur = e^{β}

  = OR ex fumeur/non = e^{β}

- Sous-entend un effet linéaire entre les catégories fumeur-ex fumeur-non fumeur

Variables explicatives continues

- OR = $e^{\beta 1}$
- OR associé à un accroissement unitaire
- Ex :
  - OR âge (1 an) =1.02
  - Pour chaque augmentation de 1 an, le risque de maladie augmente de 2%
  - OR âge (10 ans)= $(e^{\beta 1})^{10}$=1.22
  - Pour chaque augmentation de 10 ans, le risque de maladie augmente de 22%

Arbitrage du codage

- Exemple de l'âge
- Variable continu
  - Risque identique pour chaque augmentation d'1 an d'âge
- Variable en plusieurs catégories
  - Risque d'une catégorie d'âge par rapport à la catégorie de référence

Interaction

- En cas d'interaction (terme d'interaction significatif)
  - On présente les 2 OR séparément

(pas de sens de donner un OR « moyen »)

  - Exemple : effet de l'âge chez la femme et effet de l'âge chez l'homme

Voir cours sur interaction

Choix des variables du modèle :1

- Variables à inclure :
  - Variable(s) d'intérêt
  - Facteurs de confusions potentiels
    - FR connus
    - Issus de l'analyse des données (p = 20%)
  - Termes d'interaction
- Stratégie de sélection :
  - Pas-à-pas descendante
  - Méthode ascendante
  - Méthode automatique
  - Manuelle

Choix des variables du modèle : 2

- Adéquation d'un modèle par rapport à un autre ?
  - Modèle 1 : Logit [P(y=1|âge, sexe)] = β_0+β1âge+ $\beta_2$sexe
  - Modèle 2 : Logit [P(y=1|âge)] = β_0+β1âge
  - Si β_2=0, Modèle 2=Modèle 1
    - Modèle 2 est un sous-modèle du modèle 1
    - Modèle 2 est emboîté dans modèle 1
- Sur le plan statistique : comparaison de 2 modèles emboîtés
  - méthode du rapport de vraisemblance : 2ln V2/V1 suit χ2 à k ddl
  - Test H0 : β_2=0 (si un seul paramètre de différence)

Previous | Next | Last | History | Open | Bookmark | Print | Restore

Current View: **C:\Documents and Settings\cecile\Mes documents\data\MAGREDIAL\analyse données\clinique logistiq.xls:etat_clinique$**

Record Count: **1122** *Date:* **14/12/2007 11:30:34**

LOGISTIC angor = diab exfum fum SEXE

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
diab	4,3697	2,4643	7,7483	1,4747	0,2922	5,0460	0,0000
exfum (Yes/No)	2,6975	1,4742	4,9360	0,9923	0,3083	3,2189	0,0013
fum (Yes/No)	1,8207	0,5867	5,6497	0,5992	0,5778	1,0371	0,2997
SEXE (2/1)	0,3109	0,1619	0,5972	-1,1682	0,3330	-3,5077	0,0005
CONSTANT	*	*	*	-3,0482	0,2448	-12,4541	0,0000

Convergence: Converged
Iterations: 6
Final -2*Log-Likelihood: 353,4149
Cases included: 826

Test	Statistic	D.F.	P-Value
Score	63,5230	4	0,0000
Likelihood Ratio	56,1189	4	0,0000

Likelihood Ratio	49,8192	4	0,0000

Previous Dataset Results Library

LOGISTIC angor = diab exfum fum

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
diab	4,4253	2,5157	7,7845	1,4873	0,2882	5,1615	0,0000
exfum (Yes/No)	2,6666	1,4704	4,8359	0,9808	0,3037	3,2295	0,0012
fum (Yes/No)	1,9601	0,6571	5,8468	0,6730	0,5576	1,2069	0,2275
CONSTANT	*	*	*	-3,4731	0,2309	-15,0387	0,0000

Convergence: Converged
Iterations: 6
Final -2*Log-Likelihood: 367,5503
Cases included: 826

Test	Statistic	D.F.	P-Value
Score	50,5150	3	0,0000
Likelihood Ratio	41,9835	3	0,0000

Comparaison des 2 modèles

- Modèle 1 : sexe + diabète + tabac
- Modèle 2 : diabète + tabac
- Coefficient du sexe ?
  - Coeff = -1,168, SE = 0,33
  - Z-statistic = -3,507, p=0,0005
- Rapport vraisemblance ?
  - -2logM1 = 353,4
  - -2logM2 = 367,6
  - Diff = 14,1 $\chi 2$ à 1ddl, p<0,001

Choix des variables du modèle : 3

- Attention au surajustement « overfitting » :
  - « Au moins 10 outcomes positifs par covariable »
- Attention données manquantes
- Finalité du modèle :
  - Pronostic
  - Étiologique : facteurs confusions +++
  - Capacité d'interprétation

TABLE 3. ODDS RATIOS FOR CHRONIC RENAL FAILURE ASSOCIATED WITH THE LIFETIME USE OF EITHER ACETAMINOPHEN OR ASPIRIN AMONG SUBJECTS WHO DID NOT USE THE OTHER ANALGESIC REGULARLY.*

VARIABLE	ACETAMINOPHEN USE	ASPIRIN USE
	odds ratio (95% confidence interval)	
Never used	1.0	1.0
Ever used	1.3 (1.0–1.6)	1.5 (1.2–1.8)
Use or used regularly	2.5 (1.7–3.6)	2.5 (1.9–3.3)
Cumulative dose		
1–99 g	1.2 (0.9–1.5)	1.4 (1.1–1.7)
100–499 g	1.3 (0.9–1.8)	1.6 (1.2–2.1)
≥500 g	3.3 (2.0–5.5)	1.9 (1.3–2.9)

*The odds ratios have been adjusted for age, sex, level of education, smoking status, use or nonuse of other analgesics, and the interaction between aspirin use and acetaminophen use. $P<0.001$ for the trend toward greater risk with increasing cumulative doses of acetaminophen; $P=0.01$ for the trend toward greater risk with increasing cumulative doses of aspirin. Regular use was defined as the use of at least two tablets per week for a period of two months or longer.

TABLE 4. ODDS RATIOS FOR CHRONIC RENAL FAILURE ASSOCIATED WITH ISOLATED REGULAR USE OF ACETAMINOPHEN OR ASPIRIN ACCORDING TO THE TYPE OF UNDERLYING RENAL DISEASE.*

UNDERLYING DISEASE	ACETAMINOPHEN USE	ASPIRIN USE
	ODDS RATIO (95% CI)†	ODDS RATIO (95% CI)‡
Diabetic nephropathy	3.6 (2.1–6.0)	2.9 (1.9–4.5)
Glomerulonephritis	1.6 (0.9–3.0)	2.6 (1.4–4.8)
Nephrosclerosis	1.7 (0.8–3.7)	2.1 (1.3–3.5)
Hereditary renal disease	2.2 (0.8–5.9)	3.1 (1.6–6.0)
Systemic disease or vasculitis	2.8 (1.2–6.5)	1.1 (0.4–2.8)
Other renal disease	2.1 (0.9–4.6)	3.7 (1.8–7.7)

*The odds ratios have been adjusted for age, sex, level of education, smoking status, use or nonuse of other analgesics, and the interaction between aspirin use and acetaminophen use. Regular use was defined as the use of at least two tablets per week for a period of two months or longer. CI denotes confidence interval.

†The reference group was nonusers of acetaminophen without regular aspirin use.

‡The reference group was nonusers of aspirin without regular acetaminophen use.

Table 2. Factors associated with choice of PD *vs* planned HD

	Peritoneal dialysis %	Adjusted OR[a]	95% CI
Age at initiation (year)			
75–79	21.7	1	
80–84	27.3	1.3	1.0–1.6
≥ 85	37.7	2.1	1.5–2.8
Gender			
Men	25.5	1	
Women	28.0	1.1	0.9–1.4
Primary renal disease			
Glomerulonephritis	21.1	1	
Vascular nephropathy	29.3	1.5	0.9–2.4
Diabetic nephropathy	24.2	1.5	0.8–2.6
Other or unknown	26.0	1.4	0.8–2.2
Diabetes			
No	29.5	1	
Yes	25.9	0.8	0.6–1.1
Congestive heart failure			
No	23.8	1	
Yes	36.4	1.8	1.5–2.3
Malignancy			
No	28.9	1	
Yes	19.6	0.7	0.5–1.1
Severe behavioural disorder			
No	25.5	1	
Yes	39.4	2.2	1.3–3.5
Any severe disability			
No	26.5	1.0	
Yes	21.0	0.9	0.6–1.5
Mobility			
Walk without help	27.5	1	
Need assistance with mobility	26.3	0.8	0.6–1.1
Totally dependent for transfers	23.4	0.7	0.4–1.2
NA	26.0	0.9	0.7–1.3
Smoking			
Never smoker	30.0	1	
Former smoker	24.9	0.7	0.5–0.9
Current smoker	16.4	0.4	0.2–0.9
Haemoglobin (g/dl)			
≥11	34.4	1	
<11	27.2	0.8	0.6–1.0
NA	19.5	0.7	0.5–1.1
Albuminaemia (g/l)			
≥35	31.8	1	
<35	28.3	0.8	0.6–1.0
NA	23.6	0.9	0.6–1.2
Body mass index (kg/m^2)			
<18.5	30.4	1.0	0.6–1.6
18.5–25	29.8	1	
25–30	30.0	1.0	0.8–1.4
≥30	17.1	0.5	0.3–0.8
NA	21.7	1.0	0.6–1.3
Baseline eGFR (ml/min/1.73 m^2)			
<10	25.0	1	
≥10	32.3	1.4	1.1–1.7
NA	19.5	1.1	0.7–1.6

[a]OR adjusted for all variables as well as for region of residence.
NA: not available; eGFR: glomerular filtration rate estimated with the simplified MDRD equation.

Those starting PD also had an MDRD eGFR ≥10 ml/min/1.73 m^2 significantly more often than those starting HD (similar results were observed with the CG equation). In contrast, PD was chosen significantly less often for obese patients and smokers. It was also less frequent for patients with malignancies, anaemia, hypoalbuminaemia or diabetes, but the associations were of only borderline significance. The percentage of patients starting PD varied from 3% to 38% across regions, and the patients' characteristics did not explain this difference. These results were similar (data not shown) when we compared those starting PD with all patients starting HD (not simply planned cases), except that the associations with malignancy, anaemia and hypoalbuminaemia levels became statistically significant: 0.6 (0.4–0.9), 0.6 (0.5–0.8) and 0.7 (0.5–0.9), respectively. Treatment modality was not associated with chronic respiratory disease or with any type of vascular disease (of the heart, brain or peripheral vessels), regardless of the reference group.

Two-year outcome

During a 2-year follow-up, 66 patients recovered renal function, two had kidney transplantations, 57 switched from PD to HD and 51 from HD to PD and 1096 died (Table 3). Patient median survival was 26.8 months. Cardiovascular disease was the cause of death in 39% of the cases. Overall, survival was 68.5% (66.7–70.2) at 1 year, 52.7% (50.4–55.1) at 2 years and 39.3% (35.9–47.8) at 3 years, but, as expected, it decreased strongly with age, especially after 2 years (Figure 1). Death occurred after withdrawal of treatment in 17, 19 and 22% of the groups aged 75–79 years, 80–84 years and older than 85 years, respectively, in a median time of 5 months (range: 2 days–44 months) after the onset of dialysis. Treatment was withdrawn for medical reasons in 67% of the cases and at the patient's request in 25%.

Relations between comorbid conditions, treatment modality and 2-year mortality

Older age, congestive heart failure, peripheral vascular disease, malignancy, chronic respiratory disease, behavioural disorders, disability, reduced mobility, low BMI and albuminaemia were independently associated with a higher risk of death at 2 years (Table 4). Unplanned HD and PD were also significantly associated with a higher risk of death than planned HD, even after adjusting for all other risk factors. When those starting PD were compared with all those starting HD, the crude and adjusted risks at 2 years were similar for both treatment modalities (Table 5). There was no differential association between PD and either planned or all HD with diabetes or congestive heart failure.

Discussion

In this registry-based study, we found that despite a high comorbidity rate (85%), elderly patients who started dialysis between 2002 and 2005 appeared to