

**DEPARTEMENT DE MATHEMATIQUES ET DE STATISTIQUE**

**FILIERE : STATISTIQUE**

**PROJET 3 DE CONSULTATION**

Réalisé par :

SAMBIANI Paguidame

QINGYUE Li

Superviseur :

Pr. MIGUEL CHAGNON

Informaticien-Statisticien, Université de Montréal.

Session académique**:** Hiver 2024

1. **Méthodologie**

Dans ce projet, on vise initialement à concevoir un modèle prédictif pour évaluer les facteurs influençant la récupération des patients. Nous avons commencé notre travail par des statistiques descriptives pour examiner le portrait général des données, ainsi que des tests t pour comparer les moyennes des groupes, permettant d'identifier les différences significatives avant la modélisation. Ensuite, on a établi un modèle de régression logistique binaire pour identifier les prédicteurs significatifs de la récupération, tout en ajustant le seuil de classification pour optimiser le compromis entre sensibilité et spécificité via des analyses de la courbe ROC. L'exploration des odds ratios a permis de comprendre l'influence des variables indépendantes. Toutes nos procédures ont été réalisées sous SPSS version 28.0, avec un seuil de significativité de 0,05.

**Table des Tableaux et Figures**

[Table 1:Statistiques descriptives des variables catégorielles du jeu de données et Statistiques descriptives des variables continues 3](#_Toc160103227)

[Table 2 Tailles d’effet : Le D de Cohen 4](#_Toc160103228)

[Table 3 : Résultats des estimations des coefficients et odds ratio 5](#_Toc160103229)

[Table 4 : Zone sous la courbe 6](#_Toc160103230)

[Table 5: Tableau de confusion du modèle logit dichotomique 7](#_Toc160103231)

[Figure 1 Courbe ROC observée pour notre modèle 7](#_Toc160077232)

[Figure 2 Courbe ROC pour la variable age 9](#_Toc160077233)

# 

1. **Statistiques descriptives**

Le but de cette section est d’examiner le portrait de notre échantillon pour pouvoir identifier les tendances générales avant toute manipulation.

Table 1:Statistiques descriptives des variables catégorielles du jeu de données et Statistiques descriptives des variables continues

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **GROUPE** | |  |
| **Full N=27** | **Partial N=50** | Valeur P\* |
| **Sex** | Femme | 13 (48,1%) | 23 (46,0%) | 1,000 |
| **myelopathy** | Non | 26 (96,3%) | 24 (48,0%) | < 0,001 |
| **radiculopathy** | Non | 2 (7,4%) | 36 (72,0%) | < 0,001 |
| **spondylolisthesis** | Non | 22 (81,5%) | 34 (68,0%) | 0,285 |
| **spinal\_stenosis** | Non | 21 (77,8%) | 8 (16,0%) | < 0,001 |
| **foraminal\_stenosis** | Non | 19 (70,4%) | 17 (34,0%) | 0,04 |
| **hernia** | Non | 4 (14,8%) | 23 (46,0%) | 0,011 |
| **AGE** |  | 43,3213,42 | 58,56 | < 0,001 |
| **NFL** |  | 11,2520,7 |  | 0,769 |
| **CCL22** |  | 672,62290,8 | 537,38242,44 | 0,033 |
| **CRP** |  | 4358287,327847538,73 | 6361811,199556800,50 | 0,354 |

***\*Notations : Effectif (Fréquence) ; Variable Moyenne écart-type***

***\* Pour les variables continues, la valeur p est le résultat du Test T.***

**Interprétations :**

Tout d’abord, on observe qu’en moyenne le biomarqueur NFL est pratiquement le même dans les deux niveaux de la variable *recuperation.* Ceci est confirmé par le test t dont la p.value est 0,769 > 0,05 : ce biomarqueur semble donc n’avoir pas d’association significative avec notre variable d’intérêt. Ensuite, quand bien même on observe une différence de moyennes pour CRP dans les deux niveaux de la variable *recuperation*, cette différence n’est pas significative selon le résultat du test t qui a une p.value de 0,354 > 0,05 : ce biomarqueur semble également n’avoir pas d’association significative avec notre variable d’intérêt.

En revanche, on observe qu’il y a une différence de moyennes significative dans les deux niveaux de la variable *recuperation* pour les variables *age* et *CCL22,* dont les p.value associées aux tests t sont toutes inférieures au seuil de 5%. Le tableau suivant donne les tailles d’effets observés.

Table 2 Tailles d’effet : Le D de Cohen

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tailles d'effet pour échantillons indépendants** | | | | |
|  | | Estimation des points | Intervalle de confiance à 95% | |
| Inférieur | Supérieur |
| CCL22 | d de Cohen | 0,520 | 0,043 | 0,993 |
| CRP | d de Cohen | -0,223 | -0,691 | 0,248 |
| NFL | d de Cohen | 0,071 | -0,398 | 0,539 |
| age | d de Cohen | -1,300 | -1,808 | -0,784 |
| Le d de Cohen utilise l'écart type combiné. | | | | |

**Interprétations des tailles d’effet :**

La taille d’effet estimée pour la variable *CCL22* est de 0,520, un effet relativement moyen. Celle de *age* est de -1,3, un effet plutôt fort suggérant que la variable *age* a un impact significatif sur la différence entre les groupes comparés. On a un signe négatif qui peut s’expliquer par le fait que les sujets dans le groupe “*recuperation = full “* ont des valeurs d'âge significativement plus basses que ceuxdu deuxième groupe, comme remarqué précédemment dans le tableau 2. Même si les tailles d’effets paraissent faibles pour *CRP* et *NFL* (-0,223 et 0,071 respectivement*),* il faut souligner que leur interprétation pourrait dépendre du contexte d’étude ici : une petite taille d'effet peut être significative dans certains cas, surtout si elle est cohérente avec les attentes théoriques ou si elle a des implications pratiques importantes.

1. **Résultats des analyses**

Afin de fournir des interprétations intéressantes de ces effets marginaux, nous avons procédé dans cette partie à des opérations sur nos variables exogènes. Ainsi la variable *age* a été divisée par 10, *NFL* par 10, *CCL22* par 100 et *CRP* par 10000, ce qui nous donne de nouvelles variables que nous avons appelées respectivement *age\_1, NFL\_1, CCL22\_1* et *CRP\_1.*

Le tableau suivant donne les résultats de la régression logistique binaire que nous effectuons avec ces quatre nouvelles variables explicatives.

Table 3 : Résultats des estimations des coefficients et odds ratio

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variables d'équation** | | | | |
|  | odds ratios | P valeur | Intervalle de confiance à 95% pour odds ratios | |
| Inférieur | Supérieur |
| age (augmentation de 10) | 0,379 | < 0,001 | 0,227 | 0,633 |
| CCL22 (augmentation de 100) | 1,190 | 0,142 | 0,943 | 1,501 |
| CRP (augmentation de 10000) | 1,000 | 0,824 | 0,999 | 1,001 |
| NFL (augmentation de10) | 1,085 | 0,777 | 0,615 | 1,915 |
| Constante | 25,986 | 0,025 |  |  |

* **Significativité des variables exogènes du modèle**

**Interprétations :**

On observe que parmi toutes nos variables indépendantes incluses dans notre modélisation, seule l’âge apporte des informations statistiquement significatives dans la récupération des patients. Les biomarqueurs CCL22, CRP, NFL sont quant à eux non significatifs.

Aussi, on rappelle que la seule information directement utilisable ici est le signe des paramètres, indiquant si la variable associée influence à la hausse ou la baisse la probabilité de l’événement considéré.

**Signe des coefficients :**

* Le signe négatif du coefficient associé à l’âge signifie que l’âge influence à la baisse la probabilité de récupération complète d’un patient donné. En d'autres termes, plus le patient est âgé, moins il est probable qu'il récupère complètement.
* Le signe positif des coefficients associés aux trois biomarqueurs CCL22, NFL, CRP indique que ces trois biomarqueurs constituent chacun un facteur positif de récupération complète.
* **Interprétation des odds ratios :**

Toute chose étant égale ailleurs,

* Une augmentation de 10 ans sur l’âge d’un patient entraîne une diminution de son “odds“ de 62,1%.
* Une variation de 100 unités du biomarqueur CCL22 d’un patient entraîne une augmentation de son “odds“ de 19%.
* De même une variation de 10 unités du biomarqueur NFL entraîne une augmentation de son “odds“ de 8,5%.

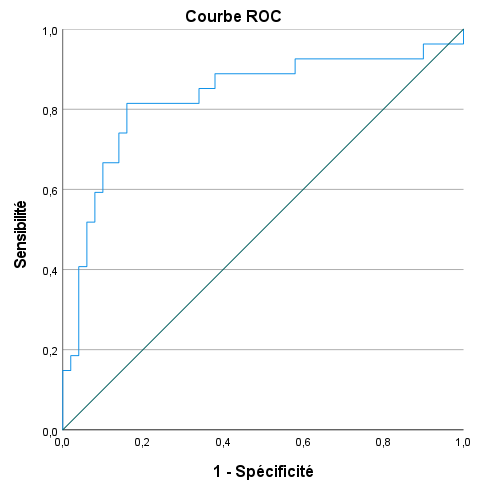
1. **Compromis entre sensibilité et spécificité**

Puisque nous travaillons ici avec des données cliniques, il est important d’arriver à un compromis entre sensibilité et spécificité de notre modèle. Nous pensons par exemple en fonction de la prévalence de la maladie étudiée, qu’il peut être grave de manquer un patient qui pourrait récupérer complètement : le test doit donc posséder une sensibilité élevée. La valeur seuil se situera au niveau de la partie supérieure droite de la courbe ROC. Inversement, en imaginant des scénarios d’une maladie possédant un traitement onéreux aux effets secondaires potentiellement graves, il convient de limiter au maximum le nombre de faux positifs pour éviter les traitements inutiles, donc de choisir une spécificité élevée. La valeur seuil sera située dans ce cas dans la partie inférieure gauche de la courbe ROC. Le seuil optimal sera donc le point de la courbe le plus éloigné de la diagonale, point correspondant également au maximum de l’indice de Youden (Se + Sp –1). Ce seuil dans notre cas est égal à 0,42 (avec indice de Youden Id= 0,655), correspondant à une sensibilité de 0,815 et d’une spécificité de 0,84. Ce qui fournit une aire sous la courbe AUC = 0,83.

Table 4 : Zone sous la courbe

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Zone sous la courbe** | | | | |
| Variable(s) de résultats tests : Probabilité prédite | | | | |
| Zone | Erreur standarda | P valeur Asymptotique b | Intervalle de confiance asymptotique à 95 % | |
| Borne inférieure | Borne supérieure |
| 0,830 | 0,055 | < 0,001 | 0,721 | 0,938 |
| a. Dans l'hypothèse non-paramétrique | | | | |
| b. Hypothèse nulle : zone vraie = 0.5 | | | | |

Figure 1 Courbe ROC observée pour notre modèle



Avec le seuil de coupure optimale de 0,42, nous obtenons la table de classification suivante :

Table 5: Tableau de confusion du modèle logit dichotomique

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table de classificationa** | | | | | | |
| Observé | | Prévisions | | | Intervalle de Confiance à 95% | |
| recuperation | | Pourcentage correct | Inférieur | Supérieur |
| partial | full |
| recuperation | partial | 42 | 8 | 84 | 0,7033946 | 0,9236204 |
| full | 5 | 22 | 81,5 | 0,6125378 | 0,9297392 |
| Pourcentage global | |  |  | 83,1 | 0.7286128 | 0.9069251 |
| a. La valeur de coupe est ,42 | | | | |  |  |

**Interprétation :**

À la lecture des résultats, nous constatons que les qualités prévisionnelles de ce modèle sont satisfaisantes car le taux d’erreur est assez faible. En effet, le tableau de classification nous montre d’une part que, sur les 27 patients qui ont complètement récupéré (pour lesquels ), le modèle indique que 22 patients ont une probabilité estimée de récupération totale supérieure à 0,42. Ainsi dans 81,5% des cas, la récupération complète est correctement prédite. D’autre part, sur les 50 patients qui ont partiellement récupéré (pour lesquels ), le modèle indique que 42 patients ont une probabilité estimée de récupération totale inférieure à 0,42, soit un pourcentage de bonnes prédictions de 84% pour la récupération partielle.

1. **Prédictions pour de nouveaux patients**

L’intérêt pratique de ce modèle mis en place va consister à pouvoir faire des prévisions de la classe de récupération (totale ou partielle) d’un nouveau patient donné à partir de ses facteurs explicatifs dont on dispose. Ainsi pour un patient i donné, le modèle logit va définir la probabilité associée à l’événement (récupération totale) comme la valeur de la fonction de répartition de la loi logistique, à savoir au point , où est le vecteur des facteurs explicatifs de l’individu i et est le vecteur des estimés des coefficients qui leurs sont associés.

Ici nous fournissons les coefficients du modèle estimé avec les vraies valeurs (sans changements d’échelle de valeurs) de *age, NFL, CCL22* et *CRP,*  . Ensuite on décidera d’une récupération partielle si la probabilité estimée est inférieure au seuil de césure de 0,42, et dans le cas contraire on décidera plutôt d’une récupération complète pour ce nouveau patient.

**Autre Alternative :**

L’âge est la seule variable significative pour prédire la récupération complète ou partielle des patients d’après les résultats de notre régression. Alors, on pourrait déterminer un seuil d’âge qui pourrait servir de point de référence pour prédire la récupération avec une règle simple : Par exemple, vous pourriez dire que les patients ayant un âge inférieur au seuil ont une probabilité plus élevée de récupération complète, tandis que ceux ayant un âge supérieur au seuil ont une probabilité moindre. Nous faisons cela par une analyse de courbe ROC de la variable *age* comme suit qui nous fournit les résultats suivants :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Zone sous la courbe** | | | | |  |
| Variable(s) de résultats tests : age | | | | |  |
| Zone | Erreur standarda | P valeur asymptotiqueb | Intervalle de confiance asymptotique à 95 % | |  |
| Borne inférieure | Borne supérieure |  |
| 0,804 | 0,063 | < 0,001 | 0,680 | 0,928 |  |
| a. Dans l'hypothèse non-paramétrique | | | | |  |
| b. Hypothèse nulle : zone vraie = 0.5 | | | | |  |

Figure 2 Courbe ROC pour la variable age

Une image contenant texte, ligne, diagramme, Tracé

Description générée automatiquement

|  |  |  |  |
| --- | --- | --- | --- |
| **Coordonnées de la courbe** | | | |
| Variable(s) de résultats tests : age | | | |
| Positif si inférieur ou égal àa | Sensibilité | 1 - Spécificité | Indice de Youden |
| 24,410958904109600 | 0,000 | 0,000 | 0,000 |
| 25,808219178082200 | 0,037 | 0,000 | 0,037 |
| 27,480821917808200 | 0,074 | 0,000 | 0,074 |
| 29,232876712328800 | 0,111 | 0,000 | 0,111 |
| 31,160273972602700 | 0,111 | 0,020 | 0,091 |
| 33,291780821917800 | 0,148 | 0,020 | 0,128 |
| 33,976712328767100 | 0,148 | 0,040 | 0,108 |
| 34,140410958904100 | 0,185 | 0,040 | 0,145 |
| 34,806164383561600 | 0,222 | 0,040 | 0,182 |
| 35,656164383561600 | 0,259 | 0,040 | 0,219 |
| 36,041095890411000 | 0,296 | 0,040 | 0,256 |
| 36,146575342465800 | 0,333 | 0,040 | 0,293 |
| 36,713698630137000 | 0,370 | 0,040 | 0,330 |
| 37,241095890411000 | 0,407 | 0,040 | 0,367 |
| 37,354794520547900 | 0,444 | 0,040 | 0,404 |
| 37,754794520547900 | 0,444 | 0,060 | 0,384 |
| 38,275342465753400 | 0,444 | 0,080 | 0,364 |
| 39,994520547945200 | 0,481 | 0,080 | 0,401 |
| 41,552054794520500 | 0,519 | 0,080 | 0,439 |
| 41,682191780821900 | 0,556 | 0,080 | 0,476 |
| 41,982191780821900 | 0,593 | 0,080 | 0,513 |
| 42,550000000000000 | 0,593 | 0,100 | 0,493 |
| 43,250000000000000 | 0,630 | 0,100 | 0,530 |
| 43,806849315068500 | 0,630 | 0,120 | 0,510 |
| 44,273972602739700 | 0,667 | 0,120 | 0,547 |
| 44,657534246575300 | 0,704 | 0,120 | 0,584 |
| 45,540410958904100 | 0,741 | 0,120 | 0,621 |
| 46,600000000000000 | 0,741 | 0,140 | 0,601 |
| 47,051369863013700 | 0,741 | 0,160 | 0,581 |
| 47,301369863013700 | 0,778 | 0,160 | 0,618 |
| 48,793150684931500 | 0,815 | 0,160 | 0,655 |
| 50,369863013698600 | 0,815 | 0,180 | 0,635 |
| 50,639726027397300 | 0,815 | 0,200 | 0,615 |
| 51,052054794520500 | 0,815 | 0,220 | 0,595 |
| 51,889041095890400 | 0,852 | 0,220 | 0,632 |
| 52,557534246575300 | 0,852 | 0,240 | 0,612 |
| 53,968493150684900 | 0,852 | 0,260 | 0,592 |
| 55,227397260274000 | 0,852 | 0,280 | 0,572 |
| 55,447945205479500 | 0,852 | 0,300 | 0,552 |
| 56,102739726027400 | 0,852 | 0,320 | 0,532 |
| 56,621232876712300 | 0,852 | 0,340 | 0,512 |
| 57,045890410958900 | 0,852 | 0,360 | 0,492 |
| 57,450684931506800 | 0,852 | 0,380 | 0,472 |
| 57,731506849315100 | 0,852 | 0,400 | 0,452 |
| 58,179452054794500 | 0,852 | 0,420 | 0,432 |
| 58,582191780821900 | 0,852 | 0,440 | 0,412 |
| 58,829452054794500 | 0,852 | 0,460 | 0,392 |
| 59,048630136986300 | 0,852 | 0,480 | 0,372 |
| 59,376712328767100 | 0,852 | 0,500 | 0,352 |
| 59,854794520547900 | 0,852 | 0,520 | 0,332 |
| 60,216438356164400 | 0,852 | 0,540 | 0,312 |
| 60,735616438356200 | 0,852 | 0,560 | 0,292 |
| 61,453424657534200 | 0,852 | 0,580 | 0,272 |
| 62,019178082191800 | 0,852 | 0,600 | 0,252 |
| 62,511643835616400 | 0,852 | 0,620 | 0,232 |
| 62,802054794520600 | 0,852 | 0,640 | 0,212 |
| 63,035616438356200 | 0,852 | 0,660 | 0,192 |
| 63,186301369863000 | 0,852 | 0,680 | 0,172 |
| 63,794520547945200 | 0,852 | 0,700 | 0,152 |
| 64,741780821917800 | 0,852 | 0,720 | 0,132 |
| 65,250000000000000 | 0,852 | 0,740 | 0,112 |
| 65,742465753424700 | 0,852 | 0,760 | 0,092 |
| 66,298630136986300 | 0,852 | 0,780 | 0,072 |
| 66,756164383561600 | 0,852 | 0,800 | 0,052 |
| 67,258904109589000 | 0,889 | 0,800 | 0,089 |
| 68,208219178082200 | 0,889 | 0,820 | 0,069 |
| 68,980821917808200 | 0,926 | 0,820 | 0,106 |
| 70,047945205479500 | 0,926 | 0,840 | 0,086 |
| 71,516438356164400 | 0,926 | 0,860 | 0,066 |
| 72,445205479452100 | 0,926 | 0,880 | 0,046 |
| 73,101369863013700 | 0,926 | 0,900 | 0,026 |
| 73,375342465753400 | 0,926 | 0,920 | 0,006 |
| 73,605479452054800 | 0,926 | 0,940 | -0,014 |
| 73,816438356164400 | 0,926 | 0,960 | -0,034 |
| 74,009589041095900 | 0,926 | 0,980 | -0,054 |
| 74,239726027397300 | 0,926 | 1,000 | -0,074 |
| 74,382191780821900 | 0,963 | 1,000 | -0,037 |
| 75,443835616438300 | 1,000 | 1,000 | 0,000 |
| a. La valeur césure la plus petite est la valeur test minimale observée moins 1, et la valeur césure la plus grande est la valeur test maximale observée plus 1. Toutes les autres valeurs césures sont les moyennes de deux valeurs tests consécutives observées. | | |  |

**Comparaison des deux méthodes :**

Sur la base de ces analyses de la courbe ROC, nous choisissons une valeur seuil de 50 ans pour l’âge, et cela nous permets de confronter les deux alternatives de prédiction précédentes toujours avec le seuil de césure de 0,42. Nous remarquons qu’il y a concordance dans la manière de prédire la récupération des patients. Les deux méthodes ont à peu près le même pouvoir de prédictions (en termes de bonnes prédictions et mauvaises prédictions). En conclusion, en optant pour des prises de décisions utilisant le seuil de 50 ans de la variable *age (*qui est la seule variable significativement associée à la récupération des patients d’après nos résultats), cela reviendrait à la même chose. On obtient ainsi une simplification dans les calculs de prédictions pour de nouveaux patients.

1. **Conclusion**

Notre étude a développé un modèle prédictif mettant en évidence l'influence significative de l'âge sur la récupération des patients, tout en notant l'absence de corrélation significative pour les biomarqueurs CRP, NFL et CCL22. Ce modèle offre un cadre pour affiner les stratégies thérapeutiques lors de l'évaluation et de la prise en charge des patients, suggérant ainsi d’adapter les stratégies de traitement et de suivi en fonction de l'âge. L'amélioration future du modèle pourra renforcer sa précision et son application clinique.

1. **Calcul de Puissance**

Il s’agit dans cette partie, d’un calcul de puissance afin de déterminer une taille d’échantillon nécessaire à l’obtention de résultats plus intéressants dans le cadre de ce travail. Puisque nous sommes dans un cadre de prédictions ici, nous avons choisi de nous intéresser à un calcul de puissance afin de déterminer la taille d’échantillon nécessaire pour atteindre une précision de plus ou moins 5% sur l’efficience du modèle, qui dans notre cas valait 83,1%. Ensuite nous discutons d’éventuelles limites dans l’interprétation de cette démarche suivie.

Pour déterminer la taille de l'échantillon nécessaire pour atteindre une précision donnée pour une proportion, nous devons utiliser la formule de l'intervalle de confiance pour une proportion :

, où :

* est l'estimation de la proportion dans l'échantillon.
* est le quantile de la distribution normale standard correspondant au niveau de confiance 1 -, nous choissisons dans notre cas ici un niveau de 95% ( c'est-à-dire = 5% ).
* est la taille d’échantillon

Ainsi pour une précision , il nous faudra une taille d’échantillon donnée par : , ce qui dans le cas de l’efficience de notre modèle se calcule :

= 216 participants environ

* **Limites de la méthode**
* Coût et ressources : La collecte de données cliniques peut être coûteuse en termes de temps, d'argent et de ressources humaines. Il est important de prendre en compte ces facteurs lors de la planification de la taille de l'échantillon et de s'assurer que les ressources nécessaires sont disponibles.
* Equilibre entre les deux groupes : Il n’est pas exclu de tomber dans un cas extrême où sur les 216 participants, on ait que seulement un petit pourcentage de patients qui récupèrent partiellement ou complètement, ce qui pourrait biaiser les résultats de modélisations.