# Anticipation de la sortie des patients dès l'admission aux urgences

**Laura Uhl [1,3], Eric Petit[2], Vincent Augusto [1], Youenn Alexandre [4], Fanny Jardinaud [3], Saber Aloui [5]**
[1] Mines Saint-Etienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158 LIMOS, F - 42023 Saint-Etienne France {l.uhl, augusto}@emse.fr
[2] Orange Labs, France, eric.petit@orange.com
[3] Enovacom, 521 avenue du Prado, Marseille, fanny.jardinaud@enovacom.co.
[4] Groupe Hospitalier Bretagne Sud, 5 avenue de Choiseul, Lorient, y.alexandre@ghbs.bzh
[5] CHU Angers, 4 rue Larrey, Angers, Saber.Aloui@chu-angers.fr

**Résumé.**
L'organisation de la sortie d'hospitalisation peut causer des prolongements de la durée d'hospitalisation. La littérature identifie de nombreux facteurs médico-sociaux associés à une durée de séjour longue ou à une sortie retardée. Dans cet article, nous nous attachons à trouver la combinaison minimale de variables disponibles dès l'admission du patient et permettant de prédire un besoin de support social pour préparer la sortie. Pour cela, nous proposons une approche originale basée sur l'inférence bayésienne qui permet de prendre en compte l'interdépendance des variables aléatoires et peut traiter de petits jeux de données. La combinaison de quatre variables permet une prédiction avec une exactitude de 88% contre 65% avec la sélection des neuf variables dépendantes selon les tests bivariés. Notre méthode peut être généralisée à d'autres problèmes de prédiction d'événements avec une vision *small data*.

**Mots clés : parcours patient, sortie d'hospitalisation, inférence bayésienne, prédiction**

## Introduction

La sortie des patients d'hospitalisation ou du service des urgences demande d'être préparée afin d'assurer un retour à domicile sécurisé ou un transfert sans délai d'attente. Cette préparation peut être plus ou moins longue en fonction de la situation familiale, sociale et médicale du patient. L'anticipation d'éventuelles difficultés de sortie peut permettre d'éviter des retards de sortie. Depuis les années 80, les hôpitaux s'interrogent sur les journées d'hospitalisation non pertinentes (JHNP) et le vieillissement de la population [Glass and Weiner, 1976 ; Hébert, 1984].
En 1989, Selker et al., ont montré que les hospitalisations prolongées à cause de la préparation insuffisante de la sortie ou de l'indisponibilité des ressources extérieures représentent respectivement 12% et 17% des causes de retard ainsi que 7% et 42% du nombre de JHNP. Ainsi, le manque de solutions de sortie était déjà responsable d'une part importante des JHNP, les autres causes étant liées à la prise en charge du patient (attente d'examens, de résultats, d'un avis ...) [Selker et al, 1989]. Des résultats comparables ont été mis en évidence par une étude sur les patients non-programmés d'une unité de chirurgie générale et vasculaire : 43,5% des JHNP sont dues aux difficultés de la sortie [Majeed et al., 2012]. Une étude se focalisant sur les patients de gériatrie trouve que les facteurs médico-sociaux expliquent les séjours de plus de 21 jours : sortie dans un établissement médico-social ou un service de soins médicaux et de réadaptation, stress de l'aidant, patient vivant seul, besoin d'une aide financière, présence d'une démence ou d'une pathologie grave [Toh et al., 2017]. Parmi tous ces facteurs, la sortie vers une structure de moyen ou long séjour a le plus haut *odds ratio* (9,22) suivi par le stress de l'aidant (3,85). Le travail de Safavi et al. [2019] se démarque des autres études en utilisant un réseau de neurones pour prédire si la sortie aura lieu dans les 24h et les barrières à la sortie. Parmi les 20 barrières les plus fréquentes, quatre sont non médicales : l'indisponibilité d'un service infirmier à domicile, le manque de support social, l'impossibilité pour le patient de se déplacer seul ou les faibles ressources financières du patient. Louis Simonet et al. ont étudié la prédiction de la sortie vers des soins de réhabilitation ou le retour à domicile. Ils ont collecté des données démographiques, médico-sociales et cliniques de patients des hôpitaux universitaires de Genève le 1er et le 3ème jour de leur admission. Leur

régression multilogistique atteint une AUC-ROC de 0,82 pour l'ensemble de test et de 0,77 pour l'ensemble de validation au 3ème jour d'hospitalisation. De cette analyse, ils ont construit un score de risque de non-retour à domicile qui prend en compte cinq critères : l'impossibilité du partenaire d'aider à la maison, le nombre de problèmes médicaux, le niveau d'autonomie pour la toilette, la prise de médicament et les transferts [Louis Simonet et al., 2008]. Ces études identifient bien les barrières à la sortie qui peuvent être dues à l'organisation interne de l'hôpital mais aussi à la difficulté de trouver une place en aval ou au refus de la famille ou du patient.
La difficulté de la gestion des sorties d'hospitalisation reste d'actualité et s'accentue car la pression sur les lits d'hospitalisation est telle qu'attendre plusieurs jours la sortie d'un patient maintient son hospitalisation pour des raisons non médicales ce qui embolise les hospitalisations.
Dans cet article, nous étudions la possibilité de prévoir dès l'arrivée d'un patient un besoin de support social pour préparer la sortie. Nous nous attachons à identifier les variables nécessaires pour anticiper ce besoin et quantifier l'information qu'elles apportent. Nous expliquons d'abord les méthodes utilisées puis nous détaillons les résultats obtenus avec notre jeu de données. Enfin, nous faisons part de nos conclusions et perspectives de recherche.

# 1 Méthodes

## 1.1 Description de l'outil ABIT

Nous avons utilisé un algorithme développé par Eric Petit d'Orange Labs : Adaptive Bayesian Inference Technique (ABIT)[1]. C'est un algorithme d'apprentissage automatique symbolique et probabiliste conçu dans une logique *small data* pour répondre au besoin d'inférences simples et robustes dans un contexte non-stationnaire. Il permet d'apprendre des règles d'inférence de façon automatique en mode supervisé, ceci à partir d'exemples en ligne et en présence d'incertitude. Pour ce faire, ABIT combine la théorie bayésienne avec celle du filtrage adaptatif. Le réseau bayésien se construit et se modifie au fur à mesure de l'ajout de séquences. Il implémente la solution bayésienne optimale calculant pour chaque hypothèse de sortie la probabilité à postériori exacte. Le calcul de cette plausibilité peut être réalisée sur la base de l'ensemble des variables explicatives ou seulement une partie (par exemple, les trois premiers attributs) donnant ainsi une grande souplesse dans son utilisation [Petit et Chêne, 2021].
Nous utiliserons la notion d'évidence pour exprimer cette probabilité *p* de façon plus intelligible sur une échelle logarithmique avec comme unité le déciban (dB) :

$$Ev(p) = 10log_{10}(\frac{p}{1-p}) \quad (1)$$

Dans cette échelle une probabilité de 50% correspond à 0 dB, de 1% à -20 dB et de 99% à 20 dB.

Dans la suite, nous appelons un attribut, une variable explicative notée ci-après $V_k$. Un attribut peut avoir plusieurs valeurs notés $v_{ki}$. Nous appelons *H* la variable dépendante ou variable à expliquer.
ABIT prend en entrée un ensemble de séquences toutes construites sur le modèle $(v_{1i}, v_{2i}, \cdots, v_{ni}, h_i)$ qui correspond à l'événement $(V_1 = v_{1i}) \cup (V_2 = v_{2i}) \cup \cdots \cup (V_n = v_{ni}) \cup (H = h_i)$ avec i ∈ ℕ.
Dans notre application, nous utilisons l'inférence bayésienne pour prédire la conclusion $H$ de la séquence à partir des valeurs des variables explicatives. En effet, ABIT exploite la structure de graphe sans cycle du réseau bayésien pour encoder la distribution de probabilité jointe de l'ensemble des variables du domaine et en déduire la probabilité à postériori $P(h_i|v_{1i}, v_{2i}, \cdots, v_{ni})$.

La règle de décision utilisée pour la classification est

$$h = \begin{cases} h_1 & \text{si } P(h_1|\cdots) > 0{,}5) \iff Ev(h_1) > 0 \\ h_2 & \text{sinon} \iff Ev(h_2) > 0 \end{cases}$$

[1] ABIT est enregistré à l'Agence pour la Protection des Programmes sous l'identifiant IDDN 1 .FR2 .0013 .2300214 .0005 .S6 .P7 .20208 .0009 .20700

ABIT fournit une mesure optimale de dépendance probabiliste entre la variable dépendante et une variable explicative, ceci au moyen de l'information mutuelle calculée selon la formule théorique :

$$I(H, V_k) = \sum_{h,v \in H,V_k} P(h,v) \times log \frac{P(h,v)}{P(h) \times P(v)} \quad (2)$$

L'information mutuelle mesure la quantité d'information qui peut être obtenue au sujet d'une variable aléatoire par l'observation des valeurs de l'autre variable (et de façon symétrique). L'information mutuelle capture n'importe quelle relation structurelle entre deux variables. C'est une mesure optimale de la dépendance. Elle est nulle si et seulement si les variables sont indépendantes, et croît lorsque la dépendance augmente. Dans notre cas, l'information mutuelle peut être normalisée par l'entropie de la variable cible *(H)*. Nous la comparerons à l'approche classique de mesure de l'indépendance ($\chi^2$ et Student).

### 1.2 Combinaison des attributs

Notre avons cherché à déterminer une combinaison minimale de variables en évaluant la capacité de prédiction du réseau bayésien. La méthode est la suivante : (1) Un jeu de données d'entraînement et un jeu de données de test sont constitués à partir du jeu de données initial. (2) ABIT est d'abord entraîné avec tous les attributs. (3) Puis les performances sont calculées avec le jeu de test. On obtient une évidence moyenne et le nombre de prédictions correctes et incorrectes. (4) On recommence l'entraînement et le calcul des performances mais en enlevant à chaque fois un attribut. (5) Si en supprimant un attribut, les performances sont identiques ou meilleures, cet attribut peut être ignoré. On met de côté l'attribut qui permet la meilleure performance en étant absent du jeu de données. (6) On recommence les étapes 4 et 5 jusqu'à ce qu'enlever un attribut n'améliore pas les performances. Cette méthode illustrée avec la figure 1, exploite le fait que diminuer le nombre d'attributs augmente mécaniquement l'occurrence des séquences et potentiellement le degré de confiance associé à la prédiction, cela étant contrebalancé par l'influence des attributs retirés. Il s'agit ici en quelque sorte de l'application du principe bayésien de parcimonie tendant à privilégier les modèles les plus simples.

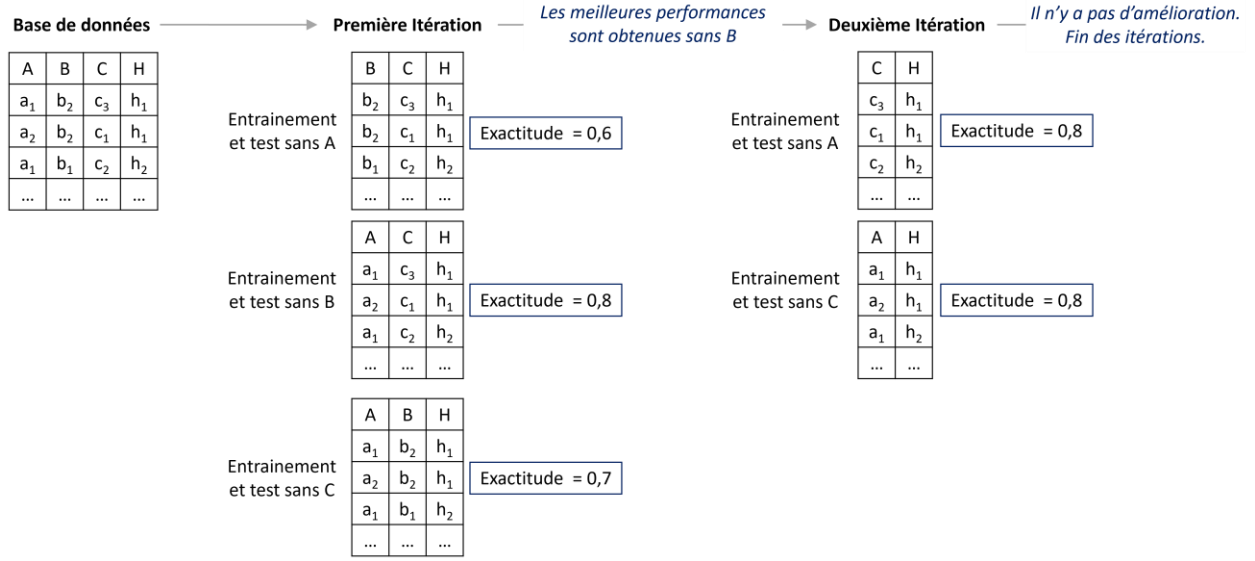


Figure 1 : Méthode itérative bayésienne pour trouver une combinaison optimale d'attributs

## 2 Résultats

### 2.1 Base de données

Les données médico-sociales sont des données sur l'environnement social d'une personne et sa capacité à réaliser les actes du quotidien. Ces données sont présentes dans les dossiers médicaux de manière éparse et lacunaire. Les informations sont dispersées dans différentes parties du Dossier Patient Informatisé (DPI) :

Tableau 1 : Analyse bivariée et test d'indépendance avec le besoin de support

| | | Besoin de support | | |
|---|---|---|---|---|
| | | Non | Oui | P-value |
| | | Nombre (%) | | |
| **Habitat** | collectif | 11 (11) | 1 (2) | 0,098 |
| | individuel | 87 (89) | 50 (98) | |
| **Altération de l'état général** | oui | 19 (19) | 26 (51) | <0,001 |
| | non | 79 (81) | 25 (49) | |
| **Autonomie** | autonome | 80 (82) | 23 (45) | <0,001 |
| | dépendant | 3 (3) | 2 (4) | |
| | partielle | 11 (11) | 12 (24) | |
| | perte | 3 (3) | 14 (27) | |
| **Addiction** | oui | 12 (12) | 11 (22) | 0,229 |
| | non | 84 (88) | 40 (78) | |
| **Démence** | oui | 4 (4) | 9 (19) | 0,010 |
| | non | 93 (96) | 39 (81) | |
| **Fragilité** | oui | 40 (43) | 47 (94) | <0,001 |
| | non | 54 (57) | 3 (6) | |
| **Entourage** | bien entouré | 63 (69) | 27 (54) | 0,164 |
| | isolé | 12 (13) | 8 (16) | |
| | peu entouré | 16 (18) | 15 (30) | |
| **Aides à domicile** | AV | 2 (2) | 1 (2) | <0,001 |
| | infirmier | 16 (18) | 19 (38) | |
| | les deux | 3 (3) | 9 (18) | |
| | non | 69 (77) | 21 (42) | |
| **Maintien à Domicile Difficile** | oui | 1 (1) | 11 (24) | <0,001 |
| | non | 94 (99) | 34 (76) | |
| **Adéquation lieu de vie** | adapté | 80 (90) | 29 (62) | <0,001 |
| | inadapté | 9 (10) | 15 (32) | |
| | insalubre | 0 (0) | 3 (6) | |
| | | **Moyenne (écart-type)** | | |
| **Age** | | 67,5 (14,4) | 75,7 (11,7) | <0,001 |
| **Nb hospitalisations précédentes** | | 0,9 (1,6) | 1,8 (2,3) | 0,006 |

certaines informations sont dans les comptes-rendus d'hospitalisation, d'autres dans les transmissions infirmières, d'autres dans le dossier social, etc. Enfin, elles sont quasiment exclusivement sous forme textuelle. Nous avons construit un jeu de données en lisant les dossiers d'une cohorte de patients.

Les critères d'inclusion sont : date de sortie en juin 2023 (un mois avant le moment du recueil de données) ; la dernière unité d'hospitalisation est oncologie ou hématologie ou rhumatologie ou maladies infectieuses ou pneumologie ou cardiologie en hospitalisation complète ; l'unité d'entrée est le service des urgences.

Les critères d'exclusion sont : âge inférieur à 18 ans ; décès du patient pendant son séjour ; patient fugueur ou sorti contre l'avis médical.

Pour déterminer les variables à recueillir, nous avons listé les facteurs discriminants selon la littérature et les facteurs identifiés à la suite de nos discussions avec le personnel hospitalier. Parmi ces facteurs, nous avons sélectionné ceux qui peuvent être connus dès l'arrivée d'un patient à l'hôpital et qui étaient mentionnés dans les dossiers patients. La variable à expliquer / prédire est le besoin d'un support médico-social. Ce besoin peut se traduire dans les dossiers par une intervention de l'équipe sociale ou la notification d'une situation nécessitant une intervention ou encore la modification du plan personnalisé de soins. Les données nécessaires à la réalisation de cette analyse ont été recueillies et saisies par un ingénieur de recherche du Groupe Hospitalier Bretagne Sud (GHBS) à l'aide de l'outil Goupile édité par InterHop. La première cohorte compte 149 séjours. Nous avons ajouté une deuxième cohorte dont les critères d'inclusion et d'exclusion

étaient identiques, excepté pour la date de sortie qui est en janvier 2024. Cette deuxième cohorte compte 29 séjours et permet de tester les performances de prédiction.

### 2.2 Analyses quantitatives

Le tableau 1 présente la répartition des valeurs en fonction du besoin de support et le résultat du test d'indépendance entre chaque variable de décision et le besoin de support social. Nous avons utilisé un test du $\chi^2$ pour les variables catégorielles et un test de Student pour les variables quantitatives (âge et nombre d'hospitalisations précédentes). Les résultats montrent que *l'addiction, l'entourage et l'habitat* sont, dans ce jeu de données, indépendants du *besoin de support*.

### 2.3 Protocole de tests pour l'approche bayésienne

Les différentes applications d'ABIT demandent d'entraîner le réseau bayésien. Afin d'avoir un réseau robuste, nous avons rejoué le jeu de données 1000 fois de suite, ceci de manière aléatoire, afin que l'ordre de présentation des données n'ait pas d'influence sur le résultat final. ABIT possède un seul hyperparamètre dont le rôle est de fixer la profondeur de mémoire de l'algorithme, assimilable à une fenêtre d'analyse glissante. À noter que l'algorithme ne stocke pas les observations, il met à jour son modèle prédictif (le réseau bayésien) à chaque nouvelle donnée, ceci de façon incrémentale et itérative. Nous avons paramétré la taille de la fenêtre à 3000, une valeur largement supérieure à la taille de l'échantillon de la base d'entrainement afin de garantir une bonne reproduction des résultats.
A partir du jeu de données récolté, nous avons exclus 39 entrées de la cohorte et 6 de la cohorte 2 qui avaient des données manquantes. Nous avons constitué deux bases d'entraînement et trois bases de test. La première base de données d'entraînement (BDE1) est constituée des 110 lignes de la cohorte 1 sans valeurs manquantes. La première base de test (BDT1) est constituée des 23 lignes de la cohorte 2 sans valeurs manquantes. Pour constituer la deuxième base d'entraînement (BDE2) et la deuxième base de test (BDT2) nous avons concaténé les deux cohortes et tiré aléatoirement 26 lignes (20%) pour former BDT2, les 80% restant forment BDE2. La troisième base de test (BDT3) rassemble toutes les lignes avec des valeurs manquantes.

### 2.4 Combinaison des attributs

Le tableau 2 présente les résultats obtenus à chaque itération. La colonne performance indique pour les colonnes encore présentes l'évidence moyenne, le nombre de prédictions correctes, le nombre de prédictions incorrectes. Par exemple, à la troisième itération les attributs *entourage* et *hospitalisations précédentes* n'étaient plus présents dans le jeu de données. En ignorant ces deux attributs, on obtient une évidence moyenne sur le jeu de test BDT1 de 0,2 dB, 12 prédictions correctes et 11 prédictions incorrectes sur 23 lignes de test. Chaque classification utilise la règle de décision énoncée plus haut. L'évidence de la plausibilité de la classe choisie sachant la séquence de test est retenue pour calculer l'évidence moyenne. Cette évidence moyenne reflète le degré de confiance moyen de l'algorithme de prédiction. Le premier critère de sélection est le nombre de prédictions correctes, l'évidence moyenne est seulement le deuxième critère.
Nous remarquons que les performances augmentent considérablement entre la première itération avec tous les attributs et la dernière avec la combinaison minimale : l'exactitude passe de 17% à 65% pour le premier jeu de données et de 35% à 88% pour le deuxième. Les résultats de chaque itération sont différents selon les jeux de données utilisés mais les résultats finaux (les attributs restants) sont très proches ; seul un attribut diffère. Il est important de noter que dans BDT1, l'*habitat* était individuel pour tous les patients. Ceci peut expliquer qu'il soit conservé avec le premier jeu de données et pas avec le deuxième. Nous pouvons donc considérer qu'une combinaison optimale est : *l'âge, la fragilité, l'altération et le nombre d'hospitalisations précédentes ou l'habitat*. Nous réduisons donc le nombre d'attributs de 12 à 4. Ce qui signifie qu'avec très peu d'information, le besoin d'un accompagnement social peut être anticipé.

Tableau 2 : résultats de la méthode itérative de sélection de la meilleure combinaison d'attributs
La colonne performance indique l'évidence moyenne en dB; le nombre de prédictions correctes ; le nombre de prédictions incorrectes.

| | BDE1 et BDT1 | | BDE2 et BDT2 | |
|---|---|---|---|---|
| **Itération** | **Attribut supprimé** | **Performance** | **Attribut supprimé** | **Performance** |
| 1 | aucun | -11 ; 4 ; 19 | aucun | -9 ; 9 ; 17 |
| 2 | entourage | -2 ; 10 ; 13 | entourage | -4,5 ; 11 ; 15 |
| 3 | hospi. prec. | 0,2 ; 12 ; 11 | autonomie | -2,1 ; 15 ; 11 |
| 4 | aides à domicile | 2 ; 13 ; 10 | madd | 8,8 ; 17 ; 9 |
| 5 | lieu de vie | 3,5 ; 14 ; 9 | aides à domicile | 10,6 ; 19 ; 7 |
| 6 | autonomie | 5,6 ; 14 ; 9 | démence | 12,7 ; 21 ; 5 |
| 7 | démence | 6,4 ; 15 ; 8 | habitat | 13,6 ; 22 ;4 |
| 8 | madd | 7,7 ; 15 ; 8 | lieu de vie | 14,1 ; 22 ; 4 |
| 9 | addiction | 10,4 ; 15 ; 8 | addiction | 7,1 ; 23 ; 3 |
| Attributs restants | âge, habitat, altération, fragilité | | âge, hospi. prec., altération, fragilité | |

### 2.5 Prédiction du besoin de support

Les parties précédentes nous ont permis d'identifier les facteurs les plus associés à un besoin de support. ABIT peut aussi être utilisé pour prédire le besoin de support à partir d'une séquence d'événements. Nous avons utilisé les deux bases d'entrainement, leur base de test respective et la base avec les valeurs manquantes (BDT3). Le tableau 3 présente les résultats de ces différentes expériences. La base de test peut contenir des relations nouvelles (absentes du jeu d'entraînement). Dans ce cas ABIT ne fournit pas de prédiction (plausibilité nulle). C'est pourquoi, nous indiquons dans le tableau, le pourcentage de lignes qu'ABIT a pu classifier. Quatre type de séquences différents ont été testés : Séquence 1 : tous les attributs ; Séquence 2 : les attributs ressortant comme corrélés au besoin de support selon le test d'indépendance statistique (*âge, hospitalisations précédentes, altération de l'état général, autonomie, démence, fragilité, aides à domicile, adéquation du lieu de vie, MADD*) ; Séquence 3 : la combinaison idéale d'attributs déduite avec BDE1 *(l'âge, l'habitat, l'altération, la fragilité*) ; Séquence 4 : la combinaison idéale d'attributs déduite avec BDE2 (*l'âge, le nombre d'hospitalisations précédentes, l'altération, la fragilité*).

Notre première observation est que la sélection des variables est effectivement un facteur d'amélioration des performances de prédiction. Les performances de prédiction sont meilleures avec les combinaisons sélectionnées. Avec la BDT3, les séquences 1 et 2 semblent prédire aussi bien le besoin de support que les séquences 3 et 4 mais il faut prendre en compte le nombre de lignes classifiées qui est fortement réduit. Le fait d'avoir un taux d'entrées classifiées élevé pour les séquences 3 et 4 (95% voire 100%) indique qu'une sélection réduite d'attributs permet de conclure sur plus de cas. Nous remarquons que la spécificité est bien meilleure que la sensibilité. Ceci s'explique par le déséquilibre entre les deux classes lors de l'apprentissage, la classe *non* est plus représentée que la classe *oui* (66% contre 34%). Un focus sur les séquences 3 et 4 montre que la sensibilité est globalement plus faible que la précision. Pour améliorer l'algorithme, il faudrait donc se concentrer sur la réduction du nombre de faux négatifs. Une nouvelle fois, les valeurs de performance sont différentes entre la BDT1 et la BDT2 sans que cela change les interprétations. En conclusion, ces premiers résultats laissent suggérer qu'il est possible de prédire le besoin d'un support social avec quelques variables que l'on peut recueillir à l'arrivée du patient aux urgences ou dans le service d'hospitalisation.

### 2.6 Comparaison des différentes approches

En plus des méthodes présentées précédemment, nous avons construit un arbre de décision avec la BDE1. Le tableau 4 permet de comparer l'interdépendance de chaque variable avec le besoin de support selon le

test du $\chi^2$ et de Student, l'information mutuelle de chaque variable par rapport au besoin de support et le coefficient de Gini selon l'arbre de décision. Les variables sont ordonnées en fonction de l'importance de l'association (information mutuelle décroissante, valeur p croissante, et coefficient de Gini décroissant).

Tableau 3 : Résultats de prédiction du besoin de support social par ABIT

Exactitude = (VP+VN)/(P+N) ; Précision = VP/(VP + FP) ; Sensibilité = TP/P ; F1-score = la moyenne harmonique (précision, sensibilité) avec VP = Vrais Positifs, VN = Vrais Négatifs, FP = Faux Positifs, P = Positifs = VP+FN, N = Négatifs = VN+FP.

| | | Séquence | Exactitude | F1 score | Précision | Sensibilité | Spécificité | % entrées classifiées |
|---|---|---|---|---|---|---|---|---|
| BDE1 | BDT1 | 1 | 0,36 | 0,22 | 0,20 | 0,25 | 0,43 | 47 |
| | | 2 | 0,56 | 0,33 | 0,33 | 0,33 | 0,67 | 78 |
| | | 3 | 0,65 | 0,56 | 0,71 | 0,45 | 0,83 | 100 |
| | | 4 | 0,64 | 0,50 | 0,67 | 0,40 | 0,83 | 95 |
| | BDT3 | 1 | 0,70 | 0,25 | 0,20 | 0,33 | 0,76 | 44 |
| | | 2 | 0,65 | 0,20 | 0,20 | 0,20 | 0,78 | 51 |
| | | 3 | 0,77 | 0,62 | 0,73 | 0,53 | 0,89 | 95 |
| | | 4 | 0,63 | 0,50 | 0,47 | 0,53 | 0,68 | 95 |
| BDE2 | BDT2 | 1 | 0,77 | 0,0 | 0,0 | 0,0 | 0,91 | 50 |
| | | 2 | 0,65 | 0,25 | 0,17 | 0,50 | 0,67 | 65 |
| | | 3 | 0,85 | 0,71 | 0,71 | 0,71 | 0,89 | 100 |
| | | 4 | 0,88 | 0,77 | 0,83 | 0,71 | 0,95 | 100 |
| | BDT3 | 1 | 0,59 | 0,31 | 0,25 | 0,40 | 0,65 | 48 |
| | | 2 | 0,58 | 0,29 | 0,25 | 0,33 | 0,67 | 53 |
| | | 3 | 0,77 | 0,62 | 0,73 | 0,53 | 0,89 | 95 |
| | | 4 | 0,60 | 0,45 | 0,44 | 0,47 | 0,68 | 95 |

Tableau 4 : Attributs classés selon trois coefficients d'association

Les variables en *italique* sont celles de la combinaison minimale.

Les variables en bleu sont les variables avec une p-value supérieure à 0,05 et en vert avec un coefficient de Gini nul.

| | Information mutuelle | Test d'indépendance | Coefficient de Gini |
|---|---|---|---|
| 1 | *Fragilité* | *Fragilité* | *Fragilité* |
| 2 | Autonomie | Autonomie | *Âge* |
| 3 | MADD | MADD | *Hospi. prec.* |
| 4 | Lieu de vie | *Altération* | *Altération* |
| 5 | Aides à domicile | Lieu de vie | Aides à domicile |
| 6 | *Hospi. prec.* | Aides à domicile | Lieu de vie |
| 7 | *Altération* | *Âge* | Autonomie |
| 8 | Démence | *Hospi. prec.* | *Habitat* |
| 9 | *Âge* | Démence | Entourage |
| 10 | Entourage | *Habitat* | Addiction |
| 11 | Addiction | Entourage | Démence |
| 12 | *Habitat* | Addiction | MADD |

Il est intéressant de noter que la combinaison idéale ne retient pas que les variables fortement corrélées au besoin de support. Cela montre que l'étude de l'association d'attributs deux à deux (information mutuelle et $\chi^2$) ne permet pas de rendre compte parfaitement de l'influence que peut avoir une combinaison de variables. Notre approche de sélection d'une combinaison d'attributs a donc un réel intérêt car elle prend en compte des interdépendances entre les variables, ignorées par les analyses bivariées. Notons que les variables de

la combinaison idéale, sont les quatre variables les plus importantes selon l'arbre de décision (excepté l'*habitat* à cause du biais de BDT1). La sélection d'attributs est donc cohérente avec la structure de l'arbre de décision : les attributs les plus importants correspondent aux premières branches.
Si nous comparons l'information mutuelle et le test d'indépendance, nous remarquons que l'ordre des variables n'est pas tout à fait identique mais que chaque attribut reste dans le même "quart" (excepté l'altération). Par exemple, les trois attributs les plus associés sont *la fragilité, l'autonomie et le MADD* et les trois les moins associés sont *l'entourage, l'habitat et l'addiction* pour les deux tests.

## 3 Discussion et conclusion

Les résultats des deux tests bivariés (l'information mutuelle et le test d'indépendance) sont cohérents entre eux. Cependant, ils ne permettent pas de capter toutes les interdépendances des variables aléatoires. Nous avons comparé l'approche classique avec l'approche bayésienne, rarement mise en œuvre, et construit une méthode itérative basée sur l'inférence bayésienne pour sélectionner la combinaison d'attributs permettant de meilleures performances de prédiction que la sélection d'attributs par les méthodes classiques de test d'indépendance. Nous l'avons implémentée avec l'algorithme ABIT car il peut fonctionner sur de petits jeux de données. La méthode peut être appliquée avec d'autres algorithmes de prédiction et elle est généralisable à n'importe quel jeu de données. Nous avons réussi à sélectionner parmi douze variables une combinaison optimale de quatre variables. Avec peu de données nous avons été capable d'anticiper un besoin de support social pour la préparation de la sortie. Ces performances de prédiction atteignent une exactitude de 88%. Ces premiers résultats laissent suggérer qu'il est possible de prédire le besoin d'un support social avec quelques variables que l'on peut recueillir à l'arrivée du patient aux urgences ou dans le service d'hospitalisation.
La limite principale de la méthode pour déterminer une combinaison minimale est la constitution du jeu d'entraînement et surtout de test. En effet, les performances de prédiction en dépendent fortement. Il est donc nécessaire de choisir aléatoirement ces deux jeux de données. Or avec une petite base de données comme la nôtre, des biais peuvent exister dans la construction de ces deux jeux de données. Une base de données plus importante serait donc nécessaire pour confirmer nos résultats. De manière général, nos résultats numériques seraient plus robustes avec plus de données. Toutefois, la taille de la base de données pourra rester raisonnable (un millier de lignes) car nos outils ne nécessitent pas d'approche *big data*. Une deuxième limite concerne le recueil via la consultation des dossiers patients. Celui-ci n'est pas optimal car les données y sont parcellaires et les informations recherchées doivent être parfois déduites. Un recueil prospectif des données éviterait ce défaut.

Dans cet article, nous nous sommes attachés à prédire le besoin de support social car une prise en charge sociale tardive induit des retards de sortie. Pour aller plus loin sur l'anticipation de la sortie, il serait intéressant d'étudier la prédiction de la destination de sortie pour un patient en général. Cela permet d'anticiper le besoin de ressources à la sortie (un lit en SMR, aides à domicile ...). Cela pourrait faire l'objet d'une autre étude où les données médico-sociales seraient précieuses et où les méthodes présentées ici pourraient être appliquées.

## Références


Eric Petit et Denis Chêne (2021). Navigation adaptative dans les systèmes interactifs: paradigme et solution. Dans : *Proceedings of the 17th "Ergonomie et Informatique Avancée" Conference*. Association for Computing Machinery, New York.
Glass, R. I., et Weiner, M. S. (1976). Seeking a social disposition for the medical patient: CAAST, a simple and objective clinical index. *Medical Care*, *14*(7), 637-641.
Hébert, R. (1984). Le vieillard à l'hôpital général: Du dumping aux lits bloqués. *Can Fam Physician*, 30, 2331-2337.
Louis Simonet, M. et al. (2008). A predictive score to identify hospitalized patients' risk of discharge to a post-acute care facility. *BMC Health Serv Res*, 8(154).
Majeed, M. U. et al. (2012). Delay in discharge and its impact on unnecessary hospital bed occupancy. BMC Health Serv Res,12 (410).
Safavi, K. C. et al. (2019). Development and Validation of a Machine Learning Model to Aid Discharge Processes for Inpatient Surgical Care. *JAMA Network Open*, 2 (12).
Toh, H. J., Lim, Z. Y., Yap, P. et Tang, T. (2017). Factors associated with prolonged length of stay in older patients. *Singapore Med J*, 58, 134–138.