1. **Méthodologie**

Dans ce projet, on vise initialement à concevoir un modèle prédictif pour évaluer les facteurs influençant la récupération des patients. On a d'abord employé des statistiques descriptives pour examiner la distribution des données. Des tests t ont été utilisés 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, ajustant le seuil de classification pour optimiser le compromis entre sensibilité et spécificité via l'analyse des courbes ROC. L'exploration des effets marginaux a permis de comprendre l'influence des variables indépendantes. Ces procédures ont été réalisées sous SPSS version 27.0, avec un seuil de significativité de 0,05.

1. **Statistiques descriptives**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | | N | Pourcentage |
| recuperation | partial | 50 | 64,9% |
| full | 27 | 35,1% |
| Total | 77 | 100,0% |

Table 1: Statistiques descriptives de la variable d’intérêt

Figure 1: Répartition graphique des patients en fonction de la recuperation

Une image contenant capture d’écran, texte, Rectangle, diagramme

Description générée automatiquement

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | **GROUPE** | | | |  |
| **Full** | | **Partial** | |  |
| Effectif | Fréquence (%) | Effectif | Fréquence (%) | Valeur P\* |
| **Sex** | 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 | 16,0% | 0,04 |
| Oui | 8 | 29,6% | 33 | 84,0% |  |
| **hernia** | Non | 4 | 14,8% | 23 | 34,0% | 0,011 |
| Oui | 23 | 85,2% | 27 | 66,0% |
| **Narrow\_canal** | Non  Oui | 19  8 | 70,4%  29,6% | 3  47 | 46,0%  54,0% | < 0,001 |
| **Nerve\_damage** | Non  Oui | 16  11 | 59,3%  40,7% | 20  30 | 40,0%  60,0% | 0,151 |
| **Smoke** | Non  Oui | 14  13 | 51,9%  40,7% | 33  17 | 66,0%  34,0% | 0,327 |
|  | **Total** | **27** | **100,0%** | **50** | **100,0%** |  |
| ***\*valeur p exacte du test de Khi deux*** | | | | | |  |

Table 2 : Statistiques descriptives

Table 3: Statistiques descriptives par modalité de recuperation

|  |  |  |  |
| --- | --- | --- | --- |
|  | **recuperation = full** | **recuperation = partial** | **Total** |
| **NFL** | 11,2520,7 |  | 10,6313,48 |
| **AGE** | 43,3213,42 |  |  |
| **CCL22** | 672,62290,8 | 537,38242,44 | 584,80266,54 |
| **CRP** | 4358287,327847538,73 | 6361811,199556800,50 | 5659276,848993306,75 |

**Notations** : Variable Moyenne écart-type

1. **Des tests de comparaison de moyennes**

Dans la suite, nous avons codé notre variable d’intérêt *recuperation* en une binaire 0 1 que nous avons nommée *recuperationC, et* nous procédons à des tests de comparaison de moyenne entre cette nouvelle variable codée et les autres variables explicatives. En effet, cela nous permettra d’avoir d’avance des intuitions sur les possibles associations de ces variables explicatives incluses dans notre modélisation ; ce que viendra confirmer/infirmer la régression logistique.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Statistiques de groupe** | | | | | |
| RecuperationC | | N | Moyenne | Ecart type | Moyenne d'erreur standard |
| age | 1 | 27 | 43,32 | 13,42 | 2,58 |
| 0 | 50 | 58,56 | 10,72 | 1,52 |
| CCL22 | 1 | 27 | 672,62 | 290,80 | 55,96 |
| 0 | 50 | 537,38 | 242,44 | 34,29 |
| CRP | 1 | 27 | 4358287,32 | 7847538,73 | 1510259,53 |
| 0 | 50 | 6361811,19 | 9556800,49 | 1351535,69 |
| NFL | 1 | 27 | 11,25 | 20,69 | 3,98 |
| 0 | 50 | 10,29 | 7,38 | 1,044 |

Table 4 : **Tests de comparaison de moyennes**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Test t pour égalité des moyennes** | | | | | |
| t | df | Signification | | Différence moyenne | Erreur standard |
| p unilatéral | p bilatéral |
| age | -5,443 | 75 | 0,000 | 0,000 | -15,24 | 2,800 |
| CCL22 | 2,176 | 75 | 0,016 | 0,033 | 135,24 | 62,15 |
| CRP | -0,932 | 75 | 0,177 | 0,354 | -2003523,87 | 2149678,53 |
| NFL | 0,295 | 75 | 0,384 | 0,769 | 0,956 | 3,24 |

**Interprétation :** Age : La différence significative dans les moyennes d'âge avec une p-valeur bilatérale de <0,001 indique que l'âge pourrait être un prédicteur important de la récupération.

CCL22 : Avec une p-valeur bilatérale de 0,033, il existe une différence significative dans les niveaux de CCL22 entre les groupes. Cela peut impliquer que le CCL22 joue un rôle dans la récupération.

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 prédictif que nous nous apprêtons à mettre en place. Cela dépendra surtout des conséquences et implications cliniques relatives à ces choix de seuil que nous ferons. Ainsi, nous pensons par exemple qu’il peut être plus grave de manquer un patient qui pourrait récupérer complètement que de classer à tort un patient comme récupérant complètement (modèle sensible pour être capable de détecter effectivement les patients qui récupèrent totalement). Mais d’un autre côté, en imaginant des scénarios où les ressources du clinicien seraient limitées, nous voulons aussi un modèle qui minimise les faux positifs pour éviter les traitements inutiles (modèle spécifique qui nous évite de prédire des patients comme récupérant totalement alors qu’ils ne le font pas effectivement).

En raison du déséquilibre de classe dans nos données (27 contre 50), notre modèle semble être biaisé vers la classe majoritaire des patients qui récupèrent partiellement avec un choix de seuil par défaut de 0.5. En fait notre modèle prédira dans la quasi-totalité la classe majoritaire (donc une haute exactitude), mais ne serait pas utile dans la pratique. Nous avons donc ajusté le seuil de classification pour équilibrer sensibilité et spécificité à travers des analyses de courbes ROC d’une première régression que nous avons faite, avec *recuperationC* comme variable dépendante.

Une image contenant diagramme, ligne, texte, Tracé

Description générée automatiquement

Figure 2 **Courbe ROC pour notre modèle**

Sur la base du tableau des coordonnées Seuil-Sensibilité- Spécificité fournie en sortie par notre analyse ROC, nous choisissons finalement un seuil de césure à 0.417, 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, qui est bien acceptable.

1. **Résultats des analyses**

* **Significativité globale du modèle :**

La probabilité associée à la statistique de Wald étant inférieur à 0,05, on peut affirmer que ce modèle est globalement significatif, avec un risque de 5%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tests composites des coefficients du modèle** | | | | |
|  | | Khi-carré | ddl | Sig. |
| Pas 1 | Pas | 26,405 | 4 | 0,000 |
| Bloc | 26,405 | 4 | 0,000 |
| Modèle | 26,405 | 4 | 0,000 |

Table 5: **Test de significativité globale de Wald**

* **Pouvoir explicatif du modèle**

Le R2 de Cox et Snell de notre modèle est de 0.29, ce qui indique un ajustement relativement faible du modèle aux données observées.

|  |  |  |  |
| --- | --- | --- | --- |
| **Récapitulatif des modèles** | | | |
| Pas | Log de vraisemblance -2 | R-deux de Cox et Snell | R-deux de Nagelkerke |
| 1 | 73,364 | 0,290 | 0,400 |

Table 6: Le R2 de Cox et Snell

* **Ajustement du modèle : test de Hosmer-Lemeshow**

Les hypothèses du test sont **:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Test de Hosmer et Lemeshow** | | | |
| Pas | Khi-carré | ddl | Sig. |
| 1 | 11,912 | 8 | 0,155 |

Table 7: Test d’ajustement du modèle

**Interprétation :**

La probabilité associée à la statistique de Hosmer-Lemeshow calculée étant supérieure à 5% (soit, prob = 0.155), il y a lieu de considérer que l’ajustement est bon. Une conclusion qui confirme bien notre appréciation des pourcentages de prédictions obtenus.

Table 8: Tableau de confusion du modèle logit dichotomique

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table de classificationa** | | | | | |
| Observé | | | Prévisions | | |
| recuperationC | | Pourcentage correct |
| 0 | 1 |
| Pas 1 | recuperationC | 0 | 42 | 8 | 84,0 |
| 1 | 5 | 22 | 81,5 |
| Pourcentage global | |  |  | 83,1 |
| a. La valeur de coupe est 0.417 | | | | | |

**Interprétation :**

Le tableau de confusion (avec un seuil de coupure de 0,417) nous montre que sur les 27 patients qui ont totalement récupéré, notre modèle arrive à prédire correctement 22 cas, et sur les 50 patients qui ont partiellement récupéré, il donne une bonne prédiction dans 42 cas. On obtient alors un taux de bonnes prédictions de 83.1%, un pourcentage nettement bon.

* **Significativité des variables exogènes du modèle**

Table 9 : Résultats des estimations des coefficients et effets marginaux

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables de l'équation** | | | | | | | | | | |
|  | | B | E.S | Wald | ddl | Sig. | Exp(B) | Intervalle de confiance 95% pour EXP(B) | |
| Inférieur | Supérieur |
| Pas 1a | CCL22 | 0,002 | 0,001 | 2,155 | 1 | 0,142 | 1,002 | 0,999 | 1,004 |
| CRP | 0,000 | 0,000 | 0,049 | 1 | 0,824 | 1,000 | 1,000 | 1,000 |
| NFL | 0,008 | 0,029 | 0,080 | 1 | 0,777 | 1,008 | 0,953 | 1,067 |
| age | -0,097 | 0,026 | 13,723 | 1 | 0,000 | 0,907 | 0,862 | 0,955 |
| Constante | 3,258 | 1,458 | 4,990 | 1 | 0,025 | 25,986 |  |  |

**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 elles non significatives.

Aussi, on rappelle que les valeurs numériques des estimations n’ont pas d’interprétation économique directe ici (ils ne correspondent aux paramètres de l’équation qu’à une constante multiplicative près), en raison du problème de la normalisation de la variance résiduelle. Ainsi, il faut retenir que la seule information directe réellement utilisable est le signe des paramètres, indiquant si la variable associée influence à la hausse ou la baisse la probabilité de l’événement considéré. Ainsi,

* Le signe négatif du coefficient associé à l’âge signifie que l’âge influence à la baisse la probabilité de récupération totale d’un patient. Autrement dit, plus on gagne en âge et moins on a de chance de récupérer totalement.
* Le signe positif des coefficients associés aux trois biomarqueurs CCL22, CRP et NFL indique que ces trois biomarqueurs constituent chacun un atout positif pour la récupération totale d’un patient donné.
* **Interprétation des effets marginaux**

Les effets marginaux mesurent la sensibilité de la probabilité de l’événement yi = 1 par rapport à des variations dans les variables explicatives xi. On aimerait dans cette partie établir quelle serait la variation de la probabilité de l’événement yi = 1 (récupération totale dans notre cas) en cas de variation d’une des variables exogènes.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| factor | AME | SE | z | p | lower | upper |
| age | -0,0754 | 0,0130 | -5,7973 | 0 | -0,1009 | -0,0499 |
| CCL22 | 0,0014 | 0,0009 | 1,5449 | 0,1224 | -0,0004 | 0,0031 |
| CRP | 0 | 0 | -0,2155 | 0,8294 | 0 | 0 |
| NFL | 0,0064 | 0,0225 | 0,2830 | 0,7771 | -0,0377 | 0,0505 |

* **Interprétation des odds ratios**

Le rapport de risques (odds ratio en anglais) est calculé comme le rapport de chances(odds) d’occurrence de l’événement étudié entre deux niveaux de la variable indépendante. Il indique combien de fois plus grande (ou plus petite) est la probabilité d’occurrence de l’événement de succès dans un groupe par rapport à un autre groupe.