APPLICATION EMPIRIQUE DE L'ANALYSE TEXTUELLE, DE L'ANALYSE DES CORRESPONDANCES ET DE LA CLASSIFICATION SUR UN EXEMPLE DE PRISE EN CHARGE DES RESIDENTS EN INSTITUTION

T Delespierre PhD, EA 4047, UFR des Sciences de la Santé Simone Veil, Université de Versailles St Quentin Montigny le Bretonneux - Institut du Bien Vieillir Korian, 75008 France
P Denormandie MD, PhD, Korian SA, Institut du Bien Vieillir Korian, 75008 Paris - France
L Josseran MD, PhD, EA 4047, UFR des Sciences de la Santé Simone Veil, Université de Versailles St Quentin 78 Montigny le Bretonneux - - Hôpital Raymond-Poincaré, APHP, 92380 Garches,

ABSTRACT

Purpose: Korian is a private European group specialized in social and medical accommodation and support for elderly and dependent people. The group has health records built through a professional data warehouse (DWH). This big data base contains lots of health unstructured data (nursing narratives) as well as layered and indexed data which allows us to connect every health report to his or her owner. We choose to farm physiotherapy care as a first use of unstructured data.

Method: The objective of this paper is to show how using SQL queries, text mining methods, multiple correspondence analysis and hierarchical clustering allows us to better understand this data and describe the residents' care.

Results: We analyzed 4051 health records coming from 1015 residents of 127 nursing homes on a 6 months period. By combining these techniques we demonstrate that different residents' clusters correspond to different health situations, having more or less mobility, and, different physiotherapy care, working more or less on limbs disabilities.

Conclusion: Other health care reporting systems could be analyzed using these techniques.

RESUME

Objectif: Korian est un groupe européen privé spécialisé dans l'accueil et l'hébergement des personnes âgées dépendantes. La société possède des dossiers de santé enregistrés dans un data warehouse (DWH) professionnel. Cette grande base de données contient de nombreuses informations de santé non structurées saisies par l'infirmière référente, au décours des soins, ainsi que des données indexées et organisées par strates qui permettent de relier chaque enregistrement à son résident. Nous avons choisi ici d'exploiter les traitements de kinésithérapie en tant que premier exemple d'utilisation de données non structurées.

Méthode: L'objectif de ce papier est de montrer comment par requêtes SQL, analyse textuelle, analyse en correspondances multiples (ACM) et classification hiérarchique, nous pouvons mieux comprendre ces informations de santé et décrire la prise en charge des résidents.

Résultats: Nous avons analysé 4051 observations provenant de 1015 résidents de 127 EHPAD (Etablissement pour Personnes Agées Dépendantes) sur une période de 6 mois. En combinant ces techniques nous démontrons qu'à des clusters différents de résidents correspondent, des situations de santé différentes, plus ou moins de mobilité, et une prise en charge kiné différente, mobiliser plus ou moins les membres.

Conclusion: D'autres systèmes d'information de santé pourraient être analysés en utilisant ces techniques.

MOTS-CLES

EHPAD, data warehouse, requêtes SQL, analyse textuelle, textmining, analyse des correspondances, classification hiérarchique, ACM, CAH, HCPC

THEMATIQUE PRINCIPALE

Analyse des données et data mining

THEMATIQUE SECONDAIRE

Apprentissage et classification

1. INTRODUCTION

La problématique d'analyse des comptes-rendus de soins infirmiers est ancienne [1]. Le développement des DWH et du cloud computing d'une part [2] et des techniques d'extraction des données [3] d'autre part ont changé la donne. Depuis près d'un an l'Institut Korian travaille en lien avec l'UFR des sciences de la santé de l'Université de Versailles St Quentin au développement d'une base de données alimentée par le dossier résident (la Base du Bien Vieillir, BBV). C'est en effet la première fois que des données issues de structures médico-sociales d'une telle volumétrie seront enregistrées au fil de l'eau, sur le long terme pour décrire l'organisation des soins.

A l'heure actuelle, le SI Korian comprend deux grands types d'informations : des données de formats fixes comme les caractéristiques sociodémographiques du résident, ses pathologies et comorbidités, ses facteurs de risque, enfin son parcours de vie: entrées – sorties, hospitalisation(s) et décès; et des données au format textuel, jusqu'à 4000 caractères, caractéristiques du parcours de soins, des facteurs de risques ou des goûts et habitudes du résident.

Les données textuelles sont utilisées aujourd'hui uniquement dans un cadre professionnel en tant qu'outil de liaison et non pas comme un véritable outil d'aide à la décision thérapeutique. L'objectif de ce travail, en étudiant la donnée textuelle *Traitement kiné* [4], est de montrer que ces données apportent un réel enrichissement dans la vision des états de santé et des soins observés de la population des résidents en EHPAD.

2. MATERIEL ET METHODE

Population:

Les résidents présents et vivants au 30 septembre 2013 avec au moins une observation *Traitement kiné* durant les 6 mois précédents.

Traitement de l'information :

1- Extraction et troncature des observations à 300 caractères par requête SQL (Standard Query Language).

2- Nettoyage et simplification des observations : suppression des dates, signes de ponctuation et des mots vides de sens dans notre contexte d'étude (oui, non etc…)

3- Analyse, tri et comptage des principales expressions par la fonction SQL LIKE et des wildcards avec %. Ces deux fonctions permettent de sélectionner des chaînes de caractères dans un texte. Construction de variables reflets de ces expressions. Par exemple une observation trouvée contenant le mot *marche* augmentait de une unité la variable *marche.*

4- Agrégation des observations non classées à l'étape 3 et des observations de plus de 30 caractères et construction d'un corpus (collection de documents textes [5]).

5- Analyse du corpus basée sur le stemming et la lemmatisation dans R avec le package de textmining *tm* [6]

6- Typologie des résidents par fusion avec d'autres tables du système d'information : âge au 30/09/2013, âge à l'entrée, sexe, nombre de chutes, gravité des chutes, antécédents et pathologies classés en 10 domaines (modèle PATHOS élaboré par la CNAMTS et le SNGC [7]).

7- Analyse des résidents suivant cette typologie par une ACP sur variables quantitatives, puis d'une ACM sur variables qualitatives suivie d'une classification ascendante hiérarchique (CAH) dans R grâce à la fonction HCPC (Hierarchical Clustering on Principal Components [8] du package FactoMineR [9].

8- Analyse descriptive des clusters obtenus à l'étape 7 puis qualitative par comparaison textuelle de 2 clusters de traitements kiné, résultats des étapes 3 à 5 à l'aide des packages R snowball [10] et wordcloud [11] pour le nuage de points et RColorBrewer [12].

Le choix des comparaisons entre les groupes sera fait en fonction de la pertinence clinique (résultat 3.1) ou des effectifs observés de chacun des groupes constitués (résultat 3.2).

Approche proposée :

Le modèle PATHOS mesure les niveaux de soins nécessaires à la prise en charge des résidents dans 8 postes de ressources représentant les huit « acteurs » des soins : médecin, psychiatre, infirmier, rééducation, psychothérapie, biologie, imagerie et pharmacie [7]. Nous y avons ajouté la gravité des chutes. Enfin, le nombre de chutes permet également de classer nos résidents par risque de chutes ultérieures [13]. Il s'agit de montrer qu'à clusters différents -suivant le profil PATHOS et le risque de chutes- correspondent des observations *Traitements kiné* différentes (résultat 3.2).

3. RESULTATS

3.1 Analyse descriptive des clusters : une extraction de la table des antécédents et pathologies, transformation en données catégorielles et fusion avec nos observations résidents ont permis de pouvoir utiliser la fonction HCPC [8] [9] après une ACM. Une seule variable qualitative illustrative : les clés résidents et trois variables quantitatives illustratives : le nombre de chutes avant le 30/09/2013 et les âges à l'entrée puis au 30/09/2013. La moyenne d'âge est de 87,7 ans, la moyenne de séjour de 2 ans et demi, le ratio femmes/hommes de 3,7. Un nombre moyen de chutes de 3,7 avec pour respectivement 79% et 85% des résidents au moins un appel du médecin et/ou une hospitalisation (voir table 1 pour la liste des variables). Le découpage proposé automatiquement en 6 classes (voir figure 1) met en évidence un fort saut inertiel entre 6 et 7 classes d'où ce choix de cut-off.

Toutes les variables qualitatives sont discriminantes pour la classification et plus particulièrement les variables région, département et EHPAD. Nous pouvons rechercher à décrire ces 6 groupes de résidents. La dernière variable du modèle HCPC contient l'information cluster. Les effectifs des 6 clusters sont les suivants : 229 pour le premier, 355 pour le second, 58 pour le troisième, 208 pour le quatrième, 147 pour le cinquième et 18 pour le dernier.

L'inertie des 2 premières composantes principales est faible (4,94%) alors qu'un calcul détaillé donne une inertie de 7,8. Les deux premiers axes fournissent donc 63% de l'inertie totale.

Une coloration des classes dans le graphique ci-dessous fait apparaître le cluster 6 très excentré. Après examen des individus, à part un seul individu, les 17 autres n'ont aucune pathologie saisie alors qu'ils possèdent tous de nombreux antécédents.

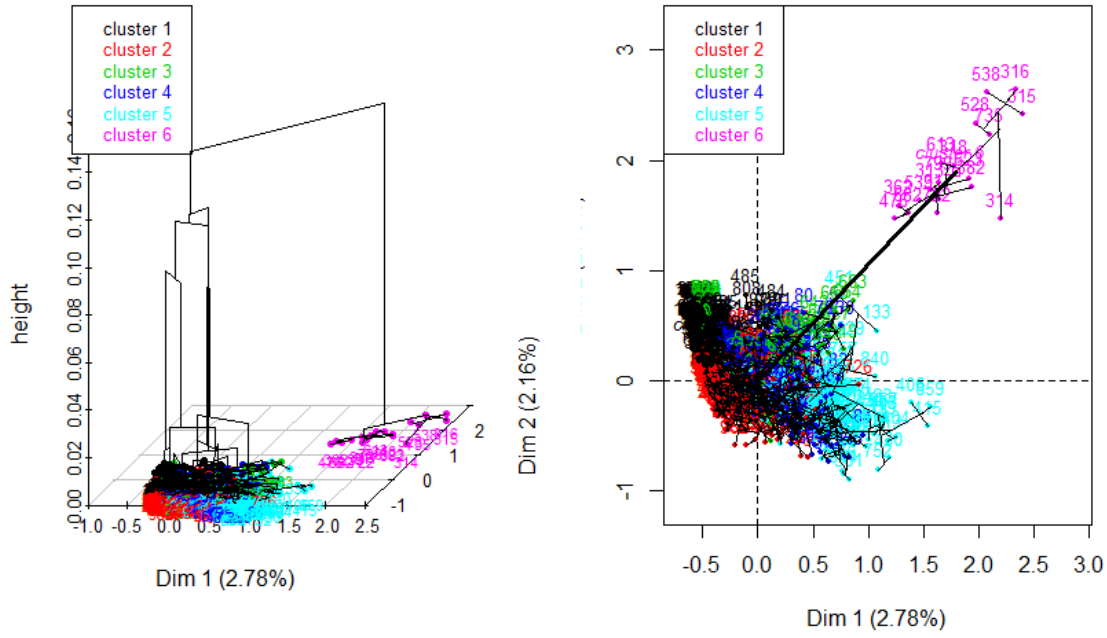

Figure 1 : Résultats de la fonction HCPC sur l'échantillon des résidents avec *Traitement kiné*

Une comparaison quantitative du cluster 6 avec les cinq autres clusters confirme ce premier examen (Tableau 1).

	cluster 6	clusters 12345
nb moyen d'atcd médicaux suivant PATHOS		
cardio-vasculaire	2,4	0,67
neuropsychiatriques	4,5	0,82
pleuropulmonaires	0,61	0,12
infections	0,61	0,07
dermatologie	0,5	0,05
ostéo-articulaire	2,11	0,53
digestif	2,72	0,39
endocrinien	0,78	0,11
uro-néphrologique	1,44	0,14
autres	1,94	0,44
nb moyen de pathologies suivant PATHOS		
cardio-vasculaire	0,17	0,67
neuropsychiatriques	0,06	1,46
pleuropulmonaire	0,06	0,14
infections	0	0,01
dermatologie	0	0,1
ostéo-articulaire	0,06	0,54
digestif	0,06	0,55
endocrinien	0	0,24
uro-néphrologique	0,06	0,28
autres	0,06	0,63
ratio femmes/hommes	5	3,62
âge moyen à l'entrée	86,67	85,15
âge moyen au 30/09/2013	88,61	87,69
nb moyen de chutes	5,67	3,63
ratio chutes + appel médecin	0,61	0,79
ratio chutes + hospitalisation	0,89	0,85

Tableau 1 : Comparaison quantitative du cluster 6 avec le reste de l'échantillon des résidents avec *Traitement kiné*.

Le cluster 6 est plus féminin (5 vs 3,62), les résidents sont plus âgés à l'entrée (+ 1,54 années), mais là depuis moins longtemps (délai entre le 30/09/2013 et la date d'entrée : 1,94 vs 2,54 années) et tombent plus (5,67 vs 3,63). Les valeurs des 3 variables quantitatives (en grisé) bien qu'illustratives sont cohérentes avec la classification et mettent en exergue la fragilité de ce groupe.

3.2 Analyse qualitative par comparaison textuelle de 2 clusters de traitements kiné : Nous choisissons de comparer le cluster 5 qui comprend 147 individus et est légèrement excentré aux quatre premiers d'effectifs globaux 886 individus.

Une analyse textuelle des corpus des clusters 5 (à gauche) et 1234 (à droite) donne les résultats suivants avec une palette qualitative de couleurs. Dans ces deux nuages la taille des mots est proportionnelle à leurs effectifs et permet de qualifier les traitements kiné.

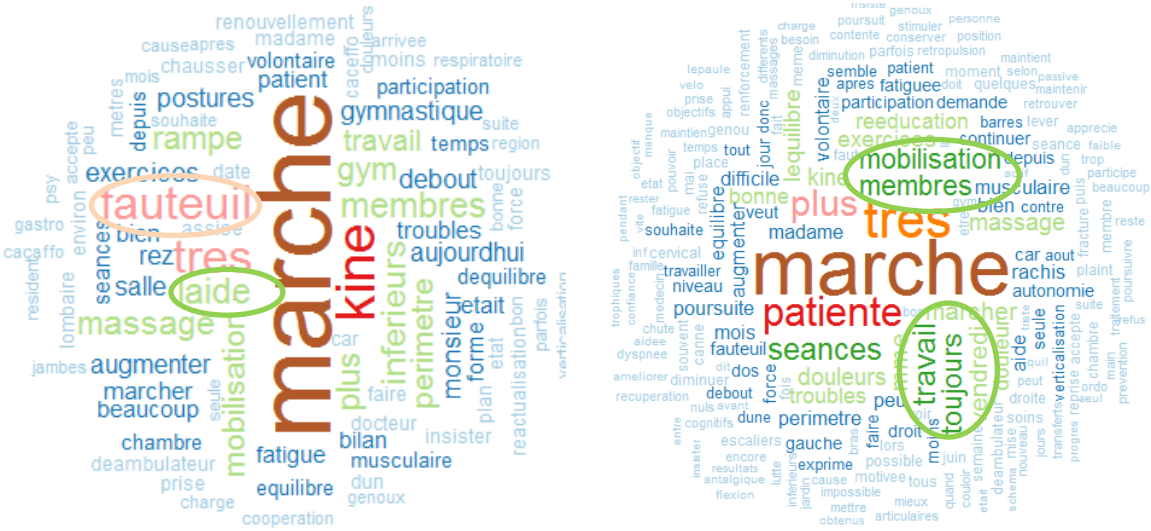

Figure 2 : Wordclouds des *Traitement kiné* des clusters 5 (à gauche) et 1234 (à droite)

On retrouve la marche en première position, fauteuil et aide (laide) ont disparu dans les clusters1234 et travail et mobilisation des membres ont pris de l'importance. Cette population semble plus active, plus mobile que la précédente. Une analyse descriptive des clusters 5 et 1234 en termes d'âge et de nombre de chutes confirme l'analyse textuelle des suivis kiné de ceux-ci : la population du cluster 5 est plus âgée (88,9 vs 87,5), plus masculine (2,9 vs 3,8 pour le ratio femmes/hommes) et tombe moins (3 vs 3,7 pour le nombre moyen de chutes), puisque plus souvent en fauteuil.

4. DISCUSSION- CONCLUSION

A partir d'une prise en charge spécifique, ici le traitement kiné, et d'une population ciblée, les résidents avec ces données, notre approche montre une certaine cohérence entre la description littérale des soins des résidents et leur état de santé validé avec le modèle Pathos (résultat 3.2). Ce résultat est important car il valide la pertinence de notre approche fondée sur l'analyse textuelle des dossiers résidents.

Cette approche souligne d'autre part que la pertinence de l'information apportée par ces outils dépend essentiellement de la qualité des informations saisies par les opérateurs (résultat 3.1). Ainsi dans notre exemple, la génération du cluster 6 est probablement liée à un défaut de saisie pour les pathologies, mais aussi due au grand nombre d'antécédents qui frappe ce groupe de résidents. En effet dans une ACM, deux individus sont d'autant plus proches qu'ils ont en commun des modalités rares, ici le nombre de pathologies égal à 0. Le fait que tous ces résidents viennent du même EHPAD renforce cette hypothèse. Un travail plus en amont sur les données brutes à l'aide de requêtes SQL ou de fusion de tables pourrait permettre de détecter et remédier à ce genre d'anomalies. Cette approche permet ainsi de suivre la qualité des informations des dossiers résidents saisies pour en améliorer l'exhaustivité et la pertinence aussi bien pour le suivi de la population que la prise en charge des résidents.

D'autres méthodes d'analyse textuelle existent, dont certaines probabilistes ou basées sur l'apprentissage automatique [5], mais il nous a semblé que notre méthode de traitement de

l'information en huit étapes, bien qu'un peu fastidieuse et ciblée sur un seul sujet, ici le *Traitement kiné*, permettait de contrôler convenablement chaque étape des calculs, de pouvoir extraire quantitativement et qualitativement l'information textuelle et d'en contrôler la cohérence avec le reste des données des dossier santé des résidents.

L'information textuelle devient une information à part entière qui peut être traitée avec les outils statistiques habituels. Ces méthodes pourront ainsi être appliquées sur des échantillons plus larges et des problématiques plus essentielles comme par exemple le cancer ou la démence.

5. REFERENCES

[1] ML Maas, C Delaney. Nursing Process Outcome Linkage Research: Issues, Current Status, and Health Policy Implications. Medical Care, 2004 ;(42(2):II-40-II-48

[2] G Zangara G, PP Corso, F Cangemi et al. A cloud based architecture to support Electronic Health Report Stud Health Technol Inform. 2014;207:380-9.

[3] M Chiarini Tremblay, DJ Berndt, SL Luther et al. Identifying fall-related injuries: Textmining the electronic health record Inf Technol Manag 2009 10:253-263 DOI 10.10007/s10799-009-0061-6.

[4] T Delespierre, P Denormandie, L Josseran. New methods to evaluate physiotherapy care in nursing homes (Oral Communication) The Journal of Nursing Home Research 2015 Vol 1.

[5] A Holzinger, I Jurisica (Eds.) Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions.In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics Springer Berlin Heidelberg 2014, Berlin 357p. (date d'accès 17/04/2016)

[6] Tutoriels tm text mining package tm: http://edutechwiki.unige.ch/fr/Tutoriel_tm_text_mining_package https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf (date d'accès 17/04/2016)

[7] JM Ducoudray, Y Eon, R Leroux et al. Le modèle PATHOS, Guide d'utilisation. Caisse Nationale de Solidarité pour l'Autonomie, 2010, Paris, 54p. http://www.cnsa.fr/documentation/guide_d_utilisation_pathos_2012.pdf. (date d'accès 17/04/2016)

[8] F Husson, J Josse, J Pagès. Principal component methods – hierarchical clustering – partitional clustering : why would we need to choose for visualizing data ? . Applied Mathematics Department, Agrocampus. September 2010. Technical Report Agrocampus. http://www.agrocampus-ouest.fr/math/ (date d'accès 17/04/2016)

[9] http://factominer.free.fr/ (date d'accès 17/04/2016)

[10] https://cran.r-project.org/web/packages/SnowballC/index.html (date d'accès 17/04/2016)

[11] https://cran.r-project.org/web/packages/wordcloud/index.html (date d'accès 17/04/2016)

[12] https://cran.r-project.org/web/packages/RColorBrewer/index.html (date d'accès 17/04/2016)

[13] A Lazkani, T Delespierre, B Bauduceau et al. Predicting falls in elderly patients with chronic pain and other chronic conditions. Aging Clinical and Experimental Research 2015; 27(5):653-61. DOI 10.1007/s40520-015-0319-2.