

**DEPARTEMENT DE MATHEMATIQUES ET DE STATISTIQUE**

**FILIERE : STATISTIQUE**

**PROJET 3 DE CONSULTATION**

Réalisé par :

SAMBIANI Paguidame

QUIGYUE 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 3](#_Toc159647535)

[Table 2: Statistiques descriptives des variables continues 4](#_Toc159647536)

[Table 3 : Tailles d’effet : Le D de Cohen 4](#_Toc159647537)

[Table 4 : Résultats des estimations des coefficients et odds ratio 5](#_Toc159647538)

[Table 5 : Aire sous la courbe 6](#_Toc159647539)

[Table 6: Tableau de confusion du modèle logit dichotomique 7](#_Toc159647540)

[Figure 2 Courbe ROC observée pour notre modèle 6](#_Toc159647545)

# 

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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **GROUPE** | | |  |
| **Full N=27** | **Partial N=50** | | Valeur P\* |
| **Sexe** | Femme | 13 (48,1%) | 23 (46,0%) | | 1,000 |
| Homme | 14 (51,9%) | 27 (54,0%) | |
| **myelopathy** | Non | 26 (96,3%) | 24 (48,0%) | | < 0,001 |
| Oui | 1 (3,7%) | 26 (52,0%) | |
| **radiculopathy** | Non | 2 (7,4%) | 36 (72,0%) | | < 0,001 |
| Oui | 25 (92,6%) | 14 (28,0%) | |
| **spondylolisthesis** | Non | 22 (81,5%) | 34 (68,0%) | | 0,285 |
| Oui | 5 (18,5%) | 16 (32,0%) | |
| **spinal\_stenosis** | Non | 21 (77,8%) | 8 (16,0%) | | < 0,001 |
| Oui | 6 (22,2%) | 42 (84,0%) | |
| **foraminal\_stenosis** | Non | 19 (70,4%) | 17 (34,0%) | | 0,04 |
| Oui | 8 (29,6%) | 33 (66,0%) | |  |
| **hernia** | Non | 4 (14,8%) | 23 (46,0%) | | 0,011 |
| Oui | 23 (85,2%) | 27 (54,0%) | |
| **Narrow\_canal** | Non  Oui | 19 (70,4%)  8 (29,6%) | 3 (6,0%)  47 (94,0%) | | < 0,001 |
| **Nerve\_damage** | Non  Oui | 16 (59,3%)  11 (40,7%) | 20 (40,0%)  30 (60,0%) | | 0,151 |
| **Smoke** | Non  Oui | 14 (51,9%)  13 (40,7%) | 33 (66,0%)  17 (34,0%) | | 0,327 |
|  | **Total** | **27 (100,0%)** | **50** | **100,0%** |  |

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

***\*valeur p exacte du test de Khi deux***

**Interprétations :**

Parmi l’ensemble des variables catégorielles du jeu de données, les résultats du test de Khi 2 indique une absence d’association statistiquement significative entre la variable *recuperation* et les variables *Smoke, Sexe, Nerve damage, spondylolisthesis, qui sont pour la plupart homogènes dans les deux modalités de* recuperation. En revanche, il y a une nette répartition inégale en termes de fréquences des autres variables catégorielles *myelopathy Narrow\_canal hernia foraminal\_stenosis spinal\_stenosis radiculopathy* dans leurs différentes modalités par rapport à la variable *recuperation. Aussi,* la p.value des tests du Khi 2 entre notre variable d’intérêt *recuperation* et ces différentes variables est inférieure à 0.05, ce qui suggère l’existence d’une certaine association significative entre elles.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **recuperation = full** | **recuperation = partial** | **p bilatéral (T-test)** |
| **NFL** | 11,2520,7 |  | 0,769 |
| **AGE** | 43,3213,42 |  | 0,000 |
| **CCL22** | 672,62290,8 | 537,38242,44 | 0,033 |
| **CRP** | 4358287,327847538,73 | 6361811,199556800,50 | 0,354 |

Table 2: Statistiques descriptives des variables continues

**Notations** : Variable Moyenne écart-type

**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 O.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 3 : Tailles d’effet : Le D de Cohen

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tailles d'effet pour échantillons indépendants** | | | | | |
|  | | Standardisationa | Estimation des points | Intervalle de confiance à 95% | |
| Inférieur | Supérieur |
| CCL22 | d de Cohen | 260,222453951784000 | 0,520 | 0,043 | 0,993 |
| CRP | d de Cohen | 9001088,687489570000000 | -0,223 | -0,691 | 0,248 |
| NFL | d de Cohen | 13,565221206357400 | 0,071 | -0,398 | 0,539 |
| age | d de Cohen | 11,725001482723100 | -1,300 | -1,808 | -0,784 |
| a. Dénominateur utilisé pour estimer les tailles d'effet.  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, indiquant un effet plutôt fort. On a un signe négatif qui peut s’expliquer par le fait que les valeurs moyennes dans le groupe “*recuperation = full “* sont plus petites que celles du 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 4 : Résultats des estimations des coefficients et odds ratio

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables de l'équation** | | | | | | | | | | |
|  | | B | E.S | Wald | ddl | Sig. | Exp(B) | Intervalle de confiance 95% pour EXP(B) | |
| Inférieur | Supérieur |
| Pas 1a | age\_1 | -0,971 | 0,262 | 13,723 | 1 | 0,000 | 0,379 | 0,227 | 0,633 |
| CCL22\_1 | 0,174 | 0,118 | 2,155 | 1 | 0,142 | 1,190 | 0,943 | 1,501 |
| CRP\_1 | 0,000 | 0,000 | 0,049 | 1 | 0,824 | 1,000 | 0,999 | 1,001 |
| NFL\_1 | 0,082 | 0,290 | 0,080 | 1 | 0,777 | 1,085 | 0,615 | 1,915 |
| Constante | 3,258 | 1,458 | 4,990 | 1 | 0,025 | 25,986 |  |  |
| a. Introduction des variables au pas 1 : age\_1, CCL22\_1, CRP\_1, NFL\_1. | | | | | | | | | | |

* **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. Autrement dit, les patients âgés ont généralement moins de chance à la récupération complète que les patients plus jeunes.
* Le signe positif des coefficients associés aux trois biomarqueurs CCL22 et NFL indique que ces trois biomarqueurs constituent chacun un facteur positif de récupération complète.
* **Interprétation des odds ratios**
* A chaque fois qu’un patient gagne 10 ans de plus sur son âge, cela entraîne une diminution de sa probabilité de récupération complète d’environ 7.54%.
* Toute variation sur la valeur du biomarqueur CRP n’a aucun effet sur la probabilité de récupération.
* Une variation de 100 unités du biomarqueur CCL22 augmente la probabilité de récupération de 0.14%, ce qui est bien négligeable.
* De même une variation de 10 unités du biomarqueur NFL augmente la probabilité de récupération complète de 0.64%.

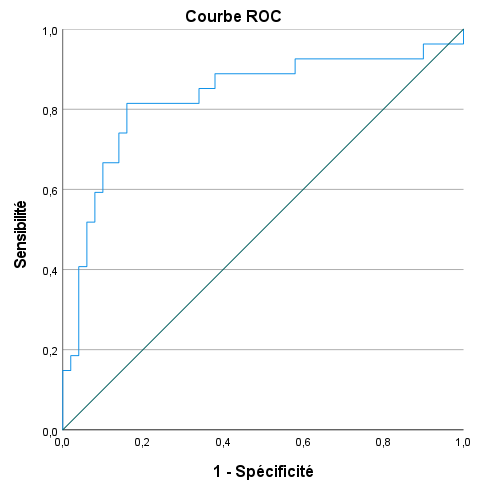
1. **Compromis entre sensibilité et spécificité**

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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Zone sous la courbe** | | | | |
| Variable(s) de résultats tests : Probabilité prédite | | | | |
| Zone | Erreur standarda | Sig. Asymptotiqueb | Intervalle de confiance asymptotique à 95 % | |
| Borne inférieure | Borne supérieure |
| 0,830 | 0,055 | 0,000 | 0,721 | 0,938 |
| a. Dans l'hypothèse non-paramétrique | | | | |
| b. Hypothèse nulle : zone vraie = 0.5 | | | | |

Table 5 : Tableau de confusion du modèle logit dichotomique

Figure 2 **Courbe ROC observée pour notre modèle**



Avec le seuil de coupure optimale de 0.42, nous obtenons la table de classification suivante :

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table de classificationa** | | | | | | | |
| Observé | | | Prévisions | | | Intervalle de Confiance à 95% | |
| recuperation | | Pourcentage correct | Inférieur | Supérieur |
| partial | full |
| Pas 1 | recuperation | partial | 42 | 8 | 84,0 | 0.7033946 | 0.9236204 |
| full | 5 | 22 | 81,5 | 0.6125378 | 0.9297392 |
| Pourcentage global | |  |  | 83,1 | 0.7249352 | 0.9035879 |
| a. La valeur de coupe est ,420 | | | | | |  |  |

**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 O.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 O.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 . Ensuite on décidera d’une récupération partielle si la probabilité estimée est inférieure au seuil de césure de 0.417, et dans le cas contraire on décidera plutôt d’une récupération complète pour ce nouveau patient.

1. **Conclusion**