

G. Post ¹, N. Assi ², A. Schmidt ², C. David ³, V. Mary ³, P. Pospieszyski ⁴, P. Lemeille ²

¹Psychiatre, consultant médical e-santé, dr.post.geoffrey@gmail.com ; ² Heva, Lyon, France ; ³ GIP SIB ; ⁴ GIP SYMARIS

Introduction

- La psychiatrie hospitalière publique est confrontée à une augmentation du volume de patients avec des moyens qui se réduisent.
- En particulier, 50 % des patients hospitalisés seront réhospitalisés dans les 2 ans. Plusieurs mécanismes sont évoqués dans la littérature scientifique pour l’expliquer : situation sociale, diagnostic clinique, comorbidités, antécédents médicaux, comportement durant l’hospitalisation, traitements reçus.
- Dans ce contexte, diminuer le nombre d’hospitalisations apparaît comme un enjeu majeur.

Objectifs

L’objectif de notre étude était de fournir un score de risque de réhospitalisation dans les 2 ans suivant une hospitalisation en psychiatrie.

Méthode

Source de données

- Base hospitalière du CH de Rouffach, issue des dossiers médicaux.
- Contient des données socio-démographiques, cliniques (tests biologiques), diagnostiques, traitements pris pendant le séjour et ceux prescrits à la sortie.

Design de l’étude

- Inclusion des patients entre le 01/06/2008 et le 31/12/2016.
- Suivi des patients jusqu’au 31/12/2018.
- Variable à prédire** = risque de réhospitalisation du patient en psychiatrie dans les 2 ans suivant une hospitalisation (par **unité patient-hospitalisation**).
- 33 633 patients-hospitalisations d’intérêt** identifiés.

Méthodologie 1 : Machine Learning (ML)

- Utilisation d’algorithmes de ML basés sur les arbres de décision : Random Forest (**RF**) , Gradient Boosting (**GB**²), Extreme Gradient Boosting (**xGB**³).
- Identification des paramètres optimaux grâce à un grid search lancé sur 5 tirages.
- Mesure des performances des algorithmes, en particulier la **justesse** et l’**aire sous la courbe ROC**, après cross-validation (jeu d’entraînement : 80 %, jeu de test : 20 %).
- Explication de l’impact des variables sur le modèle grâce aux **valeurs de Shapley**⁴.

La valeur de Shapley d’une variable pour une instance donnée est la contribution de la variable dans la différence entre le score prédit pour cette instance par le modèle et la moyenne des prédictions pour toutes les instances. Plus la valeur de Shapley est élevée plus la variable impacte la prédiction vers 1, plus elle est faible plus la variable impacte la prédiction vers 0.

Méthodologie 2 : Partial Least Squares – Discriminant Analysis (PLS-DA)⁵

- Utilisation d’une **PLS-DA** afin d’identifier des profils de patients d’intérêt.
- Sélection des composantes d’intérêt après une 10-fold cross-validation.
- Production itérative d’une série de facteurs deux à deux orthogonaux caractérisés par :
 - Des « **loadings** » : l’ensemble des poids quantifiant les contributions relatives de chacune des variables explicatives à chaque composante PLS-DA.
 - Des **scores** : obtenus en multipliant les variables originales par les poids, et quantifiant l’adhésion du patient au profil identifié.

Métriques de performances

Réhospitalisation	Prédiction		
		Non	Oui
	Non	Vrai Négatif (VN)	Faux Positif (FP)
	Oui	Faux Négatif (FN)	Vrai Positif (VP)

- Justesse** : Pourcentage de prédictions correctes : $\frac{VP+VN}{VP+VN+FP+FN}$
- Aire sous la courbe ROC** : métrique intégrant la **sensibilité** (% positives bien prédites : $\frac{VP}{VP+FN}$) et la **spécificité** (% négatives bien prédites : $\frac{VN}{VN+FP}$).
- VIP** (variable importance in the projection) : importance de la variable x_i (i=1,...,p; p étant le nombre total des variables, ici p=91), dans un modèle à k composantes définie comme :
 - Estimation quantitative du **pouvoir discriminatoire de chaque variable**.
 - Somme pondérée des **corrélations au carré** entre chaque composante PLS-DA et les variables de soins originales.

Conclusion

Cette analyse a permis de mieux appréhender l’impact des variables présentes dans la base de données de l’étude sur le score de risque de réhospitalisation en psychiatrie dans les 2 ans. Les résultats confirment les attentes des cliniciens tout en identifiant des seuils plus précis, notamment grâce aux visualisations des valeurs de Shapley. En particulier, on note l’impact de la durée du séjour, du nombre d’hospitalisations précédentes, du diagnostic du séjour et de la dose de benzodiazépines. Les analyses de PLS-DA et de GB donnent des résultats concordants ce qui permet de valider nos analyses. L’absence de données sur le décès est une limite de cette étude. L’application de ces algorithmes à des données issues d’autres centres en France permettrait de les valider afin d’envisager leurs déploiements comme outils d’aide à la décision pour les médecins.

Résultats Machine Learning

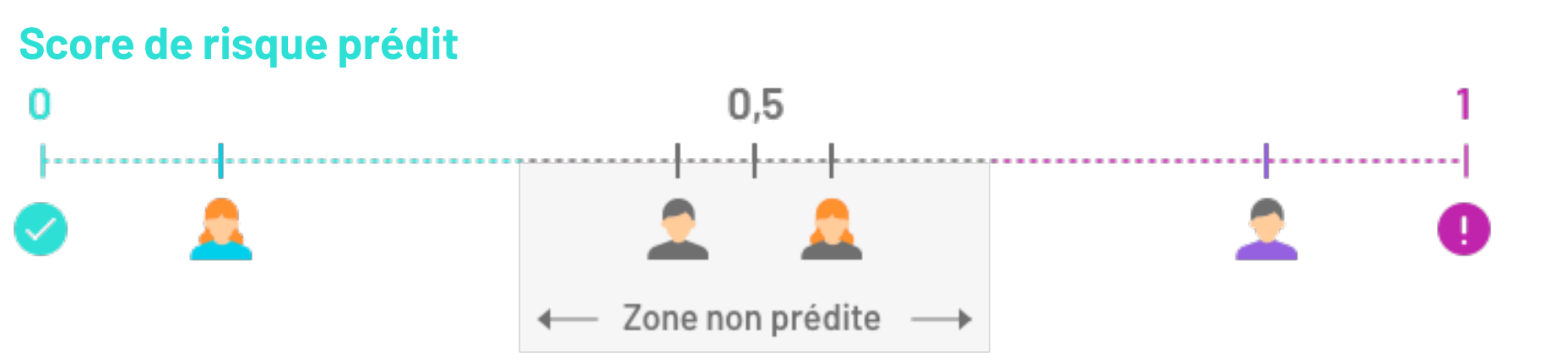
Définition des modèles

- Le **GB** était le plus performant avec 69 % de justesse et d’aire sous la courbe ROC (vs. 67 % pour le RF). Le modèle GB final était de profondeur 3, avec 500 arbres intermédiaires.

Fixation du seuil

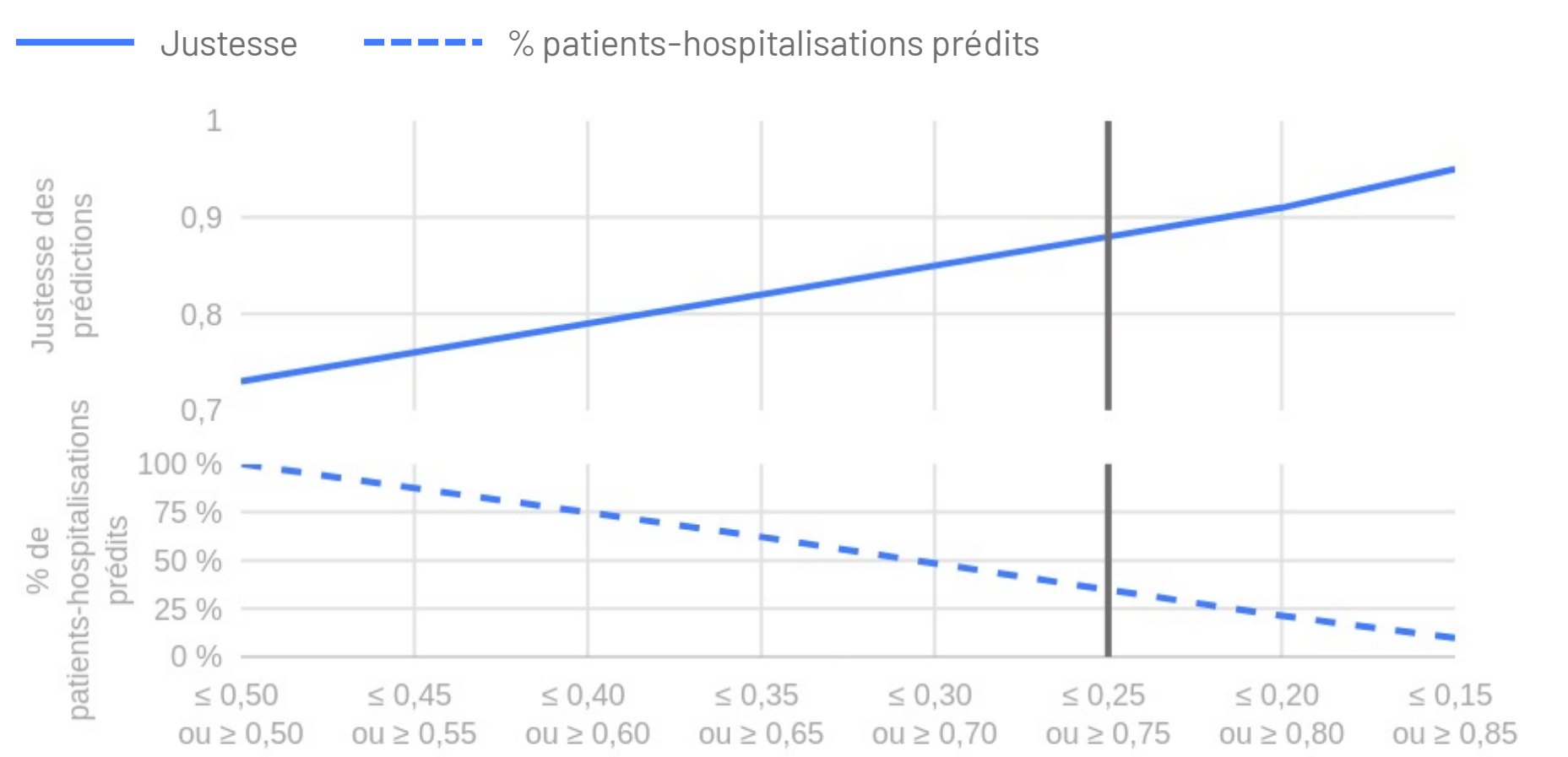
- Un score proche de 0 ou de 1 montre que l’algorithme prédictif est confiant, tandis qu’un score autour de 0,5 montre une incertitude (**Figure 1**). Il a donc été décidé de fixer une plage de valeurs au sein de laquelle ne sont pas prises en compte les prédictions estimées trop incertaines. La recherche de cette plage se fait sur le jeu de test, l’objectif étant d’obtenir de bonnes métriques de performance tout en gardant une prédiction pour un grand nombre de patients.

Figure 1. Définition des plages de prédiction : exemple sur 4 patients-hospitalisations



- Le seuil final retenu est 0,25 (Figure 2) :
 - haut risque de réhospitalisation : **score ≥ 0,75**
 - faible risque de réhospitalisation : **score ≤ 0,25**
- Prédiction pour 11 676 patients-hospitalisations (34,7 %) tout en ayant de très bonnes métriques : **88 % de justesse et d’aire sous la courbe ROC**.

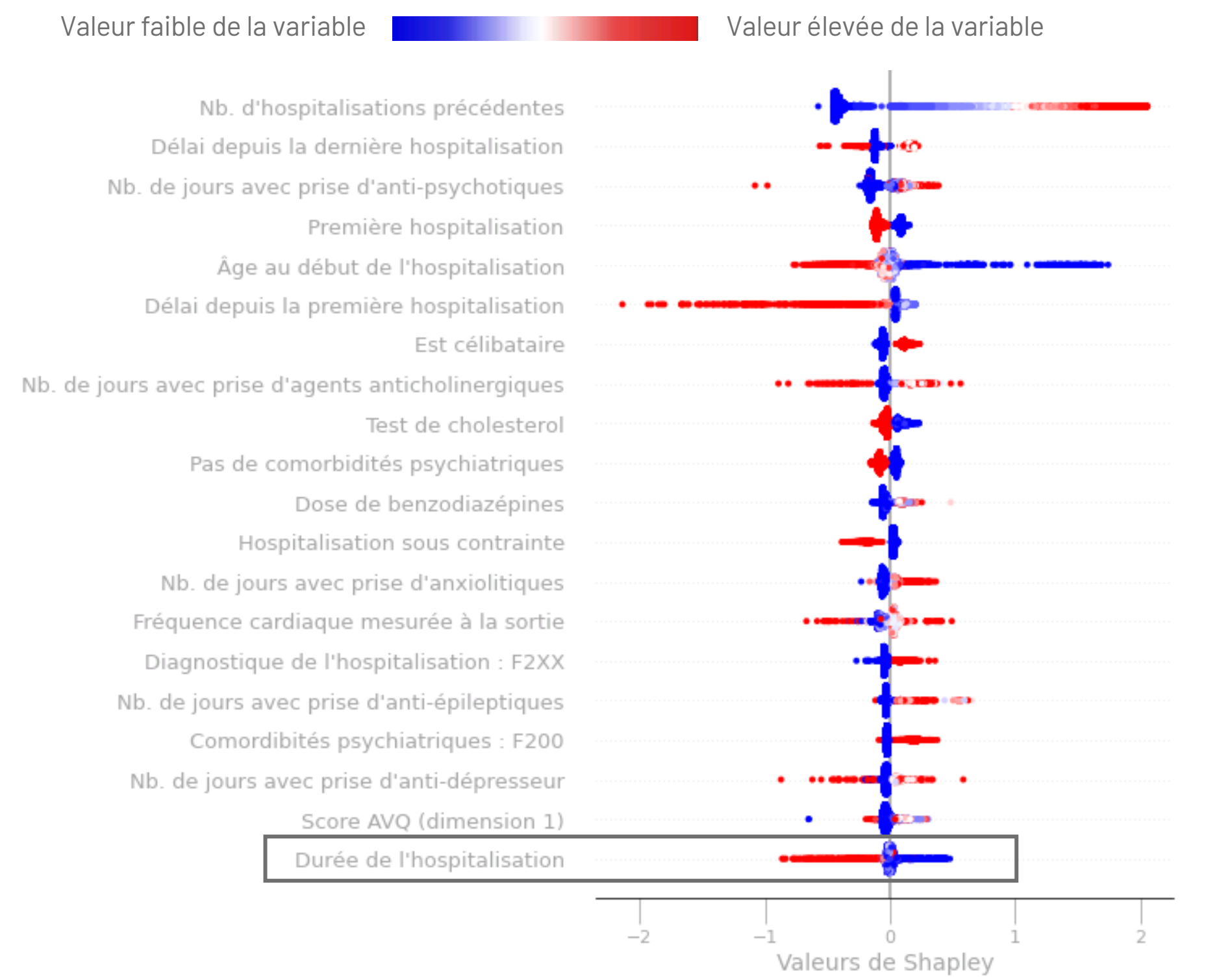
Figure 2. Choix du seuil



Résultats

- La validation clinique de l’algorithme a été réalisée grâce aux **valeurs de Shapley**.
- La **Figure 3** présente l’impact de la valeur des variables d’intérêt sur la prédiction :
 - un point par instance prédite représenté
 - la couleur correspond à la valeur de la variable
 - la position sur l’axe des abscisses correspond à l’impact de la variable sur le score prédit.

Figure 3. Valeurs de Shapley



Le fait d’avoir eu un **séjour court** fait en majorité augmenter le score prédit, tandis que le score diminue pour les patients ayant eu une **longue hospitalisation**.

- Les **Figures 4 et 5** détaillent le résultat précédent :
 - un point par instance prédite représenté
 - la valeur de Shapley de la durée du séjour ou de la fréquence cardiaque en ordonnée
 - la couleur correspond à l’âge du patient.

Être hospitalisé longtemps (> 5 jours) a un impact d’autant plus protecteur que l’âge augmente. Être hospitalisé sur un court laps de temps (< 2 jours) est d’autant plus un facteur de risque que l’âge est élevé (Figure 4).

Les valeurs de Shapley ont également permis de mettre en avant un seuil de fréquence cardiaque (~80 battements par minutes) au-delà duquel le risque de réhospitalisation est augmenté. Ce risque augmente avec l’élévation de l’âge (Figure 5).

Figure 4. Impact de la durée de séjour sur la prédiction, en fonction de l’âge

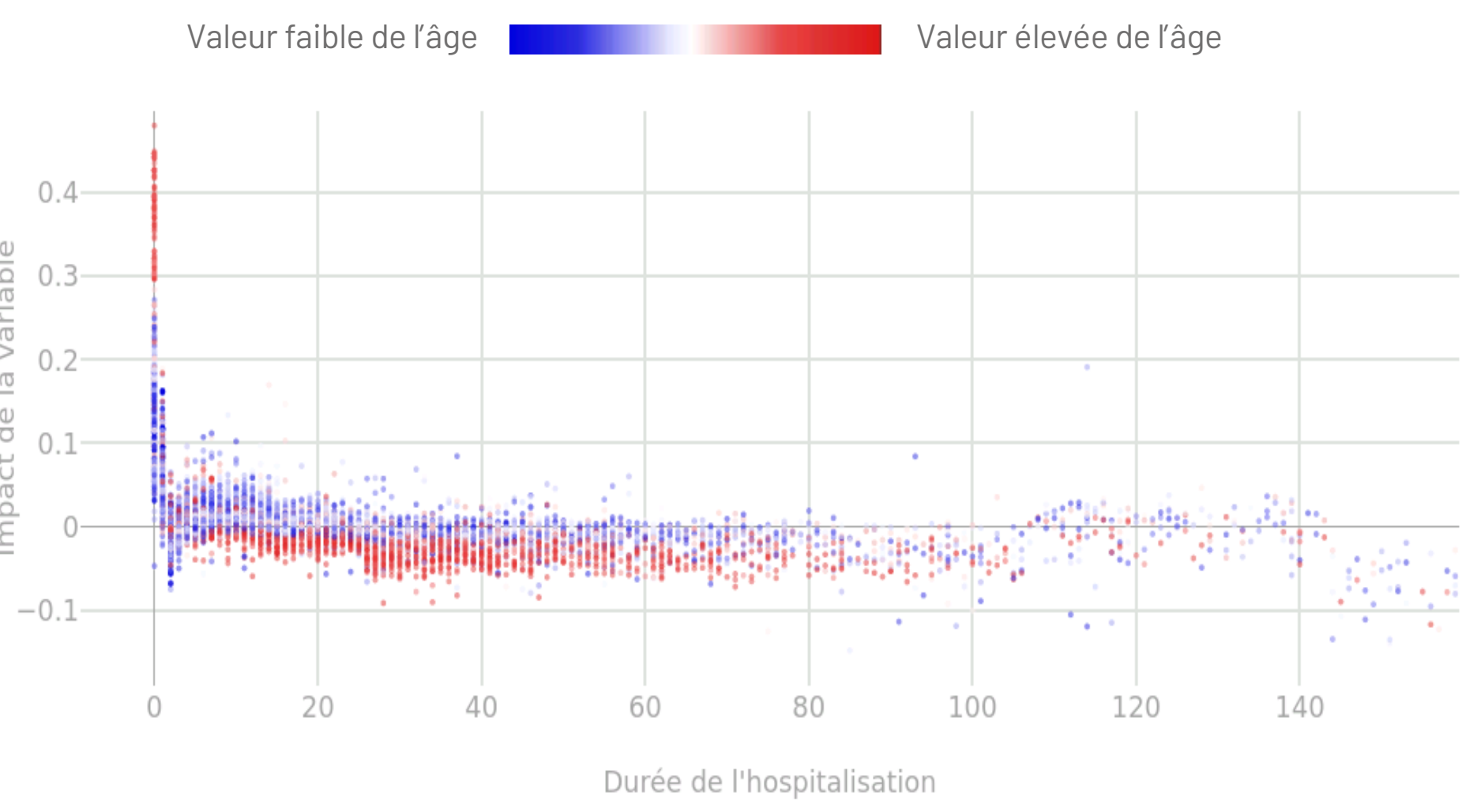
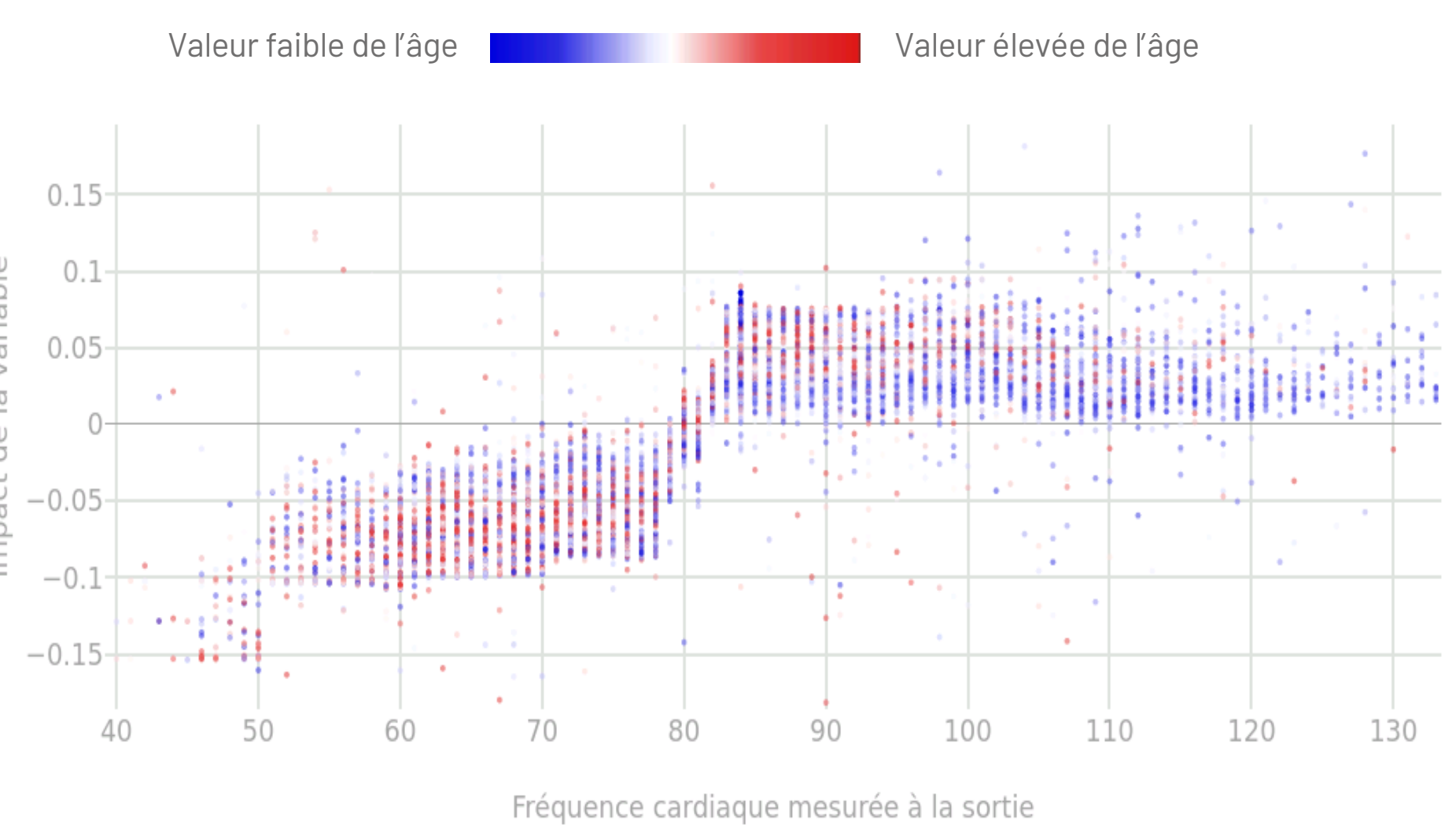


Figure 5. Impact de la fréquence cardiaque sur la prédiction, en fonction de l’âge

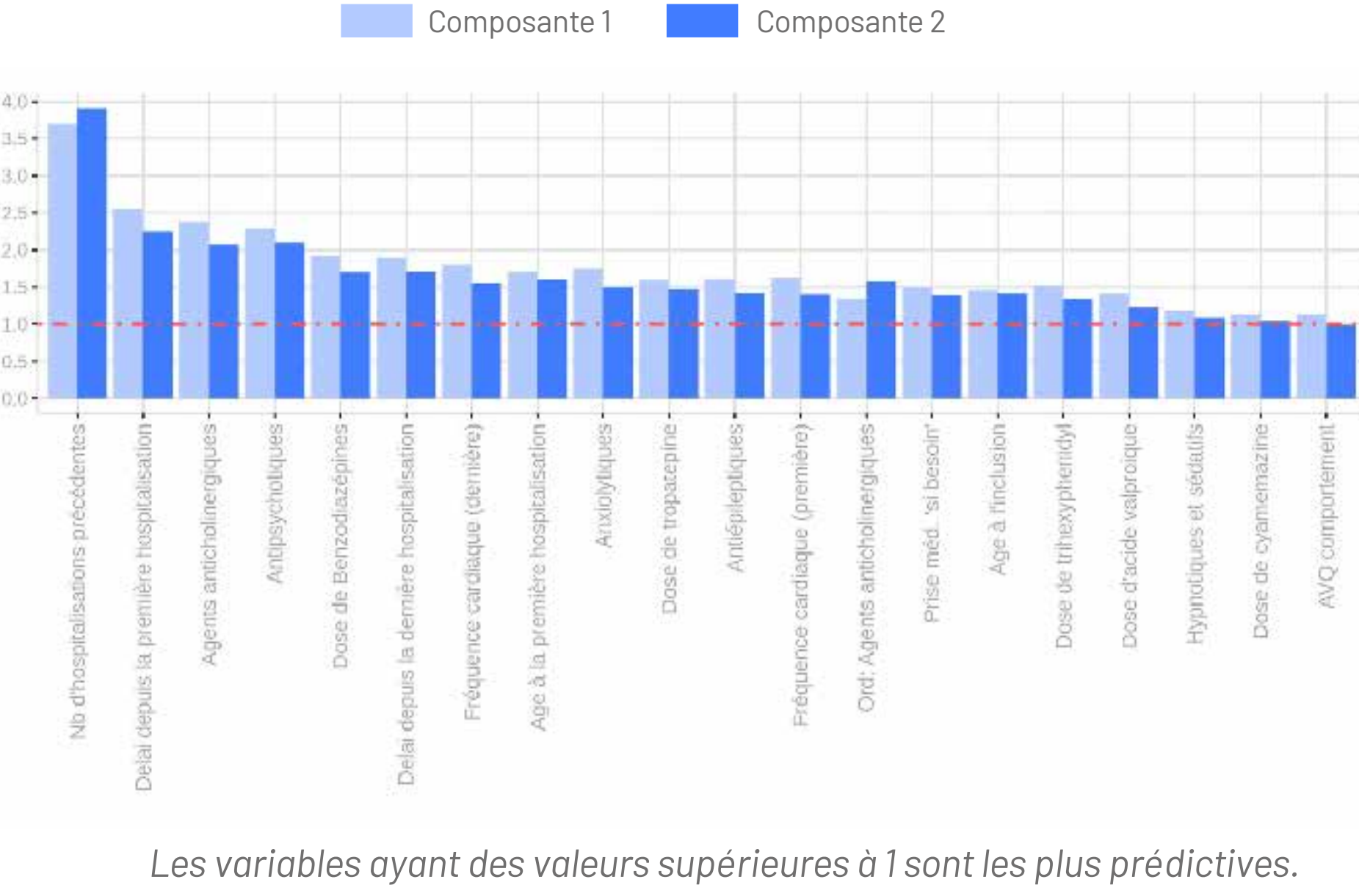


Résultats PLS-DA

Profil discriminants retenus et variables prédictives

- Très peu de discrimination observée le long des profils définis par les composantes 1 et 2 au niveau des patients (projection des scores, non montrée).
- Les variables à retenir pour la prédiction sont celles avec un **VIP > 1** (Figure 6)
 - le **nombre** d’hospitalisations précédentes,
 - le **délaï** depuis la première hospitalisation,
 - les **agents anticholinergiques**, les **antipsychotiques** et les **benzodiazépines**

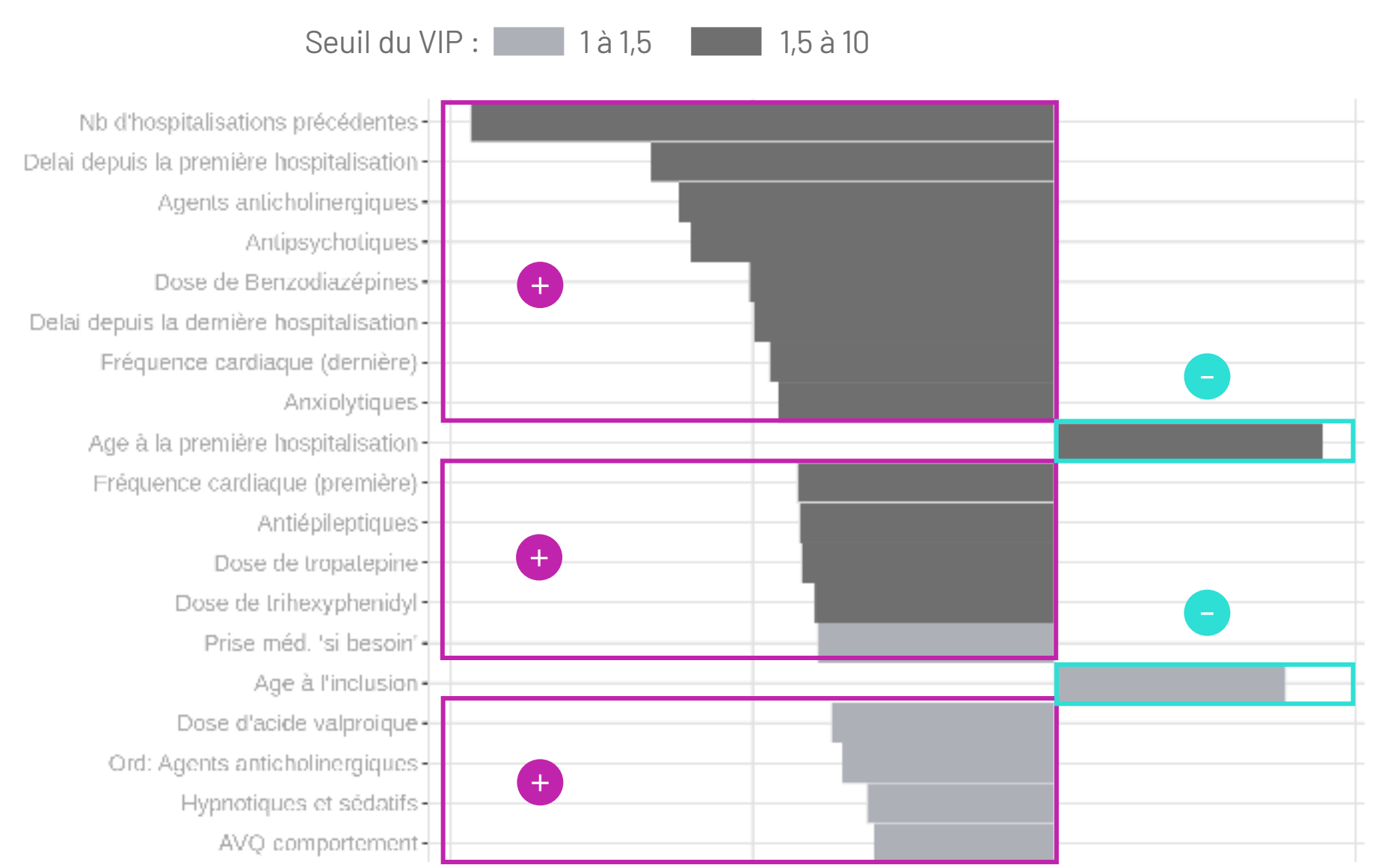
Figure 6. Importance des Variables dans la Projection (VIP) – Top 20 des variables sur 91



Les variables ayant des valeurs supérieures à 1 sont les plus prédictives.

- Mise en évidence d’un **profil des soins** expliquant 7 % de la variabilité (Composante 1, **Figure 7**) et associé à un risque diminué de réhospitalisation avec un OR = 0,76 (95 % CI : 0,75 ; 0,77, p-value < 0,0001) associé à :
 - Âge plus élevé à l’inclusion et à la première hospitalisation.
 - Plus faibles valeurs pour le nombre d’hospitalisations précédentes, le délaï depuis la dernière hospitalisation, moindre consommation d’anxiolytiques, des agents anticholinergiques et de plus faibles doses de benzodiazépines.

Figure 7. Loadings (poids) des variables de soins sur la composante 1 – Top 20 des variables ayant les valeurs absolues de poids les plus élevées (sur 91)



Références

¹RF : L. Breiman. Random forests. Machine Learning, 45(1):5–32, Oct 2001.
²GB : Friedman, J.H. (2002). <https://jerryfriedman.su.domains/ftp/stobst.pdf> Computational Statistics & Data Analysis, 38, 367–378.
³xGB : <https://doi.org/10.48550/arXiv.1603.02754>
⁴Valeurs de Shapley : Shapley, L. S., "17. A Value for n-Person Games". Contributions to the Theory of Games (AM-28), Volume II, edited by Harold William Kuhn and Albert William Tucker, Princeton: Princeton University Press, 1953, pp. 307–318.
⁵PLS-DA : Lê Cao, KA., P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics 12, 253 (2011)