Du dossier résident informatisé à la recherche en santé publique : Application des méthodes de surveillance en temps réel à des données médico-sociales de la personne âgée et exploration de données de cohorte pour la santé publique.

Tiba Delespierre

NNT : 2018SACLV030

Du dossier résident informatisé à la recherche en santé publique: application des méthodes de surveillance en temps réel à des données médico-sociales de la personne âgée et exploration de données de cohorte pour la santé publique.

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université de Versailles Saint Quentin

École doctorale n°570 **: santé publique (EDSP)**
Spécialité de doctorat : Épidémiologie

Thèse présentée et soutenue à Villejuif, le 19 juin 2018 par

Mme Tiba Delespierre

Composition du jury

M. Bruno Falissard, PUPH Université Paris-Saclay (Directeur CESP), Président du jury
M. Pascal Astagneau, PUPH Université Sorbonne Paris Cité. Rapporteur
M. François Husson, PU IRMAR - UMR 6625 du CNRS Agrocampus Ouest, Rapporteur
Me Vivianne Kovess, PU Université Sorbonne Paris Cité, Examinateur
Me Marie-Christine Boutron-Ruault, DR PH, CESP Université Paris Saclay, Examinateur
M. Loïc Josseran, PUPH, UVSQ Paris-Saclay - EA 4047, Directeur de thèse

Thèse de doctorat

Remerciements

Merci à Jean Bouyer qui m'a accordé sa confiance dans ce projet.

Merci à Loïc Josseran pour le suivi et l'encadrement de cette thèse.

Merci aux membres de mon jury d'avoir accepté d'évaluer ce travail.

Merci à Korian qui en a assuré le financement et m'a permis de pouvoir construire une cohorte de plus de 40 000 résidents ainsi que les outils algorithmiques pour pouvoir les suivre.

Merci à Sophie Boissard de m'avoir fait confiance sur ce projet de recherche et avant elle, Philippe Denormandie et Yann Coléou.

Merci à Sébastien Plasse, chef de projet chez Korian qui a permis la constitution de la plateforme d'extraction des données et m'a aidée à construire cette cohorte.

Merci à mon père qui m'a donné le goût des mathématiques et de l'effort et a fait naître en moi ce projet très humain.

Merci à ma mère qui m'a montré les vertus de la patience, de la persévérance et du travail bien fait et qui continue d'être une source d'inspiration et d'encouragement pour moi.

Merci à tous mes ami(e)s qui m'ont encouragée et à mes enfants Gabriel et Léa d'être là pour moi et avec moi.

J'ai la chance d'être passionnée par ce que je fais : explorer de nouveaux horizons pour la santé des hommes grâce à la programmation et à la modélisation. Mon souhait est que ce travail puisse faire bouger les lignes dans l'univers des maisons de retraite et changer le regard porté sur la santé et les soins apportés aux résidents, mais aussi aux personnes âgées dans leur ensemble.

J'aimerais aussi qu'il ne soit qu'une première étape.

Préambule

Korian, gestionnaire de services d'accompagnement et de soins pour les seniors, possède le premier réseau européen de maisons de retraite médicalisées, de cliniques spécialisées, de résidences services, de soins et d'hospitalisation à domicile avec 715 établissements. Présent dans quatre pays (France, Allemagne, Belgique et Italie), le Groupe dispose d'une capacité d'accueil de près de 72 000 lits.

L'Institut du Bien Vieillir IBV est une association 1901 dont les objectifs sont d'accompagner les professionnels de santé, d'améliorer le quotidien des personnes âgées, enfin d'améliorer leur prise en charge. Depuis cet automne cet institut s'est transformé en fondation dont les objectifs sont restés les mêmes.

Cette thèse a été ainsi réalisée au siège Korian, au sein de l'IBV, dans le cadre d'un partenariat Public/Privé s'inscrivant dans la logique d'évolution de la recherche académique par la collaboration avec le domaine médico-social de la personne âgée en France.

Résumé

La France connaît un vieillissement de sa population sans précédent. La part des séniors s'accroît et notre société se doit de repenser son organisation pour tenir compte de ce changement et mieux connaître cette population.

De nombreuses cohortes de personnes âgées existent déjà à travers le monde dont quatre en France (PAQUID®, GAZEL®, E3N-E4N® et 3C®) et, bien que la part de cette population vivant dans des structures d'hébergement collectif (EHPAD, cliniques de soins de suite) augmente, la connaissance de ces seniors reste lacunaire.

Aujourd'hui les groupes privés de maisons de retraite et d'établissements sanitaires tendent à s'équiper de grandes bases de données relationnelles permettant d'avoir de l'information en temps réel sur leurs patients/résidents. Depuis 2010 les dossiers de tous les résidents Korian sont dématérialisés et accessibles par requêtes. Ils comprennent à la fois des données médico-sociales structurées décrivant les résidents, leurs traitements et pathologies, mais aussi des données textuelles explicitant leur prise en charge au quotidien et saisies par le personnel soignant.

Au fil du temps et alors que le dossier résident informatisé (DRI) avait surtout été conçu comme une application de gestion de base de données, il est apparu comme une nécessité d'exploiter cette source d'informations et de construire un outil d'aide à la décision destiné à améliorer l'efficacité des soins. Ce travail de recherche et plus particulièrement cette thèse a été vue comme une formidable opportunité de mieux comprendre le potentiel informatif de ces données, d'évaluer leur fiabilité et leur capacité à apporter des réponses en santé publique. Pour cela nous avons procédé en plusieurs étapes.

- D'abord l'analyse de contenu du data warehouse DRI, l'objectif étant de construire une base de données recherche, avec un versant social et un autre de santé. Ce fut le sujet du premier article.
- Ensuite, par extraction directe des informations socio-démographiques des résidents dès leur entrée, de leurs hospitalisations et décès puis, par un processus itératif d'extractions d'informations textuelles de la table des transmissions et l'utilisation de la méthode Delphi, nous avons généré vingt-six syndromes, ajouté les hospitalisations et les décès et construit une base de données syndromique, la Base du Bien Vieillir (BBV) . Ce système d'informations d'un

nouveau type a permis la constitution d'une cohorte de santé publique à partir de la population des résidents de la BBV et l'organisation d'un suivi longitudinal syndromique de celle-ci. La BBV a également été évaluée scientifiquement dans un cadre de surveillance et de recherche en santé publique au travers d'une analyse de l'existant : contenu, périodicité, qualité des données. La cohorte construite a ainsi permis la constitution d'un outil de surveillance. Cet échantillon de population a été suivi en temps réel au moyen des fréquences quotidiennes d'apparitions des 26 syndromes des résidents. La méthodologie d'évaluation était celle des systèmes de surveillance sanitaire proposée par le CDC d'Atlanta et a été utilisée pour les syndromes grippaux et les gastro entérites aiguës. Ce fut l'objet du second article.

- Enfin la construction d'un nouvel outil de santé publique : la distribution de chacun des syndromes dans le temps (dates de transmissions) et l'espace (les EHPAD de transmissions) a ouvert le champ de la recherche à de nouvelles méthodes d'exploration des données et permis d'étudier plusieurs problématiques liées à la personne âgée : chutes répétées, cancer, vaccinations et fin de vie.

Summary

French population is rapidly aging. Senior citizens ratio is increasing and our society needs to rethink its organization, taking into account this change, better knowing this fast growing population group.

Even if numerous cohorts of elderly people already exist worldly with four in France (PAQUID®, GAZEL®, E3N-E4N® and 3C®) and, even as they live in growing numbers in nursing homes and out-patient treatment clinics, knowledge of this population segment is still missing.

Today several health and medico-social structures groups tend to invest in big relational data bases enabling them to get real-time information about their patients/residents. Since 2010, all Korian residents' files are dematerialized and accessible by requests. They contain at the same time, structured medico-social data describing the residents as well as their treatments and pathologies, but also free-textual data detailing their daily care by the medical staff.

Through time and as the computerized resident file (DRI) was mainly conceived as a data base management application, it appeared essential to mine these data and build a decision-making tool intended to improve the care efficiency. This research work and this thesis were then seen as a great opportunity to understand better these data informative potential, to assess their reliability and response to public health threats. To do this we proceeded as follows:

- First, a content analysis of the data warehouse DRI, the objective being to build a research database, with a social side and a health side. This was the first paper subject.
- Then, by direct extraction of the residents' socio-demographic information at nursing home (NH) entry, adding hospitalizations and deaths, and finally, by an iterative textual extraction process of the transmissions data and by using the Delphi method, we created twenty-four syndromes, added hospitalizations and deaths and built a syndromic data base, the Ageing Well data base. This information system of a new kind, allowed the constitution of a public health cohort for elderly people from the BBV residents'population and its syndromic longitudinal follow-up. The BBV was also scientifically assessed for surveillance and public health research through present situation analysis: content, periodicity and data quality. This cohort then gave us the opportunity to build a surveillance tool and follow the residents' population in real-time by watching their 26 daily

frequency syndromic distributions. The methodology for that assessment, Atlanta CDCs' health surveillance systems method, was used for flu and acute gastro enteritis syndroms and was the second paper subject.

- Finally, the building of a new public health tool: each syndrom's distribution through time (transmissions dates) and space (transmissions NH ids) opened the research field to new data exploration methods. I used these to study different health problems afflicting senior citizens: frequent falls, cancer, vaccinations and the end of life.

Abréviations

Abréviations	Dénomination en anglais / en français
ILI / IRA-grippe	Acute Respiratory Infection and Influenza Likely Illness / Infection Respiratoire Aiguë
AGE / GEA	Acute Gastro Enteritis / Gastro Entérite Aiguë
BBV	Ageing Well Data Base / Base du Bien Vieillir
CDC	Centers for Disease Control and Prevention
ICD-9 / CIM-9	International Classification Diseases 9 / Classification Internationale des Maladies 9
DM	Data Mining
DRI	Computerized Resident File / Dossier Résident Informatisé
DMP	Dossier Médical Partagé
DWH	Data WareHouse / Silo de données
ECDC	European CDC
EHR/ EES	Electronic Health Record / Enregistrement Electronique de Santé
EMR/EEM	Electronic Medical Record / Enregistrement Electronique Medical
EPR/EEP	Electronic Personal Record / Enregistrement Electronique Personnel
ETL	Extract Transform Load / Extraction Transformation Chargement
GIR	ISO Ressource Group / Groupe ISO Ressources
HC	Hierarchical Clustering / Classification hiérarchique
HCPC	Hierarchical Clustering on Principal Components / Classification hiérarchique sur Composantes Principales
HRA / ARS	Health Regional Agencies / Agences Régionales de Santé
IOM	Institute Of Medecine / Institut de Médecine
MCA / ACM	Multiple Component Analysis / Analyse en Composantes Multiples
PCA / ACP	Principal Component Analysis/ Analyse en Composantes Principales
NER	Named Entity Recognition / Reconnaissance d'Entités Nommées
NH / EHPAD	Nursing Homes / Etablissement pour Personnes Agées Dépendantes
NLP / TAL	Natural Language Processing / Traitement Automatique du Langage
SS	Surveillance System / Système de Surveillance
SSS	Syndromic Surveillance System / Système de Surveillance Syndromique
SQL	Standard Query Language
TM	Text Mining

Table des matières

Liste des travaux scientifiques

Articles

Empirical Advances with Text Mining of Electronic Heath Records
Delespierre T, Denormandie P, Bar-Hen A, Josseran L
Journal: BMC Medical Informatics and Decision Making
DOI: 10.1186/s12911-017-0519-0

Issues in Building a Nursing Home Syndromic Surveillance System with Textmining: Longitudinal Observational Study
Delespierre T, Josseran L
Journal: JMIR Journal of Medical Internet Research
DOI: 10.2196/jmir.9022

Présentations et Posters

Comment Passer d'un Usage Individuel du Dossier Résident à un Bénéfice Collectif ? Delespierre T, Denormandie P, Josseran L workshop EPICLIN 9 2015 à Montpellier Mai 2015

New Methods to Evaluate Physiotherapy Care in Nursing Homes Delespierre T, Denormandie P, Josseran L workshop Nursing Home Research à Toulouse Décembre 2015

De nouvelles données et de nouvelles méthodes pour évaluer les soins primaires de la population cancéreuse en EHPAD. Delespierre T, Denormandie P, Armaingaud D, Josseran L poster EPICLIN 10 2016 à Strasbourg Mai 2016

Application Empirique de l'Analyse Textuelle de l'Analyse des Correspondances et de la Classification sur un Exemple de PEC Delespierre T, Denormandie P, Josseran L journées de la Statistique 2016 à Montpellier Mai 2016

When and why do we fall in Nursing Homes? Delespierre T, Denormandie P, Josseran L workshop Nursing Home Research à Barcelone Novembre 2016

Vaccinations, a recurrent event to follow the residents' health status in nursing homes? Delespierre T, Josseran L Présentation orale SMB Paris Septembre 2017

Développement d'un système de surveillance sanitaire en temps réel dans un réseau national de maisons de retraite. Delespierre T, Josseran L Présentation orale ADELF - SFSP Amiens Octobre 2017

The end of life in nursing homes. What to expect. What to measure. Delespierre T, Letty A, Josseran L Présentation orale EPH Conference Stockholm Novembre 2017

Peut-on caractériser la fin de vie en EHPAD ? Delespierre T, Sanchez S, Armaingaud D, Letty A, Josseran L poster JASFGG Paris Novembre 2017

Liste des tables

Liste des figures

1-Introduction

1-1 Contexte

La France connaît un vieillissement de sa population. Les hypothèses les plus plausibles des démographes prévoient que notre pays « comptera 73,6 millions d'habitants en 2060, soit 11,8 millions de nouveaux habitants. Parmi eux, les plus de 60 ans représenteront à eux seuls 10,4 millions, passant de 21% de la population totale aujourd'hui à presqu'un tiers. Les plus de 75 ans, qui étaient 5,2 millions en 2007 (soit 8,9% de la population), seront alors 11,9 millions dans un demi-siècle (soit 16,2% de la population). Les 85 ans et plus, quant à eux, passeront de 1,3 à 5,4 millions, ce qui signifie qu'ils seront quatre fois plus nombreux qu'aujourd'hui. » [1- 2]

Alors que la part des séniors s'accroît, la société française doit repenser son organisation pour tenir compte de ce changement. Cette mutation doit pouvoir accorder toute leur place aux personnes âgées et les soutenir du mieux possible, notamment aux périodes de grands âges [3] et de grande vulnérabilité. « Le vieillissement s'inscrit dans la logique métamorphique de la vie qu'il nous faut assumer comme telle, avec ses difficultés mais aussi ses richesses. Il est l'occasion de renouer avec certains fondamentaux qui donnent sens à l'ensemble de l'existence humaine. » [4] C'est ainsi un devoir éthique de s'en préoccuper et de faire en sorte qu'elle se déroule dans les meilleures conditions et avec le moins d'incapacités possibles. Cette évolution démographique, de par les transformations sociales qu'elle augure, nous questionne d'ailleurs sur sa prise en compte dans nos sociétés.

Afin de remplir au mieux ce contrat moral, et anticiper des mesures d'accompagnement, notre société se doit de mieux connaître cette population. Ce besoin de connaissances et de compréhension est vrai dans de nombreux domaines couvrant des sujets aussi variés que l'emploi, les loisirs, les questions sur l'habitat ou bien la santé. Ouchi [5], met en avant cette nécessité d'agir. Il étudia le cas du Japon, l'un des pays au monde à la proportion de population

âgée la plus importante (en 2008, plus de 10% de la population avait plus de 75 ans). Il présente un pays où, notamment grâce aux avancées technologiques, à l'hygiène de vie (et la combinaison probable de facteurs culturels), la longévité sans incapacité s'accroît et questionne. Il propose ainsi des solutions de réorganisation sociale allant de la réinsertion professionnelle, à une refonte de l'organisation de la santé publique pour le grand âge. Parmi ces propositions pour la santé publique, il insiste notamment sur la nécessité de former des spécialistes médicaux du vieillissement qui soient en mesure de comprendre la complexité du processus de sénescence dans son ensemble pour pouvoir dépister, diagnostiquer et prévenir correctement ses complications.

Cette connaissance sur le vieillissement, ne peut en toute logique se structurer que grâce à une base solide d'études scientifiques fiables et cohérentes. C'est l'un des points soulignés par Washko [6] en 2012 dans une étude américaine consacrée à la question de la place des personnes vieillissantes (avec infirmité) aux Etats-Unis, dans laquelle il présente la nécessité d'une structure fédératrice qui puisse impulser, centraliser et coordonner une recherche opérationnelle nationale sur cette population. Ce point de vue rejoint les conclusions d'Ouchi, au Japon, dans la nécessité de promouvoir et structurer la recherche « clinique » sur la personne âgée. La coordination des études nationales, multicentriques aurait pour ambition d'accumuler et de diffuser de la connaissance, de fournir des recommandations afin d'améliorer les prises en charge gériatriques. Ouchi souligne ainsi l'importance du recueil de données cliniques pour chaque centre de soins participant à cet effort scientifique. Ce dernier exemple met l'accent sur les études cliniques, mais la connaissance doit s'enrichir aussi de données en population générale (de nombreuses cohortes de personnes âgées existent déjà à travers le monde dont quatre en France [7]) ou d'études du secteur médico-social aujourd'hui peu développées.

Les données médico-sociales contenues dans les grandes bases de données des groupes d'établissements pour personnes âgées dépendantes (EHPAD) et cliniques de soins de suite et de réadaptation (SSR) pourraient enrichir la connaissance car la proportion de personnes vivant en hébergement collectif (maisons de retraite, unités de soins de longue durée ou foyers

logements) a progressé aux grands âges : 21,2% des 85 ans et plus vivaient en hébergement collectif en 2004 contre 13,1% en 1962. Leur part augmente notablement à partir de 85 ans et atteint un peu plus de 40% des effectifs au-delà de 90 ans [8]. Ces données pourraient s'ajouter aux savoirs cliniques existants en fournissant des informations intégrant de nouvelles dimensions sur le vieillissement et sa prise en charge, et ainsi devenir une source importante de prévention des difficultés liées aux grands âges. En effet, l'organisation même du secteur médico-social, la médicalisation limitée, un personnel moins porté sur la dimension technique de la prise en charge et davantage tourné sur les rapports humains, le statut des personnes qui ne sont pas des patients mais des résidents, portent à penser et à aborder cette recherche d'une façon différente et transversale.

En France, la nécessité de décloisonner les activités des secteurs sanitaire et médico-social a été ressentie, et encouragée par la loi Hôpital, patients, santé et territoires (HPST) [9] en 2009 et la création des agences régionales de santé (ARS) en 2010. Ces réformes récentes de santé publique, encouragent le nécessaire rapprochement entre les secteurs sanitaire et médicosocial, et poussent ces acteurs à repenser leur fonctionnement et à collaborer pour mieux s'adapter aux besoins de la santé des séniors et assurer la coordination des soins et de la prise en charge. Cette mutualisation progressive des activités, la bonne articulation des périmètres d'intervention, appuyés par des technologies novatrices ont créé de nouvelles sources d'information sur l'état de santé de la personne résidente en structure médico-sociale, mais qui restent sous-exploitées à ce jour.

La recherche dans le secteur médico-social a ainsi une vraie légitimité dans ce contexte et est importante pour mieux comprendre et accompagner les évolutions de santé des personnes âgées. Le champ des possibles est large : une étude canadienne de 2011 menée dans 134 centres médicalisés de soins longue durée, a ainsi pu identifier 9 thématiques de recherches pour améliorer les prises en charge allant du contrôle du risque infectieux à l'amélioration de la qualité des soins en passant par la modification des conditions de résidence [10]. Le poids démographique du 3ème et 4ème âge [11] dans les décennies à venir doit nous inciter à explorer

cette voie.

A l'ère des nouvelles technologies de l'information, de nombreux établissements sanitaires ou médico-sociaux s'équipent de grandes bases de données relationnelles permettant d'avoir de l'information en temps réel sur leurs patients/résidents. Nous nous attachons ici à explorer et évaluer l'une de ces sources continues de données médico-sociales et à comprendre dans quelle mesure celle-ci peut permettre une connaissance épidémiologique [12] fiable de l'évolution de l'état de santé des personnes institutionnalisées.

Cette source de données est la base de données médico-sociale du groupe Korian créée à partir de données collectées dans les dossiers électroniques de ses résidents en France. Ce groupe privé spécialisé dans l'hébergement et l'accompagnement médico-social des personnes âgées et dépendantes est présent sur l'ensemble du territoire français où il gère en France 365 établissements pour personnes âgées (EHPAD et SSR). La volonté de s'inscrire dans une continuité de prise en charge et une intégration des secteurs sanitaires et médicaux sociaux se retrouve notamment dans la mise en place en 2010 d'un outil informatisé contenant des informations transverses. Le groupe a donc fait le choix d'informatiser les dossiers de ses résidents ce qui représente aujourd'hui une base de données d'environ 40 000 dossiers pour l'un des deux systèmes d'information. Chaque dossier comprend des données médico-sociales structurées et exploitables informatiquement. Alors que cette source de données a un potentiel informatif incontestable, il a été décidé de financer un travail de recherche au sein du groupe destiné à mieux comprendre son contenu, à évaluer sa fiabilité et sa capacité à apporter des réponses en santé publique.

Le déroulé des différentes tâches pour tenter de répondre à ces questions de recherche a été conçu de la manière suivante :

1- **Analyser le contenu** de la base de données et le scinder en domaines fonctionnels ;
2- **Construire une cohorte de personnes âgées** par extraction de tous les résidents entrés dans un des établissements à compter de la mise en place du SI ;
3- **Construire des outils de suivi de leur prise en charge en matière de soins,**

d'hospitalisations et de survie. En effet, en analysant toutes les tables de ce SI il est apparu que celle des transmissions détenait l'essentiel du suivi au quotidien et pouvait permettre de décrire, au moins en partie, l'état de santé des résidents en temps presque réel. Pour cela nous nous sommes alors appuyés sur la notion de syndrome, ensemble de signes cliniques et de symptômes qu'un sujet est susceptible de présenter lors de certaines maladies, ou bien dans des circonstances cliniques d'écart à la norme pas nécessairement pathologiques [13].

4- **Analyser ces nouveaux outils au travers de la surveillance des épidémies de grippe et de gastro-entérite aiguës** (GEA) sur plus de six années et, pour la dernière période, l'hiver 2016-2017, en temps quasi réel par les méthodes du CDC d'Atlanta.

Alors que la cohorte au point 2 a été obtenue par requêtes simples, la construction des outils de suivi évoquée au point 3 a nécessité de développer des outils s'appuyant sur l'analyse textuelle des transmissions (premier article). Vingt-six syndromes ont ainsi été définis au sein d'un thésaurus décrivant plusieurs facettes des problématiques de santé touchant les résidents en EHPAD (deuxième article). Une fois la cohorte construite et ses outils de suivi évoqués au point 3 générés, notre cohorte a pu jouer le rôle de cohorte épidémiologique à l'usage des résidents en EHPAD.

Exploitant les données de cette population nous avons cherché à modéliser les parcours de vie de résidents atteints de cancer (poster en 2016 à Strasbourg), les chutes récurrentes en institution (présentation en 2016 à Barcelone), le profil syndromique des résidents et la prédiction de leur durée de séjour dès l'entrée en établissement (présentation en 2017 à Paris), enfin la modélisation de la fin de vie (poster à Paris et présentation à Stockholm en 2017).

1-2 Objectif principal

Un objectif principal dans ce travail de recherche : comment passer d'une base de données médico-sociales à visée professionnelle, le DRI, à une base de données à visée santé publique ? Pour tenter d'y répondre nous avons procédé en plusieurs étapes :

1- Analysé le contenu du DWH DRI,
2- Analysé le contenu de l'information textuelle contenue dans les transmissions du DRI ;
3- Construit une cohorte par extraction d'une population de résidents du DRI ;
4- Construit une variable indicatrice de syndrome grippal et une variable indicatrice de GEA par extraction de l'information textuelle contenue dans les transmissions et ajouté les hospitalisations et les décès de la cohorte ;
5- Construit un outil de surveillance syndromique en assurant le suivi de notre cohorte par ces quatre indicateurs ;
6- Construit d'autres syndromes ainsi que leur suivi en temps réel et tenté de développer un outil de santé publique.

1-2-1 Analyse de contenu du DWH DRI

C'est suite à un essai clinique destiné à mesurer l'efficacité de l'hygiène de mains en maison de retraite médicalisées et l'adhérence des personnels des EHPAD aux recommandations internationales d'hygiène [14] qu'est apparu le potentiel informationnel du journal des transmissions dans le DRI. Ce journal utilisé en tant qu'outil de communication entre les différentes équipes de soins, permet à l'équipe de jour d'informer l'équipe de nuit et réciproquement, sur les soins pratiqués, les traitements administrés et les problèmes éventuels rencontrés au cours de la période couverte pour chaque résident pris en charge. C'est dans ce journal que l'on trouve par exemple des mentions des épisodes infectieux touchant certains résidents comme la grippe et les gastro-entérites aiguës et les protocoles de soins appliqués pour endiguer d'éventuelles épidémies.

1-2-2 Analyse de l'information textuelle des transmissions

A partir de cette constatation, nous avons décidé d'analyser l'information textuelle contenue dans les transmissions et d'évaluer sa pertinence. Ensuite de montrer comment cette information pouvait compléter le reste des données socio-démographiques et médicales disponibles dans le DWH et ainsi aboutir à une cohorte et à un outil de santé publique.

C'est au travers d'une analyse textuelle qualitative et quantitative d'un échantillon relativement petit et bien défini de récits cliniques relatifs aux traitements de kinésithérapie qu'il a été

possible de contrôler la précision des informations récoltées pour décrire les vies et les soins des résidents.

1-2-3 Construction d'une cohorte épidémiologique

Il a été ensuite été envisagé d'extraire la population des résidents entrés dans les différents établissements au cours du temps et alimentés dans le DRI avec leurs données socio-démographiques et d'y adjoindre des informations textuelles extraites de leurs transmissions.

1-2-4 Construction de quatre variables syndromiques

Alors que les données hospitalisations et décès peuvent être obtenues par extraction directe, les variables syndromiques grippe et GEA le seront par analyse textuelle des transmissions et le suivi de leurs distributions au fil de l'eau, ainsi que par l'étude de leur validité et leur pertinence.

1-2-5 Construction d'un outil de surveillance syndromique

Le suivi des distributions des deux syndromes pré-cités au sein de la cohorte pré-citée permettra alors de construire l'outil de surveillance grippe et GEA en institution et donc de bâtir et évaluer un système national écologique de surveillance en santé publique des maisons de retraite.

1-2-6 Construction d'une liste de syndromes

En recherchant d'autres informations dans les transmissions, reflets des problématiques majeures touchant la population âgée en institution, un ensemble de variables syndromiques pourront être construites et permettre de générer un outil d'exploration des soins et de prise en charge globale du résident tout en enrichissant la vision du suivi des syndromes grippe et GEA.

1-2-7 Construction d'un nouvel outil de santé publique ?

Pourra-t-on alors répondre à la question : la cohorte proposée en 1-2-3 et les distributions des syndromes en 1-2-4 et 1-2-6 permettent-elles la construction d'un outil de santé publique pour les personnes âgées en institution ?

1-3 Méthodes

1-3-1 Analyse de contenu du DWH DRI

Depuis 2010 les dossiers de tous les résidents Korian sont dématérialisés et accessibles par requêtes. Comprenant à la fois des données médico-sociales structurées et des données textuelles explicitant leur prise en charge au quotidien, il est apparu que ce corpus de données pouvait représenter une source de données appropriée pour comprendre et suivre l'état de santé des résidents dans leur ensemble. Les silos de données composant le DRI se présentant sous la forme d'une base de données Oracle®, nous avons exploré de manière systématique l'ensemble des fichiers composant son système d'informations (SI) par requêtes SQL.

Recherchant l'information pertinente sur le cycle de vie des résidents en EHPAD, nous avons ainsi repéré huit domaines fonctionnels listés dans la table 1 décrivant globalement toutes les facettes du dossier résident: les données administratives, les parcours de vie, le niveau de dépendance, les pathologies, facteurs de risques, médicaments, aides techniques et enfin les évaluations [15]. Seule information manquante dans cette table, la prise en charge au quotidien par les équipes de soins et contenue dans la table des transmissions.

Données administratives	Parcours de vie	Niveau de dépendance	Pathologies	Risques	Médicaments	Aides techniques	Evaluations
Age Ville d'origine Nomenclature CPAM Sexe Situation familiale & soutien	EHPAD d'origine Entrées/ sorties* Hospitalisa-tions* Congés* Décès	GIR* ALD* APA* IADL* ADL* Fragilité*	Cardio-vasculaire* Arthrose* Alzheimer* Parkinson* Cancer*	Chutes* Escarres* Dénutrition * Déshydra-tation* Dépression * Infectieux*	Classes principales* Nombre*	Lit spécialisé* Déambula-teur* Fauteuil roulant* Aide auditive* Aide visuelle* Appareil dentaire*	MMSE* NPI * GDS* EVA* DN4* Norton*

Table 1 : Les 8 blocs fonctionnels du Dossier Résident Informatisé

Dès l'entrée du résident en EHPAD, un DRI vient alimenter le SI de la société. Chaque DRI comprend des données provenant des champs social et médical. Ces données sont enregistrées et transmises en temps réel. Le DWH comprend deux grands types de données :

- Des informations stables, dites 'dures' comprenant les caractéristiques sociodémographiques du résident, ses pathologies et comorbidités, ses facteurs de risque, enfin son parcours de vie: entrées – sorties, hospitalisation(s) et décès (voir table 1);
- Des informations récoltées au fil de l'eau, au cours du parcours de soins, dites infos 'molles' telles qu'une altération de l'état général ou une dépression, une suspicion de démence ou d'AVC, des chutes, le plus souvent des remarques du personnel soignant concernant des points à surveiller ou des difficultés rencontrées lors des soins.

Alors que les premières informations sont structurées, les secondes sont essentiellement textuelles et se présentent sous la forme de mots, acronymes ou phrases de longueur variable. Néanmoins, le fait que chaque information, qu'elle que soit son type, soit indexée sur les identifiants du résident et de son établissement, devait permettre de les agréger en un tout cohérent. Grâce aux informations textuelles extraites du journal des transmissions il devenait théoriquement possible d'enrichir le dossier de chaque résident et d'en donner une image plus 'dynamique', de définir des trajectoires de vie.

De manière à valider scientifiquement la possibilité d'intégrer des informations textuelles pertinentes extraites des transmissions, nous avons entrepris une analyse textuelle qualitative et quantitative d'un échantillon relativement petit et bien défini de récits cliniques relatifs aux traitements de kinésithérapie, l'idée étant qu'il devait être alors possible de contrôler de manière systématique la pertinence des informations récoltées pour décrire les vies et les soins des résidents. C'était l'objectif du premier article.

1-3-2 Analyse de l'information textuelle des transmissions

1-3-2-1 Présentation du premier article

L'objectif : Montrer qu'au travers d'une analyse textuelle qualitative et quantitative d'un échantillon relativement petit et bien défini de récits cliniques relatifs aux traitements kiné, il était possible de construire un corpus de kinésithérapie et, par ce processus, de générer un nouveau domaine de connaissances par l'addition d'informations pertinentes pour décrire les vies et les soins apportés aux résidents.

La méthode : un traitement en deux étapes. Dans une première étape, des mots représentatifs des soins en kinésithérapie ont été extraits par Standard Query Language (SQL) avec la fonction LIKE, des caractères génériques et une recherche automatique de motifs, suivie de text mining et d'un nuage de mots à l'aide de packages R®. Ensuite, au cours d'une autre étape, des analyses en composantes principales et en correspondances multiples, suivies de classifications sur le même échantillon de résidents, enrichies enfin d'autres données de santé mesurant les différents niveaux de besoins de soins ont été successivement testées.

Le résultat : Cette étude textuelle empirique au moyen d'un traitement textuel en deux étapes a montré que les techniques de text mining et de data mining pouvaient fournir des outils accessibles pour améliorer la santé des résidents et la qualité des soins en ajoutant de nouvelles données simples et utiles aux enregistrements de santé électroniques. Au cours de cette expérimentation nous avons créé des variables textuelles qualitatives avec une signification clinique et qui acquièrent une validité bio-statistique lorsqu'elles sont utilisées pour répondre à une problématique de santé. Elles peuvent ainsi être utilisées soit pour décrire un phénomène de santé soit pour qualifier des trajectoires de santé des résidents dans le temps.

1-3-2-2 Perspectives liées à ces premiers travaux

Les travaux liés à ce premier article ont permis au travers d'une étude approfondie des mots-clés liés à la prise en charge kinésithérapeutique de montrer qu'il est possible de construire un corpus spécifique et d'ajouter de nouvelles informations permettant de décrire les vies et les soins apportés aux résidents. La génération de variables textuelles à partir de la description de ces traitements se concentrait uniquement sur une courte période, mais j'ai également montré

que l'on pouvait réellement suivre au cours du temps les occurrences de ces variables (voir ci-dessous).

Dans ce schéma, l'extraction des données textuelles traitement kiné s'est faite sur 2 années [16] et a permis de calculer les effectifs mensuels des quatre mots-clés les plus utilisés sur l'ensemble des établissements pour décrire les séances kiné au fil des mois écoulés: autonomie, équilibre, douleur et marche. Après une montée en charge pendant un peu plus d'une année, on constate une diminution de l'usage du mot autonomie. Renseignements pris auprès de la direction médicale, cela correspondait à un changement de priorités suite à la fusion avec la société Medica® et n'était pas représentatif d'une modification de priorités des kinésithérapeutes qui restait de renforcer l'autonomie des résidents.

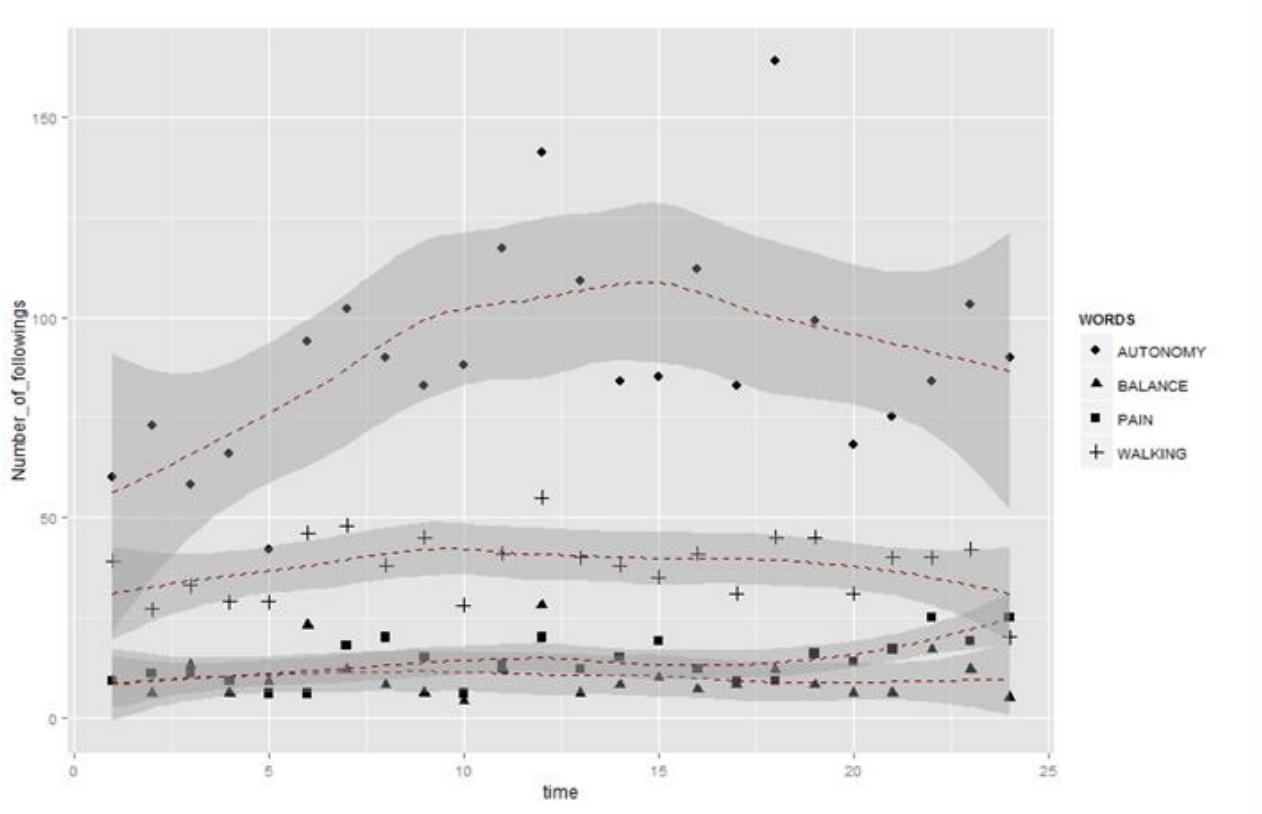


Figure 1 : Les quatre mots les plus fréquents extraits du corpus du traitement kiné sur la période [01/04/2013 – 31/03/2015]

Ce qui est illustré ici est la variabilité des occurrences en fonction du contexte. Des événements parfois complètement extérieurs au domaine ou à la problématique peuvent influer sur les fonctions de répartition des données textuelles. Il est donc nécessaire d'être relativement

prudent dans les analyses statistiques et de toujours s'appuyer sur des faisceaux de preuves chaque fois que c'est possible.

1-3-3 Construction d'une cohorte épidémiologique

1-3-3-1 Les données socio-démographiques

Pour chaque resident des données de formats fixes : date d'entrée, sexe, âge et GIR (groupe ISO ressources) à l'entrée en établissement, ainsi que l'identifiant de l'établissement. Le GIR définit en France le niveau d'autonomie des personnes âgées indexé sur des prestations gouvernementales [17]. Les données récupérées du DRI ne sont pas à ce stade à perspective temporelle. Elles permettent juste de donner un profil succint du résident lors de son entrée.

1-3-3-2 Les données syndromiques

Ensuite des données textuelles, de format libre jusqu'à 4000 caractères, produites par le personnel soignant lors de son activité quotidienne, saisies dans la table **transmissions** et utilisées en tant qu'outil de liaison. Ces informations non formatées, saisies au fil de l'eau par les équipes ont été retravaillées pour être utilisables dans un but scientifique grâce à une analyse textuelle en 3 étapes: d'abord, troncature et nettoyage des données, ensuite, requêtes SQL (Standard Query Language) dans le DWH pour bâtir les syndromes, enfin, conférence de consensus par les experts métiers et text mining, avec plusieurs allers-retours entre les deux dernières étapes pour détecter puis sélectionner les mots-clés les plus pertinents.

Parmi les données retenues disponibles dans la table **transmissions**, la description et l'évolution de la morbidité était centrale. Dans ce cadre, ont d'abord été définis les syndromes grippe et gastro-entérite aiguë en tant que premiers objets pour définir un outil de surveillance sanitaire pour les personnes âgées à partir des résidents en EHPAD. Au vu de l'intérêt terrain et de la richesse des informations saisies puis extraites, l'élaboration de ces deux syndromes a rapidement été suivie de celle de vingt-deux autres : douleurs, dépression et idées noires, comportement, troubles cardio-vasculaires, démence, vaccination, altération de l'état général,

cancer, déshydratation, dénutrition et déglutition, problèmes de transit, état cutané, problèmes bucco-dentaires, allergies, vision, chutes, audition, sommeil, troubles urinaires, excès de poids, diabète, fragilité. Enfin, hospitalisations et décès ont été agrégés au reste de l'information syndromique par extraction directe des tables **hospitalisations** et **décès** du DRI.

1-3-3-3 La construction de la cohorte des résidents

L'extraction de tous les résidents entrés dans au moins un établissement à compter du 1er novembre 2010 (date de début du DRI) et ce jusque fin février 2017 (fin de la saison grippale de l'hiver 2016-2017) a permis de générer une cohorte. Chaque semaine la cohorte est augmentée des nouvelles entrées. Les descriptions de séjours des résidents peuvent être tronquées à gauche, lorsque ceux-ci sont entrés avant le 01/11/2010, tronquées à droite s'ils ne sont pas encore décédés à la date de la dernière semaine extraite, ici le 26/02/2017, ou complètes, lorsque le séjour complet du résident est inclus dans la période d'extraction (voir figure 2 ci-dessous).

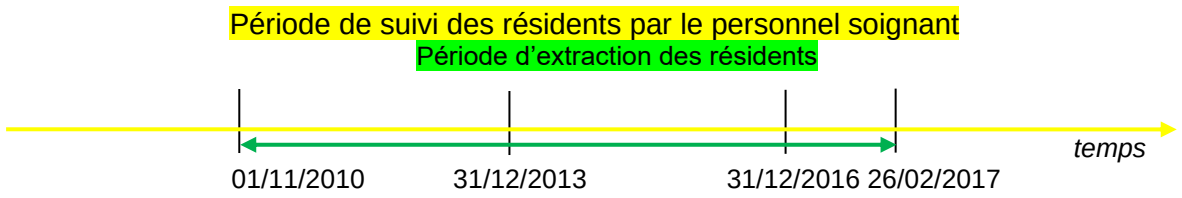


Figure 2 : Schéma temporel d'extraction et de suivi des résidents dans le Dossier Résident Informatisé

1-3-4 Construction de variables syndromiques

La construction des variables syndromiques grippe et gastro-entérites aiguës a été conçue par analyse textuelle des transmissions. Elle comprenait 3 étapes, la première consistant à tronquer l'information et à la nettoyer : les transmissions pouvant comprendre jusqu'à 4000 caractères, nous les avons tronquées à 300 caractères et supprimé de nombreux caractères spéciaux de ponctuation, des accents, le but étant de pouvoir stocker simplement, efficacement et sans erreur l'essentiel de l'information textuelle contenue dans les transmissions. Le choix de 300 caractères a fait suite au travail exposé dans le premier article sur l'analyse des transmissions kiné [18]. En effet, la majorité

des transmissions sont souvent courtes et peuvent comprendre des caractères spéciaux de saisie qui empêchent une exportation simple et correcte au format **csv** ou **xls** vers EXCEL ou R.

La seconde étape se composait d'un traitement itératif en deux sous-étapes :

- Étape 2-1 : pré-sélection de mots-clés et d'expressions-types décrivant habituellement des syndromes grippaux pour le syndrome GRIPPE, et les gastro-entérites aiguës pour le syndrome GEA. Pour cela, une revue de la littérature des diagnostics correspondant à ces deux syndromes, l'examen de nombreuses transmissions sur de longues périodes et la consultation de médecins-gériatres chez Korian, enfin, construction des requêtes SQL afférentes.
- Étape 2-2 : analyse des syndromes GRIPPE et GEA obtenus. Correction des mots-clés et expressions-types.

Enfin la troisième étape analysait de nombreuses périodes hivernales ou non hivernales (de novembre 2010 à mars 2016) et corrigeait certaines anomalies temporelles rencontrées (par exemple un surnombre de cas par rapport aux tendances sur la France entière). Nous avons ainsi abouti à des syndromes pré-construits qui ont été resoumis aux médecins travaillant chez Korian, mais aussi à des gériatres externes au groupe pour un dernier examen et avis.

Le stockage de l'information syndromique se fait dans trois tables (pour les traitements voir la figure 3, de gauche à droite) : la première table sert à stocker les effectifs syndromiques, la seconde table contient l'information booléenne correspondante : TRUE pour le syndrome GRIPPE si le syndrome est la grippe (le syndrome est à FALSE par défaut). Enfin, la troisième table contient l'expression littérale tronquée et nettoyée du syndrome. Cette manière de faire permet de définir plusieurs syndromes à partir d'une seule transmission ce qui est fréquemment le cas comme nous le verrons par la suite. Ainsi, un résident peut par exemple souffrir d'un syndrome grippal conjugué à une gastro-entérite aiguë (oui dans l'ordinogramme et stockage de 2 syndromes), mais il peut aussi ne souffrir ni de l'un, ni de l'autre (non dans l'ordinogramme et aucun stockage).

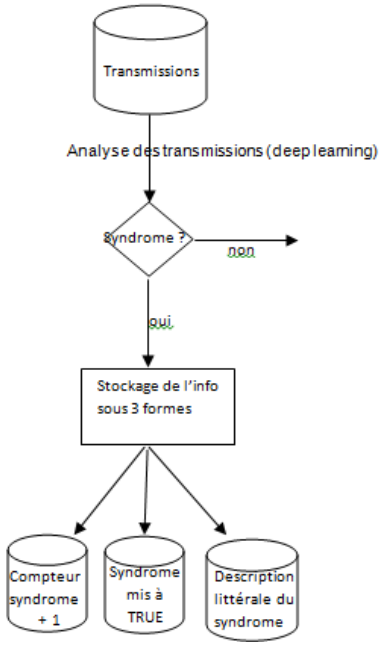


Figure 3 : Le processus de stockage de l'information syndromique dans la Base du Bien Vieillir

1-3-5 Construction d'un outil de surveillance syndromique

L'outil de surveillance en grippe et gastro-entérites aiguës en institution comprend donc quatre tables : une table **entrées** avec les résidents entrés dans un des établissements qui consitue la cohorte et les trois tables représentées de gauche à droite dans la figure 3 : la table **effectifs**, la table **syndromes** et la table **descriptions**. Les **décès** et **hospitalisations** successives sont intégrés en tant que syndromes au même titre que **grippe** et **gea** dans les 3 tables ci-dessus par requêtes simples sur les tables decess et hospitalisations du DRI.

Présentation du second article

L'objectif : Construire et évaluer un système national écologique de surveillance en santé publique des maisons de retraite.

La méthode en trois étapes: Première étape, la construction d'une cohorte de personnes âgées à l'aide d'un réseau national de 125 EHPAD et les données médicales et personnelles extraites des enregistrements électroniques de santé (EES) des résidents. Deuxième étape,

l'utilisation des effectifs syndromiques grippe et GEA définis définis par recherche de motifs avec le langage SQL et son opérateur LIKE ainsi qu'une méthode Delphi-apparentée, de novembre 2010 à juin 2016 pour construire les séries temporelles grippe et GEA. Troisième étape, l'utilisation des algorithmes de surveillance early aberration reporting system (EARS) et Bayes du package *R*® de surveillance *surveillance*® pour évaluer nos séries temporelles syndromiques et les comparer aux données syndromiques du Réseau Sentinelles, le gold-standard français pour les épidémies de grippe et de gastro-entérite, suivant les guidelines d'évaluation du CDC (Centers for Disease Control and Prevention) d'Atlanta.

Le résultat : La construction d'une large cohorte épidémiologique par extraction de toute l'information socio-démographique disponible et l'utilisation de l'algorithme EARS_C3 sur nos données grippales et de gastro-entérite aiguë ont permis d'anticiper la dernière épidémie de grippe, ainsi que son intensité et sa virulence.

1-3-6 Construction d'une liste de syndromes

Une fois les quatre syndromes construits et fonctionnels (grippe, GEA, décès et hospitalisations), nous avons imaginé pouvoir développer d'autres variables syndromiques bâties en trois étapes suivant le même principe que celui décrit au paragraphe 1-3-4. Parmi celles-ci nous trouvons vingt-deux nouveaux syndromes : douleurs, comportement (attitude de refus), démence, altération de l'état général, déshydratation, dénutrition et problèmes de déglutition, troubles d'audition, problèmes bucco-dentaires, cancer, vaccinations, excès de poids, état cutané, allergies, chutes, dépression et idées noires, troubles cardio-vasculaires, troubles du sommeil, troubles de la vision, problèmes de transit, troubles urinaires, diabète et fragilité.

1-3-7 Construction d'un nouvel outil de santé publique

Pour répondre à cette question, nous avons entrepris de nombreux travaux sur le cancer, les chutes répétées ou non, les vaccinations contre la grippe et la fin de vie qui sont venus compléter l'outil de surveillance grippe et gea. Ils sont présentés dans les résultats.

1-4 Problématique générale

Alors que nous venons de présenter les objectifs de cette thèse et les méthodes pour y accéder, toutes celles-ci s'appuient sur l'analyse textuelle des transmissions, mais qu'en est-il de leur réel contenu ? Les informations textuelles ont-elles une réelle pertinence ? Peuvent-elles remplacer de véritables diagnostics cliniques ? Et si oui, dans quel contexte ? Peut-on réellement les utiliser pour répondre à des problématiques de santé publique ? Ainsi, peuvent-elles nous aider à anticiper des épidémies, à établir des trajectoires de vie pour des groupes de résidents, à définir des typologies relativement à une question clinique, ou nous permettre de prédire les besoins en ressources dans les maisons de retraite ? Plus précisément, peuvent-elles nous permettre de générer un outil de surveillance à destination des personnes agées et une cohorte qui apportent des réponses à des questions de santé publique pour cette population fragile ?

Pour tenter de répondre à ces deux questions nous avons exploré au travers d'une revue de la littérature (§ 2), ce qui avait été fait ailleurs sur des sujets similaires et, plus spécifiquement, comment l'information médicale de nature textuelle avait été analysée, transformée et utilisée dans un contexte de santé publique.

Dans ce qui suit, nous verrons donc tout d'abord différents SI anglo-saxons cherchant à résoudre ces deux questions simultanément (§ 2-1). Nous exposerons leur méthodologie et quelques-uns de leurs résultats. Ensuite, nous présenterons plusieurs exemples de SSS en France et à l'Etranger (§ 2-2), et comment l'information médicale est extraite puis analysée. Enfin plusieurs cohortes et panels français et étrangers (§ 2-3) seront introduits avec des problématiques de santé similaires.

La partie suivante présentera nos résultats (§ 3) et comment nous avons résolu cette double problématique (outil de surveillance et cohorte de santé publique). Tout d'abord l'analyse de contenu du DRI et tous les défis liés à l'extraction de l'information textuelle soulevés dans le premier article (§ 3-1). Ensuite la conception d'un SI d'un nouveau type, la Base du Bien Vieillir (BBV) (§ 3-2), s'appuyant essentiellement sur l'analyse et le traitement des transmissions exposée dans le second article : la construction d'une cohorte (§ 3-3) et du SSS afférent pour la grippe et les gea (§ 3-4). Enfin des exemples de travaux réalisés sur la BBV et les résultats

obtenus (§ 3-5).

Ce travail se terminera par une discussion où nous comparerons notre SI aux autres SI existants (§ 4) : ce qu'il est (§ 4-2), ce qu'il n'est pas (§ 4-3), comment poursuivre, améliorer l'information extraite, enfin les perspectives scientifiques qu'il ouvre (§ 4-4).

2-Revue de la littérature

2-1 Des EHR à la santé publique

2-1-1 Les débuts

Les EES / EHR ont été conçus au départ aux Etats-Unis, suivant la vision de l'Institut de Médecine (IOM) de developper une infrastructure informationnelle destinée à soutenir les soins, améliorer la santé des consommateurs, augmenter la qualité et la traçabilité de la politique publique ainsi que celle de la recherche clinique et en santé [19]. L'EES est depuis ses débuts un concept évolutif défini en tant que collection longitudinale d'informations électroniques de santé concernant des patients considérés individuellement ou au sein d'une population. Au départ il a été défini comme un mécanisme d'intégration d'informations sur des soins collectés à la fois au format papier et au format électronique, le but étant d'améliorer la qualité des soins dispensés [20]. D'une part les EES devaient permettre de diminuer les erreurs dues à la saisie manuelle, d'autre part faciliter les besoins en stockage, enfin permettre de pister les interactions médicamenteuses pour un patient particulier ou un groupe de patients. Il s'agissait pour l'IOM de créer une collection longitudinale d'informations électroniques sur et pour les individus et les populations pour réduire les erreurs et développer des systèmes d'aide à la décision en santé. [21 - 22]

2-1-2 EMR et EHR versus EPR

Alors que les EHR ont une visée plutôt longitudinale, les Enregistrements Electroniques Médicaux (EEM / EMR) sont des dossiers électroniques patients créés par des fournisseurs pour des usages spécifiques comme à l'hôpital par exemple avec le PMSI [23] et ses tarifications à l'acte avec le système T2A ou dans des environnements de médecine ambulatoire destinés à coordonner soins de ville et soins à l'hôpital avec des logiciels d'aide à la prescription [24]. Quant aux Enregistrements Electroniques Personnels (EEP / EPR), il s'agit d'applications électroniques pour enregistrer des données médicales personnelles que la personne peut elle-même contrôler et rendre accessibles aux fournisseurs de soins lorsqu'elle le désire. C'est par

exemple l'objectif avoué du DMP, dossier médical personnel ou partagé [25 - 26]. Ainsi, alors que le DRI Korian est plutôt un EPR, la BBV est plutôt un EHR.

2-1-3 Utilisation des EHR pour la santé publique ?

La promesse des technologies de l'information: apporter de la donnée de santé organisée et disponible, instantanément, de telle sorte que l'on puisse prendre soin du patient de manière ajustée et adaptée à ses besoins, avec une réduction globale des coûts [27]. Et pour cela remédier aux disparités de santé et progresser dans les soins apportés aux malades chroniques : maladies cardio-vasculaires, asthme, diabète. Ensuite, améliorer la surveillance en santé publique en monitorant la morbidité grippale, l'efficacité des vaccins, le génotypage, en rapportant les maladies à déclaration obligatoire et chaque cas, en tenant des registres de cancer et d'autres pathologies, en communiquant à propos des soins cliniques, de l'immunisation des patients, en définissant des messages d'alertes à usage du public. Enfin encore, la dématérialisation des naissances et décès ou encore la diminution des décès dues à des surdosages médicamenteux. [28]

"We have lots of information technology. We just don't have any information."

Alors que la surveillance en santé publique a bien progressé en qualité, rapidité et précision dans les services de santé aux Etats-Unis grâce au reporting des laboratoires, il reste que la collecte d'information clinique détaillée pour rapporter les cas, confirmer les diagnostics, pour comprendre la transmission ou déterminer des facteurs de risque associés à une infection dépendent encore fortement de processus manuels [29].

De plus, par manque de ressources, les systèmes EHR actuels sont d'abord construits pour servir la pratique clinique plutôt que structurés pour être utilisés pour la santé publique [30]. Enfin, il y a très souvent des questions de politique publique concernant leur utilisation et diffusion, notamment en France au travers de la loi informatique et libertés [31 - 32].

Néanmoins, les prévalences et fonctionalités croissantes des EHR devraient accroître la qualité de la surveillance en santé publique en apportant des informations meilleures pour guider les interventions publiques et rapprocher pratique et clinique et conduire à un système de santé publique plus efficace et efficient [29-30].

2-2 La surveillance syndromique

2-2-1 Le concept de surveillance

"What gets measured gets done." —Anonymous

La surveillance en santé publique fait référence à la collecte, analyse et interprétation de données pour viser la prévention en santé. Elle est le fondement de la pratique en santé publique. Elle est un outil pour estimer les statuts et comportements en santé des populations. Comme la surveillance peut mesurer directement ce qui se passe dans la population, elle est utile à la fois pour mesurer les besoins en interventions en santé et les effets produits [33].

2-2-2 Le concept de surveillance syndromique

La surveillance syndromique est définie par le CDC comme : "une approche, dans laquelle les intervenants sont assistés par des procédures d'enregistrement automatiques des données, qui permettent la mise à disposition de données pour le suivi et l'analyse épidémiologique en temps réel ou proche du temps réel. Cela afin de détecter des événements habituels ou inhabituels plus tôt qu'il n'aurait été possible de le faire sur la base des méthodes traditionnelles de surveillance" [34]. Parmi ces événements, on trouvera par exemple un nombre important (significativement supérieur à un seuil d'alerte défini statistiquement à partir de périodes précédentes) de cas d'une maladie à déclaration obligatoire comme la grippe ou la gastroentérite aiguë, ou l'apparition de phénomènes pathologiques inexpliqués ou inconnus [35] comme lors de la canicule de 2003 [36].

Les quatre étapes nécessaires au fonctionnement d'un SSS présentées dans la figure 4 sont les suivantes [37] :

1- La **collecte des données**, signes cliniques et symptômes, auprès des hôpitaux, des services d'urgence, des laboratoires.

2- La **gestion** et l'**analyse** statistique et épidémiologique des données collectées à l'aide de différents outils : analyse descriptive des données, génération de seuils d'alertes statistiques, aide à la décision.

3- La **communication** des résultats aux autorités sanitaires à l'aide de bulletins et d'alertes.

4- Enfin, le **partage des données** et des résultats à destination du public et au sein des autorités sanitaires.

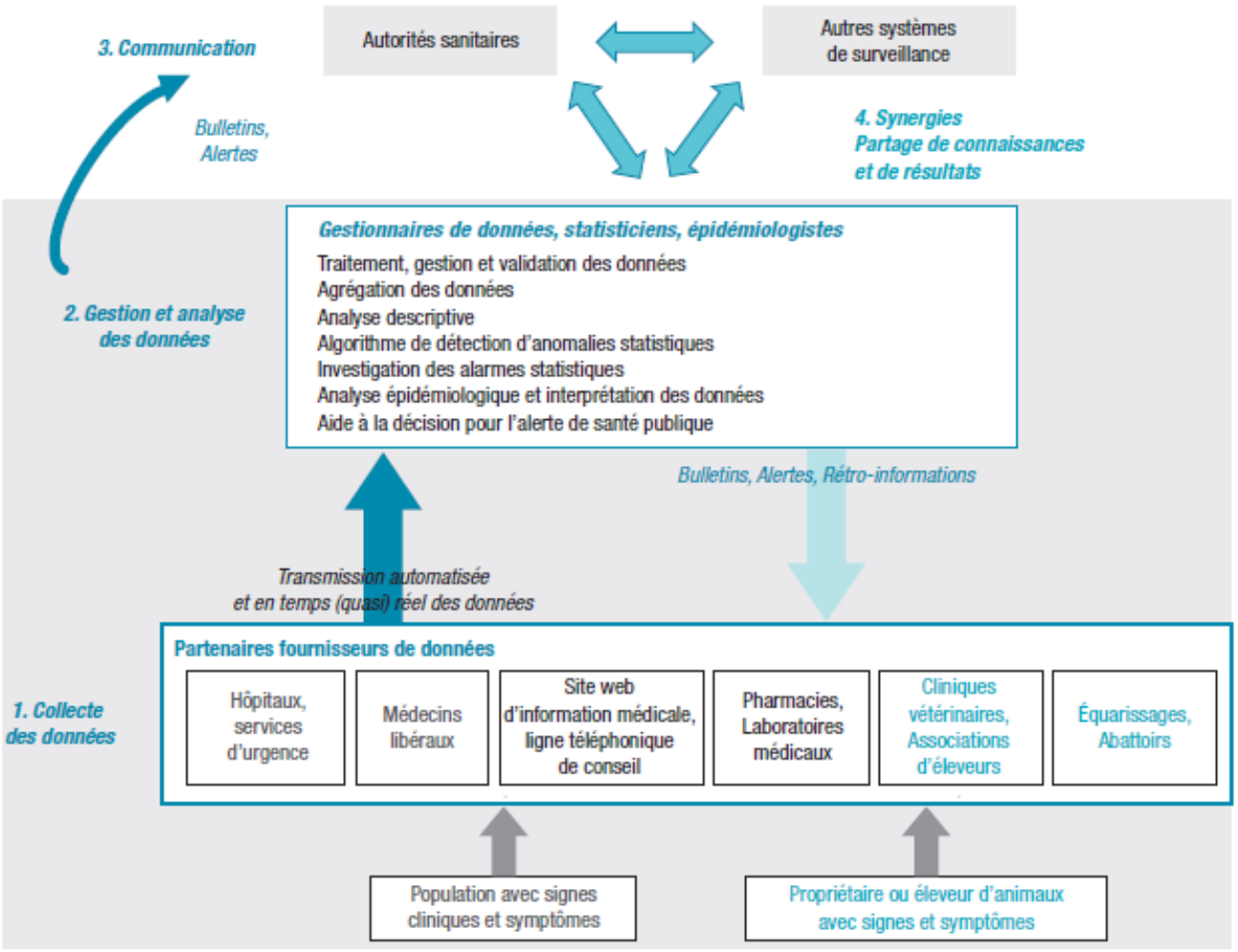


Figure 4: L'articulation des quatre composantes d'un Système de Surveillance Syndromique

2-2-3 Quatre exemples de SSS

SurSaUD:

Depuis 2003, l'Institut de veille sanitaire a développé un SSS basé sur la collecte de données non spécifiques [38]. Le système permet la centralisation quotidienne d'informations, provenant, au 1er février 2015, de :

a- un peu plus de 600 services d'urgences participant au réseau de surveillance coordonnée des urgences (OSCOUR®) ;

b- 60 associations SOS Médecins (données de médecine d'urgences de ville) ;

c- 3 000 communes, pour les données de mortalité, par l'intermédiaire de l'Institut

national de la statistique et des études économiques (INSEE).

Ci-dessous les 3 composantes de SurSaUD [39].

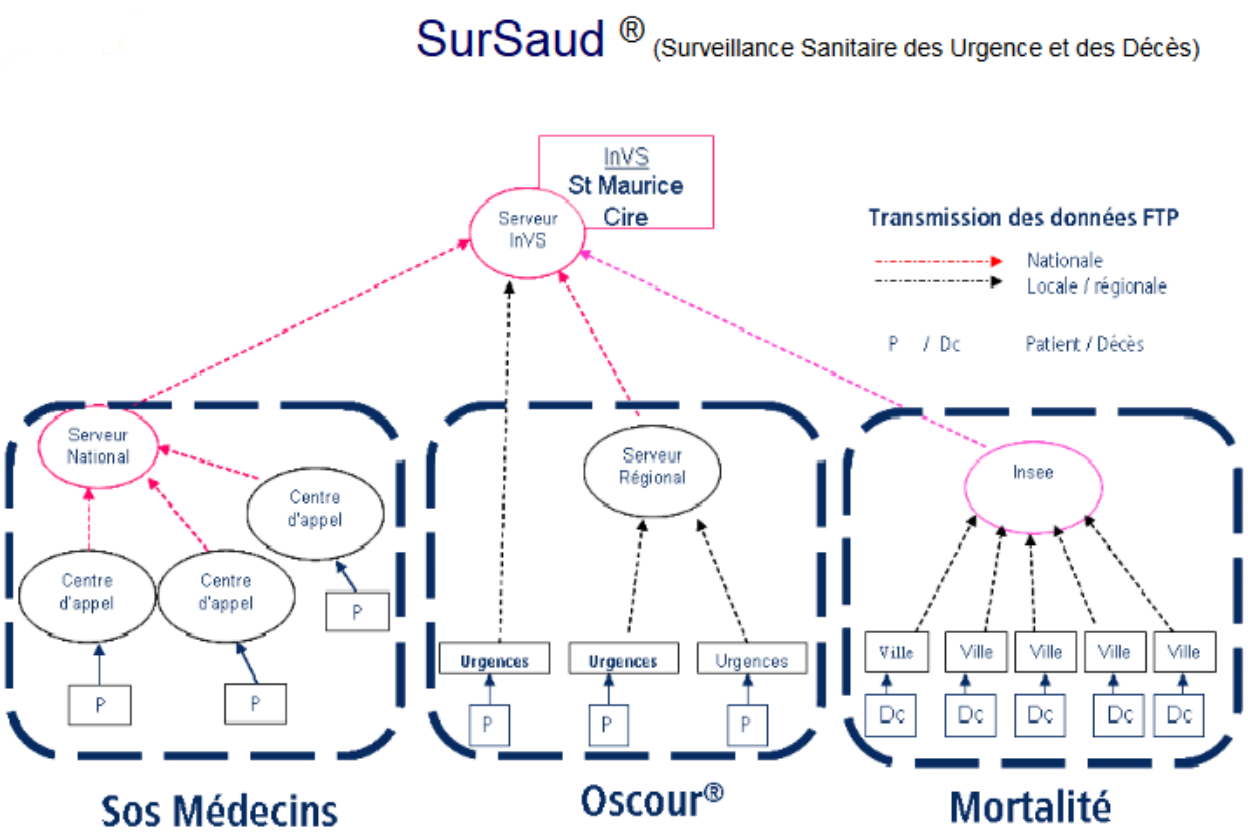


Figure 5: Architecture du Système de Surveillance Syndromique SurSaUD

Le Réseau Sentinelles:

L'INSERM et l'Université Pierre et Marie Curie ont développé un SI basé sur un réseau de médecins généralistes en France métropolitaine, appelé le réseau Sentinelles [40]. Créé en 1984, le réseau Sentinelles est composé de 1300 médecins généralistes libéraux (soit 2,2% de la totalité des MGL en France métropolitaine) et d'une centaine de pédiatres libéraux, volontaires, répartis sur le territoire métropolitain français. Les médecins Sentinelles collectent de façon continue des informations sur 9 indicateurs de santé dont les syndromes grippaux et la diarrhée aiguë.

A partir des données de leurs patients transmises chaque semaine, il est possible d'estimer le taux d'incidence hebdomadaire pour chaque indicateur et de suivre son évolution dans le temps et dans l'espace. Les axes principaux de recherche du réseau sont:

1 - la validation d'outils de détection et de prévision de la dynamique d'une épidémie ;

2 - l'estimation en temps réel de l'efficacité du vaccin antigrippal ;

3 - la modélisation des maladies infectieuses dans une optique d'aide à la décision.

The Essence II biosurveillance system:

Electronic Surveillance System for the Early Notification of Community-based Epidemics, version 2 a été développé au sein d'une collaboration entre le Department of Defense Global Emerging Infections System (DoD GEIS) et le Applied Physics Laboratory (APL) de la John Hopkins University. [41]

Fonctionnel en 2003, ESSENCE II utilise des indicateurs de santé non traditionnels par regroupements syndromiques couplés avec des techniques analytiques adancées comme le NLP (Natural Language Processing – Traitement Automatique du Langage). Parmi les syndromes traités on trouve : les décès, les problèmes d'origines gastro-intestinaux, neurologiques, dermatologiques, respiratoires, infectieux et non spécifiques. Les sources de données analysées pour la SS incluent les appels au 911, les appels aux hotlines des services infirmiers, des centres anti-poisons, les visites chez les médecins libéraux et les cliniques militaires, des requêtes auprès de laboratoires, les visites aux urgences et les prescriptions médicamenteuses. Chaque groupe syndromique est défini par un ensemble de symptômes et cas définis par la CIM 9 (Classification Internationale des Maladies – International Classification of Diseases) advenant aux prémisses de la maladie. ESSENCE II monitore quotidiennement chaque groupe syndromique pour identifier des anomalies éventuelles en fonction de la saison et de l'évolution endémique de certaines souches.

ESSENCE II est composé de plusieurs composants techniques en parallèle, mis à jour en permanence et communicant entre-eux de manière sécurisée et dé-identifiée [42] (voir la figure 6) : ESSENCE II comprend ainsi 3 entités logiques :

1 – les hôpitaux qui génèrent les données ;

2 – les SSS qui analysent les données dé-identifiées et les codifient;

3 – les utilisateurs qui récupèrent ces informations syndromiques anonymisées.

Parmi les utilisateurs on trouve les hôpitaux participants, mais également les services de santé locaux ainsi que les développeurs de cette application.

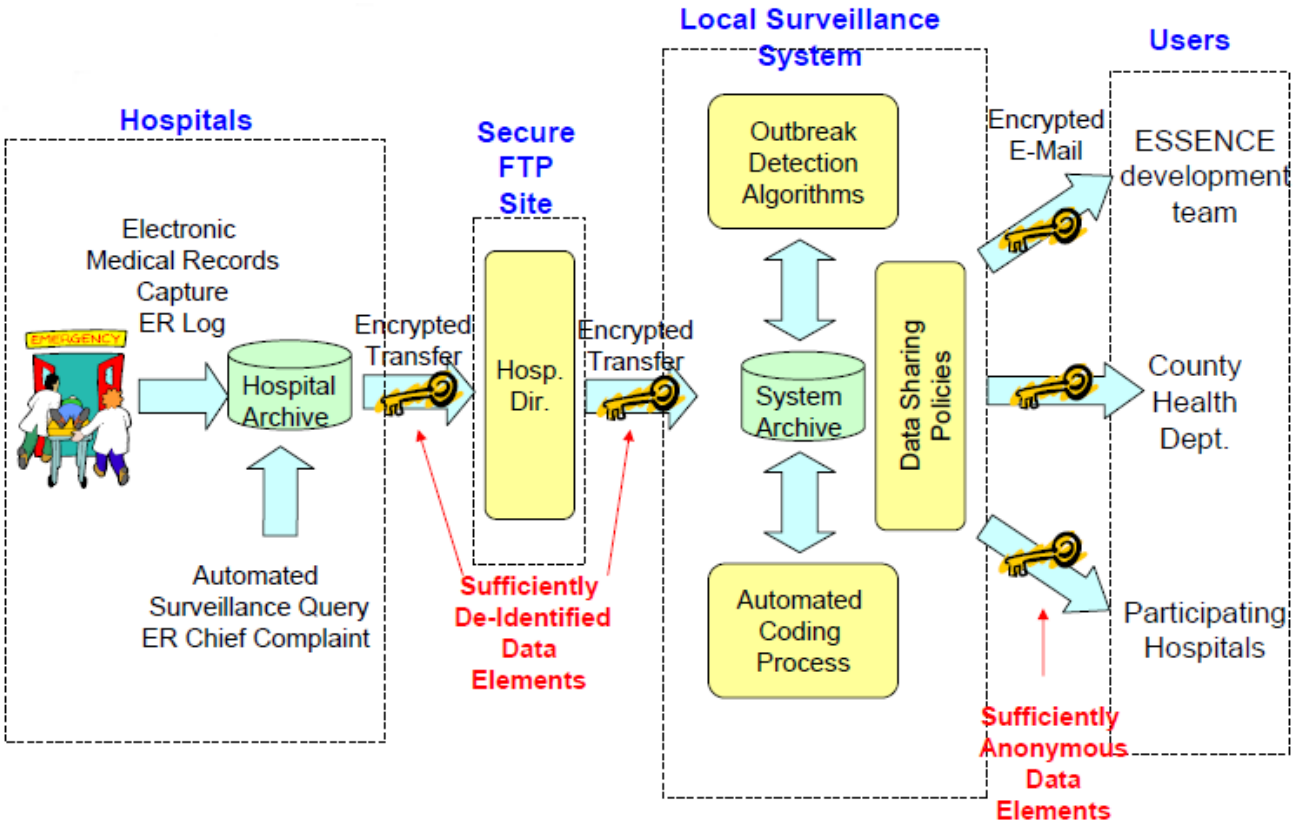


Figure 6: La collecte des données (ETL), leur transmission sécurisée (dé-identifiée), leur codification (NLP + ICD-9) et enfin leur utilisation sous forme anonymisée, aux instances de santé publique au sein du Système de Surveillance Syndromique ESSENCE II

Depuis 2003, l'APL et l'Armed Forces Health Surveillance Center (AFHSC) ont réalisé une suite logicielle flexible et opensource pour la surveillance électronique des maladies disponible pour toutes sortes d'environnements et finalisée en 2013 : SAGES, Suite for Automated Global Electronic bioSurveillance (SAGES) [43]. Le système ESSENCE est maintenant intégré dans le National Surveillance System Program du CDC avec les composants OpenESSENCE et ESSENCE Desktop Edition.

ESPnet:

La plateforme de surveillance ESPnet, Electronic Support for Public health, a été développée par le Harvard Center of Excellence in Public HealthInformatics et le Massachusetts Department of Public Health avec des financements du CDC [44 – 46]. Le système est utilisé par quatre services de soins ambulatoires dans le Massachusetts et l'Ohio avec les départements de santé associés et sert 2

millions de personnes.

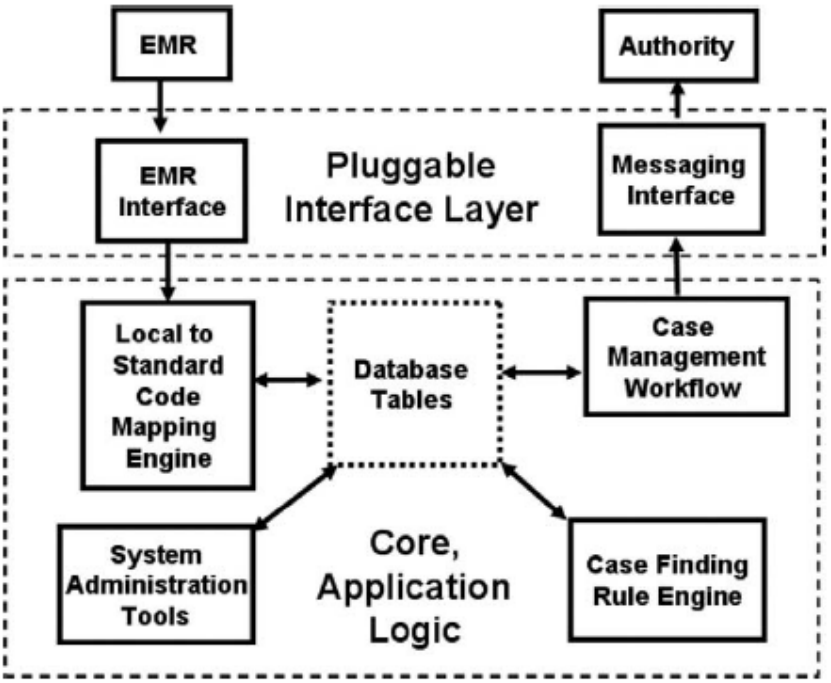


Figure 7: Architecture de la plateforme ESPnet

Le serveur EMR est programmé pour renvoyer des données structurées de chaque cas patient au réseau ESPnet soit en temps réel, soit de nuit. Chaque cas en provenance de données EMR est analysé rigoureusement : les algorithmes ESPnet utilisent conjointement la codification des diagnostics des médecins, les prescriptions, les tests des laboratoires et certains signes vitaux pour identifier des maladies et des conditions sanitaires d'intérêt.

Ainsi, par exemple détecter et notifier un cas de tuberculose (TB) nécessite :

- un code diagnostic TB,
- un test de laboratoire pour la TB,
- la prescription de 2 ou plus d'anti-tuberculeux.

Peuvent ainsi être analysés et détectés la tuberculose, Chlamydia, la gonorrhée, la syphillis, la maladie de Lyme, les hépatites A, B et C, les maladies chroniques comme l'asthme, les diabètes de types 1 et 2, l'obésité, l'hyperlipidémie et le fait de fumer, enfin la grippe et les effets indésirables liés aux vaccinations.

Le plus gros défi pour ESPnet est la mise à jour en continu des algoritmes de détection NLP lorsque de nouveaux tests sont mis en place ou lorsque les traitements ou les conditions de détection évoluent, le but étant de calibrer chaque algorithme de telle sorte qu'il maximise la sensibilité et la

valeur prédictive positive de la pathologie pour ne pas surcharger les services sanitaires avec des faux positifs [44].

2-3 Les cohortes de personnes âgées en France et à l'étranger

2-3-1 Cohortes versus panels

Une cohorte épidémiologique est un type d'enquête dont le principe est le suivi longitudinal, à l'échelle individuelle d'un groupe de sujets [47] partageant des caractéristiques semblables. Les cohortes en population générale sont souvent généralistes et analysent de nombreux problèmes de santé ainsi que leurs déterminants. L'effectif de l'échantillon à observer pour une précision donnée dépend de la fréquence du phénomène dans la population et, pour que cet effet puisse être mesuré correctement au sein d'une cohorte, il faut qu'il n'y ait pas de biais de sélection. Seule l'exhaustivité des données dans une population comme dans les bases médico-administratives (SNIIRAM [48 - 49], PMSI [23]) peut permettre d'éviter certains biais.

Dans une étude de panel l'unité d'analyse initiée à l'inclusion est répétée à intervalles de temps spécifiés à l'avance et dure souvent plusieurs années [50]. La plupart de ces études exploitent des données quantitatives et structurées et examinent les exacts changements au cours du temps, pour différentes tranches d'âge et parfois durant toute la vie.

Ci-dessous sont listées et résumées les principales cohortes 'généralistes' françaises sur le vieillissement. D'autres existent suivant des axes de recherche plus spécifiques comme par exemple la cohorte RIEHO qui étudie les conditions d'hospitalisation en urgence des personnes âgées de plus de 80 ans et leur suivi pendant une année ensuite, ou la cohorte SUVI MAX 2 qui examine les effets de l'alimentation sur le vieillissement ou encore la cohorte SAFES qui cherche à évaluer la fragilité .du sujet âgé suite à une admission aux urgences et après une évaluation gérontologique standardisée.Toutes sont présentées sur le portail Epidémiologie France dans la partie cohortes + vieillissement [51]. Ensuite sont présentées 3 exemples de panels sur le vieillissement. Le premier, HRS, la US Health and Retirement Study [52] est à l'origine des autres panels qui étudient le passage à la retraite sous les aspects physiques, mentaux, sociaux et économiques.

2-3-2 Les principales cohortes sur le vieillissement en France

Nom court	Thématique	Domaine médical	Finan-cement	Responsable ou Promoteur	Statut	Type de BD Effectif	Critères d'inclusion	Inter-ven-tion-nel	Période du recueil	Durée du suivi	Objectif principal	Objectifs secondaires	Ce qui est recherché Appariement	Données recuei-llies
EVA	Cardiologie Neurologie	Déclin, fonctions cérébrales, fonctions cognitives, événements de santé	Privé	INSERM	Public	Issues d'enquêtes 1 389	2 sexes nés entre 1922 et 1932, entre 49 et 79 ans résidant à Nantes	Non	[07/1991 – 12/2001]	9 ans	Suivi longitudinal du vieillissement vasculaire déclin des fonctions cognitives	FR biologiques liés au stress oxydatif	Evénements de santé morbidité mortalité	Examens cliniques périodiques Auto-questionnaires papier en face à face Echographie crânienne IRM Prélèvements biologiques
PAQUID	Neurologie Psychologie et psychiatrie	Vieillissement cérébral, démence sénile, entrée en institution, Alzheimer, dépression, dépendance, mortalité	Mixte	INSERM	Public	Issues d'enquêtes 3 777	2 sexes 65 ans et plus : 65 à 79 et 80 ans et plus résidant en Dordogne et Girond	Non	[01/1998 – 01/2018]	30 ans	Vieillissement cérébral et fonctionnel après 65 ans, distinguer modalités normales et pathologiques identifier les sujets à haut risque	Action préventive pour les sujets à haut risque	Evénements de santé morbidité Mortalité CepiDC, CG Gironde	Examens cliniques périodiques Auto-questionnaires papier en face à face Prélèvements biologiques ADL, IADL, MMSE médicaments consommés
GAZEL	Biologie, cancérologie, cardiologie, déficiences et handicaps endocrinologie et métabolisme hématologie, médecine du travail, neurologie, pneumologie, psycho et psychiatrie, rhumatologie, traumatologie	Cohorte généraliste, déterminants sociaux, adultes, vieillissement, risques professionnels	Public	INSERM	Public	Issues d'enquêtes 20 625	2 sexes, agents EDF 35-50 ans en 1989 France entière	Non	[1989 – 2012]	> 22 ans	Migraine, ostéoporose, maladies cardio-vasculaires, dépression, troubles musculo-squelettiques, accidents de la route Etude du vieillissement cognitif et fonctionnel	Qualité de vie, inégalités sociales de santé chez les personnes âgées	Evénements de santé morbidité Mortalité consommation de soins, qualité de vie perçue, hospitalisations CepiDC, SNIIRAM, ALD, diagnostics hospitalisations avec PMSI	Examens cliniques périodiques Auto-questionnaires papier envoyés par la poste Prélèvements biologiques, fonctions cognitives (MMSE) et physiques
E3N E4N	Allergologie, biologie, cancérologie, déficiences et handicaps, dermatologie vénérologie, endocrinologie,	Génétique, modes de vie et comportements produits de santé, vieillissement	Mixte	INSERM - IGR	Public	Issues d'enquêtes 100 000	Femmes adhérentes à la MGEN nées entre 1925 et 1950, 45 ans et plus, 3 classes	Non	[1990 – 12/2014]	En cours	Recherche de FR (modes de vie, alimentation, traitements hormonaux) des cancers		Evénements de santé morbidité Mortalité Mode de vie et état de santé CepiDC,	Examens cliniques périodiques dont un sous-échantillon de 25 000 femmes Auto-questionnaires papier envoyés par

	métabolisme, gastro-entérologie, hépatologie, gériatrie, gynécologie, hématologie, neurologie, pneumologie, psychologie et psychiatrie, radiologie et imagerie, rhumatologie, traumatologie						d'âge : moins de 65, de 65 à 79 et 80 ans et plus				et pathologies chroniques		SNIIRAM	la poste et téléphone Prélèvements biologiques, imagerie, comptes-rendus anatomopathologique pour 80% des cancers CIM9 et CIM10 pour les pathologies dont Parkinson, Endométriose, asthme, hypertension, maladie de Crohn, dépression…
3C	Neurologie, psychologie et psychiatrie, radiologie et imagerie	Génétique, mode de vie et comportements, produits de santé, plan Alzheimer	Mixte	ISPED	Public	Issues d'enquêtes 9 294	2 sexes des 3 villes : Bordeaux,Dijon et Montpellier, de 65 à 79 et 80 ans et plus	Non	[03/1999 – 08/2013]	4 ans, collecte terminée	Impact des facteurs vasculaires sur le risque de démence dans la population âgée	Incidence et FR des pathologies cardio-vasculaires, consommation de médicaments, troubles de la marche et équilibre, incidence et FR de la perte d'autonomie, dépression	Evénements de santé morbidité, consommation de santé et produits de santé Mortalité CNAMTS et PMSI	Examen clinique à l'inclusion Auto-questionnaires papier périodiques en face à face, Prélèvements biologiques, imagerie,
COPANFLU	Maladies infectieuses	Facteurs sociaux et psycho-sociaux, intoxication, mode de vie et comportements, infections, virus A (H1N1), syndromes infectieux respiratoires	Public	EHESP	Public	Issues d'enquêtes 1 451	1000 ménages volontaires en France avec consentement : 2 sexes et toutes les classes d'âge	Non	[08/2009 – 12/2012]	2 ans collecte terminée	Déterminants épidémiologiques, environnementaux, immunologiques, sociaux, génétiques et virologiques du risque d'infection par le virus A (H1N1) et swine-like (SWL)	Impact : morbidité, complications, consommation de ressources en santé, histoire naturelle et variabilité interindividuelle, modifications de comportements, niveau de perception du risque, caractériser l'immunité homotypique(à vie par contact) et hétéro subtypique (croisée)	Evénements de santé morbidité, mortalité, exposition au risque environnemental, perception du risque, qualité de vie	Auto-questionnaires papier périodiques en face à face, téléphone : saisie avec double saisie, examens biologiques

Table 2 : Les cohortes observationnelles sur le vieillissement en France

2-3-3 Quelques pannels sur le vieillissement en Europe et ailleurs

Nom court	Thématique	Domaine médical	Finan-cement	Responsable ou Promoteur	Statut	Type de BD Effectif	Critères d'inclusion	Inter ven-tionnel	Période du recueil	Durée du suivi	Objectif principal	Ce qui est recherché	Données recuei-llies
HRS	Le passage à la retraite et le processus du vieillissement dans sa globalité [52]	BMI, ADL, iADL, tabac, alcool, activité physique, autonomie, dépression, fonctions cognitives, événements de santé	Alzheimer Association, AEAR, AARP, AoA, Eldercare Locator…	National Institute on Aging	Public	Issues d'enquêtes	2 sexes âgés de 50 ans et plus aux Etats-Unis	Non	Depuis 1996	21 ans en cours	Suivi longitudinal du vieillissement et du passage à la retraite sous ses dimensions sociales, économiques et de santé	Étude des transitions sociales, familiales et économiques liées au passage à la retraite, FR sur la durée de vie et sa qualité	Entretiens tous les 2 ans jusqu'au décès Age, sexe, statut marital, santé mentale, activité professionnelle et socio-économique, bénévolat, religion, qualité de vie, espérance de vie,
ELSA	Santé, statut économique et qualité de vie au cours du temps [53 - 54]		Department of Health, Department for Transport, the Department for Work and Pensions and the National Institute on Aging (USA).	The Institute for Fiscal Studies	Public	Issues d'enquêtes	2 sexes âgés de 50 ans et plus en Grande Bretagne	Non	Recrutement depuis 2002 à partir de la cohorte anglaise BCS70	Plus de 40 ans au total	idem	idem	Entretiens tous les 2 ans jusqu'au décès, âge, sexe, handicap, santé, travail, retraite, revenus, activité sociale et culturelle, fonctions cognitives
SHARE	Le passage à la retraite et le processus du vieillissement dans sa globalité [55]	BMI, ADL, iADL, tabac, alcool, activité physique, autonomie, dépression, fonctions cognitives, événements de santé	Commission Européenne	Munich Center for the Economics of Aging Max PlanckInstitute for Social Law and Social Policy	Public	Issues d'enquêtes Plus de 120 000	2 sexes âgés de 50 ans et plus des 27 pays européens + Israël	Non	2004 vague 1 – 2017 démarrage de la vague 7	13 ans en cours	idem	Niveau socio-économique à l'entrée puis étude des transitions sociales, familiales et économiques, relations intra –familiales et inter-générationnelles, fin de vie et décès	Auto-questionnaires papier en face à fac (plus de 297 000 entretiens) sauf pour les premiers entretiens de la vague1 assistés par ordinateur tous les 2 ans Age, sexe, statut marital, santé mentale, activité professionnelle et socio-économique, bénévolat, religion, qualité de vie, espérance de vie, connaissances et utilisation de l'informatique, Prélèvements sang

Table 3 : Quelques panels observationnels sur le vieillissement dans le Monde

3- Résultats

3-1 Du DRI à la BBV

Tout le processus d'analyse de contenu du DWH DRI et de l'analyse de contenu de l'information textuelle présenté aux paragraphes 1-2-1 et 1-2-2 (objectifs), puis aux paragraphes 1-3-1 et 1-3-2 (méthodes) a donné lieu à l'étude qualitative et quantitative d'une partie des transmissions (les traitements kiné) réalisée dans le premier article.

3-1-1 Résumé des résultats du premier article

En combinant extractions par requêtes SQL, textmining, ACP et ACM, suivies de classifications, nous avons caractérisé les récits de prise en charge relatifs aux traitements kiné par une liste de mots-clés construits ainsi que les besoins en soins de kinésithérapie pour des groupes de résidents sélectionnés au préalable. Nous avons réussi également à détecter des défauts d'alimentation de certains champs ainsi qu'un groupe de résidents avec des valeurs extrêmes. Enfin, les données de kinésithérapie et les données de santé de notre échantillon de résidents ont été agrégées et des différences de situation de santé associées à des différences qualitatives et quantitatives de données ont été détectées.

Cette étude textuelle empirique au moyen d'un traitement textuel en deux étapes (voir paragraphe 1-3-2) a montré que les techniques de text mining et de data mining pouvaient fournir des outils accessibles pour améliorer la santé des résidents et la qualité des soins en ajoutant de nouvelles données simples et utiles aux EES. Le text mining, employé ici sur une liste de traitements kiné normalisés et utilisant l'extraction d'informations (information extraction IE), la reconnaissance d'entités nommées (named entity recognition NER) et le data mining (DM), peut apporter un réel avantage pour décrire les soins, tout en apportant de nouvelles informations médicales et en aidant à intégrer le système des EES dans l'environnement de travail du personnel médical et paramédical.

3-1-2 Premier article

Delespierre *et al. BMC Medical Informatics and Decision Making* (2017) 17:127
DOI 10.1186/s12911-017-0519-0

BMC Medical Informatics and Decision Making

RESEARCH ARTICLE Open Access

Empirical advances with text mining of electronic health records

T. Delespierre[1,2*], P. Denormandie[3], A. Bar-Hen[4] and L. Josseran[2]

Abstract

Background: Korian is a private group specializing in medical accommodations for elderly and dependent people. A professional data warehouse (DWH) established in 2010 hosts all of the residents' data. Inside this information system (IS), clinical narratives (CNs) were used only by medical staff as a residents' care linking tool.
The objective of this study was to show that, through qualitative and quantitative textual analysis of a relatively small physiotherapy and well-defined CN sample, it was possible to build a physiotherapy corpus and, through this process, generate a new body of knowledge by adding relevant information to describe the residents' care and lives.

Methods: Meaningful words were extracted through Standard Query Language (SQL) with the LIKE function and wildcards to perform pattern matching, followed by text mining and a word cloud using R® packages. Another step involved principal components and multiple correspondence analyses, plus clustering on the same residents' sample as well as on other health data using a health model measuring the residents' care level needs.

Results: By combining these techniques, physiotherapy treatments could be characterized by a list of constructed keywords, and the residents' health characteristics were built. Feeding defects or health outlier groups could be detected, physiotherapy residents' data and their health data were matched, and differences in health situations showed qualitative and quantitative differences in physiotherapy narratives.

Conclusions: This textual experiment using a textual process in two stages showed that text mining and data mining techniques provide convenient tools to improve residents' health and quality of care by adding new, simple, useable data to the electronic health record (EHR). When used with a normalized physiotherapy problem list, text mining through information extraction (IE), named entity recognition (NER) and data mining (DM) can provide a real advantage to describe health care, adding new medical material and helping to integrate the EHR system into the health staff work environment.

Keywords: Nursing homes, SQL query, Information extraction, Named entity recognition, Data mining, Text mining, Word cloud, Multiple component analysis, Principal component analysis, Hierarchical clustering

Background

Issues with nursing care narrative (CN) analysis [1], as well as with electronic health record (EHR) analysis [2–4], are recurrent, but data warehousing (DWH) [5] and cloud computing developments on the one hand [6] and data extraction techniques on the other hand [7, 8] have changed the way CN analysis is performed and used. Today, EHR is a valuable source of clinical information [4, 9], but the abundance of unstructured textual data in EHR presents a real challenge to realizing its full potential [8, 10, 11]. EHR, with the parallel rapid growth of CN, plus the need for improved quality of care and reduced medical errors, is a strong incentive for the development of natural language processing (NLP) [8]. The free text of the CN is a rich resource, in which health staff record events or information history as told to them by their patients [12] or the care provided linked to residents' health status. However, working with this textual material depends strongly on the availability of NLP tools and ex- pertise in using them. Much of the available clinical data from DWH are in narrative form and can be

Correspondence: tiba.baroukh@gmail.com
[1]Institut du Bien Vieillir Korian, 21-25 rue Balzac, 75008 Paris, France
[2]Research lab: EA 4047, UFR des Sciences de la Santé Simone Veil, UVSQ Université Paris-Saclay, 2 Avenue de la Source de la Bièvre, Montigny le Bretonneux 78180, France
Full list of author information is available at the end of the article

used as a convenient tool by health providers and medical staff. Thus, textual data are used most of the time through a professional data frame as a networking tool between health care professionals and not as a pertinent, thera- peutic decision-making tool.

The (Korian) group specializes in medical nursing homes and acute and subacute care clinics. Since 2010, this European group has had health records with data coming from both the medical and social dimensions in four countries (France, Germany, Italy and Belgium), built through several professional DWHs using OR- ACLE® technology [13]. For every new resident admitted to one of its nursing homes (NHs), a personal electronic resident medical file is opened to collect individual patient data at different time points of the stay: at admis- sion (admission date, medical history, marital status, birth- date, tastes and habits), on a daily basis (new pathologies, chronic disease evolution, date of death, drug prescrip- tions), or after specific medical or health care professional visits. Whereas data are recorded by different profes- sionals following the resident's situation, through struc- tured, layered and indexed data organized through a hierarchically data scheme, CNs are mostly captured along the way in a few tables and are used extensively as net- working tools. While these textual data include direct speeches, acronyms or simply words, could they actually contain relevant and reusable health information?

The objective of this study was therefore twofold: first, to illustrate how CN text mining processes could enhance electronic medical record (EMR) data; and then, to demonstrate the convergence of information between CN-extracted data and EMR data strictly speaking.

Physiotherapy care data were chosen for two main reasons: first, as a meaningful tool preserving motor function in frail, elderly people and preventing them from a long list of physical ailments, second as well-defined interventions directed to an identifiable target population inside the NH population [14]. The adjusted size and the well-defined sample allowed checking every step of the process, opening the way towards syndrome la- belling, patient stratification and improved targeting of care [15].

Methods

Aim, design and settings

Our goal was to show through this whole textual experiment (see Fig. 1) that text mining through IE, NER and DM techniques [8] could be essential to better follow resi- dents' health paths and improve their quality of care by adding new, simple, useable data, as well as valuable and matching information with the already existing EHR data.

The design to extract textual data from physiother- apy care narratives involved two information systems: the DWH, using ORACLE® queries to extract data from tables and build the physiotherapy corpus; and R Studio® for the statistics and data mining [16], as well as the textual analysis of the corpus (in black and dark blue, respectively, in Fig. 1).

Characteristics of participants and material

The data source for the corpus analysis of clinical text was built through the selection of all of the residents alive on September 30, 2013 with at least one physiotherapy narrative during the previous 6 months [17]. This corpus contained a total of 4051 physiotherapy CNs for 1015 residents from 127 nursing homes located in eleven regions of France during the period from 04/01/2013 to 09/30/2013. These records were extracted from one table, de-identified, and anonymized; see Additional file 1: Annex 1 for the physiotherapy corpus and Additional file 2: Annex 2 for the anonymization process [18–20]. Physiotherapy keywords and the most frequent corpus expressions were translated to generate all of the tables and Figs. 3 and 4. For full details of this process, see Additional file 3: Annex 3.

Data were also gathered from other tables (right side of Fig. 1): socio-demographic data with the residents' ages and sexes, medical histories and pathologies on September 30, 2013, as well as their falling histories from the beginning of their NH stay through September 30, 2013; on the building of the 1015 anonymized residents' health tables, see Additional file 4b: Annex 4.

In this sample, medical histories, pathologies and fall variables were cast from multiple event variables to synthetic counting variables to have only one record per resident (see Table 2).

Description of the processes and statistical analysis *Building the physiotherapy variables: The textual SQL plus text mining strategy*

The physiotherapy comments could take different formats (date, one or two words, sometimes even dull words, such as 'others' or No, pre-constructed small sentences or free text) depending on the professional feeding the system. By analysing several hundreds of them through the two main stages described in Fig. 1, a list of physio keywords was built, where most of the time, one physio keyword stood for one physio concept.

By combining precisely monitored data mining techniques: designing specific physiotherapy textual SQL quer- ies [21] and then going through unsupervised queries with text mining [22], a precise and easy-to-check textual de- scription of the physiotherapy treatments was built with relevant physiotherapy variables.

The first stage was done on the DWH's side and is detailed on the workflow Fig. 1:

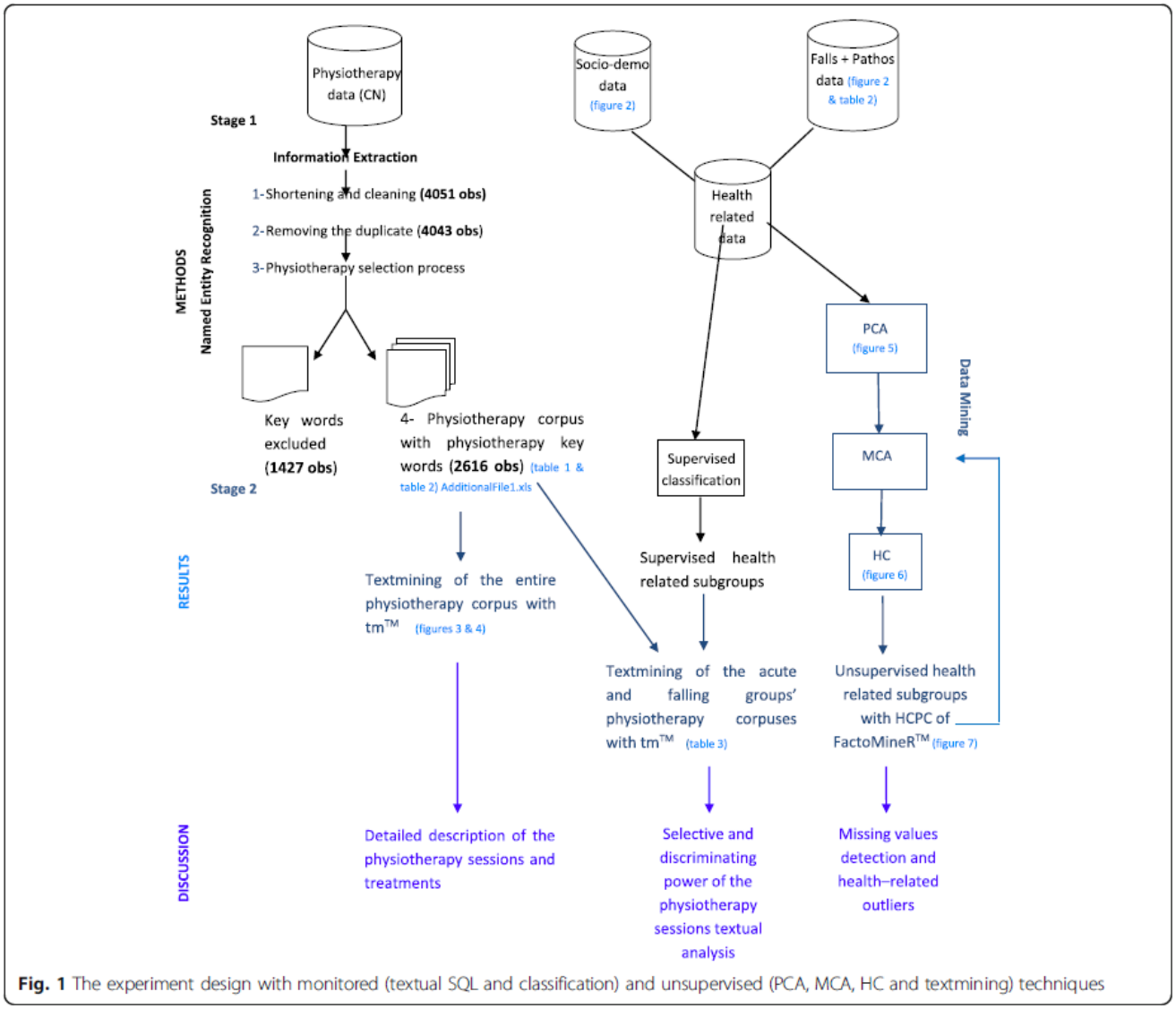


Fig. 1 The experiment design with monitored (textual SQL and classification) and unsupervised (PCA, MCA, HC and textmining) techniques

1- shortening the character fields from 4000 to 300 and removing accentuation, dates or words such as *No*;
2- removing some overlapping CN;
3- excluding meaningless words or expressions CN using ORACLE® queries with the SQL LIKE function and wildcards to perform pattern matching [21];
4- building physiotherapy variables with the same technique to detect some keywords, such as '*walk*', '*balance*' or '*autonomy*' (see Table 1), describing physiotherapy care and counting them.

Every CN with one of these words was counted as one word occurrence. These words were selected after extensively checking and listing the CNs and iteratively checking the results. Among these built-in physiotherapy variables, we met three different situations:

a- one word-one concept always alone for words such as '*useless*';
b- one word-one concept alone most of the time and sometimes combined in sentences for words such as '*partial*', '*antalgic*', '*good*', '*functional recovery*' and' *autonomy maintenance*';
c- one word-one concept always combined in sentences for words such as '*massage*', '*pain*', '*cognition*', '*balance*' and '*walk-walking*'.

Through this process, we could also analyse the defined observations' lengths greater than or equal to 30 as those belonging to the free text class, using a cut-off from the lengthiest of our classified words and expressions (26 characters in French), and the non-classified class as those not containing any of the keywords listed in Table 1. These two classes as tools helped us to count and check all the meaningful words. Finally, for the physiotherapy variables described in cases 3 and a above,

Table 1 The physiotherapy corpus built through the SQL process (stage 1) with selected one physio expression–one physio concept and their precision

word	frequency	percentage	Precision a priori	Sentence precision	Precision a posteriori
free_text	*707*	*17.45*	*N/A*	*N/A*	*N/A*
autonomy	560(524)	13.82	93.57	80.56	98.75
renewal	518	12.79	100	-	100
per_week	476	11.75	100	-	100
good	383(327)	9.45	85.37	94.64	99.22
functional_recovery	358	8.84	100	-	100
non_classified	*318*	*7.85*	*N/A*	*N/A*	*N/A*
walking	302	8	-	100	100
partial	295(294)	7.28	100	100	100
antalgic	222(208)	5.43	100	100	100
per_day	209	5.16	100	-	100
others	194	4.79	100	-	100
pain	86	1.88	-	96.51	96.51
balance	84	2.07	-	98.81	98.81
massage	74	1.83	-	100	100
participation	51	1.26	-	96.8	96.8
voluntary	47	1.16	-	100	100
motivation	37	0.91	-	100	100
stimulation	33	0.81	-	100	100
stopping_treatment	21	0.51	100	-	100
cognition	15	0.37	-	100	100
useless	11	0.27	100	-	100
modification_treatment	9	0.22	100	-	100

N/A not applicable

the precision of the one word-one concept was obviously of 100%. Then, for cases such as b, there was a precision a priori defined as the alone/together ratio, when the word- concept is used alone versus when it is used in sentences. In the other cases, c and b in sentences it had to be com- puted (two last columns in Table 1). The precision a pos- teriori aggregated the a priori and sentence precisions whenever applicable.

The second stage was conducted with RStudio®: all the non-classified CNs from the first phase (non_classified in Table 1) or the most frequent expressions describing the physiotherapy care found in the first stage, here 'walking', 'autonomy' or 'functional recovery', were aggregated in a single corpus and analysed through text mining. This technique based on stemming and lemmatization (com- bining words of the same family in the same group) used the R® package tm® [22].

After listing the words in a bar plot with the tm® R package (Fig. 3), the same corpus was analysed with three other R® packages – SnowballC® [23], wordcloud® [24] and RColorBrewer® [25] – to build a word cloud (Fig. 4) as a method of showing what comes at the forefront in the physiotherapist's concerns.

The first package uses the C libstemmer library, which implements Porter's word stem- ming algorithm for collapsing words to a common root to aid in the comparison of vocabulary, the second one builds the word clouds, and the third one chooses the words' colours. The wordcloud® function was parame- trized to show all the words appearing at least 10 times in the corpus, with 35% of them vertically for good readabil- ity. A qualitative palette for qualitative data was chosen with the RColorBrewer® package, with which the sizes and colours of the words were defined according to their frequencies.

Selecting the health-related variables

To show convergence between CN-extracted data, socio-demographic and EMR data and how they enhanced residents' health information, residents' ages, entry ages, genders, medical histories and pathologies were added, as well as the number of falls and their severities, i.e., whether the physician was called or whether the resident was hospitalized [7, 26–28] The whole idea here was to build residents' health features from health variables

already in the database but not designed for this purpose (right side of Fig. 1). These variables were chosen as describing relevant residents' features and emphasizing the problems afflicting the residents, especially those following physiotherapy programmes [28, 29]. The residents' medical histories and pathologies followed the Pathos model [30] designed by the National Health Insurance Fund and the National Union of Clinical Geriatrics to assess the NH caregiving work load. Pathos is defined as a thesaurus of 50 pathological states, classified into ten domains: cardio-vascular, neuro-psychiatry, pleuro- pulmonary, infections, dermatology, osteo-articular, digest- ive, endocrine, uro-nephrology, and others (see Table 2). In fact, Pathos is a tool used by all NHs in France and it measures the residents' care level needs through 8 resource posts: physician, psychiatrist, nurse, physio- therapy care, psychotherapy, biology, imaging and prescriptions. All these variables were well repre- sented in the IS. Finally, the geographic level was integrated because medical histories or pathologies might differ depending where the residents lived. It was the same for the NH identification because physiotherapy care or its provision might differ in its description

Table 2 The residents' Pathos variable frequencies during their NH stays on 09/30/2013

Number of medical histories per resident as of 09/30/2013	0	1	2	3	4	5	6	7
cardio-vascular	*607*	*211*	*125*	*55*	*13*	*4*		
neuro-psychiatry	*571*	*205*	*117*	*66*	*32*	*16*	*7*	*1*
pleuro-pulmonary	896	111	8					
infections	941	69	5					
dermatology	954	58	3					
osteo-articular	*655*	*232*	*92*	*31*	*4*	*1*		
digestive	*709*	*200*	*80*	*26*				
endocrine	888	116	10	1				
uro-nephrology	873	119	22	1				
others	*679*	*235*	*71*	*27*	*3*			
Number of pathologies per resident as of 09/30/2013	0	1	2	3	4	5	6	7
cardio-vascular	*417*	*319*	*191*	*68*	*16*	*4*		
neuro-psychiatry	*290*	*291*	*230*	*125*	*66*	*9*	*4*	
pleuro-pulmonary	881	131	2	1				
infections	977	37	1					
dermatology	924	87	4					
osteo-articular	*622*	*279*	*89*	*20*	*4*	*1*		
digestive	*635*	*258*	*98*	*22*	*2*			
endocrine	*794*	*201*	*20*					
uro-nephrology	*758*	*231*	*25*	*1*				
others	*573*	*294*	*110*	*35*	*3*			

In italics all medical histories or pathologies with at least 200 residents (19.7%) afflicted at least once

from one NH to another.

Matching the physiotherapy residents' narratives with their health-related subgroups through supervised classification

Having built the residents' health subgroups, we could now match the textual data with the health-related data (Fig. 1) to determine whether different health situations could be reflected by a variation in the vocabulary used. The two selected subgroups were a 'falling' group, including residents who fell at least 15 times since their NH entry, corresponding to the extreme number of falls' class (see the Number of falls histogram in Fig. 2), and an 'acute' group, including residents who fell, and there was either a physician exam or a hospitalization fol- lowing the fall. The two corresponding physiotherapy corpuses could then be built [7, 26, 27] using the 2 textual analysis stages discussed above, and they were compared qualitatively (the supervised health related subgroups analysed through text mining in Fig. 1 and detailed in Table 2).

Unsupervised data mining techniques to the residents' health data

Subsequently, three unsupervised techniques were applied

to the residents' health data, principal component analysis (PCA) [31, 32], multiple component analysis (MCA) [33–35] and hierarchical clustering (HC) on principal components (PC) with the HCPC function [36, 37] of the R® package FactoMineR® [37], to determine whether the current data knowledge could be improved and helped by building physiotherapy health subgroups.

All three methods were used as dimension reduction techniques. PCA transforms the original variables into independent linear combinations of them. While PCA processes quantitative data using mostly Euclidian distance, MCA uses qualitative data and the Ch2 distance. After the reduction step, scree plots in PCA and factor analysis are used to visually assess the components or factors that explain most of the variability in the data. Given the percentage of variation to be captured in the abridged data set, one could select the number of princi- pal components to be considered.

For PCA, as for MCA, only the initial dimension can be retained to stabilize the clustering by deleting the noise from the data, which is essentially the HCPC function's job. This function allows for the performing of HC and partitioning of the PC of several methods, choosing the best number of clusters and visualizing the tree, the partition and the principal components.

To apply MCA to our data, the residents' ages, entry ages and numbers of falls were transformed into categorical variables. The HC on the MCA then proposed

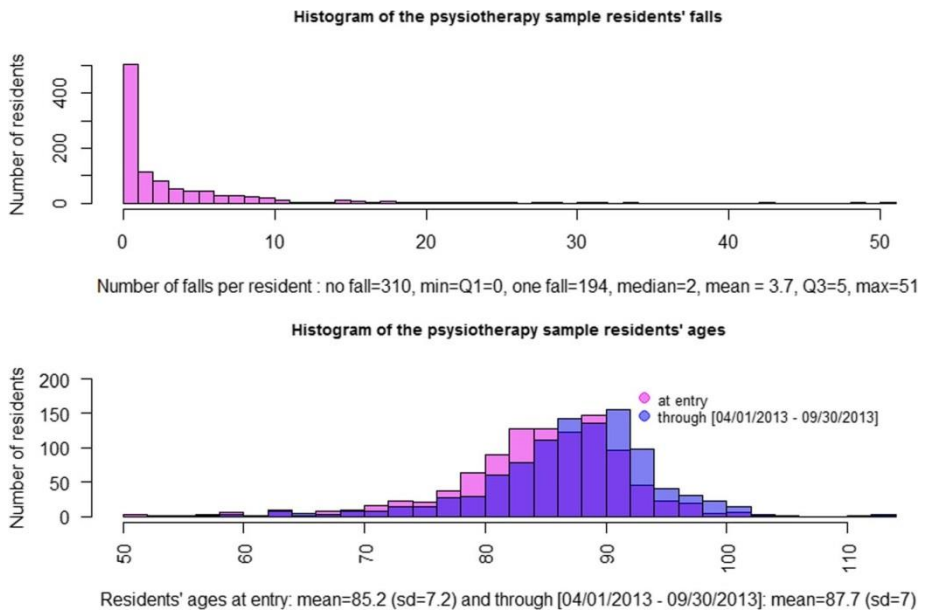


Fig. 2 The residents' number of falls and ages frequencies of the physiotherapy sample

an optimal cutting with 6 clusters (we can see below in Fig. 5 a strong inertia gain between five groups and six groups). Finally, the HCPC function [36, 37] was used to automatically retrieve every resident's cluster num- ber and identify his or her health's characteristics.

Results

There were 1015 residents (796 women and 219 men) entering NHs at a mean age of 85 years and 2 months old, staying there for an average time of 2 years and a half as of September 30, 2013, and falling an aver- age of 3.7 times since their NH entry (see Fig. 2). The corpus of 4051 physiotherapy CNs provided an average of 4 CNs per resident for the six-month period, describing essentially physiotherapy care. For example, the list of words or short expressions appearing at least 5 times included: stopping_treat- ment, modifying_treatment, walking, autonomy, balance, cognition, per_day, per_week, antalgic, pain, renewal, functional_recovery, partial, motivation, voluntary, mas- sage, stimulation, participation, and others. After cut- ting the CN at the 300th character (see above in the Building the physiotherapy variables subsection), the physiotherapy thesaurus was composed of 2165 differ- ent words with a mean length of 8.2 characters per word. Most of the CNs were composed of only one word (1588 CNs of 4051, 39.3%, with the first quartile Q1 = one word, Median = two words, Mean = 3.5 words, third quartile Q3 = 3 words, Maximum = 40 words).

Characterizing the physiotherapy treatments through a small list of constructed key-words

The two-stage process described above yielded the following results (Table 1 for the first stage and Figs. 3 and 4 for the second stage).

In Table 1, free_text observations (in italic) were those with more than 30 characters, and non_classified obser- vations were those not containing any of the classified words listed above (in grey and black, respectively). When frequency numbers were followed by another number, the second ones stood for the number of times when they were used alone.

Between the first and second stages, the textual obser- vations containing only the words 'renewal', 'n (or) m times per_week', 'n (or) m times per_day', 'others', 'stop- ping_treatment', and 'modification_treatment' were re- moved as observations not describing physiotherapy care, totalling 1427 observations (see Fig. 1), i.e., 35.23% of the corpus of 4051 observations. Hopefully, as said in building the physiotherapy variables subsection, we had a majority of one physio keyword - one physio concept expressions to describe the physio sessions and could then compute, except for free_text and non_classified, their precision as detailed here.

In the other cases, words used in sentences were most of the time rightly analysed. For example with the auton- omy concept, we excluded the negative cases: 'autonomy loss', 'significant loss of autonomy', 'resident who lost her autonomy in a wheelchair', idem for the pain concept: 'good mobility without pain', 'not painful', 'without pain'; for the good concept we removed these three cases:

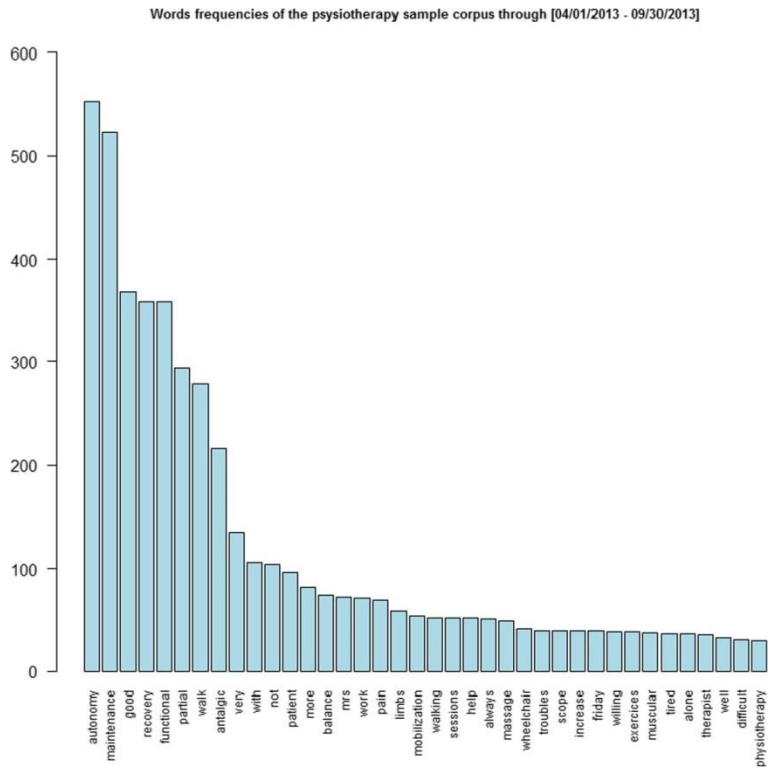


Fig. 3 Barplot of words appearing at least 30 times in the physiotherapy sample corpus (stage 2)

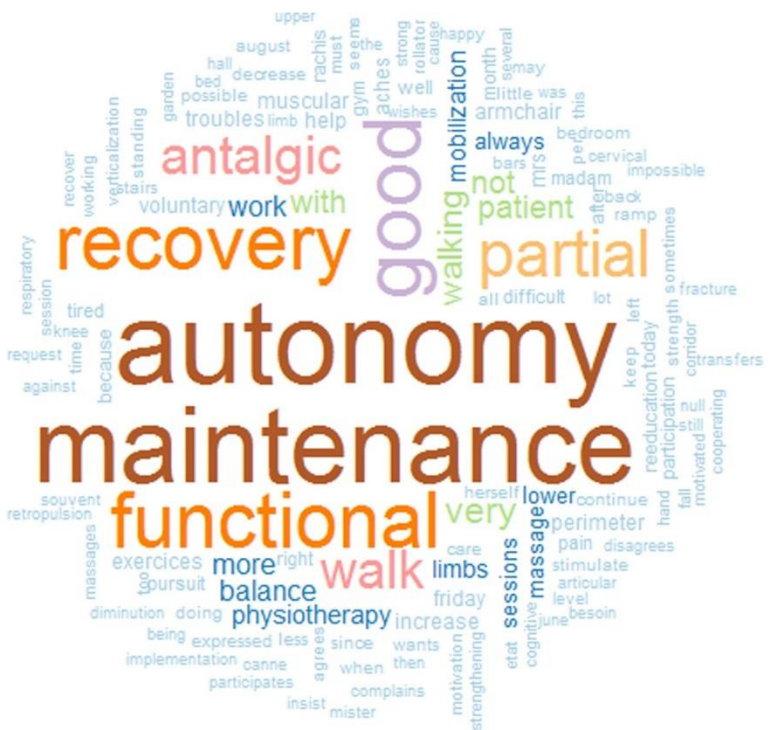


Fig. 4 Word cloud with words appearing at least 10 times in the physiotherapy sample corpus (stage 2)

'*monitor the good device installation*', '*hinder the smooth functioning of* ', '*it would be good to convince her*'. The real precision can be found on the last column table.

The bar plot in Fig. 3 lists the words appearing at least
30 times in the physiotherapy corpus. The first words were autonomy and maintenance, functional, recovery, good, partial, walking, and antalgic, as well as other words, such as work, limbs, mobilization and wheelchair.

The word cloud in Fig. 4 shows the words appearing at least 10 times in the physiotherapy corpus. Coming at the forefront were the same words – autonomy and mainten- ance, functional and recovery, good, partial, antalgic and walk – but also other words, such as anger, tired, recover and ache. Whereas the bar plot provides the precise word frequencies, the word cloud focuses on the residents' diffi- culties and the therapist's hard work with them.

The residents' health features

As detailed in the health-related variables selection subsection, gathering data from other files, the health residents' features were described according to the Pathos model and were classified into ten domains (Table 2) plus their number of falls (Fig. 2), as well as their sex and age.

Any resident could have a medical history at entry or any pathology at all on 09/30/2013 (first column) up to six or seven of them (6th or 7th columns) in up to ten domains (cardio-vascular, neuro-psychiatry, etc.). Residents entered NHs mostly with cardio-vascular, neuro-psychiatric, articular and digestive problems, which usually increased after some time (in italics all medical histories or pathologies with at least 200 residents (19.7%) afflicted at least once).

Selecting two different health-related subgroups of residents through supervised classification and comparing their physiotherapy corpuses

Text mining of two residents' health-related subgroups, built through supervised classification, found that even small differences in health situations resulted in qualita- tive and quantitative differences in physiotherapy narra- tives: 261 residents in the acute group and 58 in the falling group with, respectively, 993 and 232 textual observations. After removing the textual observations containing only the words 'renewal', 'per_week', 'per_day', 'stopping_treatment', and 'modification_treatment', (see above), there were still 513 and 109 textual observations containing 921 and 427 distinct words (Table 3), respectively.

In this table, each word frequency is followed by its sample ratio. For example, the word 'autonomy' appeared 127 times in the acute group and 29 times in the falling group. Words' ratios reflect the physiotherapy care priorities, which can differ in each group. For example, functional and recovery seem more meaningful in the acute group (37%) than in the falling group (28%).

Unsupervised health related subgroups built through data-mining techniques

As explained in the unsupervised data mining techniques on the residents' health data subsection, exploratory techniques, PCA and MCA, followed by clustering on these data, were used to visualize different subgroups with different needs. First, PCA was performed because our variables were essentially numeric (except for the NH names, the regions and the sex, not studied here) (see Fig. 5).

The eigenvalue decomposition yielded seven eigenvalues greater than one, but together, they defined only 59% of the total inertia, with the two first principal components bringing less than 30% of it (18.94% + 10.93% giving 29.87% see Fig. 5). Plus, all medical histories (variables a_xx) followed one direction and all pathologies (variables p_xx) another perpendicular to the first one, indicating that the health problems at the residents' NH entry were mostly 'independent' of their heath situations after a while. Finally, the number of falls (variable nb_falls) pointed to a third direction perpendicular to the first principal plan.

Then MCA was performed, adding the NH names, regions and departments, keeping at first the ages and number of falls as illustrative variables, after casting the numbers of medical histories and pathologies as categorical variables, and using the HCPC function. The HC on the MCA proposed an optimal cutting with 6 clusters (see in Fig. 6 the horizontal line showing the best clustering cut, plus a strong inertia gain between five and six clusters), which why, as explained in the unsupervised data mining techniques on the residents' health data sub- section, the three continuous variables – age, age at the NH entry and number of falls – were also divided into six-quantile groups to cast them into qualitative variables, using cut-off points as fairly as possible to follow the best clustering number based on the HCPC function, hoping to obtain the most discriminant residents' partition. All of these categorical variables were strongly discriminatory factors, especially the geographic variables, contributing most significantly to the global inertia (for example, the Chi2-test yielded, for the NH names, region and depart- ment variables, 3 null p-values). The HCPC plot function in 3D and 2D with colours showed 6 clusters (Fig. 7) with the HCPC modelling of the last variable providing the cluster number. There were 229 residents for the first one, 335 for the second, 58 for the third, 208 for the fourth,
147 for the fifth and n1 for the last one, which was coloured in magenta lying far from the five others. After examining the sixth cluster's residents, all of the

Table 3 The word frequencies of the acute and falling groups' physiotherapy corpuses computed with tm®

Hospitalized or followed by a physician after falling (261 residents)			Fallen at least 15 times (58 residents)		
word	frequency	ratio	word	frequency	ratio
autonomy	127	0.49	autonomy	29	0.50
functional	97	0.37	walking	23	0.40
recovery	96	0.37	functional	16	0.28
walking	64	0.25	recovery	16	0.28
antalgic	45	0.17	mobilization	12	0.21
very	42	0.16	very	11	0.19
plus	34	0.13	help	10	0.17
patient	28	0.11	plus	10	0.17
work	26	0.10	always	9	0.16
always	19	0.07	work	8	0.14
limbs	18	0.07	massage	7	0.12
massage	17	0.07	physiotherapist	6	0.10
mobilization	16	0.06	limbs	6	0.10
troubles	15	0.06	going further	6	0.10
help	13	0.05	rachis	6	0.10
pain	13	0.05	re-education	6	0.10
exercises	13	0.05	good	5	0.09
sessions	13	0.05	wheelchair	5	0.09
friday	13	0.05	fracture	5	0.09
good	12	0.05	less	5	0.09
difficult	12	0.05	new	5	0.09

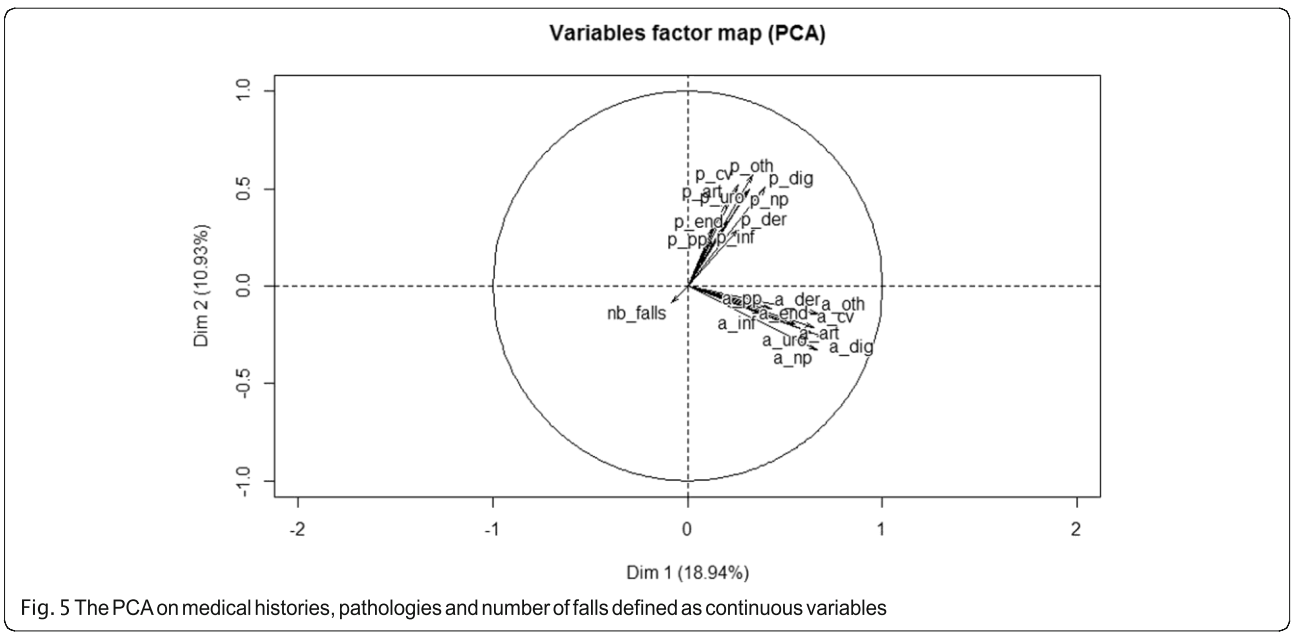


Fig. 5 The PCA on medical histories, pathologies and number of falls defined as continuous variables

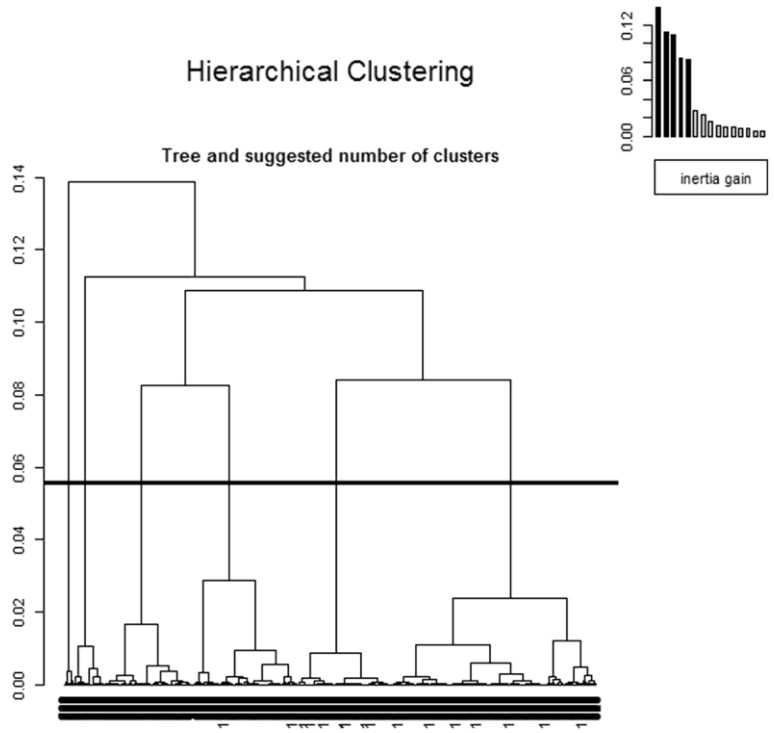


Fig. 6 HC + MCA with regions, departments, NH's names, gender, medical histories and pathologies as categorical variables

residents, except one, had no pathology at all but numerous medical histories. Furthermore, they all came from the same NH. One last MCA plus clustering was attempted, adding age, age at NH entry and number of falls, defined as categorical variables, with 6 modalities, which then yielded five clusters instead of six, the fifth corresponding to the sixth one before, and the five clusters becoming four.

Finally, by removing these n, all the remaining residents were still near the gravity centre.

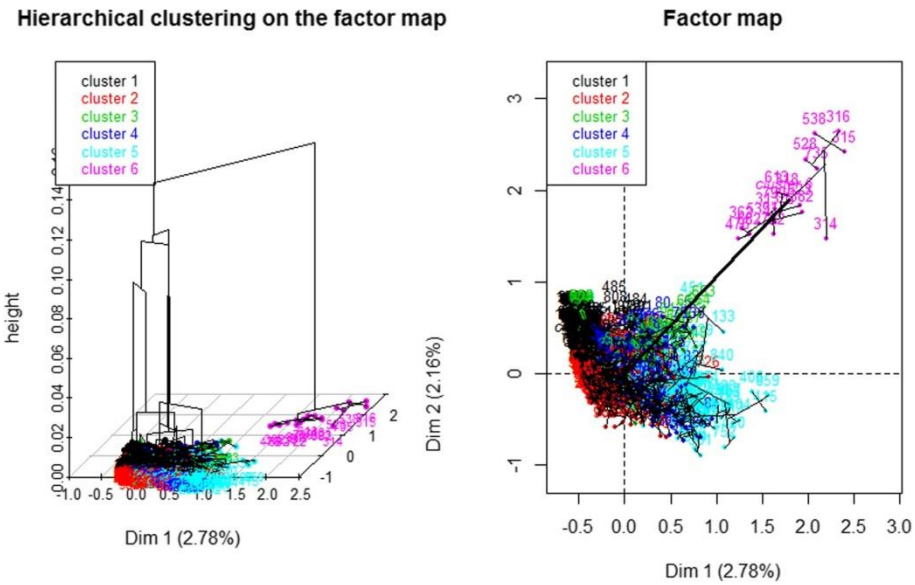


Fig. 7 The HCPC plot function of the 6 clusters in 3D and 2D

Discussion

Through this textual experiment, it was shown that text mining and data mining techniques could add new, sim- ple, useable data to EHR to improve residents' health and quality of care.

To achieve this goal and because only limited studies of strategies designed to increase the efficiency of processing CNs exist [12], it was decided to work on these textual data [14] to understand their characteristics through data mining, which is the core technology of cus- tomer relationship management [38] and one of the best tools to analyse the quality of care in this type of data frame. This physiotherapy CN sample might contain many specialized medical terms with no normalization expected, but it was believed that this textual information, which was potentially available, might be transformed into useful and understandable knowledge to facilitate professional practice and active interdisciplinary collaboration and research [39].

This textual experiment with the physiotherapy treatments, using a textual process in two stages, displayed convenient tools to improve the residents' health and quality of care by adding new, simple, useable data to the EHR. Text mining and data mining techniques on this DWH enriched the residents' health features by adding new, simple, useable data, such as physiotherapy keyword frequencies, to the EHR, opening the way to- wards more substantial textual information uses, such as patient stratification and improved targeting of care [15]. This textual approach also gives new meaning to all the time spent by medical assistants, nurses and doctors feeding the residents' EHR without changing their working habits or adding new tedious tasks.

The textual process in two steps, as presented in Table 1 (first step with SQL) and Fig. 3 (second step with tm®), offered better control of the quantitative and qualitative aspects of the textual analysis: the corpus analysed by tm® was already cleaned up of frequent and meaningless words, and conversely, meaningful words or expressions appeared at the forefront. This method in two steps attempted to resolve the research issue of detecting truly meaningful words in this corpus, as is usually done with the tf-idf technique with its word-weighting scheme, which helps to adjust for the most frequently occurring terms often not being the most meaningful ones, espe- cially for the user [40].

Shortening the physiotherapy care narratives from 4000 characters to 300, as there was a majority of short character fields, allowed us to expedite the textual ex- tractions and ease the textual analysis [3]. Examining the difference when selecting the first 400 characters, in- stead of the first 300, we found only 34 CNs, less than 1% longer than 300 char, and we compared the two sets' 45 most frequent words to find no significant difference between the two sets (see Additional file 5: Annex 5). Other strategies with NLP methods to increase efficiency and usefulness, such as extracting concepts or filtering high word counts, can be found in [12, 17].

Figure 3 shows that few terms were used frequently, such as autonomy, maintenance, recovery or functional (the left portion of Fig. 3), while many terms were used infrequently, such as exercises, troubles, re-education, difficult, pain or participation (the bottom/right portion) [17]. This finding is reminiscent of Zipf 's law [41], which describes the empirical frequency distribution of words in general language as having a large peak and a heavy one-sided tail. As in [17], the most frequent terms were general rather than specific and reflected the domain from which they arose.

This study had several strengths. First, by comparing the numbers of keywords and expressions of both steps, the tm process could be checked, and the whole text mining process could be iteratively improved to provide a detailed description of the physiotherapy sessions and treatments. Even after choosing to stop at the second step, it remained possible to refine the SQL queries gradually, adding new expressions and using wildcards to extract more complex textual data, as will be done in the future. Additionally, by working like this, a physiotherapy care keyword list was defined, which could be enriched and improved. It worked like a problem list, as was done in the UMLS-CORE project [42, 17]. This technique will allow us to reuse it later with other IS in our group and ease their overlapping. Also by design and as shown in Table 1, this list was built using simple SQL queries. In each query, what was searched for was precisely what was identified: for instance, the word balance in the physio CN described with an almost 100% precision what these physio-sessions were all about: improving the resident's balance and nothing else (except for one case with balance, 'unaware of his weak balance'). It was mostly the case for walking (improve walking), pain (alleviate pain), massage, motivation and so on. Finally, much better visualization of the words describ- ing physiotherapy care was available, with this 'filtered' word cloud, which is good at communicating qualitative data and can be a simple tool to identify the focus of written material [43] as well as capture risk perception [44].

Second, by selecting medical history, pathologies and number of falls and by adding the regions, departments and NH identifications, it became possible to describe all of the residents and their NH locations following physiotherapy sessions over a six-month period and to detect some data discrepancies.

By performing PCA of the residents' health groups, one interesting result was found: all of the medical histories followed one direction and all of the pathologies another, perpendicular to the first, suggesting that

the residents' health groups changed greatly over time. However, the number of falls pointed to a third direction perpendicular to the first principal plan, while being one of the main risk factor in which we were interested [26, 7, 27]. Nevertheless, by checking with another PCA method using the built-in R stats® package prcomp function [45], adding this time age and entry age, we found even less inertia (27%), with the first two PCs defined as before by medical histories and pathologies and the third one being essentially defined by age and entry age. With MCA, rather than PCA, followed by HC, we found why one of the clusters was so far from the other five: in MCA, 2 individuals are nearer if they have common rare modal- ities for one or more variables: here all ten p_xx variables, standing for pathologies, had a zero value for all except one resident of the last cluster. Looking further at Fig. 7, the distance between the 6th cluster and the five others might have come essentially from the absence or presence of pathologies. It is likely that, because medical histories and pathologies are fed into the same table, there was an omission or oversight when feeding the database. How- ever, this group was also found to be older (1 year and a half) and to have fallen more (2 times more). Finally, by removing these n residents, all of the remaining were still near the gravity centre as before, and we concluded that this sample was very homogeneous with our criteria: age, entry age, sex, medical histories and pathologies in 10 domains, gravity and number of falls. With more hetero- geneity in the data, it might have been possible to find more disjointed health subgroups and interesting differences to describe.

Third, using text mining [22] on two slightly different residents' samples showed that even small differences in health situations yielded to easy to detect qualitative and quantitative differences in physiotherapy narratives (Table 3). Autonomy was the main objective in both groups, as it was for the whole sample (see the words autonomy maintenance in Figs. 3 and 4) in respectively, 49% and 50% of physiotherapy care narratives; then, the words functional recovery (37% as seen in Figs. 3 and 4) for the acute group versus walking (40% idem) for the falling group; and finally, the words antalgic (17% idem) and pain (5% idem) in the acute group and not in the falling group.

Nevertheless, this study had several limitations. First, it was observational and explored only a se- lected sample of this CN database on a restricted subject, physiotherapy comments, and with a short follow-up period of 6 months, but throughout this whole experiment, we could better know the free text content of our CN and how to use it by building dummy variables defining the physiotherapy treatment and checking their values. Second, while word stemming was used when performing the SQL queries with the LIKE function and wildcards [21], this option was not selected when using package tm® for two reasons. First, it did not work well in French; for example, even if it cut the word marche (walking) into march, the words marche and marcher (to walk) were not combined. It was the same thing for the words bon (good) and bonnes (good): bonnes was cut into bonn but was still different from bon and so was counted as another word. Second, when the words are mapped into word clouds, it is more difficult to read the words' stems than the words. We attempted to define the best trade-off between data visualization and accuracy of the data.

Finally, matching residents with their CNs relied on physiotherapy care observations, and physiotherapy care discrepancies were found between NHs. For example, physiotherapy care descriptions varied greatly in frequency from one region to another: there were 201 physiotherapy observations for the North-West region with 17 NHs, whereas there were 754 of them for the South region with the same number of NHs, showing that, without normalization, the textual data are highly dependent on the style, precision and depth of physiotherapy care descriptions and the people feeding the IS. Using here a normalized physiotherapy problem list systematically for every physiotherapy session could help to solve this problem, but even without normalization, additional information is often available in unstructured free text, as shown with the experiment in the UK with rheumatoid arthritis (RA) in primary care of the general population [46, 47].

Conclusion

Through text and data mining techniques, an empirical approach to integrating health narratives into the existing information system was illustrated. Thanks to this physiotherapy data textual analysis in two stages (Fig. 1), new health variables describing residents' autonomy, functional recovery and pain or walking difficulties were built. This textual data could help define health subgroups, such as at-risk or recurrent fallers, through classification or could predict future health problems, such as hospitalizations or deaths, through logistic regression machine learning algorithms.

These new semantic technologies could improve the residents' follow-up over time and their health paths and could offer well-adjusted solutions to their multiple health problems. As with incentive programmes of the Centers for Medicare and Medicaid Services in the United States [48], it should be possible to answer questions about the meaningful use of EHR and quality of health care in the near future.

As was just said, combining structured data (age, gender, health groups) and unstructured data (physiotherapy care narratives) together provided a richer resident description [49].

MCA plus clustering could help differentiate residents, for example, those not having any pathologies, and could enable greater patient stratification, for example, through NH indexation or regions.

Nevertheless, to be truly useful, all of these health observations must be normalized to be reused later or be compared with other data samples, qualitatively and quantitatively. As explained in the Handbook on Research on ICT (Information and Communication Technology) for Human-Centered Healthcare and Social Care Services [50], the critical bottleneck today is, namely, information handover and reuse, and only interactively validated and semantically processed texts can be helpful to all health parties. To achieve this goal, problem lists must be defined for every type of health domain, as has been done in an-other DWH of the (Korian) group. There, health problem lists are in one table and the narratives in another one, pointing to the first one through indices. Another experiment in the UK used a complex comprehensive process of developing code lists building the clinical entity RA (rheumatoid arthritis) [46] through indicators and markers, and a third one, in the Boston area [51, 52], examined clinician and health care providers' attitudes towards problem lists in EHR and found that a common approach, completeness and standardization were necessary.

We showed here that CN could add valuable health information to the residents' health data in our database. We are confident in using it further through a clinician's finely tuned keywords and problems list, describing many geriatric ailments in the whole resident population, as was done for RA in the UK general adult population [46].

As explained in [52], future work on CN data should optimize the use of key functions to improve health providers' time efficiency, as well as data quality, integrity and usefulness [48, 53]. The next step will then be to fully integrate the CN data, with the problem list being the main tool defining health key functions. This goal will be met with finely tuned health data textual extraction analysing the whole textual generation process and relying on real health data content and staff uses. We hope being able to improve preventive health by better characterizing residents' falls following influenza vaccinations as well as through better management of chronic diseases, such as cancer, chronic pain, diabetes or dementia.

Endnotes

[1] $n < 30$. We removed the real number for NH de-identification and for data harmonization.

Additional files

Additional file 1: Empirical Advances with, Annex 1, The physiotherapy corpus, content: the 2616 remaining de-identified textual clinical narratives of the 1015 residents' sample in French (XLS 233 kb)

Additional file 2: Empirical Advances with, Annex 2, The physiotherapy corpus anonymization, content: describes the physiotherapy corpus de-identification and anonymization process (DOC 39 kb)

Additional file 3: Empirical Advances with, Annex 3, The physiotherapy corpus translation, content: describes the physiotherapy corpus translation process from French to English. (DOC 26 kb)

Additional file 4b: Empirical Advances with, Annex 4, The 1015 residents' health table, content: the 1015 de-identified residents' medical histories and pathologies on September 30th 2013, as well as their falling history and 10-anonymized NHs (XLS 256 kb)

Additional file 5: Empirical Advances with, Annex 5, Comparing the two sets of key words with 300 char and 400 char, content: compares the physiotherapy narratives' corpuses cut after 300 characters and after 400, in fact the most frequent words in the two corpuses, with a Chi2 test. (DOC 48 kb)

Abbreviations

CN: Clinical narrative; DM: Data mining; DWH: Data warehouse; EHR: Electronic health record; EMR: Electronic medical record; HCPC: Hierarchical clustering on principal components; IE: Information extraction; IS: Information system; MCA: Multiple correspondence analysis; NER: Named entry recognition; NH: Nursing home; PCA: Principal components analysis; RA: Rheumatoid arthritis; SQL: Standard Query Language

Acknowledgments

Authors would like to thank Sebastien Plasse, project manager from the Korian group Information Systems Direction who gave them details about the IS structure and how best extract data.

Funding

Institut du Bien Vieillir (Institute of Well Ageing) inside the Korian group is funding Tiba Delespierre's public health thesis and financing this manuscript as well as every scientific result the main author may publish.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and can be found in the additional files section (Additional file 1: Annex 1 and Additional file 4b: Annex 4). There will be no public availability of data and materials for this study. Accordingly, this data will not be reposited for public use. Requests for an already de-identified version of the dataset may be made to the authors and will be considered on an individual basis.

Authors' contributions

All authors contributed to project conception. TD and LJ were responsible for initial article drafting. TD was in charge of data gathering and computer programming for all data analysis, as well as of the study design. PD, head of the Institut du Bien Vieillir participated in the redaction of the article. ABH, professor at the CNAM, reviewed the whole data processing and the study design. LJ, professor at UVSQ, reviewed and improved the study design. All authors read and approved the final manuscript. They all gave final approval of the manuscript and agreed to be accountable for its integrity.

Ethics approval and consent to participate

The use of this database in the frame of epidemiological studies has been authorized by the French National Commission for Data protection and Liberties (CNIL). The Institut du Bien Vieillir filed a declaration of conformity to a baseline methodology which received in March 2017 an agreement number: 2.041.050, in accordance with the Act n°78–17 of 6 January 1978 on Data Processing, Data Files and Individual Liberties. All residents are informed at their NH entry about their EHR and their right to oppose its use. While the primary purpose of this medical research was to generate new knowledge, this goal didn't take precedence over the rights and interests of the NH residents. All the new generated information was extracted from already

existing data and was de-identified and anonymized when necessary to protect their health and rights. There were no images and no identifying details on individuals reported within this manuscript

Consent for publication
Not applicable.

Competing interests
The authors do not have any financial or non-financial competing interests to report. Funding to support TD's work is reported above.

Publisher's Note

Author details

[1]Institut du Bien Vieillir Korian, 21-25 rue Balzac, 75008 Paris, France. [2]Research lab: EA 4047, UFR des Sciences de la Santé Simone Veil, UVSQ Université Paris-Saclay, 2 Avenue de la Source de la Bièvre, Montigny le Bretonneux 78180, France. [3]MNH Group, 185 rue de Bercy, 75012 Paris, France. [4]UFR de Mathématiques et Informatique, Université de Paris Descartes, 45 rue des Saints-Pères, Paris 75006, France.

Received: 7 December 2016 Accepted: 4 August 2017
Published online: 22 August 2017

References

1. Maas ML, Delaney C. Nursing process outcome linkage research: issues, current status, and health policy implications. Med Care. 2004;42(2):II-40–8.
2. Ventres W, Kooienga S, Vuckovic N, et al. Physicians, Patients, and the Electronic Health Record: An Ethnographic Analysis Annals of Family Medecine, n°2 March/April. 2006;4:124–32. www.annfammed.org.
3. Mc Ginn CA, Grenier S, Duplantie J, et al. Comparison of user groups' perspectives of barriers and facilitators to implementing electronic health records: a systematic review. BMC Med. 2011;9:46.
4. Cebul RD, Love TE, Jain AK, et al. Electronic health records and quality of diabetes care. N Engl J Med. 2011;365:825–33.
5. Genes N, Chandra D, Ellis S, et al. Validating emergency vital signs using a data quality engine for data warehouse. Open Med Inform J. 2013;7:34–9.
6. Zangara G, Corso PP, Cangemi F, et al. A cloud based architecture to support electronic health report. Stud Health Technol Inform. 2014;207:380–9.
7. Tremblay MC, Berndt DJ, Luther SL, et al. Identifying fall-related injuries: Textmining the electronic health record. Inf Technol Manag. 2009;10:253–63. doi:10.10007/s10799-009-0061-6.
8. SM Meystre, GK Savova, KC Kipper-Schuler et al. Extracting information from textual documents in the electronic health record : a review of recent research IMIA yearbook of medical informatics 2008.
9. Hornberger J. Electronic health records: a guide for clinicians and administrators. Book and media review. JAMA. 2009;301:110.
10. Savova GK, Masanz JJ, Ogren PV. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17:507e513. doi:10.1136/jamia.2009.001560.
11. Middleton B, Bloomrosen M, Dente MA, et al. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. J Am Med Inform Assoc. 2013;20: e2–8. doi:10.1136/amiajnl-2012-001458.
12. Gundlapalli AV, Redd A, Carter M, et al. Validating a strategy for psychosocial phenotyping using a large corpus of clinical text. J Am Med Inform Assoc. 2013;20:e355–64. doi:10.1136/amiajnl-2013-001946.
13. https://www.oracle.com/fr/index.html. Accessed 3 July 2017.
14. Delespierre T, Denormandie P, Josseran L. New methods to evaluate physiotherapy care in nursing homes. JNHR the Journal of Nursing Home Research International Working Group December 2-3, 2015 Toulouse, France Vol 12015OC36p30.
15. Min Song Opinion: Text Mining in the Clinic. The Scientist (April 1, 2013).
16. https://www.rstudio.com/. Accessed 3 July 2017.
17. ST Wu, H Liu, D li et al. Unified medical language system term occurrences in clinical notes: a large-scale corpus analysis J Am Med Inform Assoc 2012; 19:e149-e156, DOI 10.1136/amiajnl-2011-000744.
18. Biro S, Williamson T, Leggett JA, et al. Utility of linking primary care electronic medical records with Canadian census data to study the determinants of chronic disease: an example based on socioeconomic status and obesity. BMC Med Inform Decis Mak. 2016;16:32. doi:10.1186/s12911-016-0272-9.
19. Nguyen B. Techniques d'anonymisation. Statistique et société, Vol. 2, N° 4 décembre 2014. http://www.benjamin-nguyen.fr/papers/ss.pdf. Accessed 11 Aug 2017.
20. http://drees.social-sante.gouv.fr/IMG/pdf/5_test_anonymisation_donnees_pmsi.pdf. Accessed 11 Aug2017.
21. http://www.tutorialspoint.com/sql/sql-like-clause.htm. Accessed 11 Aug 2017.
22. http://edutechwiki.unige.ch/fr/Tutoriel_tm_text_mining_package. Accessed 11 Aug 2017. https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf. Accessed 11 Aug 2017.
23. https://cran.r-project.org/web/packages/SnowballC/index.html. Accessed 11 Aug 2017.
24. https://cran.r-project.org/web/packages/wordcloud/index.html. Accessed 11 Aug 2017.
25. https://cran.r-project.org/web/packages/RColorBrewer/index.html. Accessed 11 Aug 2017.
26. Lee TT, Liu CY, Kuo Y-H, et al. Application of data mining to the identification of critical factors in patient falls using a web-based reporting system. Int J Med Inform. 2011;80(2):141–50. Special Issue: Security in Health Information Systems. February2011
27. Lazkani A, Delespierre T, Bauduceau B, et al. Predicting falls in elderly patients with chronic pain and other chronic conditions. Aging Clin Exp Res. 2015;27(5):653–61. doi:10.1007/s40520-015-0319-2. Epub 2015.
28. Leemrijse CJ, de Boer ME, van den Ende CHM, et al. Factors associated with physiotherapy provision in a population of elderly nursing home residents; a cross sectional study. BMC Geriatr. 2007;7:7. doi:10.1186/1471-2318-7-7.
29. Office of Inspector General J G Brown Physical And Occupational Therapy in Nursing Homes Medical Necessity and Quality of Care. Department of Health and Human Services OEI-09-97-001211999.
30. JM Ducoudray, Y Eon, C Le Provost et al. Le modèle PATHOS, Guide d'utilisation 2017 rédigé par la CNAMTS (Caisse Nationale d'Assurance Maladie des Travailleurs Salariés) et le SNGC (Syndicat National de Gérontologie Clinique).
31. Krefis AC, Schwarz NG, Nkrumah B, et al. Principal component analysis of socioeconomic factors and their association with malaria in children from the Ashanti region. Ghana Malar J. 2010;9:201.
32. Ahmed SA, Siddiqi JS, Quaiser S. Principal component analysis to explore climatic variability that facilitates the emergence of dengue outbreak in Karachi. Pak J Meteorol. 2014;11(21):1.
33. http://maths.cnam.fr/IMG/pdf/CHIENS2012_cle0f5221.pdf. Accessed 11 Aug 2017. http://maths.cnam.fr/IMG/pdf/Epose-Pages-Dec09.pdf. Accessed 11 Aug 2017.
34. Ayele D, Zewotir T, Mwambi H. Multiple correspondence analysis as a tool for analysis of large health surveys in African settings. Afr Health Sci. 2014; 14(4):1036.
35. P Soares Costa, N Correia Santos, P Cunha et al. The Use of Multiple Correspondence Analysis to Explore Associations between Categories of Qualitative Variables in Healthy Ageing Hindawi Publishing Corporation Journal of Aging Research Volume 2013, Article ID 302163, 12 pages http://dx.doi.org/10.1155/2013/302163 (Accessed 28 July 2016).
36. F Husson, J Josse, J Pagès. Principal component methods – hierarchical clustering – partitional clustering : why would we need to choose for visualizing data ? Technical report Agrocampus 2010.
37. http://factominer.free.fr/. Accessed 11 Aug 2017.
38. Cheng B-W, Chang C-L, Liu I-S. Enhancing care services quality of nursing homes using data mining. Total Qual Manage Bus Excell. July 2005;16(5):575–96.
39. Holzinger A, Jurisica I. Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions. LNCS 8401 pp 1–18, 2014 Springer Verlag.https://link.springer.com/chapter/10.1007/978-3-662-43968-5_1. Accessed 11 Aug 2017.
40. J Beel, S Langer, B Gipp (2017). TF-IDuF: A Novel Term-Weighting Scheme for User Modeling based on Users' Personal Document Collections (Accessed 6 June 2017) (PDF). iConference.
41. Piantadosi ST. Zipf's word frequency law in natural language: a critical review and future directions June 2, 2015. https://colala.bcs.rochester.edu/papers/piantadosi2014zipfs.pdf. Accessed 11 Aug 2017.
42. Fung KW, McDonald C, Srinivasan S. The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institution. J Am

Med Inform Assoc. 2010;17:675e680. doi:10.1136/jamia.2010.007047.
43. Atenstaedt R. Word cloud analysis of the BJGP. Br J Gen Pract. 2012;62(596): 148. doi:10.3399/bjgp12X630142.
44. Dressel K, Schüle S. Using Word Clouds for Risk Perception in the Field of Public Health – the Case of Vector-Borne Diseases. In: Planet@Risk, Davos: Global Risk Forum GRF Davos. 2014;2(2):85-88
45. http://www.sthda.com/english/wiki/principal-component-analysis-in-r- prcomp-vs-princomp-r-software-and-data-mining. Accessed 11 Aug 2017
46. A Nicholson, E Ford, KA Davies et al. Optimising Use of Electronic Health Records to Describe the Presentation of Rheumatoid Arthritis in Primary Care: A Strategy for Developing Code Lists. PLOS One February 2013, Volume 8 Issue 2 e54878 http://dx.doi.org/10.1371/journal.pone.0054878 (Accessed 13 Sept 2016).
47. Ford E, Nicholson A, Koeling R, et al. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? BMC Med Res Methodol 2013; 13: 105. Volume 8 Issue 2 e54878
48. Kern LM, Edwards A, Kaushal R. The meaningful use of electronic health records and health care quality. Am J Med Qual. 2015;30(6):512–9. doi:10.1177/1062860614546547.
49. MM Cruz-Cunha, IM Miranda, P Gonçales. Handbook on Research on ICT for Human-Centered Healthcare and Social Care Services 2013, IGI Global.
50. Wright A, Maloney FL, Feblowitz JC. Clinician attitudes toward and use of electronic problem lists: a thematic analysis. BMC Med Inform Decis Making. 2011;11:36. https://doi.org/10.1186/1472-6947-11-36.
51. Holmes C, Brown M, Hilaire D St. Healthcare provider attitudes towards the problem list in an electronic health record: a mixed-methods qualitative study. BMC Med Inform Decis Making. 2012;12:127. https://doi.org/10.1186/1472-6947-12-127.
52. Makam AN, Lanham HJ, Batchelor K, et al. Use and satisfaction with key functions of a common commercial electronic health record: a survey of primary care providers. BMC Med Inform Decis Making. 2013;13:86. https:// doi.org/10.1186/1472-6947-13-86.
53. Bowman S, Rhia MJ, Fahima CCS. Impact of electronic health record systems on information integrity: quality and safety implications. Perspect Health Inf Manag. 2013 Fall; 10(Fall): 1c

3-1-3 Perspectives liées à l'analyse textuelle d'un échantillon des transmissions

A l'issue de l'étude faite sur les traitements kiné, nous pouvons conclure que d'une part, **l'information textuelle issue des transmissions enrichit la vision de la prise en charge et des soins** en apportant de nouvelles informations médicales et en aidant à intégrer le système des EES dans l'environnement de travail du personnel médical et paramédical. D'autre part, **les comptes-rendus de soins présentent une grande variabilité suivant les établissements**, provenant à la fois du profil de santé des résidents et du personnel référent qui saisit les comptes-rendus.

Une fois cette mise en garde faite quant à la fiabilité des données, nous pouvons maintenant bâtir notre base de données à visée de santé publique que nous appellerons Base du Bien Vieillir (BBV) en intégrant l'information textuelle au reste des informations socio-démographiques extraites du DRI.

3-2 La construction de la BBV

Au moyen de la procédure en trois étapes décrite ci-dessous, extraction des données du DRI, transformation et formattage des données, chargement, la BBV est alimentée au gré des besoins et à la demande.

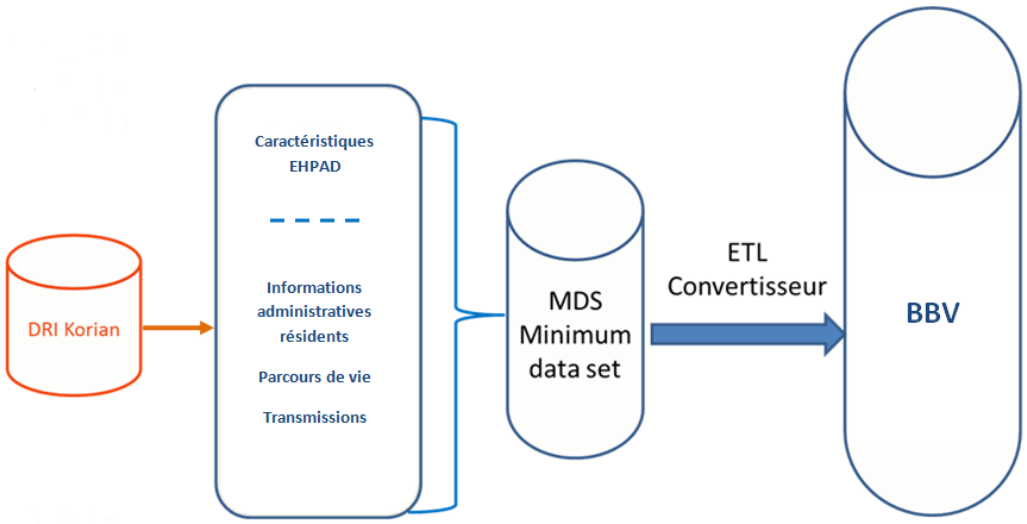


Figure 8 : La génération des données de la Base du Bien Vieillir

Deux types de données sont extraits :

- des données aux formatx fixes : les identifiants de l'établissement ainsi que les profils socio-démographiques des résidents à l'entrée dans une des structures : date de naissance, sexe, date d'entrée et GIR (groupe ISO ressources) [17] d'entrée.
- des données textuelles extraites de la table des transmissions comprenant hospitalisations et décès.

Ainsi, par extraction, transformation et chargement dans un silo de données spécifique, la BBV donne une deuxième vie aux données: après la production pour le soin, sans travail supplémentaire pour les équipes sur le terrain et sans modification du SI existant.

3-2-1 Les données socio-démographiques

Pour chaque résident des données de formats fixes : date d'entrée, sexe, âge et GIR à l'entrée en établissement [17].

3-2-2 Les données syndromiques

Ensuite des données textuelles, de format libre jusqu'à 4000 caractères, produites par le personnel soignant lors de son activité quotidienne, saisies dans la table transmissions et utilisées en tant qu'outil de liaison. Ces informations non formattées, saisies au fil de l'eau par les équipes ont été retravaillées pour être utilisables dans un but scientifique grâce à une analyse textuelle en 3 étapes: *premièrement,* troncature et nettoyage des données, *deuxièmement,* requêtes SQL (Standard Query Language) dans le DWH pour bâtir les syndromes, *troisièmement*, conférence de consensus par les experts métiers et text mining, avec plusieurs allers-retours entre les trois étapes.

Parmi les données retenues disponibles dans la table transmissions, la description et l'évolution de la morbidité était centrale. Dans ce cadre, ont d'abord été définis les syndromes grippe et gastro-entérite aiguë en tant que premiers objets pour définir un outil de surveillance sanitaire

pour les personnes âgées à partir des résidents en EHPAD. Au vu de l'intérêt terrain et de la richesse des informations saisies puis extraites, l'élaboration de ces deux syndromes a rapidement été suivie de celle de vingt-deux autres listés § 1-3-6. Enfin, *hospitalisations* et *décès* ont été agrégés au reste de l'information syndromique par extraction directe des tables hospitalisations et décès du DRI.

3-3 La construction de la cohorte des résidents

L'extraction de tous les résidents entrés dans au moins un établissement à compter du 1er novembre 2010 (date de début du DRI) et ce jusque fin février 2017 (fin de la saison grippale de l'hiver 2016-2017) a permis de générer une cohorte de 41 061 personnes âgées comprenant 12 983 hommes (31,61%) et 28 083 femmes (68,38%) d'âges moyens respectifs à l'entrée 84,33 et 85,82. Chaque semaine la cohorte est augmentée des nouvelles entrées. Sa construction suit le processus décrit § 1-3-3-3, figure 2.

3-4 Le système de surveillance syndromique

3-4-1 Résumé des résultats du second article

Une cohorte de personnes âgées a été construite à l'aide d'un réseau national de 125 EHPAD par extraction de l'information socio-démographique, de l'autonomie et d'un ensemble de vingt-six syndromes dont la grippe et les gea, de novembre 2010 à juin 2016. Les algorithmes de surveillance early aberration reporting system (EARS) et Bayes du package *R*® de surveillance *surveillance*® ont ensuite été utilisés pour évaluer nos données syndromiques grippe et gea et les comparer aux données syndromiques du Réseau Sentinelles, le gold-standard français pour les épidémies de grippe et de gastro-entérite, suivant les guidelines d'évaluation du CDC (Centers for Disease Control and Prevention) d'Atlanta.

Avec ces techniques nous avons réussi à construire une cohorte de 41 061 résidents par extraction de toute l'information socio-démographique disponible. L'algorithme EARS_C3 sur nos données grippales et de gastro-entérite aiguë a donné des sensibilités respectives de 0,482 et 0,539 et des spécificités de 0,844 et 0,952 sur une période de 6 années, permettant d'anticiper la dernière épidémie de grippe en détectant des signaux grippaux précoces et de caractériser son intensité et sa virulence. Nous avons également évalué la qualité des syndromes IRA-grippe et GEA lors de la dernière saison épidémique en calculant leurs précisions au cours de leurs pics épidémiques respectifs (semaines 03-2017 et 01-2017) et trouvé des valeurs de 0,.98 et 0,96 alors que les précisions étaient légèrement plus faibles durant le creux de la vague de cet été : 0,95 et 0,92 (semaine 33-2016).

Cette étude a confirmé que l'utilisation des données de transmissions des soins pouvait permettre de développer un véritable système de surveillance en santé dédié aux personnes âgées pour la grippe et les GEA. L'accès à des données de santé séniors répond à un enjeu de santé publique : la surveillance de cette population fragile. Cette base de données devrait permettre d'améliorer dans un avenir proche la prévention, les soins et apporter une réponse rapide en cas d'alertes sanitaires.

3-4-2 Second article

Issues in Building a Nursing Home Syndromic Surveillance System with Textmining: Longitudinal Observational Study

Tiba Delespierre[1], Loïc Josseran

[1:] Korian siège 21-25 rue Balzac 75008 Paris
(+33) 6 58 37 65 03
tiba.baroukh@gmail.com

Abstract

Background: New Nursing Homes (NH) data warehouses fed from residents' medical records open the way for monitoring health elderly population on a daily basis. Elsewhere syndromic surveillance has already shown that professional data can be used for public health (PH) surveillance but not during a long-term follow-up of the same cohort.

Objective: The goal of this study was to build and assess a national ecological NH PH surveillance system (SS).

Methods: Using a national network of 126 NH, a residents' cohort was built. Medical and personal data were extracted from their electronic health records (EHR) and transmitted through internet to a national server almost in real time. Socio-demographic, autonomy and syndromic information were recorded. A set of twenty-six syndromes was defined using pattern matching with the standard query language LIKE operator and a Delphi-like method, between November 2010 and June 2016. Early aberration reporting system (EARS) and Bayes surveillance algorithms of the R surveillance® package were then used to assess our influenza and acute gastro enteritis (AGE) syndromic data against the Sentinelles network data, French epidemics gold-standard, following Centers for Disease and Control (CDC) surveillance system assessment guidelines.

Results: By extracting all socio-demographic residents' data, a cohort of 41 061 senior citizens was built. *EARS_C3* algorithm on NH influenza and AGE syndromic data gave respectively sensitivities of 0.482 and 0.539 and specificities of 0.844 and 0.952 over a 6 year-period, forecasting the last influenza outbreak by catching early flu signals. Also, assessing influenza and AGE syndromic data quality, precisions during last season epidemic weeks' peaks (weeks 03-2017 and 01-2017) were of 0.98 and 0.96, whereas during last summer epidemic weeks' low (week 33-2016) were of 0.95 and 0.92.

Conclusions: This study confirmed that using syndromic information gives a good opportunity to develop a genuine French national PH SS dedicated to senior citizens. Access to seniors' free-text validated health data on influenza and AGE responds to a PH issue for the surveillance of this fragile population. This database will also make possible new ecological research on other subjects that will improve prevention, care and rapid response when facing health threats.

Keywords:
Centers for Disease Control and Prevention (CDC), Nursing Homes (NH), Syndromic Surveillance System (SSS), SQL pattern matching, Delphi method, Sentinelles network, Influenza like Illness (ILI), Acute Respiratory Infection (ARI), Acute Gastro Enteritis (AGE), Information System (IS)

Introduction

Population in developed countries is aging [1] and French population follows this trend. In France, in 2050, 22.3 million people will be 65 years old and more compared to 12.6 million in 2005, an increase of 80% in 45 years. Between 2013 and 2050, the senior population will grow more than the population as a whole. Similarly, life expectancy at birth in France, one of the highest in the world, is projected to surpass 86 years for men and 90 for women [2].

This increase will then have to be anticipated and will affect care and related costs [3]. It is then essential to improve our knowledge of this senescence process, to help prevent pathologies increase and improve quality of life at extreme ages.

In spite of this major population expected evolution, ecological research on this aged population is still limited [4]. Case or ad hoc studies do not consider individual variability and cannot analyze health issues as a whole. Data then need to be recorded for quite a long time and on a daily basis, helping to address this lack of knowledge. This has to be done in a natural way, in a professional environment with caregivers and medical staff [5].

As until now, follow-up studies on senior citizens realized by using costly to set up and follow cohorts [6-8]. Data are occasionally stored, even if the follow-up is long and based on auto-questionnaires or planned interviews with health professionals. This approach does not allow describing in detail the daily life of this population and storing health evolutions in the long run.

On the contrary, nursing homes (NH) offer this possibility of tracking and recording them daily as health professionals feed these information for their proper use and, this time, without any memory bias [9]. These new data, as well as their uses suggest innovative approaches to improve health knowledge.

Korian (Paris, France) as the first private NH European group has these kinds of data. This enterprise holds 290 NH and approximatively 4% (290/7394 [10]) of the French NH network, distributed all over the country, mostly in urban areas (see Multimedia Appendix 1). A professional data warehouse (DWH) set up in 2010 hosts half the company French residents' population data. Their health follow-ups are recorded daily from 126 NH. For every new resident admitted in one of the NH, a personal electronic resident medical file (PERMF) is set up. Data is collected at various times: at admission (admission date, medical history, marital status, birthdate, tastes and habits), on a daily basis (new pathologies, chronic disease evolution, date of death, drugs prescriptions), or just after specific medical or health care professional visits. Items include diagnosis, outcomes, as well as socio-demographic information.

Elsewhere and a little earlier, at the beginning of the 2000 years, syndromic surveillance (SS) [11-19] showed that professional data could also be used for health and alert surveillance [20-26]. Here professional data use for SS was only done using point-data

analysis (going to the emergency, 911, web queries) [19, 25- 26] and not during a long term follow-up of the same people, and even more, not dedicated to senior citizens.

As we have just seen, data gathered by different NH professionals offer the opportunity of following the residents' situations on the flow and on a daily basis, and through this process, of building syndromic surveillance data. The objective of this study was then to build and assess an influenza and acute gastro enteritis national ecological NH public health (PH) SS: describing and validating the BBV (Base du Bien Vieillir that is, Aging Well Data Base) architecture. Thus, and through a new health data building paradigm we engineered a NH syndromic surveillance system (SSS) based on already validated criteria [11, 13], hopefully opening the way to new research and knowledge about the senescence process.

Methods

Data Collection

All data are transmitted from 126 NH in real time to a national server using the group intranet. Records collected from the PERMF server are anonymized (see Multimedia appendix 2) when sent to the BBV server, keeping track of every resident even when moving from one NH to another. After this first step, health and socio-demographic data are extracted, transformed and loaded (ETL) to build the BBV data (see Multimedia appendix 3 for details). Following this second step, all residents have 2 types of data:

1- gender, age and GIR (Iso Resource Group), a French autonomy level rating indexed to government benefits [27-32] at the NH entry;
2- daily care information fed on the flow by the caregivers and the medical staff, whenever deemed useful that is, their syndromic information and finally hospitalizations and/or death.

At the same time, every Sunday, all residents' daily care information is aggregated to count the weekly number of syndromes per NH.

By extracting all residents of the PERMRF data base from its inception, from November 1st, 2010 till mid-February 2017 and adding every new resident entering one of the NH network every week, a *one-week moving* cohort of residents followed during their entire NH life course was built opening the way for our SSS. Even if most residents of this cohort were followed during their entire NH life course, syndromic data could be left truncated, for people entered before the information system (IS) inception or right truncated, for people entered lately.

At IS core, the data transmissions' table containing key information about the residents' care fed on a daily basis. Data take the form of big size character fields (of up to 4,000 characters). By extracting these and using residents' and NHs' indexes and data transmissions' dates (see Multimedia appendix 3 for a complete example), all residents can be tracked during time - every day with syndromic data - and space - every NH with syndromic data - with queries and text mining, building their syndromic life course, beginning at their date of entry and ending with their last available data transmission or death. The BBV has then two nested time frames: by day for every resident and by week for every NH.

Building the ARI-ILI and AGE Syndromes

With a multi-step learning and textmining (MSL – TM) process (see Multimedia appendix 3 for the four phases process) of the data transmissions' file similar to what was experimented in [33], using problems' list logic [34-36] and pattern matching with the SQL LIKE operator [37], twenty-four syndromes were implemented [38-49], following the Sursaud® SSS method [16]. Starting with ARI-ILI (Acute Respiratory Infection and Influenza like Illness) and AGE (Acute Gastro Enteritis) syndromes (see Multimedia appendix 3 for two examples and the syndromes' list), extracting directly hospitalizations and deaths, this NH IS kept for every resident, every day, in every NH, from none to twenty-six daily syndromes whenever appropriate (see again Multimedia appendix 3 for full details of the whole process [50-61]).

The Surveillance Tools Framework

Syndromic Systems attempt to detect outbreaks through statistical analysis of aggregated cases data to improve on competent clinicians in detecting early-stage or small outbreaks [62]. It focuses on data collected prior to clinical diagnosis or laboratory confirmation [63]. Statistical laws are then defined to give an answer to the question "knowing the average number of expected events during a period of time, what is the probability to observe the current situation?" [62].

The SSS generation was designed using a Pentaho® extraction platform for all the ETL process [64] and is described figure 1. It follows the CDC Working Group recommendations [11, 13].

The whole process was done in four steps: first, the ILI and AGE syndromes built through the MSL - TM process [65], second, the weekly ILI and AGE syndromic data aggregation and the time series (TS) generation with their statistical alerts using the R *surveillance*® package [66-67], third, the *Sentinelles* data joining, the ARI-ILI and AGE French surveillance gold standard [68], and last, the alerting system interfacing the *surveillance*® package [66-67] statistical alerts with the NH General Practitioners (GP) coordinators signals, eventually reporting to the Health Regional Agencies (HRA).

It is only after that last step that epidemiologists in the national public health agency's regional units (HRA in figure 1) are asked to choose an alert level for the regions they are in charge of: non-epidemic, pre/post epidemic or epidemic [69]. A public health alert will then be defined as such by the public health agency Santé Publique France (SPF) after every signal has been verified and validated [70] (for further details see Multimedia appendix 4 [22, 67, 69-77]). Relevant information for French epidemiologists includes since January 2016 at a regional level, the Sentinelles (2.1% of French private GP), as well as the OSCOUR (88% of French hospitals ED visits) and SurSaUD (95% of French emergency GP consultations) data but also, local specific surveillance data such as NH ARI clusters' surveillance.

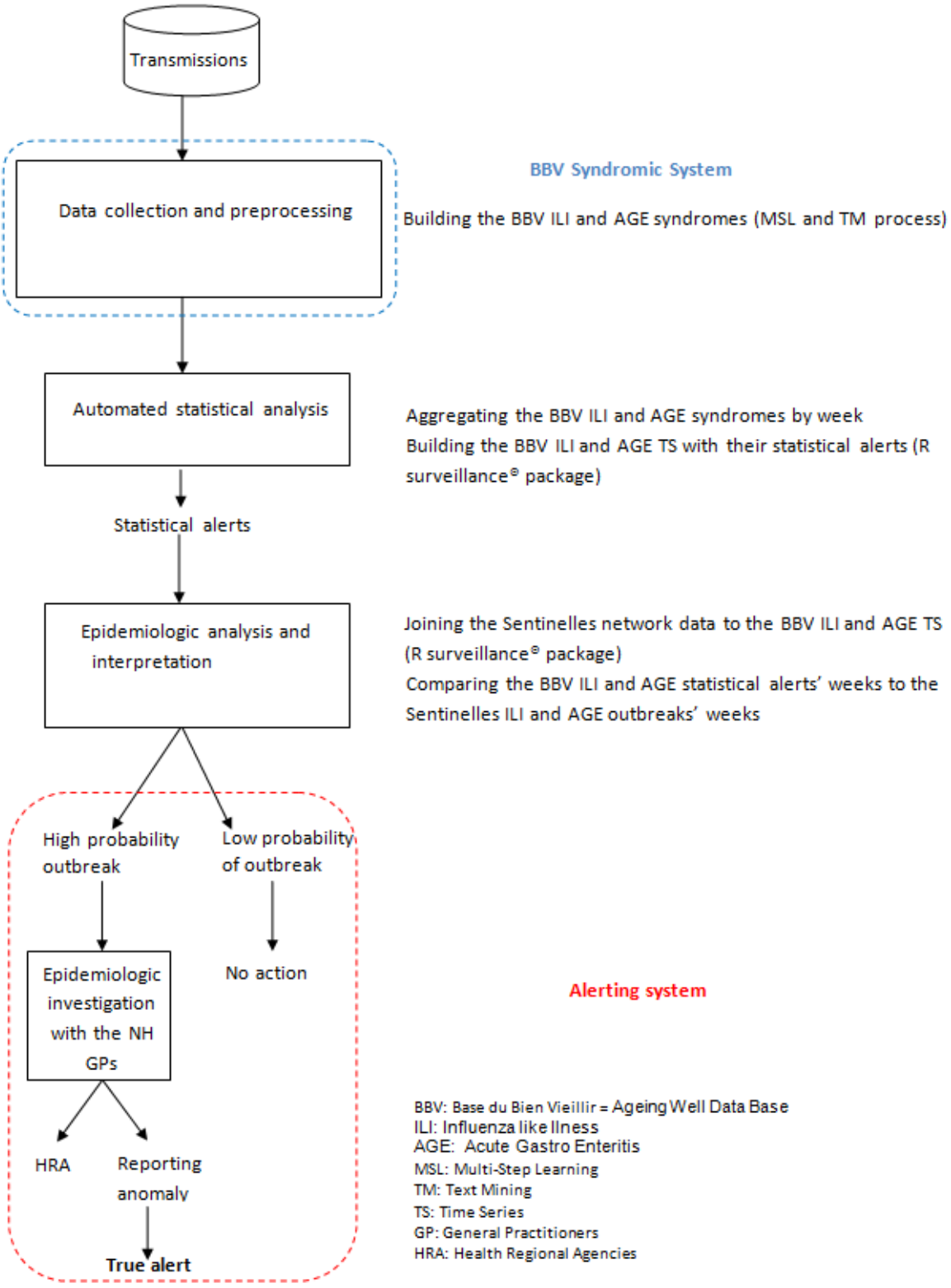


FIGURE 1: THE WHOLE NH ARI-ILI AND AGE SURVEILLANCE TOOLS FRAMEWORK

Syndromic Data Analysis

Data Flow Build Up and Stabilization

As explained above, we computed weekly counts of ILI and AGE cases, as well as hospitalizations and deaths, as with *Sentinelles*, avoiding week and weekend days' heterogeneity [78]. Then, with the *ggplot* function of the *ggplot2* R® package [79] used with

local regression curves fitted to the NH data (figure 2) [80], we were able to track yearly tendencies as well as inconsistent data not reflecting the seasonal wintry spikes.

Assessing the syndromic data flow over time, by computing the summary statistics of deaths, hospitalizations, ARI-ILI and AGE weekly syndromes' counts during the 3 following periods [11/01/2010 – 11/01/2011[, then [11/01/2011 – 11/01/2012[, and finally [11/01/2010 – 02/26/2017] we chose to exclude the first year's data and covered the period from November 1, 2011 to February 26, 2017.

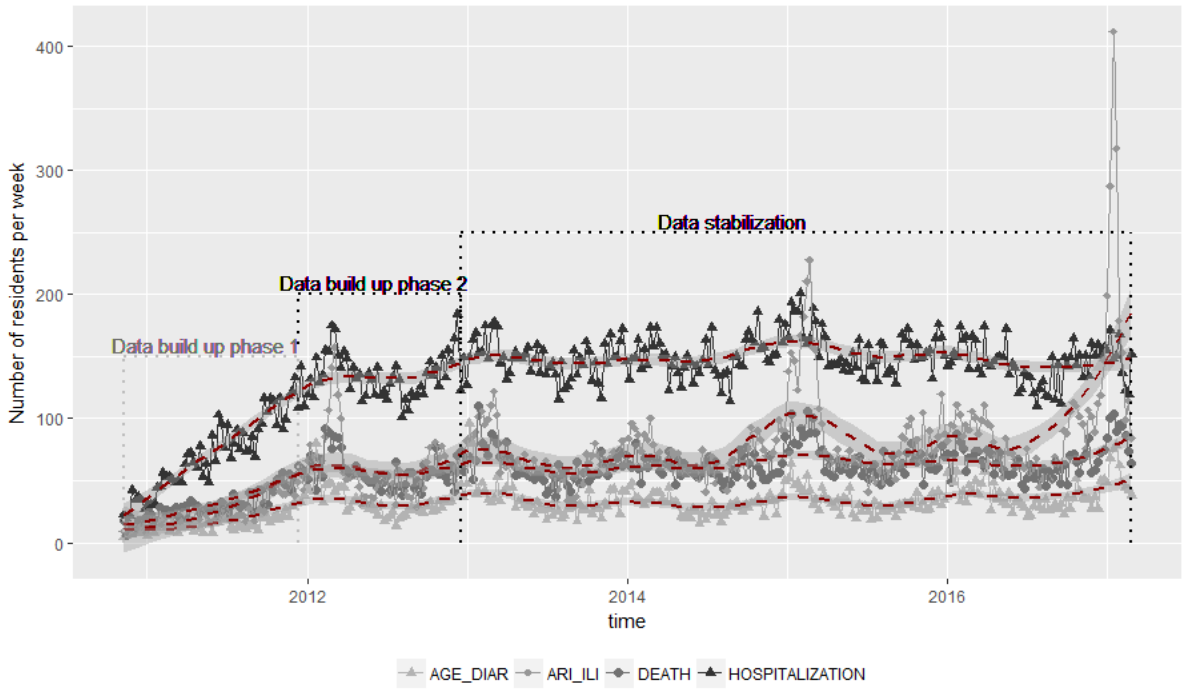


FIGURE 2: ACUTE GASTRO ENTERITIS, ACUTE RESPIRATORY INFECTIONS, DEATHS AND HOSPITALIZATIONS DATA FLOW BUILD UP AND STABILIZATION IN 11 REGIONS COVERING FRANCE BETWEEN 11/01/2010 AND 02/26/2017

Building the ARI-ILI and AGE TS

The ARI-ILI and AGE TS were built by aggregating all NH weekly syndromes counts. The choice of a statistical method to analyze these then rested on definitions of statistical alerts, adapting the BBV SSS data to fit French public health infrastructure and available SS data sources [81], here data of the ARI-ILI and AGE Sentinelles network [22]. Whereas the Sentinelles network used the Serfling method [71-72] relying on disease incidence levels of preceding years, we used the CDC steady favorite, the CUSUM methods not drawing on data from preceding years but just from preceding weeks and one recent method, the Bayes method allowing fine tuning [73]. For further details, see again Multimedia appendix 4.

Quality - Precision

Following the process described in the *Building the ARI-ILI and AGE Syndromes* subsection, the whole procedure was reviewed over 3 weeks of data transmissions: one in mid-August 2016 when there was no epidemic and two in January 2017, respectively, at the ILI and AGE epidemics weeks' peaks, according to the *Sentinelles* network [82], computing the

percentage of miss-coded ARI-ILI and AGE syndromes among the extracted data transmissions defined as such [26].

Stability

The idea here was to check the syndromic data transmission flow stability in quantity (the syndromes counts) and quality (several different recurring syndromes) during the complete period and for all 126 NH, computing the weekly syndromes frequencies for every NH.

The syndromic data flow stability was traced by designing 3 chronic diseases and one often-chronic ailment indexes [83] built as follows: whenever a resident had diabetes or a cardio-vascular problem or depression or fell, the resident's transmission date and syndromic event type were set apart. Then a similar event during a 200 days period after this resident's syndromic event was searched for, defining 4 syndromic ratios for the 6 years from year 2011 up to February 27, 2017. For further details, see Multimedia appendix 5.

Flexibility - Timeliness - Representativeness - Usefulness

Adaptability and reactiveness of the system were evaluated during outbreak and routine periods according to the CDC surveillance systems guidelines [18, 15, 26]. Representativeness, completeness and usefulness were assessed using the distribution description of ILI cases by time and origin during this last flu season, as well as by rating sex, age and GIR at entry and age at illness missing data [13].

Surveillance Algorithms' Quality

Except for the lag, the four algorithms were compared using the *algo.quality* function for *Bayes* [73] and rebuilding it for the *EARS* algorithms. This quality is defined by 4 numbers -- the number of True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) -- and 4 criteria -- the *sensitivity* [84] sometimes called recall [85] as the ratio of epidemic weeks correctly identified, the *specificity*, as the ratio of non-epidemic weeks correctly identified, the Euclidean distance between the perfect method with *specificity=sensitivity=1* and ours (*distance = sqrt((1-spec)^2 + (sens-1)^2)*), and finally the *precision* or *positive predictive value* (PPV), as the ratio of epidemic weeks correctly identified among the weeks defined as epidemic (with statistical alarm) [13].

Results

The Cohort

As explained in the *Data Collection* subsection by extracting all residents already there on November 1, 2010 and then by adding those entered every week in one of the 126 NH, a cohort of 41 061 residents (figure 2) was built with 12 983 men (32%) of mean age 84.33 and 28 083 women (68%) of mean age 85.82.

The ARI-ILI and AGE Syndromes and the Surveillance inside Korian

As described in *Building the ARI-IRI and AGE Syndromes* subsection, the BBV syndromic algorithm extracted all ARI-ILI and AGE cases, plus hospitalizations and deaths, every week, from November 1, 2010 until mid-February 2017 and built the four TS. Using the BBV ARI-ILI

and AGE syndromic TS we were able to track the last flu season (winter 2016-2017) early on, even before the epidemic and compare our syndromic counts with the Korian GPs' number of cases. The first ones were usually much greater than the second ones, as several syndromic cases could identify the same resident over time, but both of them were always strongly correlated.

Syndromic Data Analysis

Checking the Data Flow during Time

We managed to highlight 3 different phases in the NH data flow as shown in figure 2 and table 1 with two build-up phases during the first two years of the IS implementation. As seen below, between the first and the second year, the median and mean weekly syndromes counts more than doubled. For that reason, the first year of data was excluded from the syndromic data analysis.

	Min.	Q1	Median	Mean	Q3	Max
Phase 1	4	16	**24.5**	**33.45**	42.25	118
Phase 2	13	40.5	**55**	**70.92**	109.25	184
Stabilization	15	48	**67**	**82.72**	124	412

TABLE 1: ASSESSING THE BBV 4 SYNDROMES' WEEKLY COUNTS

The ARI-ILI and AGE TS

In the *surveillance* package, both *Bayes*' (see figures 3 and 5) and *EARS_C3s'* algorithms with alpha=0.025 (see figures 4 and 6) used the 12 former ARI_ILI (figures 3 and 4) and AGE (figures 5 and 6) NH weeks' counts to define *alarm weeks* (red triangles). *Outbreak weeks* (green vertical lines) were defined according to the ILI and AGE *Sentinelles* data during the same period [01/01/2011 – 01/16/2017]. Finally, the blue dotted lines were the upper limits at which alarms were triggered with both algorithms.

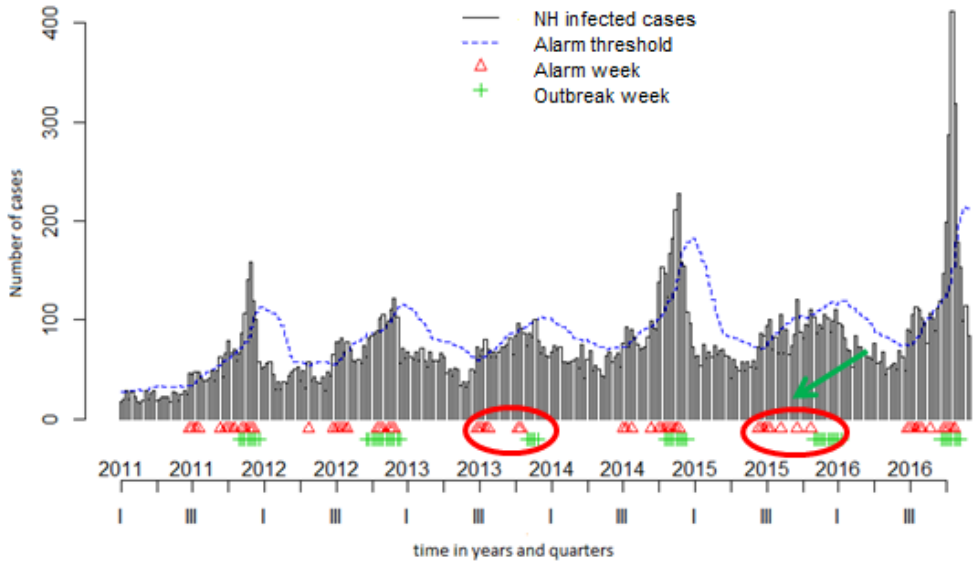


FIGURE 3: THE ILI BBV TIME SERIES (TS) USING THE *BAYES'* ALARM ALGORITHM WITH 12 WEEKS UPSTREAM AND THE ILI *SENTINELLES* OUTBREAKS

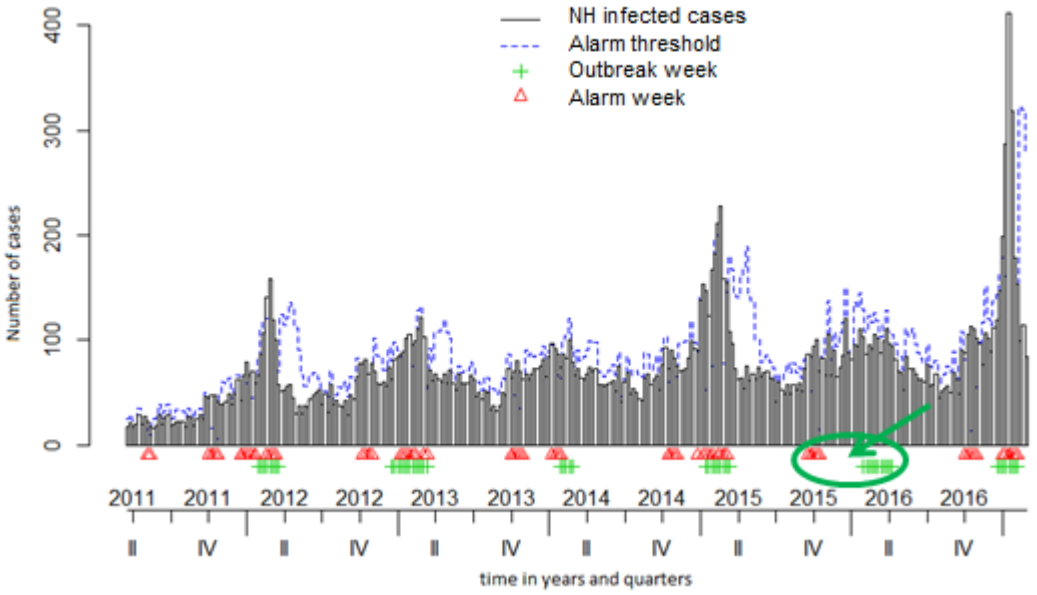


FIGURE 4: THE ILI BBV TS USING THE *EARS_C3'S* ALARM ALGORITHM WITH 12 WEEKS UPSTREAM AND THE ILI *SENTINELLES* OUTBREAKS

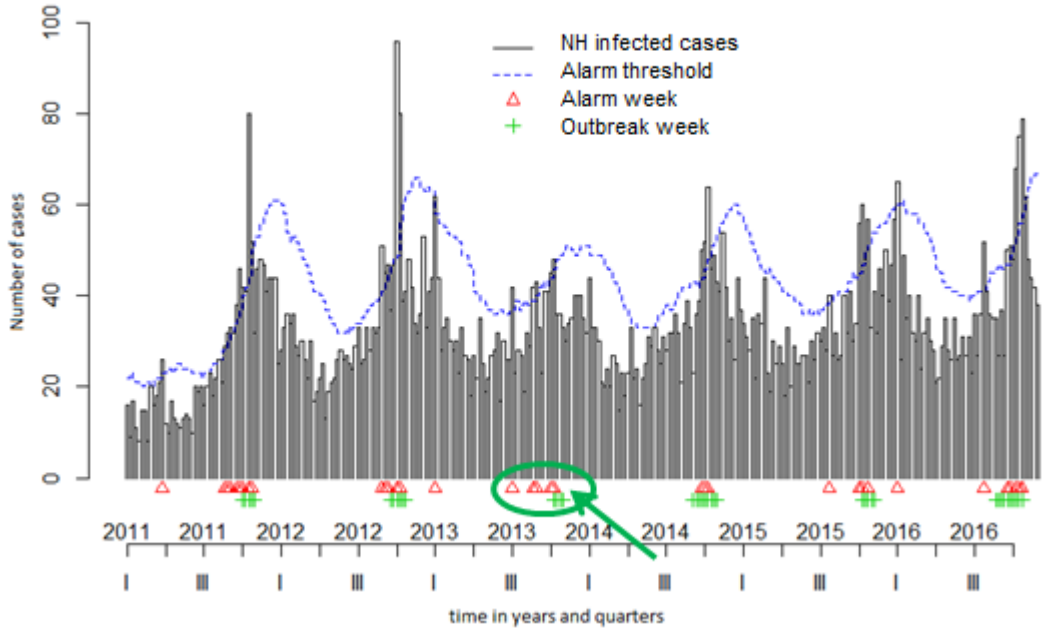


FIGURE 5: THE AGE BBV TS USING THE *BAYES'* ALARM ALGORITHM WITH 12 WEEKS UPSTREAM AND THE AGE *SENTINELLES* OUTBREAKS

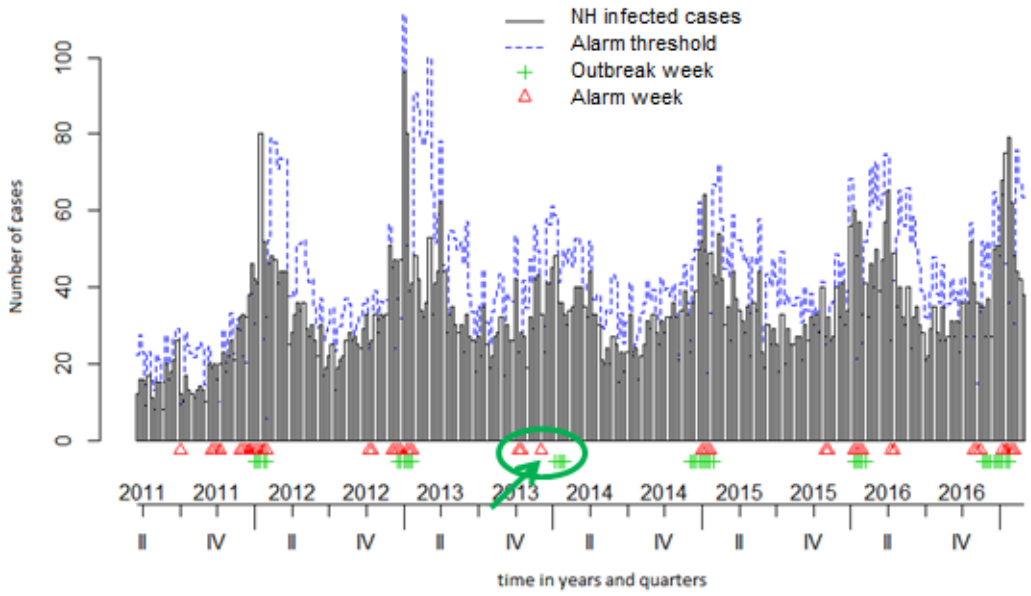


FIGURE 6: THE AGE BBV TS USING *THE EARS_C3'S* ALARM ALGORITHM WITH 12 WEEKS UPSTREAM AND THE AGE *SENTINELLES* OUTBREAKS

Senior citizens suffer much more of either ARI-ILI or AGE than the general population all year long and even in summer. This often results mechanically in triggering statistical alerts long before the general population epidemics. It can be seen in figures 3 and 4 for ARI_ILI and figures 5 and 6 for AGE where the red triangles (the SSS alarm weeks) appear always before the green bars (the Sentinelles' network outbreak weeks). It is especially true for ARI-ILI during the 2013-14 and 2015-16 winters and for AGE during the 2013-14 winter with both algorithms.

Quality - Precision

To compute the percentage of miss-coded ARI-*ILI* and *AGE* syndromes, all ARI-ILI and AGE syndromic data were extracted during 3 weeks, one in mid-August 2016, the thirty-third week (third column) when there was neither flu, nor AGE epidemic, and two in January 2017, the third and first weeks, at the ILI and AGE epidemics weeks' peaks, according to the Sentinelles network (second column). Then, each ILI and AGE syndromic data transmission was examined, rating it as correct, adding to TP, or incorrect, adding to FP (see table 2).

Disease	Epidemic period	Week of study	Number of residents with syndromes	Number of residents with syndromes x transmission days	Number of residents with ILI or AGE	FP	TP	Precision or PPV
ILI	**no**	sem33-2016	5399	10631	56	3	53	**0,95**
AGE	**no**	sem33-2016	5399	10631	26	2	24	**0,92**
ILI	**yes**	sem03-2017	6013	12215	318	5	313	**0,98**
AGE	**yes**	sem01-2017	5862	11933	68	3	65	**0,96**

Precision=TP/ (TP+FP)

TABLE 2: ASSESSING BBV ILI AND AGE SYNDROMES DATA TRANSMISSIONS PRECISION DURING 3 PERIODS: LAST SUMMER (2016-33RD WEEK), THE LAST ILI (2017-3RD WEEK) AND AGE (2017-1ST WEEK) EPIDEMICS WEEKS' PEAKS

The precision was best during epidemic weeks' peaks: 98% and 96% respectively, as ILI and AGE, versus 95% and 92% in summer and there were very little FP. For example, for ILI FP, were excluded *'his son has flu', 'emergencies overloaded with flu cases'*, '*no flu symptoms'*, *'could they take care of my girl who has flu?',* and finally, *'serrure dégrippée'* which means *unjammed* in French but has the same word stem '*gripp'* as flu. We had already excluded the word *'grippé'* in this context, which means *jammed* for a lock. In addition, by checking the flu cases we found 67 flu tests mentions using nasal swabs, adding 21% new cases. For AGE FP, were excluded *'vomited without diarrhea'* (2 times) as both words are needed to classify as an AGE syndrome and '*diarrhea protocol if fever*'.

Stability

As detailed in *Syndromic Data Analysis* subsection- *Stability* sub-subsection, the data transmission stability was evaluated by studying the weekly syndromes frequencies for every NH from November 1 2010 as well as the ratio of weeks with syndromic data transmissions:

- The weekly 26 syndromes frequencies averaged over the number of NH (126) ranged from 21.45 to 180.57 (mean=88.5 standard deviation=22.8).
- The ratio of weeks of data transmissions per NH was built by computing the number of data transmissions weeks versus the data transmissions weeks span and ranged from 89% to 100% (mean= 100%, standard deviation=1%). Only one NH had a ratio of less than 95%.

Finally, the syndromic data flow over time studied with 3 chronic illnesses and falls syndromes distributions showed great stability (see Multimedia appendix 5 for further details).

Flexibility - Timeliness - Representativeness - Usefulness

Table 2 showed the flexibility and reactiveness of our syndromic system during epidemic periods, following at the same time the increasing number of cases, going from a weekly population of 5 399 to 6 015 without losing any precision, from 95% to 98%. As shown this winter, by using the NH indexes and regions, the flu epidemic was followed geographically, week after week, finding from the beginning where the epidemic was most intense, tracking the most severe cases and related hospitalizations and deaths.

Great geographic heterogeneity was detected between regions in terms of ratios of infected, the Rhone valley (Rhone Alpes) and South (Sud) being the most afflicted with 24% and 19 % and the Southeast (Sud-Est), being the less with only 9%. There was also great variability in terms of population characteristics for example, those from the South-West were the oldest afflicted at a mean age of 89.7 whereas the South residents were the youngest at 87.6, 2 years and a month difference being quite a lot given the mean duration of residents' stay. Finally, among 1 800 residents with a flu transmission, only one had no personal data.

Surveillance Algorithms' Quality

The 3 surveillance package EARS's algorithms were compared for both diseases, with two different confidence interval levels, alpha=0.001, the EARS_C3 default level, and alpha=0.025, the *Bayes*' algorithm default level (see the algorithm quality in table 3). The EARS_C3 with alpha=0.025, gave the best results for both diseases. Nevertheless, the *Bayes*' algorithm seemed better to define alarm weeks when epidemics were less intense as for ILI 2015-16 and AGE 2013-14 seasons, as there were less lag weeks between the

Bayes' alarm weeks and the Sentinelles outbreak weeks (green arrows were added in figures 4-6 to highlight this trend).

Disease	Algorithm	Number of weeks	TP	FP	TN	FN	Sensitivity or recall	Specificity	Distance	Precision or PPV
ILI	***Bayes***	12	22	44	210	32	0.407	0.827	0.617	0.333
	EARS_C1	7	6	12	248	48	0.111	0.954	0.890	0.333
	EARS_C2	9	13	26	232	41	0.241	0.899	0.766	0.333
	EARS_C3[a]	12	19	31	225	35	0.352	0.879	0.659	0.380
	EARS_C3[b]	***12***	***26***	***40***	***216***	***28***	***0.482***	***0.844***	***0.542***	***0.394***
AGE	***Bayes***	12	16	18	251	23	0.410	0.933	0.594	0.471
	EARS_C3[a]	***12***	***16***	***13***	***258***	***23***	***0.410***	***0.952***	***0.607***	***0.592***
	EARS_C3[b]	***12***	***21***	***26***	***245***	***18***	***0.539***	***0.904***	***0.471***	***0.447***

[a]with alpha=0.001, [b]with alpha=0.025
Distance = sqrt $((1\text{-spec})^2 + (\text{sens-}1)^2)$ is the Euclidean distance of (specificity, sensibility) from (1, 1)
Precision=TP/ (TP+FP) is the True Positives ratio among the positives.

TABLE 3: COMPARING THE SURVEILLANCE ALGORITHMS' QUALITY ON BBV ILI AND AGE TS TO ILI AND AGE SENTINELLES' OUTBREAKS DETECTION

Either with ILI or AGE TS, mostly coherence between NH data and the Sentinelles data could be witnessed. Also only 12 weeks of data (see table 3) were needed to detect outbreaks, most of the time several weeks ahead of Sentinelles' outbreaks. This was especially true for the last flu season (winter 2016-17 in figures 3-4).

Discussion

Principal Findings

We built and assessed a national ecological NH PH SS dedicated to senior citizens. By using a national network of 126 NH and extracting all socio-demographic as well as daily medical data from EHR, a cohort of 41 061 residents was built. Through textual analysis of clinical narratives (CN), we implemented ARI_ILI and AGE syndromes. We also engineered related TS by computing weekly headcounts. Alarms with *EARS_C3* and *Bayes* algorithms on these, over a 6 year-period, allowed us to forecast the 2016-17 influenza outbreak by more than 2 weeks as can be seen in figure 4: our statistical alarms were triggered in December whereas the influenza epidemic according to SPF started only in January.

With just four tables, this IS of a new kind showed that it is possible to follow almost every resident every day, where he or she is, during his or her entire NH life, hopefully selecting most of his or her ARI-ILI and AGE health events, from NH entry until death or exit. Furthermore, each relevant syndrome is defined by two syndromic representations: either a simple additive syndromic image that is, its four Boolean [65] syndromic components allowing whatever filtering, or its literal expression for further textual analysis or in-depth health questioning. By this whole process, free textual information extracted from CN was shaped into numerical data for further statistical or machine learning analysis.

We engineered here a real NH SSS on qualitative data, offering immediate accessibility without adding any extra work to medical staff [11]. By using SQL LIKE pattern matching [37] and Delphi like experts' consensus [57-58] on the data transmissions' file, we followed last season ARI-ILI and AGE epidemics and found almost in real-time that the flu reached

dramatically NH residents, tracking them geographically and timely, searching for flu related hospitalizations and deaths. Preventing disruptions of medical tasks and medical and paramedical staff turnover by predicting even one or two weeks ahead, the epidemic intensity could greatly improve the NH human resources management over time and help preventing sanitary disasters by strengthening hygiene measures for example.

As explained in [12], early detection of outbreaks can be achieved in three ways: first, by prompt recognition and reporting of disease case reports. Here, we could find most of flu and AGE cases by syndromic descriptions fed in the data transmissions table. Second, by improving the ability to recognize patterns indicative of a possible outbreak early in its course, using analytic tools, counting syndromes by NH and building time series with the surveillance® package. Third, by exploiting data that can signify an outbreak earlier in its course. More specifically, adding hospitalizations and deaths syndromes to the ARI_ILI, AGE syndromes, allowed us to assess the flu and AGE outbreaks intensities as well as their severities long before French health authorities this last season and follow precisely and locally the residents' syndromic population thanks to the NH and residents' indexes.

This framework with its three components wholly described in figure 1, has then showed its efficacy as a Public Health SS for early detection of outbreaks. By bringing to light new data data not available elsewhere when needed, this SSS improves NH ARI-ILI epidemics' knowledge. Its tools' efficacy could even be quantified by assessing syndromes' precision, stability, flexibility, timeliness, representativeness and finally algorithms' quality [12, 66].

For the AGE data, even with lots of cases, a good correlation could be found for every winter season between the NH alarm weeks and Sentinelles outbreak weeks, (as shown last row in table 3 by the small distance value of 0.471), the first ones, almost always preceding the latter by several weeks, except for the 2014-2015 winter where the AGE epidemic reached essentially senior citizens in NH [86]. During last winter, the AGE outbreak started at the same time as in other NH in France.

Limitations

This SSS using mostly the transmissions' qualitative data, is nor exhaustive as some syndromes may still not be described in the SSS, neither complete, as medical staff may not have fed all syndromic information at some day for whatever reason. So ILI and AGE syndromic data recall, what proportion of cases in classes were correctly assigned to their classes [65], could not be assessed. It was the same for the F score, the harmonic mean of both recall and precision. At this moment, the syndromic information depends essentially on the medical staff available time and dedication to feed the system as showed in the *Results* section *Syndromic Data Analysis* subsection *Stability* sub-subsection, where one NH had a ratio of data transmissions weeks of 89%, 293 weeks of data transmission over a total span of 329 weeks.

As soon as the cold season begins, elderly people may get respiratory syncytial virus (RSV) just as the very young children. In fact, RSV is a common cause of acute respiratory illness in older adults, the risk of serious respiratory infection increasing with age [87 – 88]. Usually, RSV spreads quickly just before flu or at the same time and is largely indistinguishable from influenza based on clinical presentation alone [87, 50-51]. It is rather a recurring problem in older adults causing 2 to 5% of adult community-acquired pneumonias [89]. Triggering an alarm even for RSV would allow to quickly organizing care to the residents.

Then, by following our syndromic ARI_ILI data two trends could be traced, one starting in early November, may be the RSV, followed by another one later, starting usually in December as this year or later as last year. Depending on the flu epidemic characteristics

and as ARI, ILI and RSV could not be distinguished in our textmining algorithm, a flu threshold could be detected whenever appropriate or several weeks ahead. As can be seen, during the 2013-2014 and 2015-2016 winters, between the first alarm weeks and the outbreak weeks, quite long times elapsed [90-91], but as not really reaching elderly people, there was not something clear to find. On the opposite, during the 2014-2015 and 2016-2017 winters, a much better correlation could be found, the first ones, probably because of RSV, always preceding the latter by approximatively 8 weeks (figures 3-4), thus often triggering alarms before those of the Sentinelles network.

At the same time proportionately much more ILI new cases with our SSS were found than with the Sentinelles network especially for this last influenza season (see the last ARI-ILI surge at the beginning of 2017 in figures 3-4). As a type *A* influenza virus, it reached people older than 75 much more than the rest of the French population [82]. Then, as soon as clusters of NH ARI cases appeared, many flu tests had to be done to label residents as flu positive or negative. Moreover, even as some tests were negative they derived from the flu epidemic health protocol and were mandatory to HRA hygienic safety measures [92-96], increasing the number of cases still more.

Nevertheless, as detailed above, less lag weeks were found with the *Bayes*' algorithm and even an overlap of alarm weeks and outbreak weeks for the ILI 2015-16 (figure 3) and AGE 2013-14 (figure 5) epidemics and nothing like that with the EARS_C3 algorithm (figures 4 and 6). It could then be tried in the following years to mix both algorithms as was done in [97] for Salmonella and decide triggering an alarm whenever one of the two algorithms reaches its alarm threshold, probably improving both sensitivity and specificity. Or again, as in the new MASS (Module for the Analysis of SurSaUD and Sentinelles' data) system [69] designed by Santé Publique France, combining 3 statistical methods and 3 different data sources, used since January 2016 to define the public health alerts.

Finally, the epidemiologic analysis and interpretation steps (figure 1) were not fully automated. Some work still needs to be done, especially the whole Sentinelles data extraction process. Some similar job was done before on another project [97-98].

Conclusion

Outbreak alerts are more reliable when systems focus on specific syndromes that reflect high-probability events such as influenza [62] as could be seen in this real-life experiment. However, there is always room for improvement, as the aggregation of ARI and ILI as well as RSV constraint shows. Nevertheless, this IS gives already a rich and detailed *syndromic* image of these residents. Plus, as syndromes are modular and the *Pentaho*® platform [64] allows extraction from different data silos, it will be possible to add new syndromes, may be RSV, whenever needed and also to adapt them to the new IS twice as big due next year.

This study follows another work on CN using textual analysis and clearing the way for this syndromic health IS design [99]. Tracking flu and AGE epidemics seasons almost in real time and following their impacts especially during this last year acute flu season has helped to show this SSS usefulness. What's more, the [November 2010 – June 2016] syndromic data were used to build ARI_ILI and AGE algorithms and nothing had to be added or retrieved to follow these last season epidemics trends, so these algorithms exhibited as well flexibility, adaptability, stability and timeliness.

This study highlights some differences between the NH residents' population and the general population, which hampers a better correspondence between NH alarm weeks and

Sentinelles outbreak weeks. The main challenges here are extending the syndromic IS, improving the syndromes descriptions, as well as better taking into account NH residents' distinctiveness. Monitoring flu and AGE using the BBV IS could give way to a real SS for all senior citizens in France. For example, there are incoming discussions between Korian and HRA about targeting RVS besides flu and handling what differentiates them.

Korian NH are already working with HRA at a local level, exchanging clinical data with them whenever outbreaks are detected. These data sharing could then be extended with syndromic data integration, resulting in HRA reactiveness improvement [100]. Indeed, syndromic data are always available before, even if less precise. NH residents as a whole are a frail and captive population functioning as an ever-increasing reservoir for any contagious illness [101-102]. It is then essential being able to prevent with all possible disposable tools any health catastrophe in the near future.

This syndromic IS offers a real opportunity finding new ways to seniors' functioning modelization and opens hopefully the path towards specific clinical hypotheses formulation. Other works included studying the use of this IS applied to other public health problems such as frequent falls or falls with casualties [103], but also working towards a better life ending with cancer [104]. Ultimately, the aims are removing all preventable deaths and improving the residents' end of life with more autonomy, less pain and an improved quality of life, translating this new knowledge into health benefits for seniors everywhere..

Acknowledgments
Authors would like to thank Sebastien Plasse, project manager from the Korian group Information Systems Direction who gave Tiba Delespierre details about the IS structure and how best extract data. He also helped building the Syndromic Surveillance System.

Authors' contributions
TD and LJ had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors contributed to the study concept and design, critically revised the manuscript for important intellectual content, and supervised the conduct of the study. TD oversaw the data extraction and analysis as well as statistical analysis, and LJ obtained funding.

Funding
Foundation Korian of Well Ageing inside the Korian group is funding Tiba Delespierre's public health thesis and financing this manuscript as well as every scientific result the main author may publish.

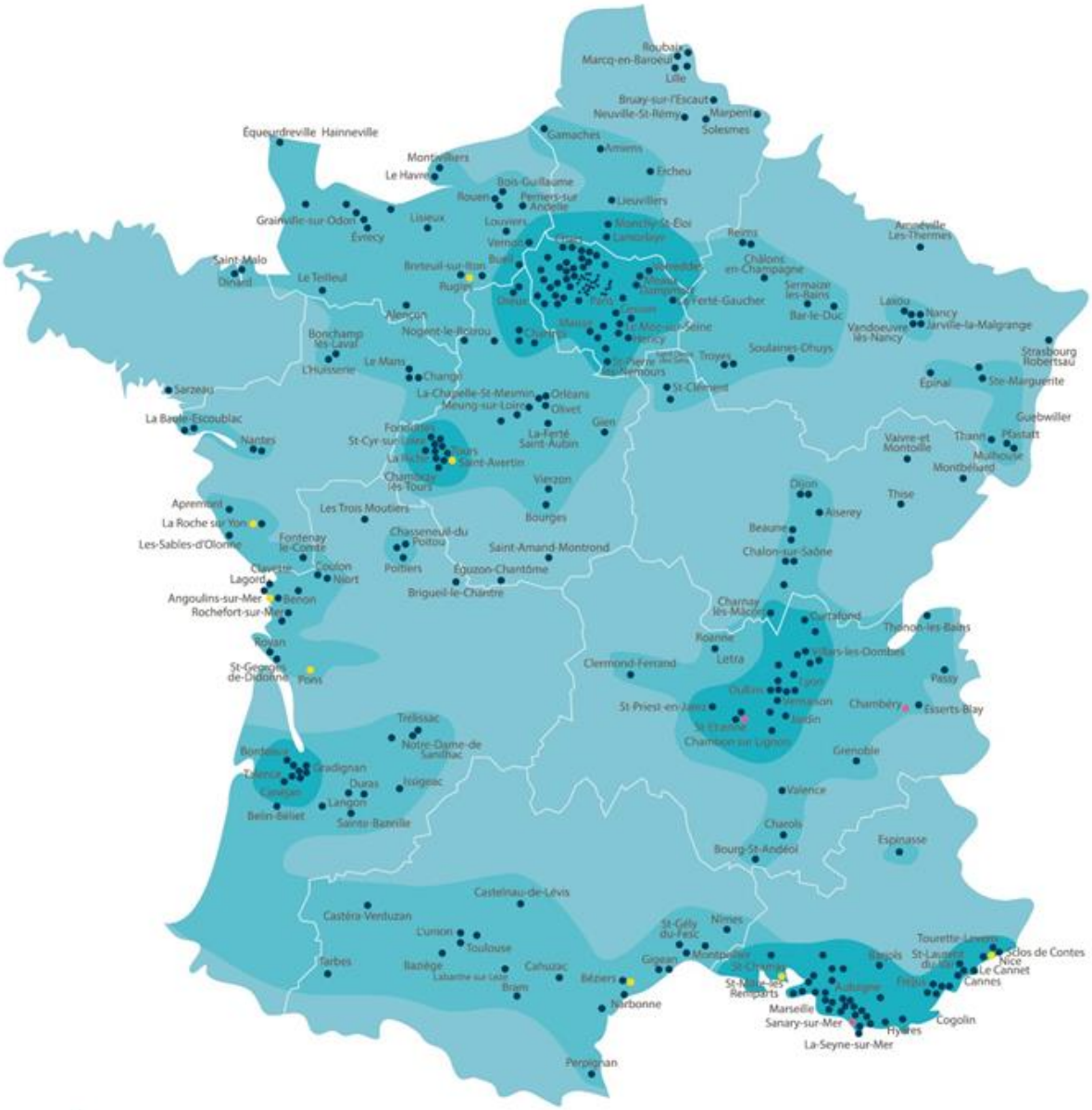
Conflicts of interest
LJ does not have any financial competing interests to report but he's a member of the Korian Aging Well Committee. Funding to support TD's work is reported above.

Ethics approval and consent to participate
The use of this database in the frame of epidemiological studies has been authorized by the French National Commission for Data protection and Liberties (CNIL). The Institut du Bien Vieillir, which became the Foundation Korian of Well Ageing, filed a declaration of conformity to a baseline methodology, which received in March 2017 an agreement number: 2.041.050, in accordance with the Act n°78–17 of 6 January 1978 on Data Processing, Data Files and Individual Liberties. All residents are informed at their NH entry about their EHR and their right to oppose its use. While the primary purpose of this medical research was to generate

new knowledge, this goal did not take precedence over the rights and interests of the NH residents. All the new generated information was extracted from already existing data and was de-identified and anonymized when necessary to protect their health and rights. There were no images and no identifying details on individuals reported within this manuscript.

Appendices

Multimedia Appendix 1
The Korian NH network in France



THE KORIAN NURSING HOMES NETWORK IN FRANCE

Multimedia Appendix 2
The anonymization process

We re-indexed all the residents' and NHs' index in order to protect personal privacy but will always be able to re-link them later by a trusted party if later needed. The new indexes are computed through piecewise linear increasing functions smoothed on the original indexes, functions that can be redefined periodically. Without an access on the Korian group original database it is then nearly impossible to find the matching between old and new indexes.

Multimedia Appendix 3
The BBV ARI-ILI and AGE syndromic information building process in four phases

Phase 1

Internal medical experts and PERMF users were interviewed on how to best define ARI-ILI and AGE syndromes with words [36]. Two lists were selected, one defining the first syndrome with words such as *acute*, *fever*, *flu*, *coughing*, *asthenia*, *pneumonia*, and a second with words such as *diarrhea*, *vomiting*, *with blood*, *abdominal pain*, *fever.* Among these words designed for the ARI-ILI syndrome, several of them describe non-specific respiratory diseases' symptoms. For example, *fever* and *coughing* can also come from other respiratory problems such as the Respiratory Syncytial Virus [50-51].

Phase 2

Phase 2 involved an algorithm of SQL queries automatically analyzing all health descriptions and picking the right sets of words describing each syndrome [52]. This data processing was then tried through small periods of winter, when there were numerous cases and through small periods of summer, when there were much less (especially for flu), analyzing every result, true positive (TP), true negative (TN) cases and each health description, to see what words to add and what words to remove to improve results. Each new iteration of this process resulted in better-defined cases. Through this process, the words' meaning and shape were also checked to exclude erroneous interpretations [55-56].

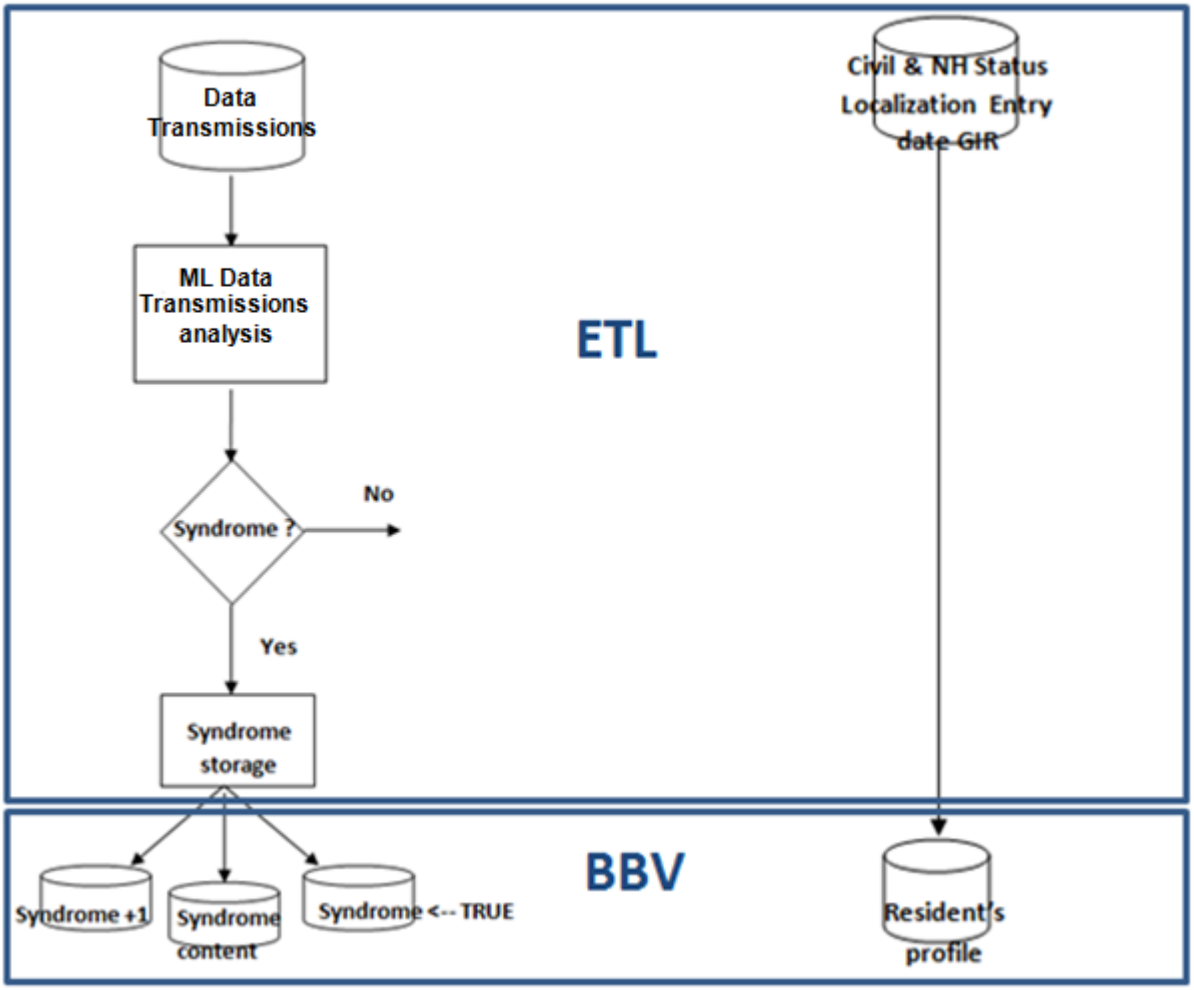
Phase 3

After this second step, medical experts were asked again to check the syndromes, adding or removing some words to describe better them in a Delphi-like process [55-56] until reaching an agreement [57]. As in [58], the ARI_ILI and AGE syndromic definitions construction followed a process based on consensus and current use.

Phase 4

Finally, through whole past seasons, week after week, our syndromic numbers were compared with national trends defined elsewhere (*Sentinelles* network data [22]), the aim being to have concordant data and being able to use the CDC surveillance algorithms on them to detect or even better predict NH epidemics. Whenever an anomaly was found, the discrepancies periods were re-analyzed either to find an explication or to adjust the words' lists and/or combinations. For example, for the ARI-ILI syndrome, several wrong cases came from influenza and pneumonia vaccines. To conclude, all this process could be assimilated to supervised learning [59-60], defining the method with a small test sample and refining it with a bigger learning one, except that in this study, we had many data, enough data to test and learn on numerous samples, not fearing any over adjustment and not needing any cross-validation [61].

The four BBV tables



THE BBV SYNDROMIC INFORMATION SYSTEM BUILDING PROCESS

The whole syndromic building process is described on the left side of the figure.

Whenever machine learning data transmission analysis led to syndromic data and described one or several of the 26 syndromes (**Yes**), it was stored in three outcomes tables (bottom left of the figure). From left to right, the first table added one to each of the syndromes distributions needed to describe the data transmission, the second table recorded the syndromic sentence and the last one built the Boolean filters to use for subsequent SQL requests.

Otherwise, if data transmission analysis did not lead to syndromic data (**No**), nothing was stored.

Example 1 of an ARI-ILI syndrome literal description and how it is processed:

"D: ASTHENIC ++ THIS MORNING. FEVERISH 101DEG. PALE, CONGESTED, WHEEZING. 1G PARACETAMOL GIVEN”

The first step involves analyzing this short sentence. The process catches the words *feverish, congested* and *wheezing* as describing an ARI-ILI syndrome. Then this ARI-ILI syndrome is stored in the BBV database through 3 different ways:

1- one is added to the NH's ARI-ILI syndromes' weekly count
2- the short sentence is fed in the literal syndromes' description table;
3- ARI-ILI syndrome is fed with TRUE and the 25 others fed with FALSE in the Boolean syndromes' description table.

Example 2 of a combined AGE / ARI-ILI syndrome literal description and how it is processed:

"ASTHENIC +++ VOMITING FOOD, FEVERISH RECORDING VITAL SIGNS: TA =11/7, PULSE 97R, SAT 97%, TDEG=102 PARACETAMOL SUPPOSITORY GIVEN AT 10:30PM TDEG=99 , SAT = 93 AT 4:30AM CHECK WITH MG MONITORING +++ (FLU SYNDROME ??)"

Here the process catches the words *vomiting* and *feverish* as describing an AGE- syndrome and the words *feverish* and *flu* as describing an ARI-ILI syndrome. Both syndromes are stored in the BBV database through 3 different ways:

1- one is added to the NH's AGE syndromes' weekly count;
2- one is added to the NH's ARI-ILI syndromes' weekly count;
3- the short sentence is fed in the literal syndromes' description table;
4- AGE and ARI-ILI syndromes are fed with TRUE and the 24 others fed with FALSE in the Boolean syndromes' description table.

With this method, syndromic data is split in 3 parts:

1- The twenty-six syndromes numbers traced every week in every nursing home, generating the syndromic surveillance tool (first table from the left);
2- The literal syndromic description of every impacted resident (second table from the left). These descriptions, which sometimes directly refer to the resident or resident's relations, are loaded separately with an exclusively restricted access.
3- The Boolean syndromic description of every impacted resident (third table on the left);

Finally, added to these syndromic data, we added the residents' profile with gender, age and Iso Resource Group at the NH entry and NH localization loaded on a fourth table (bottom right of the figure).

The BBV 26 syndromes list

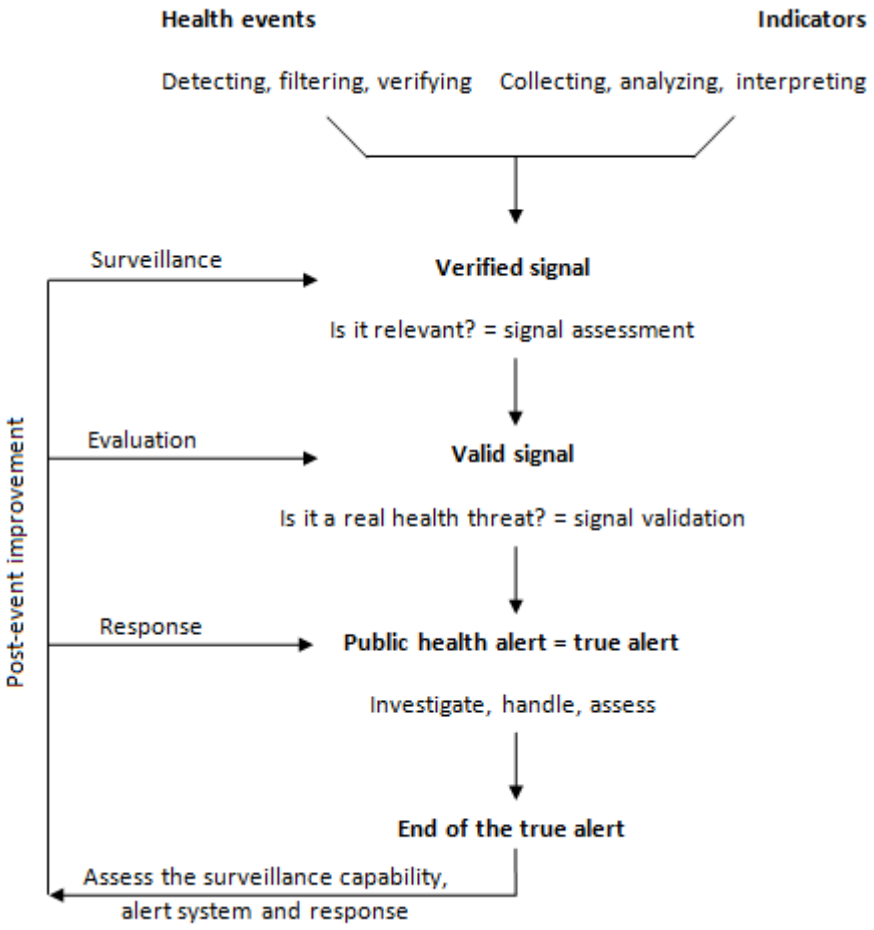
Collected data covered elderly key concerns and health priorities. The complete twenty-six syndromes' list was:

1- ARI-ILI and AGE built through the 4 phases described above;
2- Hospitalizations and deaths directly built from the hospitalizations and deaths tables' extractions;
3- Twenty-two remaining syndromes built through the 3 phases described above as the fourth phase was not applicable to them: 1-pain, 2-behavior, 3-dementia, 4-general state alteration, 5-dehydration, 6-denutrition and swallowing, 7-cutaneous state, 8-allergies, 9-falls, 10-depression and dark thoughts, 11-cardio-vascular symptoms, 12-

audition, 13-oral health, 14-cancer, 15-sleeping problems, 16-vaccination, 17-vision, 18-intestinal transit, 19-urinary track, 20-frailty, 21-overweight and 22-diabetes.

Multimedia Appendix 4
How public health alerts are defined:

The French public health agency triggers a public health alert after the following 3 steps are fulfilled: first, signals' reception (health events and/or statistical indicators), second, signals' validation after local, regional and national epidemiologists' verifications and third, assessing the health threat by public health authorities [70] (Santé Publique France SPF).



THE SURVEILLANCE, ALERT AND RESPONSE CONCEPTUAL FRAME PROCESS ACCORDING TO SANTÉ PUBLIQUE FRANCE *[70]* *added*

For all surveillance methods, building time series (TS) and statistical indicators from them or from other available data is always the first step. Here in this experiment as reference material, we chose to use the Sentinelles network [22] outbreaks data.

Statistical alerts according to the Sentinelles network

Alerts are triggered whenever outbreak levels are exceeded. The outbreak detection level with the Sentinelles network corresponds to the estimated baseline prediction confidence interval upper-limit

computed with the Serfling method [71-72]. For the ARI-ILI and AGE Sentinelles data, when weekly incidence crosses this threshold, there is a high probability of being in an epidemic period, especially when the weekly cases' numbers cross this threshold twice in the row [72].

True alerts according to the MASS system

The MASS system ((Module for the Analysis of SurSaUD and Sentinelles' data) combines 3 different statistical methods on 3 different data sources [69], including the Sentinelles network data to define its signals, verify and validate them. These outbreaks are defined by SPF following the whole process described in the figure above and are then **true, genuine alerts**.

How the surveillance® package works:

TS with the surveillance® package [67] involved defining them as disease progress objects containing two vectors: first, the observed number of weekly counts and second, a Boolean vector state indicating whether there was an outbreak that week or not [73]. Here in this experiment, we chose to use the Sentinelles outbreaks and so in the influenza case, the vector of outbreak states contained the ILI Sentinelles' epidemic weeks. Then we selected 4 algorithms to build our statistical indicators: Bayes and the 3 CDC CUSUM methods.

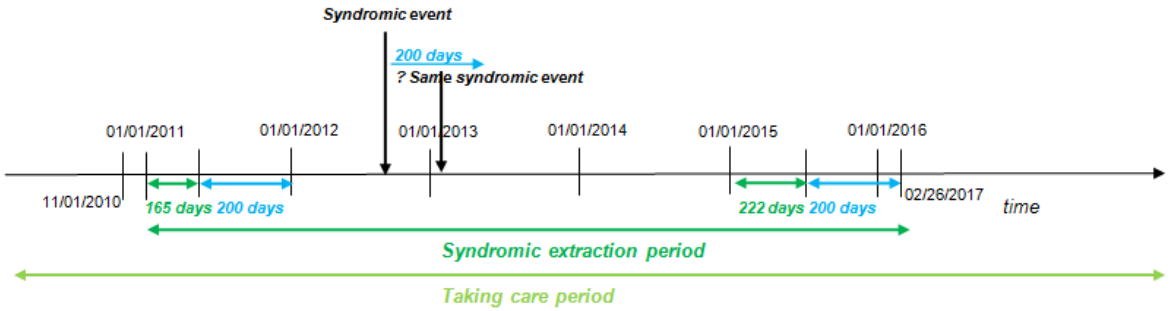
1- **The Bayes surveillance algorithm:** This algorithm, using the 12 former weeks, assumes that the reference values are identically and independently Poisson distributed where a Gamma-distribution is used as Prior distribution for the Poisson parameter. Within this framework, quantiles of the predictive posterior distribution are used as a measure for defining alarm thresholds and alarm weeks [12, 73].
2- **The 3 EARS' (early aberration reporting system) surveillance algorithms:** These algorithms, using 7, 9 or 12 preceding weeks [74-77] are non-historical methods based on positive 1-sided cumulative sums (CUSUM) calculations [74]. For C1 and C2, a warning is generated when the current count is greater than the baseline mean plus 3 standard deviations. For C3, a warning is generated when the last three CUSUMs' mean is greater than the baseline mean plus 2 standard deviations [75]. Each warning generates an alarm week. Further details can be found in [67] at the earsC section.

After these two steps, the signal is verified and the surveillance process starts as such. The last step is then comparing *alarm weeks* built with the *Bayes* or *EARS_*C_i, *i=1, 2, 3* algorithms to *epidemic weeks* (*outbreak weeks*, here *epidemic weeks* recognized as such by the *Sentinelles* network), the aim being the best possible overlapping between the two.

Multimedia Appendix 5
Syndromic Data-Flow Stability

Syndromic data-flow stability was defined here by 4 syndromes indexes: diabetes, cardio-vascular problem, depression and falls.

For the 2016 year, as 200 days-reference periods were needed, only residents with the first 222 (165+57 = 61% of 365) days of 2016 (denominators counts) could encounter these same syndromic events (numerators counts) through whole 200 days periods thereafter (reaching the end of February 2017: see figure below). To counter this, we then checked stability by comparing our four 2016 ratios to the 2015 ones, computing four 2015-2016 syndromic events relative ratios.



THE SYNDROMIC DATA FLOW STABILITY ASSESSMENT PROCESS

The 4 syndromes' data flow stability ratio (see table below) was computed as follows:

1- on denominator, for every year and every syndrome, the transmissions headcounts;
2- on numerator, the same syndromic events headcounts, for the same people on a 200 days period thereafter.

The small arrows under each attrition ratio showed the 2011-2015 trends. All four syndromic ratios through time were highlighted. The last column computed the 2015-2016 transmissions headcounts differences, allocated to 2015 transmissions headcounts and as explained above, showed some attrition. Every first transmission was followed on average by another one of the same type through a 200 days period for about 40% of diabetics, 50% of cardiac patients, 76% of depressives and 66% of frequent fallers for the 6 years period.

Syndrome headcounts, ratios & trends	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2015-2016 Relative ratios**
Diabetes	818/1574	765/1786	716/1736	724/1759	654/1658[a]	564/1496[a]	**162[a]/1658**
↘	**52%**	**43%**	**41%**	**41%**	**39%**	**38%**	**10[b]%**
Cardio-vascular	1871/3682	2227/4533	2074/4223	2002/3893	1980/3953	1510/3327	470/3953
→	**51%**	**49%**	**49%**	**51%**	**50%**	**45%**	12%
Depression	6413/8254	5379/7073	4496/5900	4265/5407	4358/5458	3496/4623	862/5458
→	**78%**	**76%**	**76%**	**79%**	**80%**	**76%**	16%
Frequent falls	4458/6410	3942/6036	3249/5039	3190/4810	3126/4683	2746/4238	380/4683
↘	**70%**	**65%**	**65%**	**66%**	**67%**	**65%**	8%

[a] 162 = 1658 – 1496
[b] 10% = 162/1658

RATING SYNDROMIC DATA FLOW STABILITY DURING TIME WITH FOUR SYNDROMES FREQUENCIES: DIABETES, CARDIO-VASCULAR PROBLEMS, DEPRESSION AND FALLS

References

1 The 2018 Ageing Report Underlying Assumptions and Projection Methodologies. Joint Report prepared by the European Commission (DG ECFIN) and the Economic Policy Committee (AWG) https://ec.europa.eu/info/sites/info/files/economy-finance/ip065_en.pdf. Accessed: 2017-12-28. (Archived by WebCite® at http://www.webcitation.org/6w377yoQH)

2 Desrivierre D. D'ici 2050, la population augmenterait dans toutes les régions de métropole. Insee Première n° 1652 ; juin 2017. https://www.insee.fr/fr/statistiques/2867738#graphique-figure1A. Accessed: 2017-12-28. (Archived by WebCite® at http://www.webcitation.org/6w38s0ipq)

3 Delbès C, Gaymu J. La Population en EHPAD en France Qui vit en institution ? Gérontologie et société 1/2005 (n° 112), p. 13-24 https://www.cairn.info/revue-gerontologie-et-societe1-2005-1-page-13.htm. Accessed: 2018-01-03.

4 Monaghan P, Charmantier A, Nussey DH, et al. The evolutionary ecology of senescence. Functional Ecology 2008, 22. 371-378 doi:10.111/j.1365-2435.2008.01418.x http://www.webcitation.org/6w3Bji2IC

5 Berge GT. Drivers and barriers to structuring information in Electronic Health Records (2016). PACIS 2016 Proceedings. 18. URL:http://aisel.aisnet.org/pacis2016/18/. Accessed: 2017-12-28. (Archived by WebCite® at http://www.webcitation.org/6w3AQiskr)

6 Dépendance 4 cohortes - Rapport final INSERM Projet Dépendance 4 cohortes épidémiologiques Haute Normandie, Paquid, 3Cités et AMI - Novembre 2011 http://www.cnsa.fr/documentation/_projet_dependance_4_cohortes_cnsa_version_finale_nov2011_.pdf Accessed: 2018-01-13. (Archived by WebCite® at http://www.webcitation.org/6wRmUQsvL)

7 Banks J., Batty G.D., Nazroo J. et al. The dynamics of ageing.: Evidence from the English Longitudinal Study of Ageing 2002-15 (Wave 7) October 2016. http://www.elsa-project.ac.uk/publicationDetails/id/8696 Accessed: 2018-01-13. (Archived by WebCite® at http://www.webcitation.org/6wRmtF4UA)

8 The Survey of Health, Ageing and Retirement in Europe http://www.share-project.org/home0.html Accessed: 2018-01-13.

9 Urban Institute Final Report. A study funded by the Office of the National Coordinator for Health Information Technology of the U.S. Department of Health and Human Services https://www.healthit.gov/sites/default/files/hit_lessons_learned_lit_review_final_08-01-2013.pdf. Accessed: 2017-12-28. (Archived by WebCite® at http://www.webcitation.org/6w3D5k2iW)

10 Tableaux de l'économie française. Edition 2017. Personnes âgées dépendantes. https://www.insee.fr/fr/statistiques/2569388?sommaire=2587886 Accessed: 2018-01-13. (Archived by WebCite® at http://www.webcitation.org/6wRoVTec3)

11 CDC MMWR Recommendations and Reports Updated guidelines for evaluating public health surveillance systems: recommendations from the Guidelines Working Group. MMWR Morb Mortal Wkly Rep. 2001; 50(RR13):1-35 https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5013a1.htm. Accessed: 2017-12-28. (Archived by WebCite® at http://www.webcitation.org/6w3EbepiT)

12 CDC MMWR Supplement Overview of Syndromic Surveillance What is Syndromic Surveillance? September 24, 2004 / 53(Suppl);5-11 https://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a3.htm. Accessed: 2017-12-28. (Archived by WebCite® at http://www.webcitation.org/6w3F1H5NA)

13 CDC MMWR Recommendations and Reports. Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks

https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5305a1.htm. Accessed: 2017-12-28. (Archived by WebCite® at http://www.webcitation.org/6w3FR4fq3)

14 Katz R, May L, Baker J, et al. Redefining syndromic surveillance. Journal of Epidemiology and Global Health (2011) 1, 21– 31 doi:10.1016/j.jegh.2011.06.003 Accessed: 2017-12-28. (Archived by WebCite® at http://www.webcitation.org/6w3GlHmUG) PMID: 23856373

15 Josseran L, Fouillet A. La Surveillance Syndromique : bilan et perspective d'un concept prometteur, Rev Epidemiol Sante Publique (2013), http://dx.doi.org/10.1016/j.respe.2013.01.094 Accessed: 2017-12-28

16 Fouillet A, Medina S, Medeiros H, et al. La Surveillance Syndromique en Europe : le Projet Européen Triple-S BEH 3-4 | 21 janvier 2014, La surveillance syndromique en France en 2014 http://invs.santepubliquefrance.fr//beh/2014/3-4/index.html. Accessed: 2017-12-28. (Archived by WebCite® at http://www.webcitation.org/6w3LByIfI)

17 Soulakis ND. Syndromic Surveillance for Bioterrorism Related Inhalation Anthrax in an Emergency Department Population. Public Health Thesis University of Pittsburgh 2012 https://pdfs.semanticscholar.org/1fe3/3be571d665c70d545183e9aaa05aed7ca674.pdf. Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w54b0a6h)

18 Flamand C, Larrieu S, Couvy E, et al. Validation of a Syndromic Surveillance System using a General Practionner House Calls Network Bordeaux, France EUROSURVEILLANCE Vol . 13 · Issues 4–6 · Apr–Jun 2008 · http://www.eurosurveillance.org/content/10.2807/ese.13.25.18905-en. Accessed: 2017-12-28. (Archived by WebCite® at http://www.webcitation.org/6w3IGGCvG)

19 Josseran L, Caillère N, Brun-Ney D, et al. Syndromic surveillance and heat wave morbidity: a pilot study based on emergency departments in France. BMC Medical Informatics and Decision Making 2009, 9:14 doi:10.1186/1472-6947-9-14 http://www.webcitation.org/6w3IlJgaG PMID: 19232122

20 Van Ganse E, Belhassen M. SNIIRAM: Primary and Secondary Case Resource Use in France. January 23, 2015. https://fr.slideshare.net/RespiratoryEffectivenessGroup/sniiram-primary-and-secondary-care-resource-use-in-france. (Accessed: 2018-01-03).

21 Moulis G, Lapeyre-Mestre M, Palmaro A, et al. Review - French health insurance databases: What interest for medical research? _ Les bases de données de l'assurance maladie française : quel intérêt pour la recherche médicale ? La Revue de médecine interne 36 (2015) 411–417 Accessed: 2017-12-28.(Archived by WebCite® at http://www.webcitation.org/6w3LwK4Pr)

22 the Sentinelles network : https://websenti.u707.jussieu.fr/sentiweb/ (Accessed: 2017-12-28)

23 Caillère N, Fouillet A, Henry V, et al. Le système français de Surveillance sanitaire des urgences et des décès(SurSaUD®). Saint-Maurice: InVS; 2012. 12 p http://invs.santepubliquefrance.fr/Publications-et-outils/Rapports-et-syntheses/Autres-thematiques/2012/Le-systeme-francais-de-Surveillance-sanitaire-des-urgences-et-des-deces-SurSaUD-R (Accessed: 2017-12-28)

24 Sur quel dispositif de surveillance syndromique se base la veille sanitaire en Normandie? Bulletin de veille sanitaire - N° 14 / Février 2015 CIRE/InVS (Accessed: 2017-12-28)

25Smith S, Smith GE, Olowokure B, et al. Early spread of the 2009 infuenza A(H1N1) pandemic in the United Kingdom – use of local syndromic data, May–August 2009. Euro Surveill. 2011;16(3):pii=19771. Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4UrEPc5) PMID: 21262185

26 Josseran L, Fouillet A, Caillère N, et al. Assessment of a Syndromic Surveillance System Based

on Morbidity Data: Results from the Oscour Network during a Heat Wave. PLoS ONE August 2010 Volume 5 Issue 8 e11984 https://doi.org/10.1371/journal.pone.0011984 Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4VynDFZ) PMID: 20711252

27 Ducoudray JM, Eon Y, Le Provost C, et al. Le modèle PATHOS, Guide d'utilisation 2017 rédigé par la CNAMTS (Caisse Nationale d'Assurance Maladie des Travailleurs Salariés) et le SNGC (Syndicat National de Gérontologie Clinique). http://www.cnsa.fr/documentation/modele_pathos_2017.pdf Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4ZRwZMi)

28 Ministère des solidarités et de la santé. Portail national d'information pour l'autonomie des personnes âgées et l'accompagnement de leurs proches. Publié le 17 novembre 2016, mis à jour le 31 mars 2017 http://www.pour-les-personnes-agees.gouv.fr/beneficier-daides/lallocation-personnalisee-dautonomie-apa/comment-le-gir-est-il-determine Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4a5soTN)

29 Closon MC, Habimana L, Laokri S, et al. Le modèle AGGIR PATHOS SOCIOS : un instrument potentiel pour le financement, la programmation et la gestion interne des services de gériatrie et de réadaptation. La Revue de Gériatrie 2006; 31: pp 31-39. http://www.medcomip.fr/region/region-gir-pathos/rapport-novella-cs-2012.pdf Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4b80vvj)

30 Neirynck I, Closon MC, Swine C. Epreuves de validation du modèle AGGIR PATHOS SOCIOS dans les services gériatriques et de réadaptation. La Revue de Gériatrie 2006; 31: pp 13-20. http://www.medcomip.fr/region/region-gir-pathos/rapport-novella-cs-2012.pdf Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4b80vvj)

31 Dain L. Que peut apporter la psychomotricité aux personnes âgées dépendantes. Illustration avec 2 études de cas. Mémoire en vue de l'obtention du Diplôme d'Etat de Psychomotricien Université Paul Sabatier 2011 http://www.psychomot.ups-tlse.fr/Dain2011.pdf. Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4c1bd5J)

32 Hazif-Thomas C, Reber G, Bonvalot T, et al. Syndrome dysexécutif et dépression tardive. Annales Médico Psychologiques 163 (2005) 569–576 doi:10.1016/j.amp.2005.07.005 (Accessed: 2018-01-03).

33 Cohen R, Elhadad M, Elhadad N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies BMC Bioinformatics 2013, 14:10 http://www.biomedcentral.com/1471-2105/14/10 Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4dxZqi9) PMID: 23323800

34 Bui AA, Taira RK, El-Saden S, et al. Automated medical problem list generation: towards a patient timeline. Stud Health Technol Inform. 2004;107(Pt 1):587-91. https://www.ncbi.nlm.nih.gov/pubmed/15360880. Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4eIMIJy) PMID: 15360880

35 Burton MM, Simonaitis L, Schadow G. Medication and Indication Linkage: A Practical Therapy for the Problem List? AMIA 2008 Symposium Proceedings:86-90 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655999/ Accessed: 2017-12-29 PMCID: PMC2655999

36 Campbell J. R. Strategies for Problem List Implementation in a Complex Clinical Enterprise 1998 AMIA, Inc. Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4fqjKLK)

37 Tutorial about the LIKE function (Accessed July 28, 2016) http://www.tutorialspoint.com/sql/sql-like-clause.htm Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4gGA4em)

38 Adam S, Bonsang E, Grotz C, et al. Occupational Activity and Cognitive Reserve Implications in Terms of Prevention of Cognitive Aging and Alzheimer's Disease. Clinical Interventions in Aging 2013:8 377–390 2013 http://dx.doi.org/10.2147/CIA.S39921 Accessed: 2018-01-03. (Archived by WebCite® at http://www.webcitation.org/6wCA9ZXyu) PMID: 23671387

39 Lebert F, Leroy M, Pasquier F, et al. Problématique des malades Alzheimer « jeunes » en UCC : enquête nationale en France. Geriatr Psychol Neuropsychiatr Vieil 2016 ; 14 (2) : 194-200 doi:10.1684/pnv.2016.0607 Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4h2EdgC)

40 Abadie R, Voisin T. Procédure de dépistage et de prise en charge des Symptômes Psycho Comportementaux de la Démence en EHPAD. February 6, 2014. http://www.medcomip.fr/region/region-outils/outils-egs/troubles-cpmt/procedure-spcd-ehpad.pdf Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4hHAHmN)

41 WHO Mental health and older adults Fact sheet Updated April 2016 http://www.who.int/mediacentre/factsheets/fs381/en/ Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4lVpnZM)

42 de Villiers L. Frailty. CME Continuing Medical Education Vol 31, No 10 (2013) http://www.cmej.org.za/index.php/cmej/article/view/2868/3235 Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4mPVDFL)

43 OMS Les chute Aide-mémoire N°344 Septembre 2016 http://www.who.int/mediacentre/factsheets/fs344/fr/ Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4mjnzJ2)

44 Rubenstein LZ, Josephson KR, Robbins AS. Falls in the Nursing Home Ann Intern Med. 1994;121:442-451. Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4n8HAwK) PMID: 8053619

45 Arai H, Ouchi Y, Yolode M, et al. Toward the Realization of a Better Aged Society Geriatrics and gerontology international 2012 Jan;12(1):16-22. doi: 10.1111/j.1447-0594.2011.00776.x Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w4nP0efV) PMID: 22188494

46 Zaslavsky O, Thompson H, Demiris G. The Role of Emerging Information Technologies in Frailty Assessment. Research in gerontological nursing 2012 Jul;5(3):216-28. doi: 10.3928/19404921-20120410-02. Epub 2012 Apr 25. https://www.ncbi.nlm.nih.gov/pubmed/22533942. Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w54pw9CY) PMID: 22533942

47 Repérage et maintien de l'autonomie des personnes âgées fragiles. Livre blanc. International Association of Gerontology and Geriatrics - Société Française de Gériatrie et de Gérontologie www.fragilite.org/livreblanc Accessed: 2017-12-29.

48 Fried LP, Tangen CM, Walston J, et al. Frailty in Older Adults: Evidence for a Phenotype. J Gerontol A Biol Sci Med Sci. 2001 Mar;56(3):M146-56 https://www.ncbi.nlm.nih.gov/pubmed/11253156. Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w55Uov37) PMID: 11253156

49 Meystre S, Haug PJ. Automation of a Problem List Using Natural Language Processing. BMC Medical Informatics and Decision Making 2005, 5:30 DOI: 10.1186/1472-6947-5-30 Accessed: 2018-01-01. PMID:16135244

50 Lindsay K. Managing Adult Respiratory Syncytial Virus. Physician's Weekly. Sep 2, 2016. Accessed: 2018-04-04 https://www.physiciansweekly.com/managing-adult-respiratory-syncytial-virus/

51 Respiratory Syncytial Virus Infection (RSV) Accessed: 2018-04-04 https://www.cdc.gov/rsv/index.html

52 Chapman WW. Natural Language Processing for Biosurveillance Handbook of Biosurveillance ISBN 0-12-369378-0 Elsevier Inc https://link.springer.com/chapter/10.1007%2F978-1-4419-6892-0_13. Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w57XO6CB)

53 Fieschi M, Bouhaddou O, Beuscat R, et al. L'informatique au service du patient. Comptes rendus des huitièmes Journées Francophones d'Informatique Médicale, Marseille 30 et 31 mai 2000, Informatique et Santé, Collection dirigée par P.Degoulet et M.Fieschi http://www.springer.com/us/book/9782287596919. Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w57nEk6H)

54 Hanauer DA, Zheng K, Mei Q, et al. FULL-TEXT SEARCH IN ELECTRONIC HEALTH RECORDS: CHALLENGES AND OPPORTUNITIES In: Internet Policies and Issues, Volume 7 ISBN: 978-1-61668-745-8 Editor: B.G. Kutais © 2009 Nova Science Publishers, Inc. https://www.novapublishers.com/catalog/product_info.php?products_id=13932. Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w58WOgCp)

55 Dick B. Building agreement from disagreement: the anatomy of dialectical processes. Chapel Hill, Qld: Interchange. International Congress of Action Research and Process Management, Griffith University, Brisbane, 10 to 13 July, 1990 http://www.aral.com.au/DLitt/DLitt_P24delphi.pdf. Accessed: 2017-12-30.

56 Debin M, Souty C, Turbelin C, et al. Determination of French Influenza Outbreaks Periods Between 1985 and 2011 through a web-based Delphi method, BMC Medical Informatics and Decision Making 2013, 13:138 Accessed: 2017-12-30. http://www.webcitation.org/query?url=https%3A%2F%2Fwww.ncbi.nlm.nih.gov%2Fpubmed%2F14680664&date=2017-12-30 PMID: 24364926

57 Graham B, Regehr G, Wright JG. Delphi as a method to establish consensus for diagnostic criteria. Journal of Clinical Epidemiology 2003 http://dx.doi.org/10.1016/S0895-4356(03)00211-7 https://www.ncbi.nlm.nih.gov/pubmed/14680664. Accessed: 2017-12-30. PMID: 14680664

58 Chapman WW, Dowling JN, Baer A, et al. Developing Syndrome Definitions Based on Consensus and Current Use. J Am Med Inform Assoc 2010;17:595e601. doi:10.1136/jamia.2010.003210 Accessed: 2018-01-01 PMID: 20819870

59 Liu F, Chen J, Jagannatha A, et al. Learning for Biomedical Information Extraction Methodological Review of Recent Advances http://www.webcitation.org/6w9Bmjq3t Accessed: 2018-01-01

60 Tellier I. Introduction au TALN et Ingénierie Linguistique, Université de Lille3 http://www.webcitation.org/6w9CUH0tQ Accessed: 2018-01-01

61 Zhai C, Massung S. Text Data Management and Analysis. A Practical Introduction to Information Retrieval and Text Mining ACM Books #12 Association for Computing Machinery and Morgan & Claypool Publishers http://www.webcitation.org/6w9Csma5Q Accessed: 2018-01-01 doi: 10.1145/2915031 ISBN: 978-1-97000-118-1

62 Chretien JP, Tomich NE, Gaydos JC, et al. Real-Time Public Health Surveillance for Emergency Preparedness. Commentary. American Journal of Public Health | August 2009, Vol 99, No. 8 doi: 10.2105/AJPH.2008.133926 Accessed 2018 04-04 PMID: 19542047

63 Andersson MG, Faverjon C, Vial F, et al. Using Bayes' Rule to Define the Value of Evidence from Syndromic Surveillance. PLOS ONEwww.plosone.org November 2014 | Volume 9 | Issue 11 |

e111335 https://doi.org/10.1371/journal.pone.0111335

64 http://community.pentaho.com/projects/data-integration/ Accessed: 2017-12-29

65 Gibbons C, Richards S, Valderas JM. Supervised ML Algorithm can classify Open-Text Feedback of Doctor Performance with Human-Level Accuracy J Med Internet Res 2017;19(3):e65 doi:10.2196/jmir.65332017 https://www.ncbi.nlm.nih.gov/pubmed/28298265. Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w56Q5PCk) PMID: 28298265

66 Höhle, M. An R package for the Monitoring of Infectious Diseases Computational Statistics (2007) 22: 571. doi:10.1007/s00180-007-0074-8 https://link.springer.com/article/10.1007/s00180-007-0074-8. Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w56gEos9)

67 Surveillance package https://cran.r-project.org/web/packages/surveillance/surveillance.pdf Accessed: 2018-01-01

68 Flahault A, Blanchon T, Dorléans Y, Toubiana L, Vibert JF, Valleron AJ. Virtual surveillance of communicable diseases: a 20-year experience in France. Stat Methods Med Res. 2006. 15(5):413-21 https://www.ncbi.nlm.nih.gov/pubmed/17089946. Accessed: 2017-12-29. (Archived by WebCite® at http://www.webcitation.org/6w56xOvba) PMID: 17089946

69 Pelat C, Bonmarin I, Ruello M, Fouillet A, Caserio-Schönemann C, Levy-Bruhl D, Le Strat Y, the Regional Influenza study group. Improving regional influenza surveillance through a combination of automated outbreak detection methods: the 2015/16 season in France. Euro Surveill. 2017;22(32):pii=30593. DOI: http:// dx.doi.org/10.2807/1560-7917.ES.2017.22.32.30593 Accessed: 2018-05-14 https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2017.22.32.30593

70 La veille et l'alerte sanitaires en France. Rapport rédigé en 2011 par le groupe de travail conjoint InVs et Ministère du travail, de l'emploi et de la santé. URL:http://solidarites-sante.gouv.fr/IMG/pdf/Rapport_veille_alerte_sanitaire_France.pdf. Accessed: 2018-05-14. (Archived by WebCite® at http://www.webcitation.org/6zPFnbyLa)

71 Serfling RE. Methods for current statistical analysis of excess pneumonia-influenza deaths. Public health reports. 1963;78(6):494-506 Accessed: 2018-01-01 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1915276/pdf/pubhealthreporig00078-0040.pdf

72 Costagliola D, Flahault A, Galinec D, et al. A routine tool for detection and assessment of epidemics of influenza-like syndromes in France. Am J Public Health. 1991;81(1):97-9. https://www.ncbi.nlm.nih.gov/pubmed/1983924. Accessed: 2018-01-01. (Archived by WebCite® at http://www.webcitation.org/6w9EBtO6a) PMID: 1983924

73 Höhle M, Riebler A, Paul M. Getting Started With OutBreak Detection https://cran.r-project.org/web/packages/surveillance/vignettes/surveillance.pdf Accessed: 2018-01-01 http://www.webcitation.org/6w9Em7JTU

74 Cowling BJ, Wong IOL, Ho L, et al. Methods for Monitoring Influenza Surveillance Data International Journal of Epidemiology 2006;35:1314–1321 doi:10.1093/ije/dyl162 Accessed: 2018-01-01 PMID: 16926216

75 CDC Resource Center https://www.cdc.gov/surveillancepractice/ Accessed: 2018-01-01

76 Yang P, Duang W, LvM, et al. Review of an Influenza Surveillance System Beijing People's Republic of China. Emerging Infectious Diseases • www.cdc.gov/eid • Vol. 15, No. 10, October 2009 DOI: 10.3201/eid1510.081040 https://wwwnc.cdc.gov/eid/article/15/10/08-1040_article. Accessed: 2018-01-01. (Archived by WebCite® at http://www.webcitation.org/6w9I0gZTN) PMID: 19861053

77 Jung N. Surveillance sanitaire à partir de données des services d'urgence : modélisation de séries temporelles et analyse automatique. Méthodologie [stat.ME]. 2010. https://dumas.ccsd.cnrs.fr/dumas-00516268 Accessed: 2018-01-01

78 Tokars JI, Burkom H, Xing J, et al. Enhancing Time Series Detection Algorithms for Automated Biosurveillance. Emerging Infectious Diseases • www.cdc.gov/eid • Vol. 15, No. 4, April 2009 DOI: 10.3201/eid1504.080616 https://wwwnc.cdc.gov/eid/article/15/4/08-0616_article. Accessed: 2018-01-01. (Archived by WebCite® at http://www.webcitation.org/6w9DdO9iV) PMID: 19331728

79 ggplot2 package https://cran.r-project.org/web/packages/ggplot2/index.html Accessed: 2018-01-01

80 Cleveland WS, Grosse E, Shyu WM. Local regression models. Chapter 8 of Statistical Models in S. eds Chambers JM and Hastie TJ, Wadsworth & Brooks/Cole1992 http://www.webcitation.org/6w9E07zjA Accessed: 2018-01-01

81 Cameron W, Neu AL, Murray EL, Soetebier K, Cookson ST. Responding to Syndromic Surveillance Alerts: An Adaptable Protocol for Georgia Health Districts. Advances in Disease Surveillance 2007;2:95 Accessed 2018-04-04 (Archived by WebCite® at http://www.webcitation.org/6yQttWi9M)

82 Bulletin épidémiologique grippe, semaine 9. Saison 2016-2017, Santé Publique France, 8 mars 2017 http://www.webcitation.org/6w9W2JEgY Accessed: 2018-01-01

83 Lee DC, Long JA, Wall SP, et al. Determining Chronic Disease Prevalence in Local Populations Using Emergency Department Surveillance American Journal of Public Health September 2015, Vol 105, No. 9 Research and Practice http://ajph.aphapublications.org/doi/pdf/10.2105/AJPH.2015.302679 Accessed: 2018-01-01 PMID: 26180983

84 Gault G, Larrieu S, Durand C, et al. Performance of a Syndromic System for Influenza Based on the Activity of General Practionners. Journal of Public Health pp. 1–7 doi:10.1093/pubmed/fdp020 Accessed: 2018-01-01 PMID: 19269992

85 Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation Journal of Biomedical Informatics 39 (2006) 589–599 doi:10.1016/j.jbi.2005.11.004 http://www.webcitation.org/6w9S4jADc Accessed: 2018-01-01 PMID: 16359928

86 Septfons A, Barataud D, Chiron E, et al. Surveillance des GEA en collectivités pour personnes âgées, Bilan national de cinq saisons de surveillance hivernale (Novembre 2010-Mai 2015) BEH 18-19 21 juin 2016 http://www.webcitation.org/6w9SI7s0v Accessed: 2018-01-01

87 Respiratory Syncytial Virus in Older Adults: A Hidden Annual Epidemic A Report by the National Foundation for Infectious Diseases Accessed: 2018-04-06 http://www.nfid.org/publications/reports/rsv-report.pdf

88 Falsey AR, Hennessey PA, Formica MA et al. Respiratory Syncytial Virus Infection in Elderly and High-Risk Adults. New England Journal of Medecine 352;17 http://www.nejm.org/doi/full/10.1056/NEJMoa043951 April 28, 2005

89 Falsey AR, Walsh EE. Respiratory Syncytial Virus Infection in Adults. Clinical Microbiolo- gy Reviews, 0893-8512/00/$04.0010 July 2000, p. 371–384 Accessed 2018-04-06 PMID: 10885982

90 Surveillance de la Grippe en France Métropolitaine Saison 2013-2014 Bulletin Epidémiologique

Hebdomadaire n°28 2014 http://invs.santepubliquefrance.fr/Publications-et-outils/BEH-Bulletin-epidemiologique-hebdomadaire/Archives/2014/BEH-n-28-2014 http://www.webcitation.org/6w9SXCYlh Accessed: 2018-01-01

91 Surveillance de la Grippe en France Métropolitaine Saison 2015-2016, Bulletin Epidémiologique Hebdomadaire n°32_33 2016 http://invs.santepubliquefrance.fr/Publications-et-outils/BEH-Bulletin-epidemiologique-hebdomadaire/Archives/2016/BEH-n-32-33-2016 http://www.webcitation.org/6w9T4F6WY Accessed: 2018-01-01

92 http://social-sante.gouv.fr/IMG/pdf/Plan_Pandemie_Grippale_2011.pdf http://www.webcitation.org/6w9WFMtRW Accessed: 2018-01-01

93 http://www.iledefrance.paps.sante.fr/Episode-infectieux-dans-une-collectivite-de-personnes-agees-IRA.37364.0.html http://www.webcitation.org/6w9WWZopk Accessed: 2018-01-01 http://www.iledefrance.paps.sante.fr/fileadmin/ILE-DE-FRANCE/PAPS/Informations_pratiques/MDO/PA_en_IRA/FICHE_DE_SIGNALEMENT__IRA.pdf http://www.webcitation.org/6w9WmJOay Accessed: 2018-01-01

94 http://www.cpias-ile-de-france.fr/ACTU_DIVERS/Synerpa_Guide_grippe2016.pdf. Accessed: 2018-01-01. (Archived by WebCite® at http://www.webcitation.org/6w9XLF0Lk)

95 https://www.pasteur.fr/fr/centre-medical/fiches-maladies/grippe Accessed: 2018-01-01. (Archived by WebCite® at http://www.webcitation.org/6w9XmBiHP)

96 http://france3-regions.francetvinfo.fr/centre-val-de-loire/epidemie-grippe-ars-renouvelle-ses-recommandations-prevention-1170701.html Accessed: 2018-01-01. (Archived by WebCite® at http://www.webcitation.org/6w9Y4tLVr)

97 Baroukh T. Développement d'un système de détection d'événements inhabituels dans la surveillance des salmonelles d'origine non humaine par des méthodes statistiques d'analyse des séries temporelles. Rapport de stage M2 Université de Paris Sud Septembre 2008 doi : 10.13140/RG.2.2.28143.33444 Accessed: 2018-01-02

98 Danan C, Baroukh T, Moury F, et al. Automated early warning system for the surveillance of Salmonella isolated in the agro-food chain in France. *Epidemiology and Infection 2011 May, 139*(5), 736-741. Accessed: 2018-01-02 PMID: 20598207

99 Delespierre T, Denormandie P, Bar-Hen A, et al. Empirical advances with text mining of electronic health records. BMC Medical Informatics and Decision Making (2017) 17:127 DOI 10.1186/s12911-017-0519-0 Accessed: 2018-01-02. (Archived by WebCite® at http://www.webcitation.org/6wAVPe1xb) PMID: 28830417

100 Simon N. Why Does Clinical Health Data Exchange Remain Such A Struggle? Accessed: 2018-04-13-2018 (Archived by WebCite® at http://www.webcitation.org/6ye6oD6KT)

101 Kissling E, Rondy M. Les premiers résultats de l'enquête I-MOVE d'Eurosurveillance Early 2016/17 vaccine effectiveness estimates against influenza a(h3n2): I-MOVE multicentre case control studies at primary care and hospital levels in Europe, I-MOVE/I-MOVE+ study team, Euro Surveill. 2017 Feb 16; 22(7): 30464. doi: 10.2807/1560-7917.ES.2017.22.7.30464 Accessed: 2018-01-02

102 Dorrington MG, Bowdish DM. Immunosenescence and novel vaccination strategies for the eldely. Frontiers in IMMUNOLOGY June 2013|Volume 4|Article 171. doi: 10.3389/fimmu.2013.00171 https://www.frontiersin.org/articles/10.3389/fimmu.2013.00171/full. Accessed: 2018-01-02. (Archived by WebCite® at http://www.webcitation.org/6wAXf4GCy) PMID: 23825474

103 Delespierre T, Denormandie P, Josseran L. When and why do people fall in nursing homes? JNHR the Journal of Nursing Home Research Vol 2 2016 Oral Communication 6-10 November, Barcelona Spain. http://www.jnursinghomeresearch.com/all-issues.html?a=2016&n=01. Accessed: 2018-01-02. (Archived by WebCite® at http://www.webcitation.org/6wAYRitWM)

104 Delespierre T, Denormandie P, Armaingaud D, et al. De nouvelles données et de nouvelles méthodes pour évaluer les soins primaires de la population cancéreuse en EHPAD poster EPICLIN 10 2016 à Strasbourg https://www.researchgate.net/publication/302482190_De_nouvelles_donnees_et_de_nouvelles_methodes_pour_evaluer_les_soins_primaires_de_la_population_cancereuse_en_Ehpad. Accessed: 2018-01-02. (Archived by WebCite® at http://www.webcitation.org/6wAZURk3l)

Abbreviations

AGE: acute gastro enteritis
ARI: acute respiratory infection
BBV: base du bien vieillir =. aging well data base
CDC: centers for disease control and prevention
CUSUM: cumulative sums
EARS: early aberration reporting system
EHR: electronic health record
ETL: extract transform load
GIR: groupe ISO ressources ISO resource group
GP: general practitioners
HRA: Health Regional Agencies
ILI: influenza like illness
IS: information system
PERMF: personal electronic resident medical file
PH: public health
NH: nursing homes
SQL: standard query language
SS: surveillance system
SSS: syndromic surveillance system
TS: time series

3-4-3 Perspectives liées au développement de ce nouveau SSS

Les travaux dans ce deuxième article liés à la surveillance de la grippe et la gastro-entérite aiguë ont montré la pertinence de développer un nouveau système d'information entièrement dédié à la description des soins au quotidien des résidents en EHPAD et, au travers de ce nouvel outil, d'avoir des données au fil de l'eau et en continu sur leur état de santé. De plus, hospitalisations et décès y jouent le rôle d'indicateurs synthétiques de suivi et de contrôle de notre cohorte en cas d'alerte sanitaire, comme lors de la dernière saison de grippe (hiver 2016-2017) où nous avons détecté une augmentation brutale du nombre d'hospitalisations et une sur-mortalité dans certaines régions où la moyenne d'âge était plus élevée que dans le reste des établissements du groupe.

3-5 La construction d'un nouvel outil de santé publique

3- 5-1Travaux sur le cancer

3-5-1-1 Résumé du poster

L'objectif était de montrer que l'on pouvait retracer des parcours de soins dans leur intégralité et ici sur une population de résidents atteints de cancer.

Méthodes : La table transmissions contient l'essentiel de l'information textuelle alimentée séquentiellement. Grâce aux index résident et établissement et aux dates de transmission, on peut a posteriori, par requêtes SQL, suivre un résident au cours du temps lorsque les données ont été saisies par le personnel médical.

Un algorithme s'appuyant sur ces requêtes a permis de bâtir 21 syndromes décrivant l'essentiel des pathologies et problèmes touchant les résidents. Son principe s'appuie sur le deep learning [56] : un apprentissage supervisé en plusieurs couches intelligentes de traitement textuel destiné à identifier des syndromes prédéterminés. Une première couche exclut les mots vides de sens et simplifie les expressions, une seconde aiguille vers l'un des 21 syndromes par analyse syntaxique et textmining, enfin la troisième intègre expertise médicale et textmining

pour raffiner la définition des syndromes dont le cancer. Un processus itératif d'apprentissage des couches 2 et 3 en collaboration avec les médecins coordonnateurs Korian a permis d'améliorer la définition des syndromes et la logique du système.

Nous avons extrait toutes les transmissions du syndrome cancer sur la période d'étude [01/10/2010 - 01/04/2016]. Ces transmissions *cancer* ont permis d'extraire un échantillon de résidents que nous avons croisé avec l'ensemble des transmissions syndromiques. Les délais [date d'entrée - date de première transmission 'cancer'] et les durées de séjour ont été calculés et distribués en tertiles a priori représentatifs de la gravité des symptômes et de leur état de santé. L'idée de départ était de différencier les cancers notifiés dès l'entrée souvent avec prise en charge immédiate, des cancers notifiés au fil de l'eau, mais aussi peut-être de mettre en exergue une différence de prise en charge. Enfin, il a été tenté de rapprocher ces transmissions avec les antécédents et pathologies des résidents.

Résultats : Au 17 janvier 2016, 39 561 résidents étaient présents dans 125 EHPAD. Notre analyse a permis d'identifier 4 400 résidents qui avaient au moins une information cancer transmise pendant la période d'étude. Le découpage de la population en trois tertiles de durée de séjour, a conduit aux effectifs suivants 1 468, 1 465 et 1 467 résidents (table 4). Les durées moyennes de séjour des 3 tertiles étaient respectivement de 52, 434 et 1 822 jours. A noter que pour 48.3% des résidents du premier tertile durée de séjour, il n'y avait ni mention d'une pathologie, ni mention d'un antécédent cancer dans la table des antécédents et pathologies.

Pour illustrer la grande hétérogénéité des profils des résidents avec cancer et des informations saisies, nous avons choisi de présenter les parcours de soins de 2 résidentes, appartenant aux premier et troisième tertiles de délais de déclaration de cancer avec la date de leur entrée en EHPAD, d'abord sous la forme de deux extraits textuels bruts (figures 9 et 10), véritables condensés de leurs vécus, ensuite sous la forme de deux nuages de mots (figure 11), cette fois avec l'ensemble des syndromes analysés textuellement.

Discussion – Conclusion : Notre technique d'extraction d'informations textuelles relatives à la prise en charge des résidents permet de retracer les parcours de santé et de vie et d'obtenir une image fine des problématiques de soins des résidents lorsque ces informations sont alimentées. Une comparaison des effectifs des syndromes (table 4), triés par ordre décroissant et par tertiles

de durées de séjour, montre que leur fréquence augmente avec la durée de séjour. Néanmoins, douleurs, dépression et idées noires sont prégnantes pour tous les résidents avec cancer. Les problèmes de comportement et l'altération de l'état général vont croissants avec la durée du séjour et le vieillissement concomitant de cette population.

Tenter de définir une typologie des résidents en reliant les informations syndromiques avec les autres tables du DWH s'avère difficile car le remplissage des tables est souvent lacunaire. Cependant, au vu de la richesse des informations extraites, et à la condition de contrôler quantité et qualité des données (volume, contenu et périodicité) une typologie définie exclusivement à partir des transmissions semble possible et judicieuse.

3-5-1-2 Quelques résultats

TERTILE 1 (1468 résidents) [0 - 150 jours]			*TERTILE 2 (1465 résidents) [151 - 811 jours]*			*TERTILE 3 (1467 résidents) [812 jours et +]*		
SYNDROME	NB_RESIDENTS	RATIO	SYNDROME	NB_RESIDENTS	RATIO	SYNDROME	NB_RESIDENTS	RATIO
DOULEURS	1177	0,80	DOULEURS	1447	0,99	COMPORTEMENT	1460	1,00
DEPRESSION_IDEES_NOIRES	1174	0,80	COMPORTEMENT	1445	0,99	DOULEURS	1459	0,99
COMPORTEMENT	1141	0,78	DEPRESSION_IDEES_NOIRES	1429	0,98	DEPRESSION_IDEES_NOIRES	1458	0,99
ETAT_CUTANE	1082	0,74	ETAT_CUTANE	1419	0,97	ETAT_CUTANE	1452	0,99
AEG	959	0,65	AEG	1372	0,94	AEG	1433	0,98
DENUTRITION_DEGLUTITION	834	0,57	CHUTES	1290	0,88	CHUTES	1377	0,94
CHUTES	825	0,56	TRANSIT	1225	0,84	TRANSIT	1372	0,94
TRANSIT	727	0,50	DENUTRITION_DEGLUTITION	1198	0,82	DENUTRITION_DEGLUTITION	1310	0,89
DECES	698	0,48	CARDIO_VASCULAIRE	1065	0,73	DEMENCE	1164	0,79
CARDIO_VASCULAIRE	596	0,41	DEMENCE	1027	0,70	CARDIO_VASCULAIRE	1160	0,79
DEMENCE	590	0,40	***HOSPITALISATION***	***982***	**0,67**	***HOSPITALISATION***	***1133***	**0,77**
HOSPITALISATION	***537***	0,37	TRB_URINAIRES	865	0,59	TRB_URINAIRES	1100	0,75
TRB_URINAIRES	463	0,32	DECES	758	0,52	VISION	889	0,61
DESHYDRATATION	261	0,18	VISION	636	0,43	BUCCO_DENTAIRE	857	0,58
VISION	227	0,15	BUCCO_DENTAIRE	614	0,42	GRIPPE_IRA	691	0,47
SOMMEIL	198	0,13	DESHYDRATATION	519	0,35	DECES	686	0,47
GRIPPE_IRA	185	0,13	GRIPPE_IRA	513	0,35	VACCINATION	680	0,46
BUCCO_DENTAIRE	144	0,10	SOMMEIL	459	0,31	DESHYDRATATION	676	0,46
ALLERGIES	122	0,08	VACCINATION	418	0,29	SOMMEIL	646	0,44
AUDITION	107	0,07	GEA_DIARRHEES	276	0,19	GEA_DIARRHEES	455	0,31
GEA_DIARRHEES	97	0,07	ALLERGIES	266	0,18	ALLERGIES	373	0,25
VACCINATION	74	0,05	AUDITION	253	0,17	AUDITION	311	0,21
SOINS PALLIATIFS OU MORPHINE	***352***	**0,24**		***396***	**0,27**		***379***	**0,26**
DUREE MOYENNE SEJOUR	**52 jours**			**434 jours**			**1822 jours**	

Table 4 : Effectifs syndromiques des résidents avec au moins une transmission cancer découpés en tertiles durée de séjour.

Parcours de soins de la résidente A (3ème tertile du délai avant mention d'un cancer): 165 observations, tronqué à gauche le 01/11/2010

Durée de séjour: 1437 jours, soit 3 années et 11 mois. Décès probable des suites d'une mauvaise chute.

Date d'entrée le 26/12/2009 – pas d'observation entre cette date et le 07/11/2010. Pas d'info cancer à cette date.

Refus de douche le 07/11/2010.

Angoissée les 02-03 et 13/12/2010.

Anxiété, fatigue les 27-25/01/2011, désorientation et refus de douche les 11-13/02/2011.

Plaie suintante à la jambe le 04/03/2011.

Problèmes psys avec sa fille le 06/04/2011.

Pleure, dépressive, angoisse, agressivité.

Plaies au niveau du sein qui ne cicatrise pas le 27/07/2011

Mycose sous le sein ? le 26/01/2012

Problèmes d'agressivité, attitude de refus….

Œdèmes des membres inférieurs le 10/01/2013

Bilan cardio le 11/01/2013.

Ablation de carcinomes sous l'œil droit et le sein le 01/08/2013.

Suspicion de phlébite le 06/08/2013.

Se plaint, mange peu, jambes très enflées le 01/09/2013.

Œdèmes et très fatiguée le 26/10/2013.

Chute le 22/10/2013, fracture du tibia détectée le 07/11/2013 suite à douleurs intenses.

Hospitalisation le 08/11/2013.

Retour d'hospitalisation le 14/11/2013.

Pleure, agitée, mange très peu.

En régression +++ et altération de l'état général le 28/11/2013.

Arrache sa perf et ne veut plus manger, AEG le 30/11/2013.

Décès le 02/12/2013.

Figure 9 : Condensé du parcours de soins d'une résidente A atteinte d'un cancer, entrée 9 mois avant l'implantation du DWH (troncature à gauche) et décédée le 02/12/2013.

Parcours de soins de la résidente B (1er tertile du délai avant mention d'un cancer): 226 observations, tronqué à droite le 01/04/2016

Durée de séjour au 01/04/2016 : 1977 jours, soit 5 années et 4 mois.

Date d'entrée le 02/11/2010 – pas de date de fin, les observations du parcours de soins sont tronquées à droite au 01/04/2016.

Cancer du sein prévalent, déclaré à la date d'entrée, transfert depuis une clinique.

Dépression, angoissée, problèmes de constipation, douleurs : PEC avec de la morphine

Hospitalisation le 18/07/2011

Retour d'hospitalisation le 20/07/2011

Traitement chimio, problèmes d'alimentation, douleurs : compléments alimentaires et PEC de la douleur avec ACTISKENAN, problèmes dermatologiques.

Problèmes d'asthénie et de constipation : PEC avec NORMACOL

Hospitalisation le 22/05/2012, IRM le 05/07/2012

Problèmes d'asthénie et de douleurs PEC avec ACTISKENAN

Première chute le 22/10/2013 suite à malaise vagal

Chute le 22/11/2013 et hospitalisation le même jour.

Problèmes d'asthénie et dermatologiques

Vacances du 01/08/2014 au 08/08/2014

Atelier psychomoteur et atelier équilibre et prévention des chutes le 29/09/2014

Atelier escrime le 24/10/2014

Problèmes de douleurs, de perte de poids, de transit les 24 et 25/09/2015.

De plus en plus de chutes.

Atelier psychomoteur et atelier équilibre et prévention des chutes le 04/01/2016 et escrime le 15/01/2016

Annonce du décès de Mme H le 01/04/2016.

Figure 10 : Condensé du parcours de soins d'une résidente B entrée à la date d'implantation du DWH et toujours vivante au 01/04/2016 (troncature à droite).

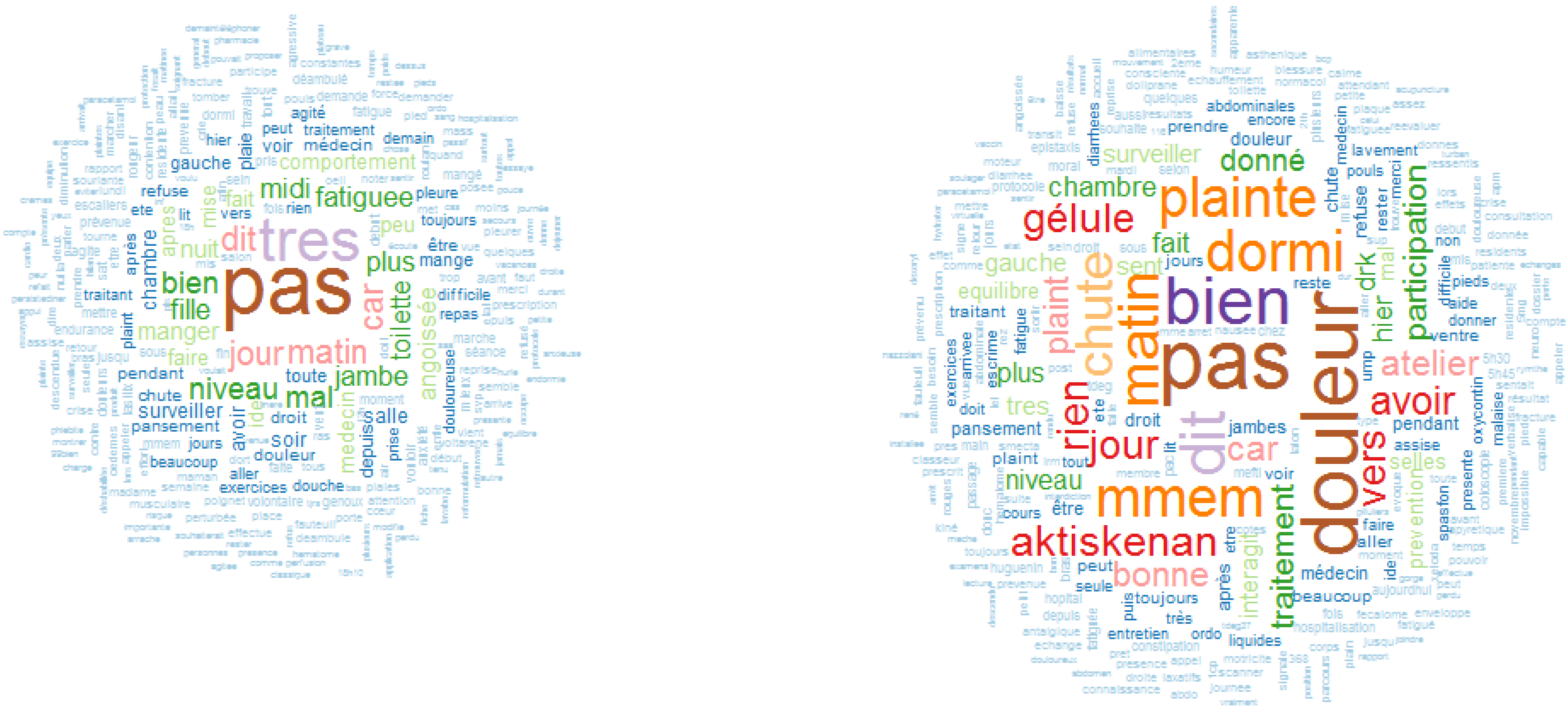


Figure 11 : 2 nuages de mots construits par analyse textuelle des parcours de soins complets des résidentes A et B.

3-5-2 Travaux sur les chutes

3-5-2-1 Résumé de la présentation

Les chutes sont des événements fréquents dans la vie des résidents et relativement bien alimentés dans les EHPAD Korian. De plus, elles constituent une préoccupation majeure de santé publique en France [57 - 58] et presque partout dans le monde [59 - 60].

L'objectif était donc cette fois de montrer que l'on pouvait retracer les chutes des résidents dans leur intégralité en analysant l'information textuelle contenue dans les transmissions.

Méthodes : Grâce aux index résidents et établissements de la table des transmissions, nous avons pu suivre chaque résident au cours du temps par requêtes et text mining et ainsi étudier ses chutes. Nous avons choisi de collecter toutes les données syndromiques provenant de 125 EHPAD pendant la période se déroulant du 01/11/2010, date d'initiation du DWH, au 01/05/2016, semaine courante. Tous les résidents avaient leurs 21 syndromes alimentés, chutes incluses, plus leurs hospitalisations et décès, leur sexe, l'âge à l'entrée et à la première chute et finalement le mois de la première chute. Nous avons choisi de découper la distribution du nombre annuel de chutes en quatre quartiles (Q1 versus Q4 dans la table 5), le but étant de comparer les 2 quartiles extrêmes à la population générale d'un point de vue syndromique, à l'aide du test du Chi2 et ainsi de déterminer des facteurs de risque de chutes, enfin de modéliser le fait de chuter suivant différents critères.

Résultats : Nous avons trouvé 41 717 résidents, 69.7% d'entre-eux tombés au moins une fois et 18 181 tombés durant une période de 12 mois. Les résidents sont tombés en moyenne une année après leur entrée en établissement. Le nombre de chutes en une année s'étendait d'une à cinquante-quatre fois, avec respectivement 6 771 résidents tombés une fois (Q1), 3 801 résidents tombés deux fois (Q2), 3 844 résidents tombés trois ou quatre fois (Q3) et 3 765 résidents tombés au moins cinq fois (Q4). Les traits les plus frappants caractérisant les chutes en comparant Q1 et Q4 étaient la démence (40.7% vs 71.3%), les hospitalisations (36.8% vs 64%), les problèmes cardio-vasculaires (39.9% vs 63%), la malnutrition (55.1% vs 80.2%) et une altération de l'état général (64.3% vs 92.2%), avec des p-values < 2.2x10-16. Nous avons aussi trouvé que les résidents tombaient plus au printemps (1 824 et 1 788 chutes durant les

mois de mars et avril) et moins en septembre (1 307 chutes durant ce mois) et que les personnes du quatrième quartile avaient une durée de séjour globalement plus longue et vivaient plus longtemps après leur dernière hospitalisation que celles du premier quartile (9 vs 10 mois et 30 vs 13 jours).

Discussion - Conclusion: Alors que le nombre de chutes saisies dans le système augmente avec le temps, nous avons restreint notre étude aux chutes intervenues sur une durée d'une année et avons trouvé que les conséquences des chutes uniques semblaient en moyenne plus graves que celles des chutes répétées. Il existe une grande variabilité des chutes, mais en agrégeant hospitalisations, décès et circonstances des chutes, nous serons en mesure, à l'aide du textmining, de les qualifier à la fois qualitativement et quantitativement, offrant une image vivante et précise des chutes des résidents. Pour cela il suffirait d'une part, de croiser les syndromes chutes avec les syndromes décès et hospitalisations, d'autre part, d'analyser textuellement les syndromes chutes. Par exemple retrouver les mots les fréquents décrivant les circonstances de la chute ou encore retrouver les effectifs syndromiques des résidents chutés le jour de leur chute.

3-5-2-2 Quelques résultats

La durée moyenne du journal des transmissions par résident (intervalle de temps s'écoulant entre la première et la dernière transmission) est de 24 mois alors que la médiane n'est que de 10 mois. Devant ce déséquilibre nous avons filtré les transmissions suivant ce critère en ne sélectionnant que les résidents avec un parcours compris entre 10 mois et 38 mois de transmissions. La durée moyenne de séjour passe alors à 21.7 mois et sa médiane à 21 mois. On trouve alors11 359 résidents, 3 507 (30.9%) hommes and 7 852 femmes (69.1%), dont 5 260 décédés (46.3%) d'âge médian à l'entrée 87 ans.

Les résidents forment ainsi un groupe plus homogène, avec plus de décédés (+5.9%), un peu plus âgé (+ 1 an) et un peu plus féminin (+1.2%) que la population non filtrée.

Saisonnalité des chutes et associations entre chutes et syndromes: Les résidents tombent en moyenne 6.35 fois durant leur séjour. Le mois de la première chute est corrélé au fait d'être entré en EHPAD le trimestre précédent. Les 20 syndromes sont associés positivement avec le fait de chuter (voir figure 12).

Les résidents entrés en octobre sont ceux qui tombent le plus en moyenne durant leur séjour (6.9 fois). Ils sont affectés principalement par des problèmes de comportement (agressivité, attitude de refus). Les résidents entrés en février sont ceux qui tombent le moins en moyenne durant leur séjour (5.6 fois). Ils sont affectés principalement par des problèmes de dépression et des idées noires.

Analyse multivariée des chutes : La probabilité de tomber durant la durée de séjour globale est de 84.7%, le sexe n'est pas significatif après ajustement sur l'âge d'entrée, la date d'entrée et les identifiants des établissements.

La probabilité de chuter de manière répétée suivant le critère de la HAS -- 2 chutes espacées de moins d'une année -- est de 68.5%. Les femmes tombent moins après ajustement sur tous les syndromes sauf les allergies et troubles cardio-vasculaires qui sont non significatifs.

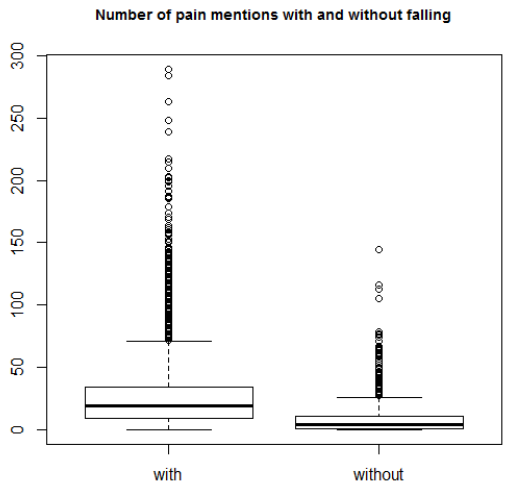
La probabilité de chuter avant 100 jours est de 35% et de 45.37% avant 200 jours.

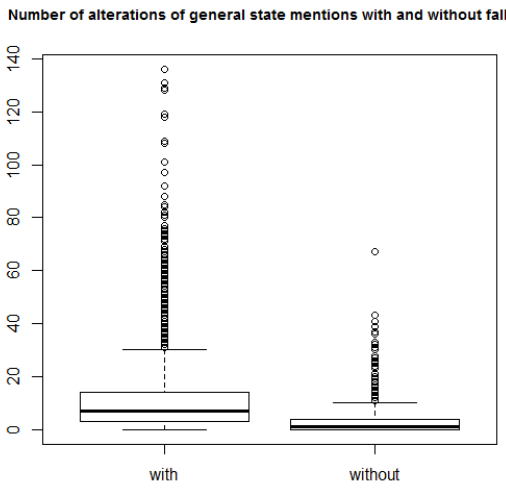
Le fait de tomber précocément est associé à des problèmes de comportement, des problèmes cutanés, des hospitalisations, des décès, des douleurs, une déshydratation et une altération de l'état général. Globalement c'est un indicateur de santé précaire (voir figure 12).

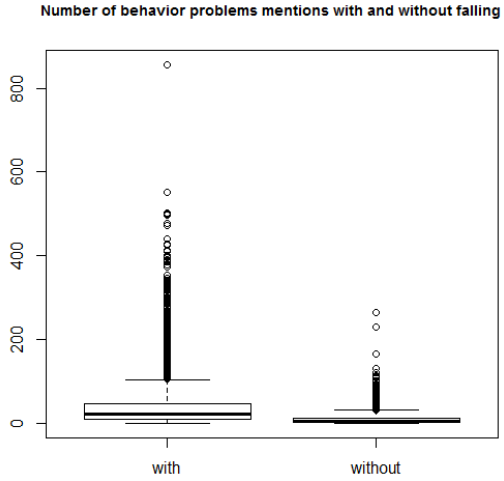
Dans tous les modèles la date d'entrée augmente significativement la probabilité de chuter. Cela signifie que la prévision du nombre de chutes est très dépendante de la quantité et de la qualité des données saisies et qu'en moyenne les transmissions sont de mieux en mieux alimentées: Tous les syndromes ainsi que la plupart des identifiants des établissements sont associés significativement à la date d'entrée en institution. Les prochaines étapes seront d'intégrer le GIR dans les analyses et de contrôler l'association entre GIR moyen par EHPAD et le fait de chuter.

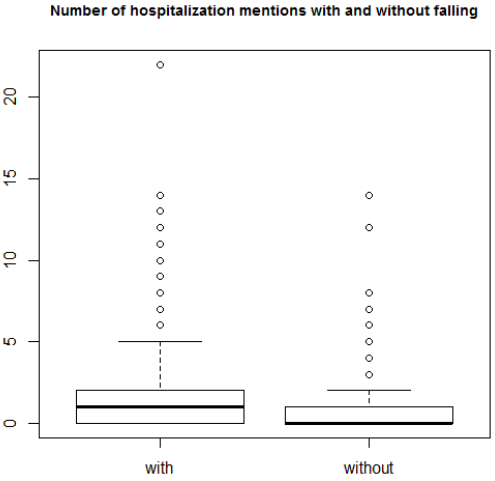
Number of falls Q1 quartile				Number of falls Q4 quartile				Resident population entered before 01/05/2016			
mean (number of falls through one year)			1				8,5				NA
total numbers			6 771				3 765				41 717
mean(age) at NH entry			85,3				85,8				85,1
mean(age) at first fall			86,3				86,5				NA
men/women ratio			31,8%				40,7%				32,1%
last hosp-death delay in days			13,4				30,6				39,1
SYNDROME	EFFECTIFS	Qi	RATIO	SYNDROME	EFFECTIFS	Qi	RATIO	SYNDROME	EFFECTIFS	Qi	RATIO
FALLS	6771	Q1	100	FALLS	3765	Q4	100	BEHAVIOR	34613	TS	87,1
BEHAVIOR	5695	Q1	84,1	PAIN	3729	Q4	99,0	PAIN	33997	TS	85,6
PAIN	5686	Q1	84,0	BEHAVIOR	3724	Q4	98,9	DEPRESSION_DARK_THOUGHTS	33210	TS	83,6
DEPRESSION_DARK_THOUGHTS	5428	Q1	80,2	CUTANEOUS_STATE	3665	Q4	97,3	CUTANEOUS_STATE	32284	TS	81,3
CUTANEOUS_STATE	5244	Q1	77,4	DEPRESSION_DARK_THOUGHTS	3659	Q4	97,2	ALTERED GENERAL STATE	28716	TS	72,3
ALTERED GENERAL STATE	4352	Q1	64,3	ALTERED GENERAL STATE	3471	Q4	92,2	FALLS	27703	TS	69,7
MALNUTRITION_SWALLOWING	3732	Q1	55,1	INTESTINAL_TRANSIT	3087	Q4	82,0	INTESTINAL_TRANSIT	25748	TS	64,8
INTESTINAL_TRANSIT	3724	Q1	55,0	MALNUTRITION_SWALLOWING	3019	Q4	80,2	MALNUTRITION_SWALLOWING	25195	TS	63,4
DEMENTIA	2759	Q1	40,7	DEMENTIA	2686	Q4	71,3	DEMENTIA	20966	TS	52,8
CARDIO_VASCULAR	2702	Q1	39,9	HOSPITALIZATION	2411	Q4	64,0	CARDIO_VASCULAR	19632	TS	49,4
DEATH	2538	Q1	37,5	CARDIO_VASCULAR	2371	Q4	63,0	HOSPITALIZATION	18832	TS	47,4
HOSPITALIZATION	2489	Q1	36,8	URINARY_TRACK_TROUBLES	2124	Q4	56,4	URINARY_TRACK_TROUBLES	18037	TS	45,4
URINARY_TRACK_TROUBLES	2471	Q1	36,5	DEATH	1940	Q4	51,5	DEATH	15994	TS	40,3
VISION	1499	Q1	22,1	DEHYDRATION	1347	Q4	35,8	VISION	12082	TS	30,4
DEHYDRATION	1381	Q1	20,4	SLEEP	1302	Q4	34,6	ORAL_HEALTH	11138	TS	28,0
ORAL_HEALTH	1221	Q1	18,0	ORAL_HEALTH	1277	Q4	33,9	DEHYDRATION	10889	TS	27,4
ILI	1167	Q1	17,2	VISION	1269	Q4	33,7	SLEEP	9619	TS	24,2
SLEEP	1055	Q1	15,6	ILI	994	Q4	26,4	ILI	9466	TS	23,8
VACCINATION	773	Q1	11,4	VACCINATION	810	Q4	21,5	VACCINATION	7710	TS	19,4
ALLERGIES	688	Q1	10,2	HEARING	618	Q4	16,4	ALLERGIES	5142	TS	12,9
CANCER	639	Q1	9,4	ALLERGIES	538	Q4	14,3	AGE	4980	TS	12,5
AGE	566	Q1	8,4	CANCER	537	Q4	14,3	CANCER	4463	TS	11,2
HEARING	473	Q1	7,0	AGE	515	Q4	13,7	HEARING	4363	TS	11,0

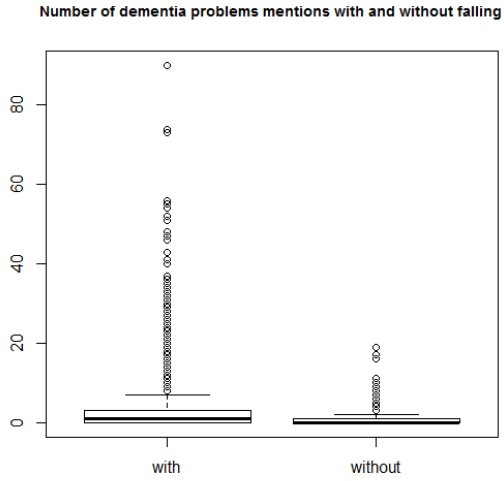
Table 5 : Effectifs syndromiques des résidents avec au moins une chute annuelle durant la période [01/11/2010 – 01/05/2016] découpés en quartiles Q1 (une seule chute) et Q4 (8 chutes et plus) du nombre annuel de chutes et comparaison avec la population générale

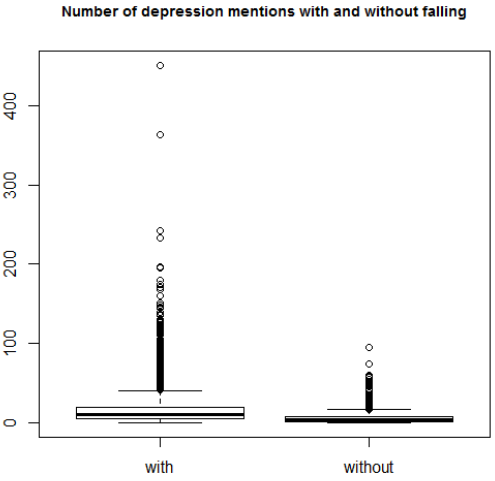

Number of pain mentions with and without falling
0
50
100
150
200
250
300
with
without


Number of alterations of general state mentions with and without falling
0
20
40
60
80
100
120
140
with
without


Number of behavior problems mentions with and without falling
0
200
400
600
800
with
without


Number of hospitalization mentions with and without falling
0
5
10
15
20
with
without


Number of dementia problems mentions with and without falling
0
20
40
60
80
with
without


Number of depression mentions with and without falling
0
100
200
300
400
with
without

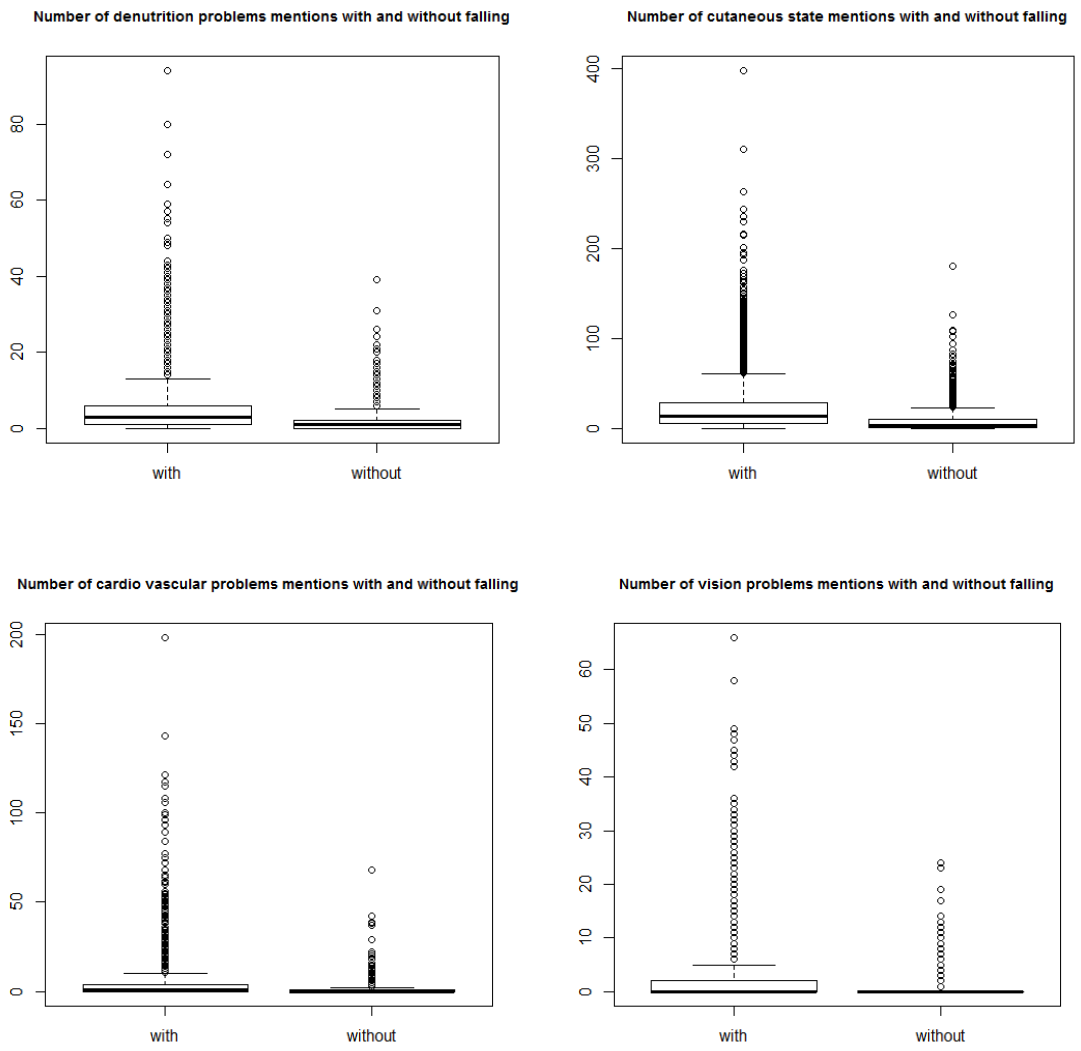


Figure 12 : Boxplots (moyenne – médiane - quartiles) des effectifs syndromiques comparés des résidents chuteurs et non chuteurs (avec de gauche à droite et du haut vers le bas) : la douleur, l'altération de l'état général, le comportement, l'hospitalisation, la démence, la dépression et idées noires, la dénutrition, les problèmes cutanés, cardio-vasculaires et la vision.

3-5-3 Travaux sur les vaccinations

3-5-3-1 Résumé de la présentation

La dernière saison hivernale a vu beaucoup de discussions et de débats autour des vaccinations contre la grippe, tout particulièrement pour les personnes âgées, comme chaque année [61 - 65]. Alors que suivre l'état de santé dans les établissements des résidents en temps réel reste difficile, grâce aux index des résidents et des établissements ainsi qu'aux dates de transmissions, il devient possible de suivre chaque résident au cours du temps et de construire

un SSS. Très peu d'événements récurrents permettent de suivre le statut de santé des résidents à des périodes prédéfinies. La vaccination est l'une d'entre elles.

L'objectif de cette étude était d'évaluer l'intérêt de périodes récurrentes et prédéfinies comme les vaccinations pour suivre le statut de santé de sous-groupes de résidents sélectionnés.

Méthodes: Dans cette étude toutes les vaccinations syndromiques contre la grippe ont été extraites de 125 établissements français entre le 01/11/2010 et le 26/02/2017. Une seconde étape a ensuite extrait tous les autres événements syndromiques sur des périodes de 100 jours autour de chaque mention de vaccination ou décès.

Résultats: Nous avons trouvé 4 192 résidents avec une mention de vaccination contre la grippe mais seulement 928 pour une seconde mention, 2 666 hospitalisés et 1 861 décédés parmi ceux-ci. Les traits syndromiques les plus saillants comparant les périodes de première mention de vaccination et avant le décès étaient la dépression et la déshydratation.

Discussion - Conclusion: Les critères nécessaires pour établir un calendrier plausible de vaccinations à compter de la première mention syndromique de vaccination, d'une part (voir figure 13), et le caractère facultatif de la mention des vaccinations dans les transmissions (la mention de celles-ci peut figurer par exemple dans le silo des prescriptions) d'autre part, ont entrainé une grande déperdition de l'information vaccinale à cet endroit. Observer néanmoins les 26 distributions syndromiques au travers de cinq périodes de vaccinations successives nous a permis de déterminer des sous-groupes de résidents et de prédire leur santé future.

3-5-3-2 Quelques résultats

Le design et les effectifs du suivi :

Après avoir sélectionné toutes les mentions de vaccinations contre la grippe, nous sommes remontés dans le temps jusqu'à la date d'entrée des résidents, définissant leur profil syndromique au travers de leurs 100 premiers jours. Puis, retournant aux périodes de premières vaccinations, nous avons de même examiné les distributions syndromiques sur 100 jours, 25 jours avant et 75 jours après la mention vaccination (voir figure 13).

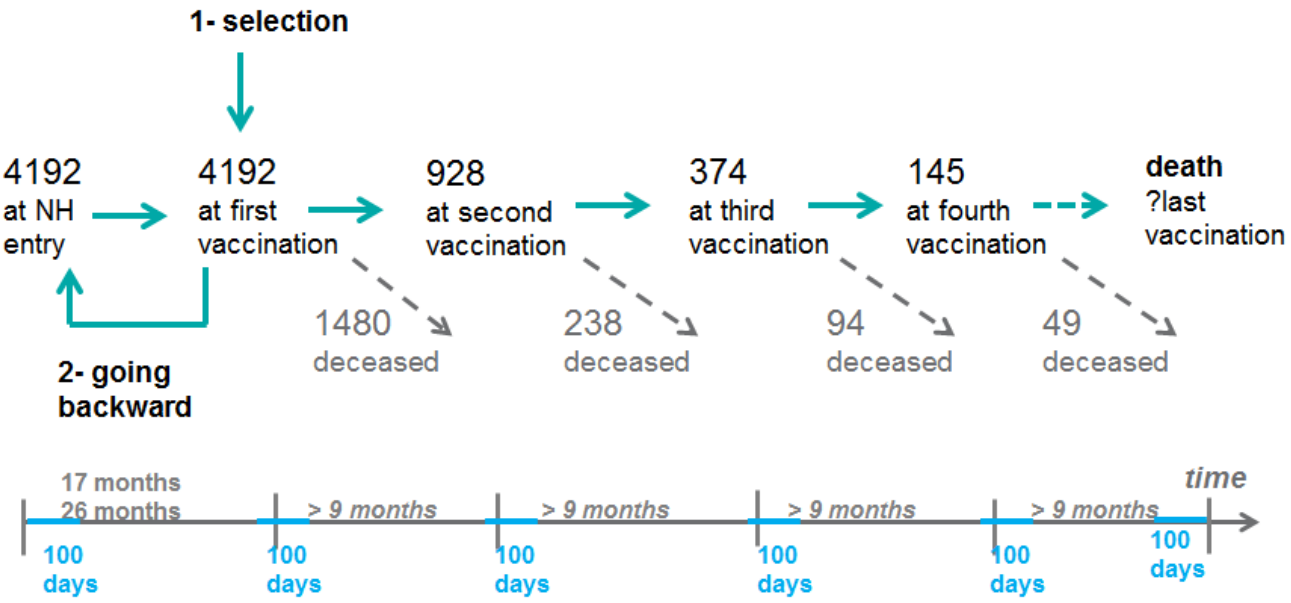


Figure 13 : Schéma temporel d'extraction des vaccinations syndromiques et de suivi des résidents dans le Dossier Résident Informatisé.

Pour la seconde période de vaccination, nous avons recherché une seconde mention de vaccination contre la grippe au moins 9 mois plus tard et défini la même période de 100 jours et ainsi de suite jusqu'aux 100 derniers jours avant le décès.

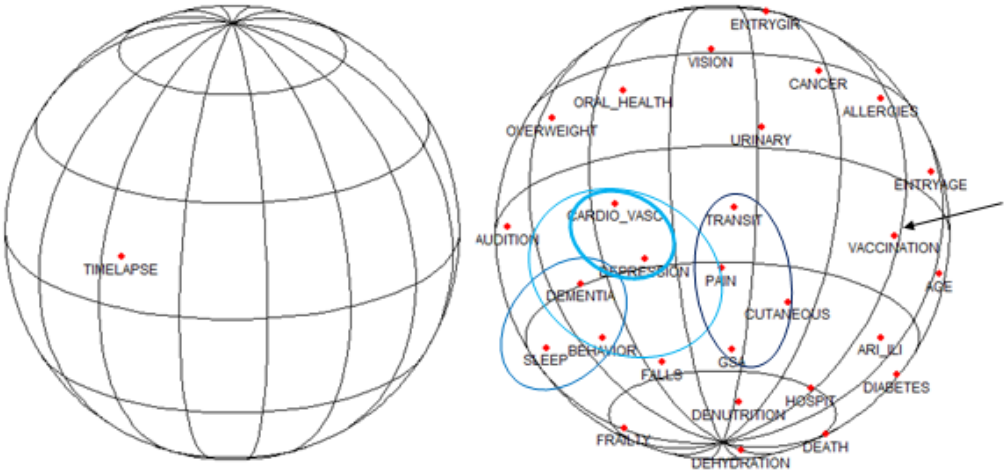
Bien sûr, comme on peut le voir ici, nous perdons un grand nombre de résidents entre les premières et secondes mentions car le personnel médical n'est pas tenu d'alimenter automatiquement le fichier des transmissions. Souvent, cela est fait lorsqu'il y a un problème, une anomalie ou lorsqu'une équipe de soins a quelque chose d'important à dire à la suivante. Avec ce design nous avons défini 6 périodes de 100 jours (voir ci-dessus).

Suivant les ratios syndromiques des hommes au cours du temps, nous constatons que : la déshydratation est un signal vital indicateur de la fin de vie; dépression et idées noires ainsi qu'état cutané sont prégnants, enfin que, même avec les vaccinations contre la grippe, les fonctions immunitaires décroissent avec le temps tandis qu'hospitalisations et syndromes grippaux croissent. Nous retrouverons ce résultat lors de l'étude sur la fin de vie.

Les femmes sont encore plus touchées par la déshydratation en fin de vie (7.7 versus 5.7 pour le ratio de fréquences syndromiques entre entrée et fin de vie). Dépression et idées noires, ainsi qu'état cutané sont également prégnants. Enfin, alors que les femmes séjournent habituellement plus longtemps en EHPAD, le processus de fragilité peut devenir plus aigu en fin

de vie (3.4 pour les hommes versus 3.7 pour les femmes), alors que les syndromes grippaux sont moins prégnants (3.4 pour les hommes versus 2.8 pour les femmes).

La situation syndromique à l'entrée :



DEPRESSION & CARDIO_VASC:0.677
DEPRESSION & PAIN: 0.484
DEPRESSION & BEHAVIOR:0.474

PAIN & GSA: 0.445
PAIN & DEPRESSION: 0.484
PAIN & CUTANEOUS: 0.463
PAIN & TRANSIT: 0.402

BEHAVIOR & DEMENTIA:0.449
BEHAVIOR & DEPRESSION:0.474
BEHAVIOR & SLEEP:0.415

Figure 14 : Analyse en Composantes Principales en 3D des données syndromiques et affichage des corrélations supérieures à 0.4.

L'ACP en 3 dimensions sur la sphère unité [66] a été faite à l'aide du package R psy®. Le syndrome comportement (BEHAVIOR) correspond à une attitude de refus. Nous avons regroupé les corrélations supérieures à 0.4 en 3 domaines : celui concernant la dépression, celui relatif à la douleur et celui relatif au comportement (voir figure 14, GSA = General State Alteration, CUTANEOUS = problèmes cutanés, SLEEP = problèmes de sommeil, TRANSIT = problèmes de transit intestinal).

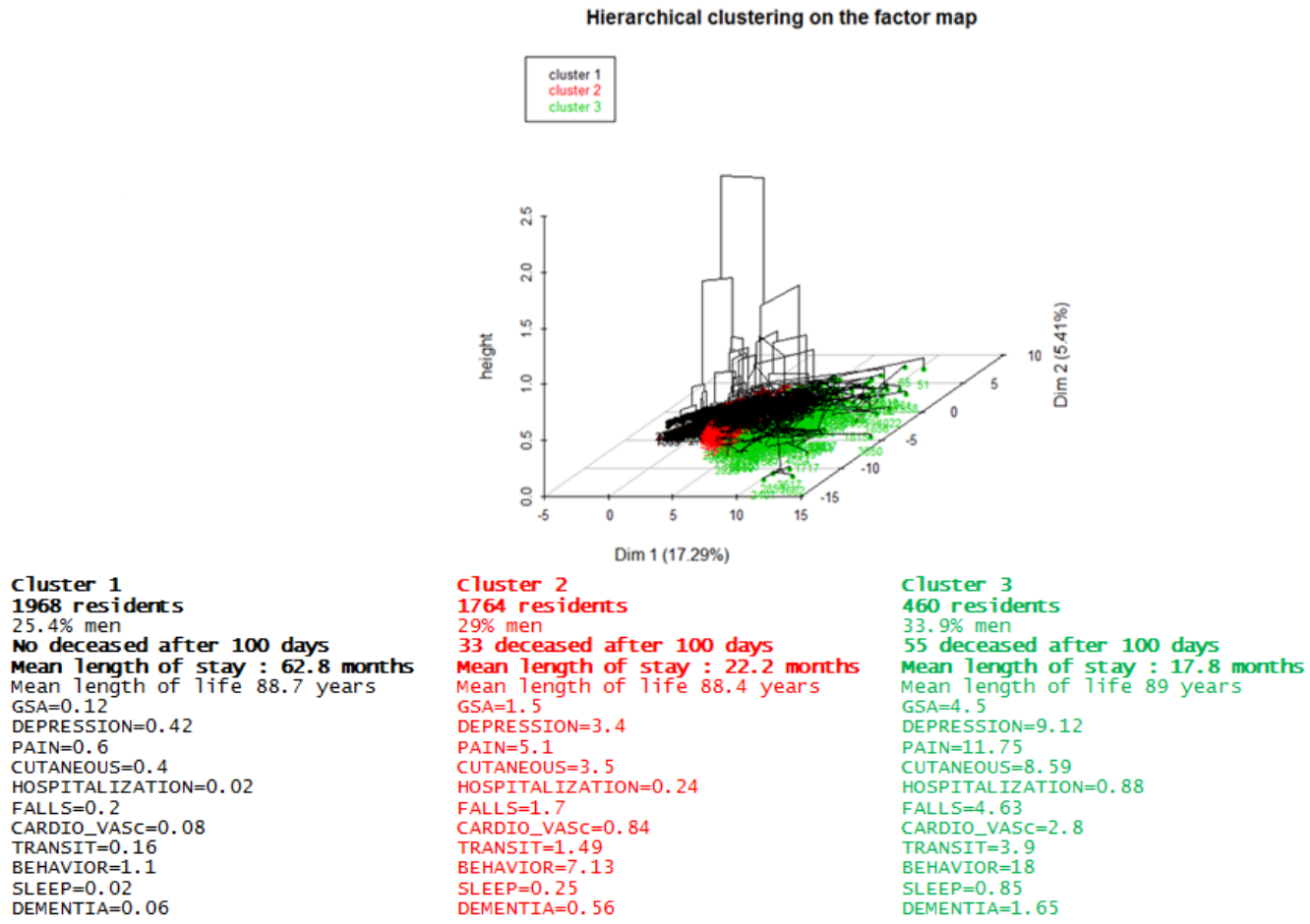


Figure 15 : Analyse en Composantes Principales sur le premier plan principal suivie d'une clusterisation en 3 groupes des données syndromiques.

Nous avons ensuite réalisé une ACP en 2 dimensions suivie d'une clusterisation hiérarchique sur le premier plan principal avec la fonction HCPC [67] au moyen du package R FactoMineR® et trouvé une bonne classification en 3 groupes (voir figure 15), chaque classe étant assez différente des 2 autres. Le premier cluster décrit le groupe le plus fort, séjournant le plus longtemps en institution (62.8 mois), avec moins d'hommes et moins de syndromes. Le troisième cluster décrit les résidents les plus faibles, avec la durée de séjour la plus faible (17.8 mois), et le second, la situation intermédiaire.

Modélisation de la durée de séjour à l'entrée en institution :

Comparant les durées moyennes de séjour entre les 3 groupes, nous avons défini une valeur seuil de 36 mois en tant que valeur moyenne de séjour séparant de manière nette les résidents du premier groupe des deux autres groupes. Nous avons alors trouvé 1 822 résidents avec une

durée de séjour supérieure à 36 mois et modélisé la probabilité de séjourner plus de 36 mois par régression logistique au vu des fréquences de distributions de chaque syndrome à l'issue des 100 premiers jours. La précision du modèle a été calculée par tirage aléatoire de 300 échantillons [apprentissage – test] de tailles respectives [75% - 25%].

Malheureusement, dans notre modèle présenté ci-dessous, les vaccinations semblent influer négativement sur la durée de séjour et la relation est très significative (voir la p-value associée aux vaccinations). Cela provient de la méthode de génération de certaines informations syndromiques et tout particulièrement des vaccinations. La précision du modèle reflète également cet état de fait : IC95% = [63.7% - 69.3%]. Néanmoins, deux résultats que nous retrouverons plus loin avec les travaux sur la fin de vie : plus on entre tard, moins on séjourne longtemps et les femmes séjournent en moyenne plus longtemps même après avoir ajusté sur l'âge à l'entrée, le fait d'être vaccinée et plusieurs autres variables syndromiques.

trainglm3.fit = glm(MORETHAN36MONTHS ~ VACCINATION + ENTRYAGE + BEHAVIOR + CUTANEOUS + SEX + HOSPIT + GSA + DEPRESSION + ARI_ILI + PAIN, data=train, family=binomial(link=logit))

```
Deviance Residuals:
   Min       1Q   Median       3Q      Max
-2.1921  -0.8577  -0.2418   0.8559   3.5415
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
  (Intercept)  4.900114   0.538666   9.097  < 2e-16 ***
  VACCINATION -1.705602   0.132639 -12.859  < 2e-16 ***
  ENTRYAGE    -0.053086   0.006325  -8.393  < 2e-16 ***
  BEHAVIOR    -0.045430   0.007329  -6.199 5.69e-10 ***
  CUTANEOUS   -0.099008   0.016114  -6.144 8.03e-10 ***
  SEX2         0.586985   0.099674   5.889 3.88e-09 ***
  HOSPIT      -0.434781   0.118884  -3.657 0.000255 ***
  GSA         -0.099944   0.030897  -3.235 0.001218 **
  DEPRESSION  -0.051140   0.015915  -3.213 0.001313 **
  ARI_ILI     -0.476772   0.156647  -3.044 0.002338 **
  PAIN        -0.025874   0.012420  -2.083 0.037223 *

AIC: 3257.3, Number of Fisher Scoring iterations: 5
```

En essayant de prédire l'efficacité des vaccinations contre la grippe chez les résidents en institution, nous avons pu constater que les mentions de vaccinations semblaient d'une part, souvent liées à certaines fragilités des résidents et donc à des inquiétudes du personnel soignant, d'autre part, dépendantes des politiques de saisie de l'information : par exemple, lors de la dernière épidémie de grippe, la vigilance avait été renforcée et donc également, la saisie d'informations médicales augmentée.

3-5-4 Travaux sur la fin de vie

3-5-4-1 Résumé de la présentation

Les données issues de la littérature montrent une prévalence élevée et une grande diversité de symptômes physiques et psychiques en fin de vie. Au premier rang des symptômes figurent les signes généraux mais également les symptômes psychiques tels que l'angoisse ou la tension psychique. Il existe aujourd'hui assez peu d'études s'intéressant aux symptômes en fin de vie des résidents en EHPAD et à leur prise en charge.

L'objectif de cette étude était d'une part, de comparer résidents en fin de séjour avec ceux en début de séjour et d'autre part, de caractériser la fin de vie des résidents au travers des soins prodigués.

Méthodes : La période d'étude sélectionnée était de 100 jours, 8 semaines étant un délai souvent trop court [68 - 69] pour évaluer les meilleures mesures à prendre [70 - 71].

Deux échantillons ont donc été constitués. Le premier a inclus les résidents entrés entre le 01/11/2010 et le 26/02/2017 dans un établissement Korian avec au moins une transmission syndromique sur les 100 premiers jours de leur séjour. Le second a inclus tous les résidents décédés entre le 01/11/2010 et le 26/02/2017 avec également au moins une information syndromique dans les100 jours précédant le décès, autre que la mention du décès et différent du premier. La comparaison des effectifs syndromiques a été réalisée par tests t sur les effectifs des 24 syndromes et hospitalisations sur les 2 échantillons.

Nous avons ensuite sélectionné toutes les transmissions à l'origine d'un ou plusieurs syndromes, de manière à établir une typologie de la fin de vie. Les distributions de chaque type de transmissions ont cette fois été évaluées sur des périodes de 100 jours. Il ne s'agissait plus cette fois de variables booléennes, absence/présence, mais de variables de comptage.

Résultats : Le *groupe 100 derniers jours* comprenait 17 782 résidents décédés avec une moyenne de 22.8 transmissions (écart-type=13.9 donc une grande variabilité) durant leurs derniers jours, 11 652 femmes (d'âge moyen 89.7 ans), 6 130 hommes (d'âge moyen 87.5 ans), avec des durées moyennes de séjours de respectivement 36 et 23 mois. Parmi ceux-ci, 1 229 avaient séjourné moins d'un mois dans la structure et 5 923 avaient séjourné plus de 3 années. Là encore, on retrouvait une grande variabilité.

Le groupe *100 premiers jours* comprenait 20 504 résidents vivants avec une moyenne de 14.5 transmissions (écart-type=12.2) durant leurs premiers jours, 14 250 femmes (d'âge moyen 85.8 ans), 6 257 hommes (d'âge moyen 83.6 ans), avec des durées moyennes de séjours de respectivement 36.6 et 23 mois. Le GIR à l'entrée était respectivement de 2.8 (+/- 1) pour les hommes et 2.7 (+/- 1) pour les femmes. Les traits syndromiques les plus saillants opposant débuts et fins de séjours étaient l'état cutané, la douleur, une attitude de refus, la dépression et les idées noires, enfin une altération de l'état général touchant respectivement 87.8%, 86%, 85.4%, 84% et 77.7% des résidents en fin de vie.

Discussion - Conclusion : Examinant les 26 distributions syndromiques au travers des 100 derniers jours, nous espérions caractériser la fin de vie des résidents et déterminer des groupes homogènes de résidents.

3-5-4-2 Quelques résultats :

Dans la table 6 ci-dessous, nous présentons les effectifs syndromiques moyens et leurs écart-types dans la population des résidents décédés entre les 100 premiers et les 100 derniers jours de séjours. Les traits syndromiques les plus saillants sont listés par ordre décroissant d'importance de gauche à droite, en fin de séjour pour la première table, et en début de séjour pour la seconde table. Toutes les comparaisons significativement différentes, calculées par un

test-t, entre les 2 échantillons de début et de fin de séjour sont colorées dans la même gamme de couleurs, une couleur par syndrome. Les traits non saillants sont affichés en gris clair.

Ensuite nous avons détaillé la distribution des décès en fonction de la durée de séjour (figure 16), pour des durées de séjour variant de moins d'un mois à 48 mois. Il reste 4 403 résidents, soit 24.62% ayant séjourné plus de quatre années, non représentés sur le graphique. Alors qu'une partie non négligeable des résidents vit moins d'une année, ici approximativement 40% de notre population (voir figure 16), chiffre à rapprocher des 47% des résidents en 2011 vivant une année et demie selon une étude de la DREES [72] il nous a semblé opportun de rechercher leur profil.

.

Nombre moyen de syndromes (écarts-types) par résident en fin de vie sur une période de 100 jours par sexe

SEXE	COMPOR-TEMENT	ETAT_CUTANE	DOULEUR	DEPRESSION	AEG	DENUTRITION	TRANSIT	CHUTES	DESHYDRATATION	CARDIO_VASC	DEMENCE
masculin	7,3 (8)	5,9 (6)	5,7 (6)	3,4 (4)	3,4 (4)	2,3 (3)	2 (3)	1,9 (3)	1 (2)	,7 (1)	,7 (1)
féminin	5,9 (7)	6 (6)	6 (7)	3,5 (4)	3,4 (4)	2,3 (3)	2,5 (4)	1,3 (2)	1 (2)	,9 (2)	,6 (1)

SEXE	URINARY_TRACK	HOSPITALISATION	DIABETE	GRIPPE_IRA	SLEEP	VISION	BUCCO_DENTAIRE	CANCER	GEA_DIARRHEE	VACCINATION
masculin	,7 (2)	,8 (1)	,5 (2)	,4 (1)	,2 (1)	,2 (1)	,2 (1)	,1 (1)	,1 (0)	,1 (0)
féminin	,7 (1)	,6 (1)	,4 (2)	,2 (1)	,2 (1)	,2 (1)	,2 (1)	,1 (1)	,1 (0)	,1 (0)

Nombre moyen de syndromes (écarts-types) par résident en début de séjour sur une période de 100 jours par sexe

SEXE	COMPOR-TEMENT	DOULEUR	ETAT_CUTANE	DEPRESSION	CHUTES	AEG	TRANSIT	DENUTRITION	DEMENCE	CARDIO_VASC	DIABETE
masculin	6,1 (9)	2,9 (4)	2,4 (4)	2,2 (3)	1,5 (2)	1,1 (2)	1 (2)	,8 (1)	,6 (1)	,5 (1)	,5 (2)
féminin	5,1 (8)	3,5 (5)	2,2 (4)	2,8 (4)	1,3 (2)	1 (2)	1,1 (2)	,7 (1)	,6 (1)	,8 (2)	,3 (2)

SEXE	URINARY_TRACK	VISION	SLEEP	HOSPITALISATION	DESHYDRATATION	BUCCO_DENTAIRE	CANCER	GRIPPE_IRA	VACCINATION	GEA_DIARRHEE
masculin	,5 (1)	,3 (1)	,3 (1)	,2 (1)	,2 (1)	,2 (1)	,1 (0)	,1 (0)	,1 (0)	0 (0)
féminin	,5 (1)	,3 (1)	,2 (1)	,2 (0)	,1 (1)	,2 (1)	,1 (0)	,1 (0)	,1 (0)	0 (0)

Table 6 : Comparaison des effectifs syndromiques moyens des fins et débuts de séjours des 17 784 résidents décédés durant la période [01/11/2010 - 26/02/2016] (11 652 femmes et 6 130 hommes)

Distribution de la durée de séjour :

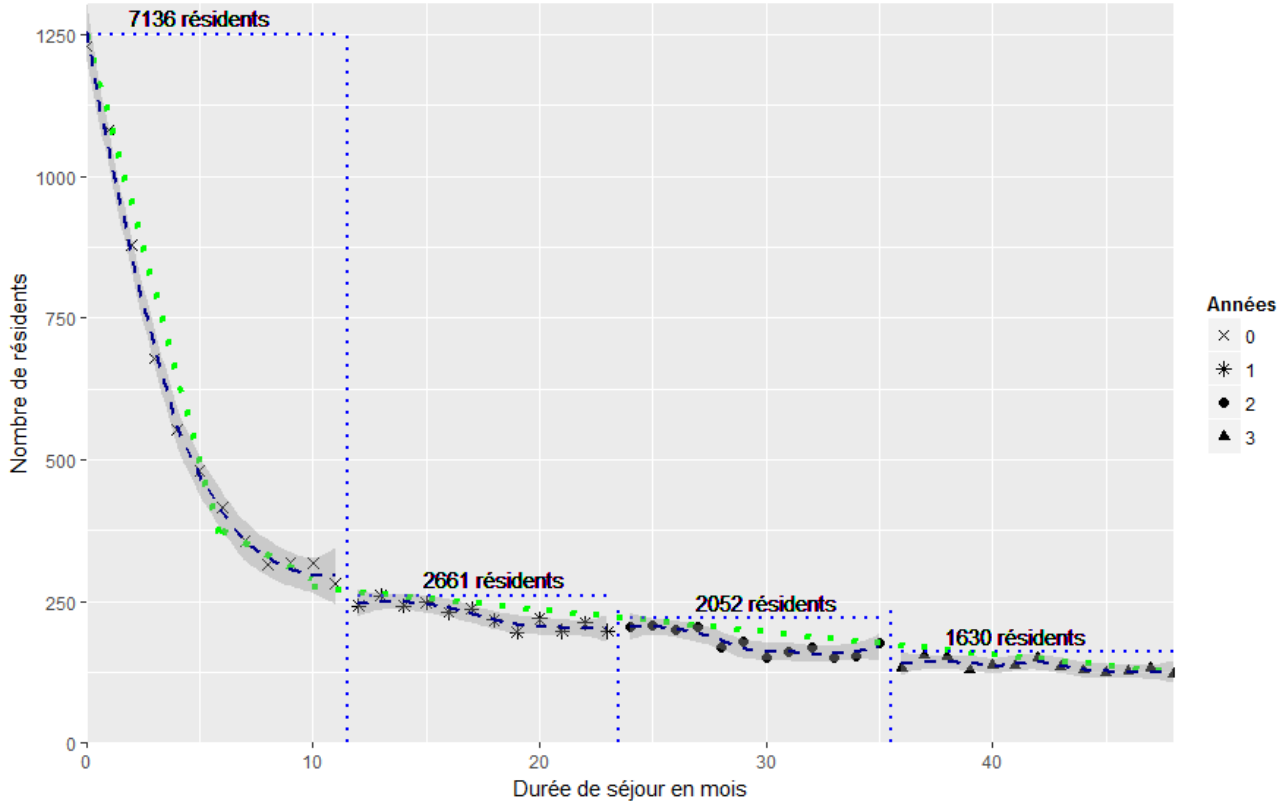


Figure 16 : Fréquence de durée de séjour des 17 882 résidents Korian en France décédés entre le 01/11/2010 et le 26/02/2017 pour les 48 premiers mois.

Caractérisation de la fin de séjour :

Après avoir enlevé un résident avec une date de naissance probablement erronée nous avons cherché à caractériser la fin de séjour des 17°881 résidents au moyen d'une classification non supervisée. Cette méthode d'analyse multivariée permet d'identifier des sous-groupes homogènes ou clusters basés sur des similarités ou caractéristiques communes partagées au sein d'une population a priori hétérogène [73]. Cette fois, nous n'avons pu utiliser le package FactoMineR [67], mais le package fastcluster [74] car notre effectif était trop volumineux. Nous avons alors calculé la matrice des distances euclidiennes entre les individus et réalisé une classification automatique avec le critère de Ward qui consiste à minimiser l'inertie intra-classes ou maximiser l'inertie inter-classes. Le nombre optimal de classes s'est décidé en traçant le screeplot (figure 17) et en examinant différents découpages possibles (figure 18).

Ensuite nous avons cherché à déterminer les principaux facteurs discriminants pour le découpage en quatre classes et pour cela nous avons étudié les résumés statistiques de toutes les variables utilisées dans la clusterisation. Ce fut l'objet de la table 7.

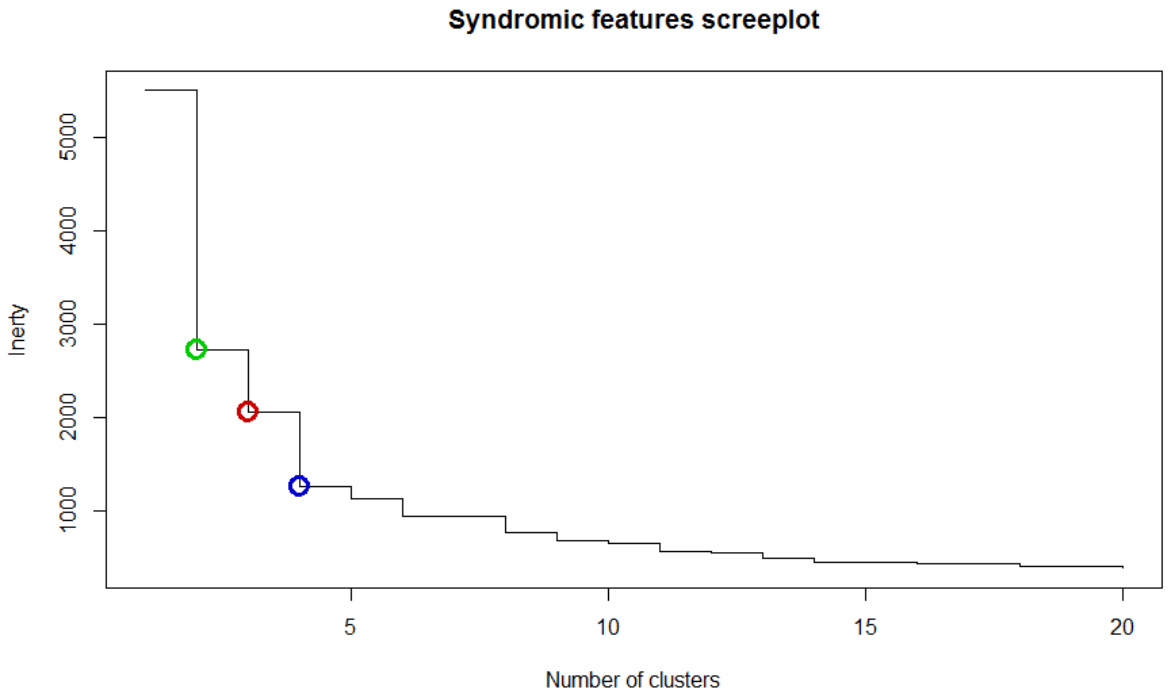


Figure 17 : Screeplot des 26 traits syndromiques des 100 derniers jours des 17 881 résidents Korian en France décédés entre le 01/11/2010 et le 26/02/2017.

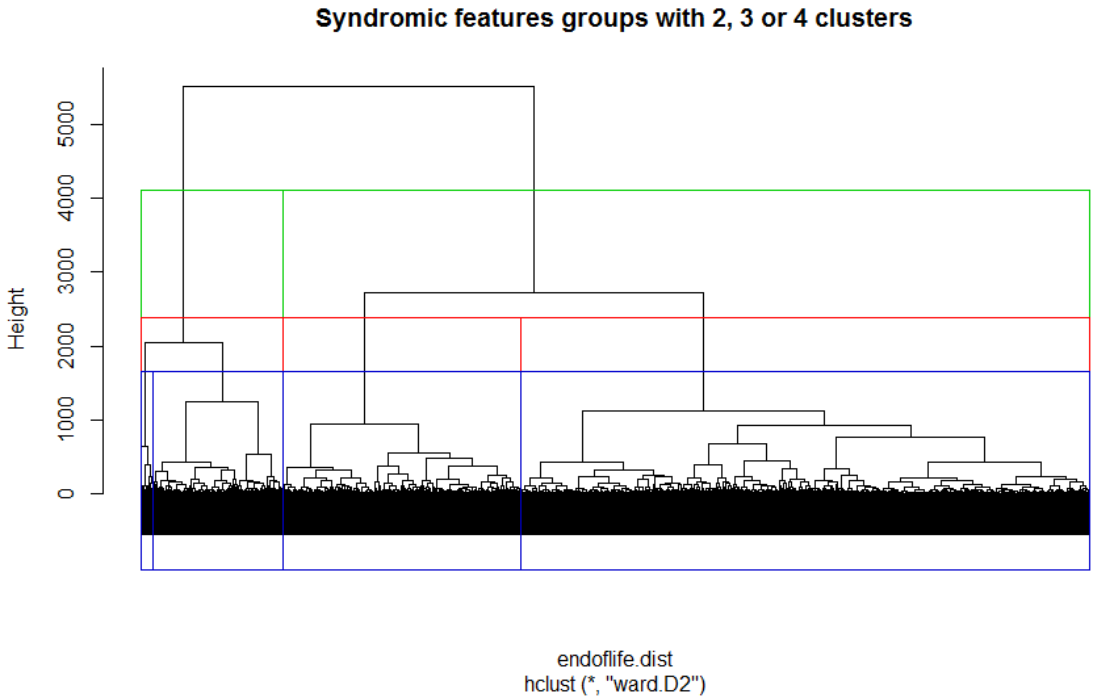


Figure 18 : Dendrogramme des 26 traits syndromiques des 100 derniers jours des 17 881 résidents Korian en France décédés entre le 01/11/2010 et le 26/02/2017 et visualisation des 3 découpages proposés par le screeplot

GR3 ***Entered late*** **N=10 659** **59.47% women** **87.8 at entry** **88.7 at death**	**GR2** ***Staying at least 8 months*** **N=4 469** **69.79% women** **84.7 at entry** **88.4 at death**	**GR1** ***The fitted*** **N=2 427** **82.82% women** **83 at entry** **90. 8 at death**	**GR4** ***Entered early*** **N=226** **81.05% women** **73.8 at entry** **90 at death**
range LENGTH-STAY: **[0 – 46] months** **mean=10 months**	**range LENGTH-STAY:** **[8 – 87] months** **mean=3 years 8 months**	**range LENGTH-STAY:** **[45 – 165] months** **mean=7 years 9 months**	**range LENGTH-STAY:** **[150 – 336] months** **mean=16 years 2 months**
Q1_Q3 at entry: **[84 – 92] years**	**Q1_Q3 at entry:** **[81 – 89] years**	**Q1_Q3 at entry:** **[80 – 88] years**	**Q1_Q3 at entry:** **[68 – 80] years**
GSA Min. : 0.000 1st Qu.: 1.000 Median : 2.000 Mean : 3.442 3rd Qu.: 5.000 Max. :44.000	GSA Min. : 0.000 1st Qu.: 1.000 Median : 2.000 Mean : 3.635 3rd Qu.: 5.000 Max. :35.000	GSA Min. : 0.000 1st Qu.: 1.000 Median : 2.000 Mean : 2.976 3rd Qu.: 4.000 Max. :35.000	GSA Min. : 0.000 1st Qu.: 0.000 Median : 1.000 Mean : 2.465 3rd Qu.: 3.000 Max. :34.000
CARDIO_VASCULAR Min. : 0.0000 1st Qu.: 0.0000 Median : 0.0000 Mean : 0.9216 3rd Qu.: 1.0000 Max. :32.0000	CARDIO_VASCULAR Min. : 0.0000 1st Qu.: 0.0000 Median : 0.0000 Mean : 0.7362 3rd Qu.: 1.0000 Max. :17.0000	CARDIO_VASCULAR Min. : 0.0000 1st Qu.: 0.0000 Median : 0.0000 Mean : 0.5229 3rd Qu.: 1.0000 Max. :17.0000	CARDIO_VASCULAR Min. :0.0000 1st Qu.:0.0000 Median :0.0000 Mean :0.4602 3rd Qu.:0.7500 Max. :5.0000
FALLS Min. : 0.00 1st Qu.: 0.00 Median : 1.00 Mean : 1.68 3rd Qu.: 2.00 Max. :28.00	FALLS Min. : 0.000 1st Qu.: 0.000 Median : 1.000 Mean : 1.471 3rd Qu.: 2.000 Max. :34.000	FALLS Min. : 0.000 1st Qu.: 0.000 Median : 0.000 Mean : 1.013 3rd Qu.: 1.000 Max. :23.000	FALLS Min. :0.0000 1st Qu.:0.0000 Median :0.0000 Mean :0.7655 3rd Qu.:1.0000 Max. :9.0000
BEHAVIOR Min. : 0.000 1st Qu.: 2.000 Median : 4.000 Mean : 6.719 3rd Qu.: 9.000 Max. :82.000	BEHAVIOR Min. : 0.000 1st Qu.: 1.000 Median : 4.000 Mean : 6.101 3rd Qu.: 8.000 Max. :66.000	BEHAVIOR Min. : 0.000 1st Qu.: 1.000 Median : 3.000 Mean : 5.471 3rd Qu.: 7.000 Max. :72.000	BEHAVIOR Min. : 0.000 1st Qu.: 1.000 Median : 2.000 Mean : 4.752 3rd Qu.: 6.000 Max. :33.000
DEMENTIA Min. : 0.0000 1st Qu.: 0.0000 Median : 0.0000 Mean : 0.6659 3rd Qu.: 1.0000 Max. :25.0000	DEMENTIA Min. :0.0000 1st Qu.:0.0000 Median :0.0000 Mean :0.5612 3rd Qu.:1.0000 Max. :9.0000	DEMENTIA Min. : 0.00 1st Qu.: 0.00 Median : 0.00 Mean : 0.56 3rd Qu.: 1.00 Max. :14.00	DEMENTIA Min. :0.0000 1st Qu.:0.0000 Median :0.0000 Mean :0.4292 3rd Qu.:1.0000 Max. :9.0000
DEPRESSION Min. : 0.000 1st Qu.: 1.000 Median : 3.000 Mean : 3.664 3rd Qu.: 5.000 Max. :42.000	DEPRESSION Min. : 0.000 1st Qu.: 1.000 Median : 2.000 Mean : 3.296 3rd Qu.: 5.000 Max. :41.000	DEPRESSION Min. : 0.000 1st Qu.: 1.000 Median : 2.000 Mean : 2.885 3rd Qu.: 4.000 Max. :21.000	DEPRESSION Min. : 0.000 1st Qu.: 1.000 Median : 2.000 Mean : 2.814 3rd Qu.: 4.000 Max. :26.000

GR3 *Entered late*	GR2 *Staying at least 8 months*	GR1 *The fitted*	GR4 *Entered early*
DEHYDRATION Min. : 0.000 1st Qu.: 0.0000 Median : 0.0000 **Mean : 0.9332** 3rd Qu.: 1.000 Max. :46.0000	DEHYDRATION Min. : 0.000 1st Qu.: 0.000 Median : 0.000 **Mean : 1.251** 3rd Qu.: 2.000 Max. :41.000	DEHYDRATION Min. : 0.000 1st Qu.: 0.000 Median : 0.000 **Mean : 1.146** 3rd Qu.: 1.000 Max. :27.000	DEHYDRATION Min. : 0.0000 1st Qu.: 0.0000 Median : 0.0000 **Mean : 0.8805** 3rd Qu.: 1.0000 Max. :12.0000
DENUTRITION Min. : 0.000 1st Qu.: 0.000 **Median : 1.000** **Mean : 2.206** 3rd Qu.: 3.000 Max. :29.000	DENUTRITION Min. : 0.000 1st Qu.: 0.000 **Median : 2.000** **Mean : 2.578** 3rd Qu.: 4.000 Max. :34.000	DENUTRITION Min. : 0.000 1st Qu.: 0.000 **Median : 2.000** **Mean : 2.335** 3rd Qu.: 3.000 Max. :26.000	DENUTRITION Min. : 0.000 1st Qu.: 0.000 **Median : 1.000** **Mean : 2.199** 3rd Qu.: 3.000 Max. :22.000
PAIN Min. : 0.000 1st Qu.: 2.000 **Median : 4.000** **Mean : 6.151** 3rd Qu.: 9.000 Max. :61.000	PAIN Min. : 0.000 1st Qu.: 1.000 **Median : 4.000** **Mean : 5.555** 3rd Qu.: 8.000 Max. :54.000	PAIN Min. : 0.000 1st Qu.: 1.000 **Median : 3.000** **Mean : 5.372** 3rd Qu.: 7.000 Max. :60.000	PAIN Min. : 0.000 1st Qu.: 1.000 **Median : 2.000** **Mean : 4.212** 3rd Qu.: 5.000 Max. :31.000
CUTANEOUS Min. : 0.000 1st Qu.: 2.000 **Median : 4.000** **Mean : 5.531** 3rd Qu.: 7.000 Max. :65.000	CUTANEOUS Min. : 0.00 1st Qu.: 2.00 **Median : 5.00** **Mean : 6.52** 3rd Qu.: 9.00 Max. :42.00	CUTANEOUS Min. : 0.000 1st Qu.: 2.000 **Median : 5.000** **Mean : 6.759** 3rd Qu.: 9.000 Max. :63.000	CUTANEOUS Min. : 0.000 1st Qu.: 2.000 **Median : 5.000** **Mean : 6.319** 3rd Qu.: 9.000 Max. :40.000
FRAILTY Min. :0.0000 1st Qu.:0.0000 Median :0.0000 **Mean :0.1221** 3rd Qu.:0.0000 Max. :7.0000	FRAILTY Min. :0.0000 1st Qu.:0.0000 Median :0.0000 **Mean :0.1123** 3rd Qu.:0.0000 Max. :4.0000	FRAILTY Min. :0.00000 1st Qu.:0.00000 Median :0.00000 **Mean :0.08447** 3rd Qu.:0.00000 Max. :3.00000	FRAILTY Min. :0.0000 1st Qu.:0.0000 Median :0.0000 **Mean :0.0708** 3rd Qu.:0.0000 Max. :1.0000
AGE Min. : 0.0000 1st Qu.: 0.0000 Median : 0.0000 **Mean : 0.1113** 3rd Qu.: 0.0000 Max. :15.0000	AGE Min. : 0.0000 1st Qu.: 0.0000 Median : 0.0000 **Mean : 0.1065** 3rd Qu.: 0.0000 Max. :10.0000	AGE Min. :0.00000 1st Qu.:0.00000 Median :0.00000 **Mean :0.09312** 3rd Qu.:0.00000 Max. :6.00000	AGE Min. :0.00000 1st Qu.:0.00000 Median :0.00000 **Mean :0.09735** 3rd Qu.:0.00000 Max. :2.00000
ARI_ILI Min. : 0.000 1st Qu.: 0.000 Median : 0.000 **Mean : 0.3013** 3rd Qu.: 0.0000 Max. :10.0000	ARI_ILI Min. :0.0000 1st Qu.:0.0000 Median :0.0000 **Mean :0.2766** 3rd Qu.:0.0000 Max. :7.0000	ARI_ILI Min. :0.0000 1st Qu.:0.0000 Median :0.0000 **Mean :0.2295** 3rd Qu.:0.0000 Max. :7.0000	ARI_ILI Min. :0.0000 1st Qu.:0.0000 Median :0.0000 **Mean :0.1903** 3rd Qu.:0.0000 Max. :5.0000
HOSPITALIZATION Min. :0.0000 1st Qu.:0.0000 **Median :1.0000** **Mean :0.7528** 3rd Qu.:1.0000 Max. :7.0000	HOSPITALIZATION Min. :0.0000 1st Qu.:0.0000 Median :0.0000 **Mean :0.6424** 3rd Qu.:1.0000 Max. :5.0000	HOSPITALIZATION Min. : 0.0000 1st Qu.: 0.0000 Median : 0.0000 **Mean : 0.4981** 3rd Qu.: 1.0000 Max. :17.0000	HOSPITALIZATION Min. :0.0000 1st Qu.:0.0000 Median :0.0000 **Mean :0.4956** 3rd Qu.:1.0000 Max. :3.0000

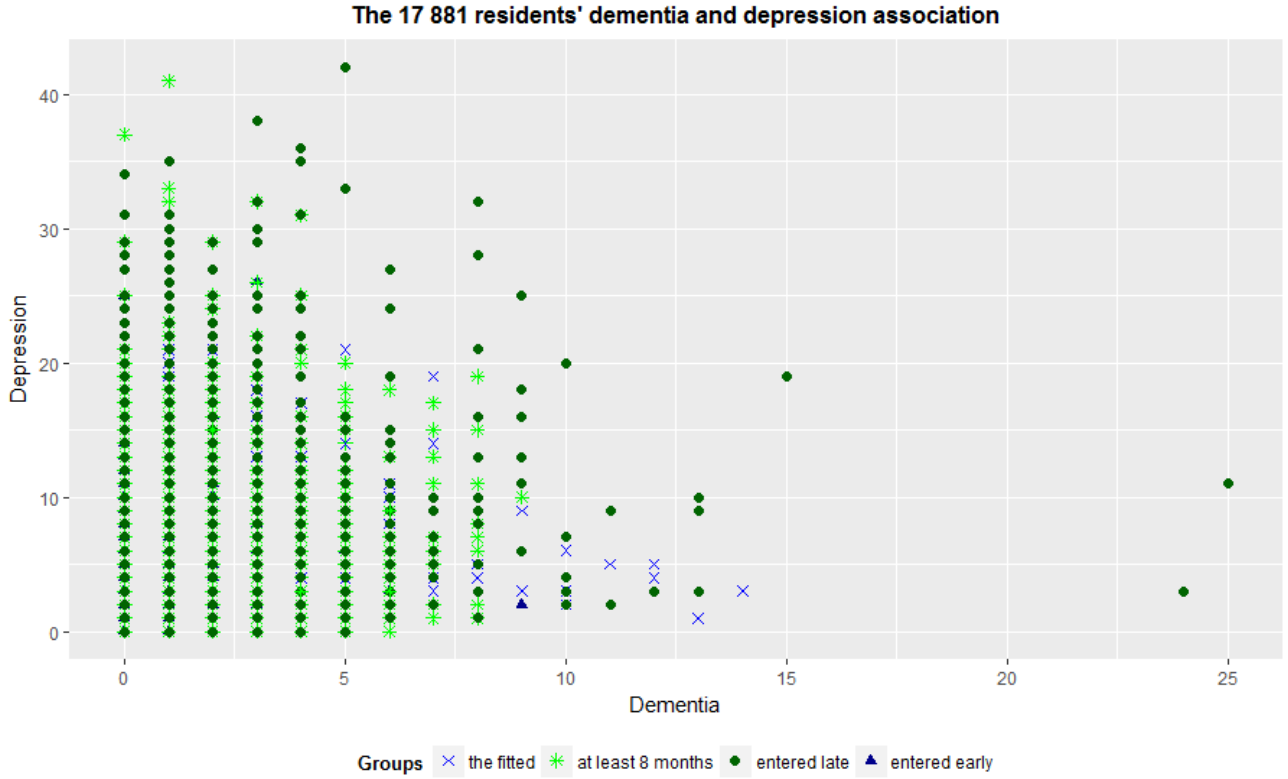
Table 7 : Résumé statistique des variables syndromiques saillantes en fin de vie pour la partition en 4 clusters proposée dans le dendrogramme des 17 881 résidents Korian en France décédés entre le 01/11/2010 et le 26/02/2017.

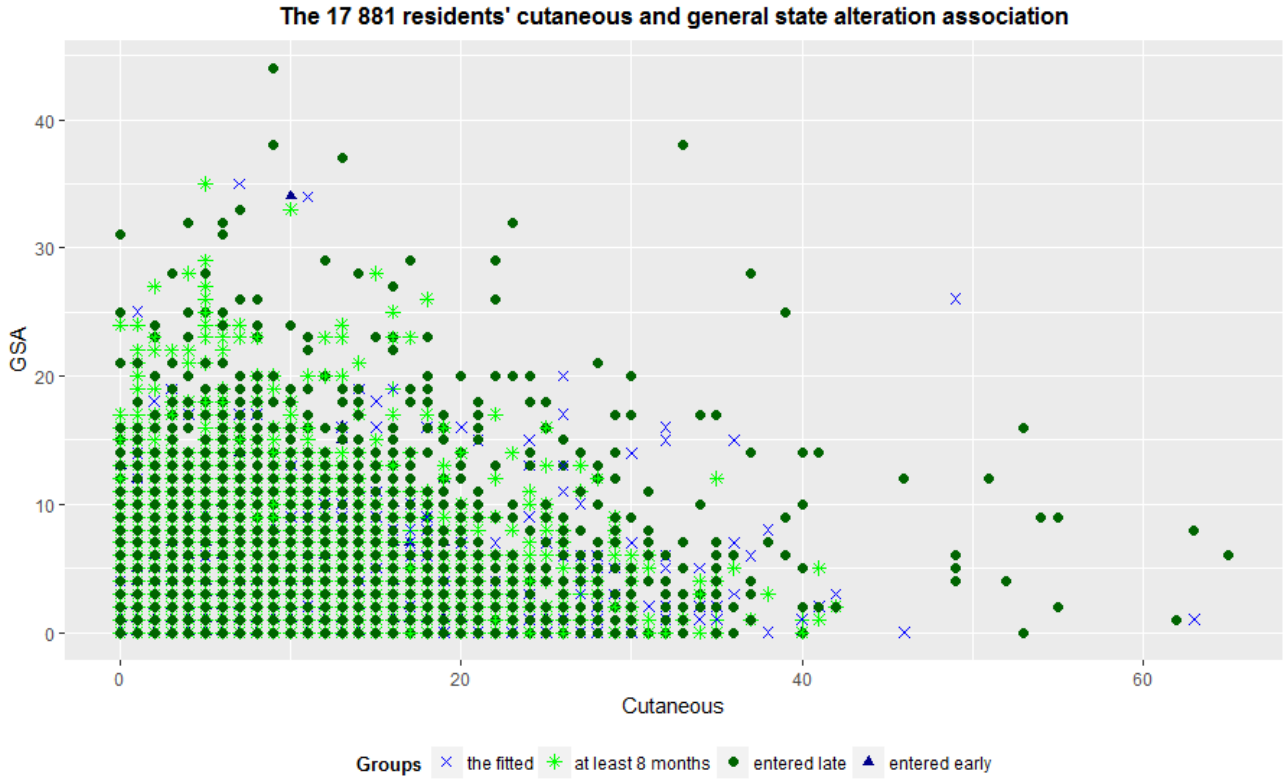
Un examen rapide des statistiques des variables syndromiques saillantes en fin de vie a permis de caractériser et dénommer ces quatre groupes : âge d'entrée et sexe semblent réellement avoir présidé à la constitution de ce découpage. De manière à confirmer ou infirmer cette hypothèse, nous avons représenté graphiquement le nuage des 17°881 résidents suivant les âges d'entrée et de décès. C'est la figure 19.

Figure 19: Nuage des âges à l'entrée et au moment du décès pour la partition en 4 clusters proposée dans le dendrogramme des 17 881 résidents Korian en France décédés entre le 01/11/2010 et le 26/02/2017.

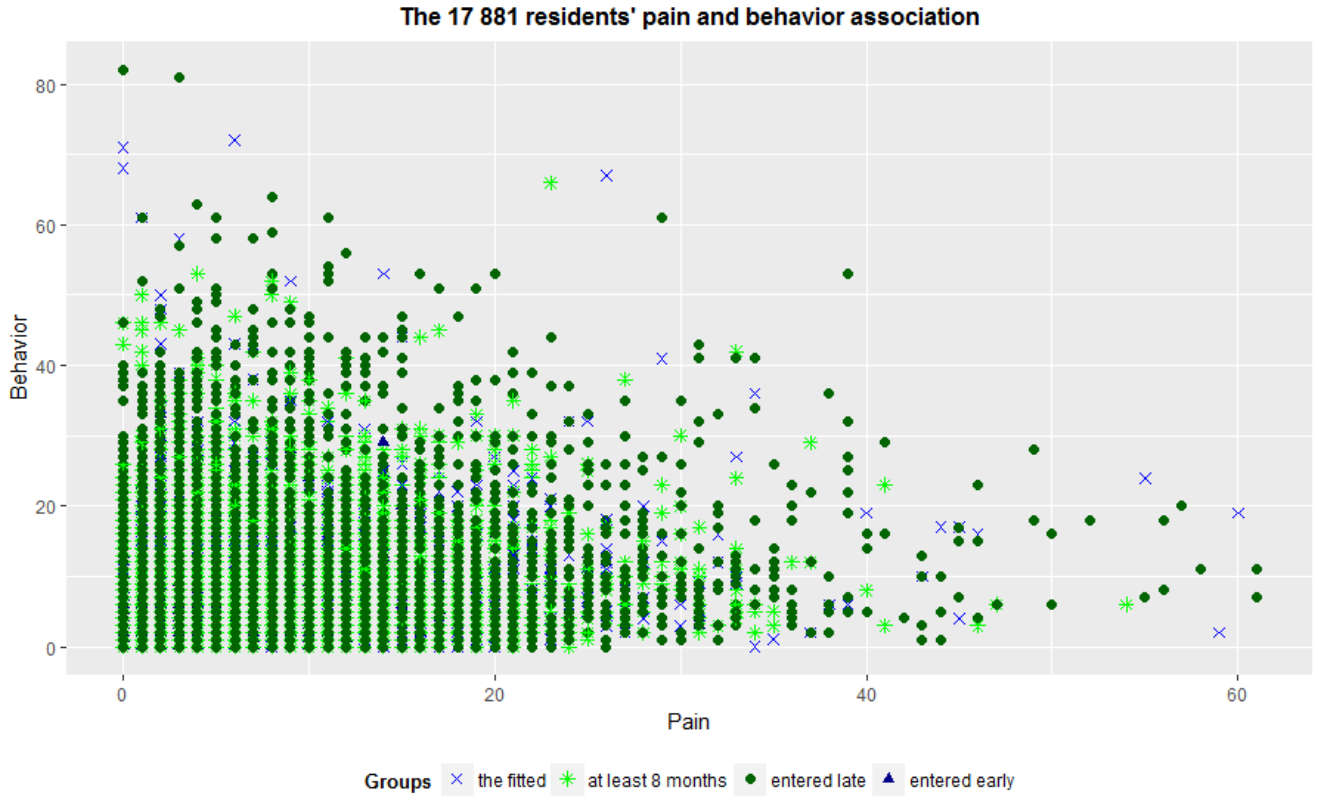
L'âge d'entrée semble hautement prédictif de l'âge du décès. Les quatre groupes se déploient parallèlement suivant un axe linéaire. Occupant la partie basse, la situation la moins favorable avec le groupe des entrés tardivement qui s'agglutinent tous le long de l'axe (losanges vert sombre). Occupant la partie haute, les résidents entrés en majorité plus tôt et donc, paradoxalement avec des problèmes de santé précoces (triangles bleu sombre), cependant en bien meilleure santé finale (voir la table 7). D'autres graphiques présentant quelques

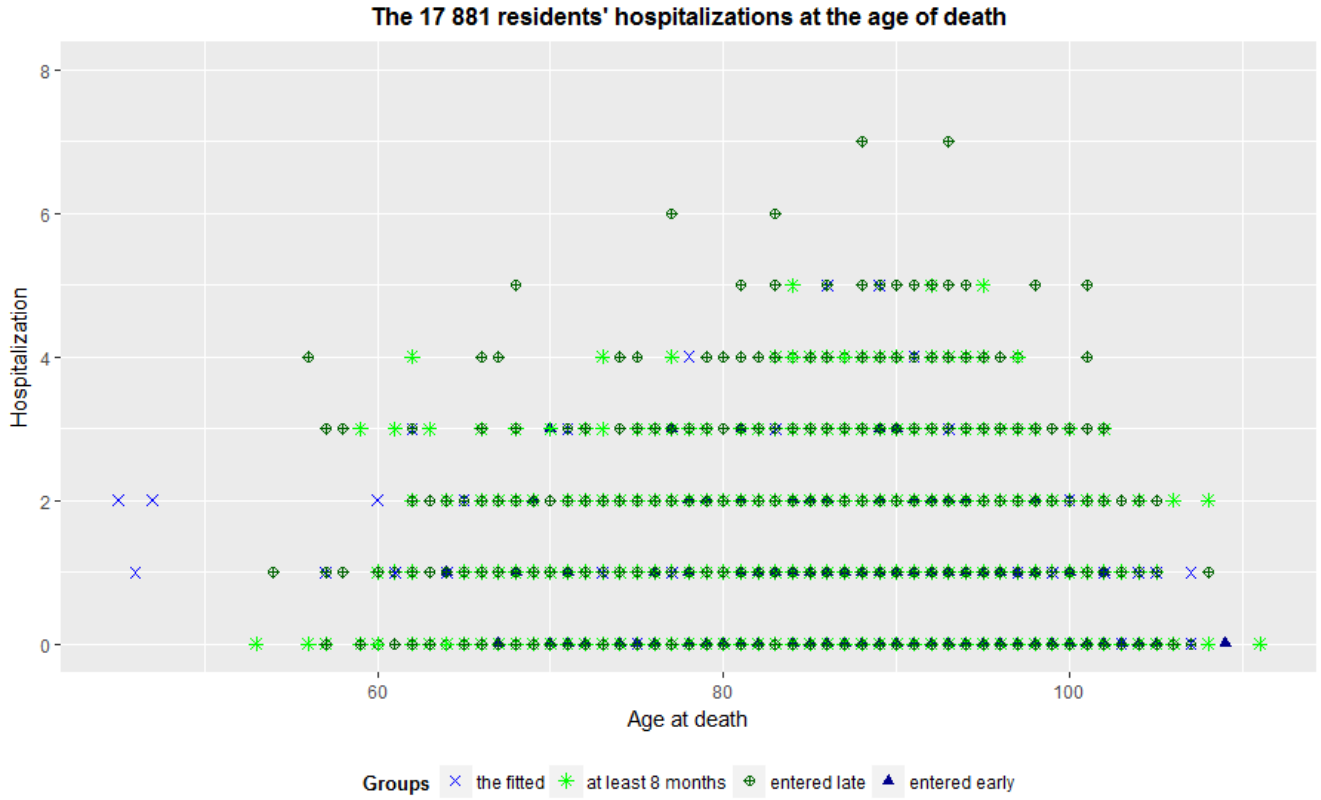
associations de syndromes particulièrement saillants en fin de vie confirment cette analyse (figures 20 – 23).





Figures 20 - 21: Associations syndromiques (démence, dépression) et (problèmes cutanés, altération de l'état général) des 17 881 résidents Korian en France décédés entre le 01/11/2010 et le 26/02/2017.





Figures 22 - 23: Associations syndromiques (douleurs, attitude de refus) et (hospitalisations et âge du décès) des 17 881 résidents Korian en France décédés entre le 01/11/2010 et le 26/02/2017

Quels que soient les syndromes saillants examinés, on retrouve toujours une majorité de résidents du groupe entrés tardivement.

Il semble donc que les tendances actuelles, à savoir rester le plus longtemps possible chez soi avant de partir en institution, ne soient pas forcément les plus judicieuses d'un point de vue bien-être ni d'efficacité des soins [75]. Ce travail vient enrichir celui déjà amorcé lors de l'examen de la fin de vie des résidents vaccinés de manière répétée contre la grippe.

4-Comparaison avec d'autres systèmes existants - Perspectives

4-1 De l'EHR DRI à la BBV

Alors que le DRI était organisé en silos, un silo avec les dossiers administratifs, un silo avec les prescriptions (non traité ici), un silo avec les transmissions et la prise en charge au quotidien, nous avons choisi de n'extraire que les informations à la fois bien alimentées et à visée de santé publique pour les personnes âgées, c'est-à-dire capable d'apporter de nouvelles données pour mieux protéger la communauté des résidents et améliorer leur santé [76]. Par exemple le SI Medissimo contenant les prescriptions n'était ni alimenté dans tous les établissements, ni de manière homogène. Nous avons donc pris le parti de ne pas utiliser ces données. Ensuite, pour la partie administrative, nous nous sommes concentrés sur l'essentiel d'un point de vue sanitaire, ainsi nous n'avons pas exploité les données socio-économiques qui figuraient dans un autre silo. A partir du SI DRI nous avons donc pu construire un nouveau SI à la fois beaucoup plus souple, léger et versatile. Cela nous a permis, dans la foulée, de bâtir une large cohorte de personnes âgées. Une fois réglés les problèmes de censure des informations à gauche, pour les résidents entrés avant la mise en place du DRI, et à droite, pour les résidents entrés récemment et donc avec une durée de séjour incomplète (voir figure 2 § 1-3-3-3), nous avons réussi, pour une large part de cette population de résidents en EHPAD, à retracer leur séjour complet et leur trajectoire de santé dans sa globalité. Chaque résident était donc défini :

1. - d'abord par son statut à l'entrée : date d'entrée et date de naissance, sexe, GIR et index d'EHPAD ;
2. - ensuite par ses syndrômes au fil de l'eau : dates et syndromes;
3. - enfin, par ses hospitalisations successives et décès éventuels.

Le fait d'extraire les hospitalisations et les décès des tables *hospitalisations* et *deces* alors que l'information était disponible dans les transmissions, nous a permis de récupérer des informations 'dures' et fiables avec des dates précises. En effet, dans les transmissions, il peut

arriver que l'information soit omise ou plus souvent répétée à mauvais escient.

Enfin, l'information quotidienne au niveau résident, soit le niveau 'micro', est complétée par les effectifs syndromiques hebdomadaires par établissement, soit le niveau 'macro' : chaque EHPAD appartenant à l'une des onze régions, permet des niveaux d'agrégation à la fois temporels (par mois, par an ou pour une période prédéfinie) et géographiques (par établissement, par région ou pour une zone prédéfinie).

4-2 Ce que la BBV a permis

A l'image des cohortes de naissance, les cohortes de personnes âgées permettent d'adopter une approche dynamique de l'évolution d'une population. Alors que les premières s'intéressent au développement de l'enfant, les secondes étudient les processus d'altération de la santé des personnes âgées. Et les deux essaient de mieux comprendre les trajectoires de vie à long terme de leurs populations [77]. C'est ainsi donc que la BBV a été imaginée, puis conçue.

Par exemple dans l'étude sur les vaccinations (voir figure 13 § 3-5-3-2), nous avons extrait par requêtes, à partir d'un événement syndromique particulier, ici la vaccination contre la grippe, un sous-groupe de résidents vaccinés à une date donnée. De là, nous avons retrouvé leurs caractéristiques socio-démographiques : sexe, tranche d'âge, puis nous les avons suivis, d'abord rétrospectivement jusqu'à leur entrée en institution, puis prospectivement, à intervalles réguliers, c'est-à-dire à chaque mention de vaccination détectée. Avec cette manière de faire, à l'entrée, puis à chaque vaccination et en définissant des fenêtres de 100 jours, nous avons obtenu des 'images syndromiques' de l'état sanitaire de nos résidents vaccinés, sortes de 'flash-santé'.

Au travers de l'étude sur le cancer, nous avons montré qu'il était possible de suivre des trajectoires de vie complètes de résidents, de retracer leurs événements de vie, comme les hospitalisations ou les chutes, de retrouver certaines prescriptions ainsi que de nombreuses comorbidités et les circonstances de leur décès (voir figures 9 et 10 § 3-5-1-2). Toutes ces données devraient permettre à terme de qualifier des trajectoires de vie en construisant par

exemple des indicateurs synthétiques. Parmi les indicateurs envisagés, un profil syndromique la semaine d'entrée, puis à l'issue des 100 premiers jours, destiné à prédire la durée de séjour ou les hospitalisations futures. Ces indicateurs à leur tour pourraient fournir les premiers outils pour envisager une véritable médecine personnalisée. Comme expliqué dans [78] la médecine personnalisée peut être considérée comme une extension des approches traditionnelles pour comprendre et traiter les maladies mais avec une plus grande précision, permettant aux médecins de définir des plans de prévention et de monitoring. Actuellement le focus concerne essentiellement les prescriptions médicamenteuses [78 – 79] mais les techniques non-médicamenteuses sont de plus en plus utilisées chez Korian et ailleurs en particulier chez les personnes atteintes de démence [80 – 82].

Ensuite, bien que nos 26 syndromes (hospitalisations et décès exclus) ne soient pas à proprement parler des cas au sens médical, il a été possible en utilisant des algorithmes bâtis sur la reconnaissance de motifs et un processus assimilable à une conférence de consensus, sur les données textuelles des cinq saisons épidémiques précédentes de grippe et de gastro-entérites aiguës, de suivre la dernière saison grippale en temps quasi-réel. Nos résultats, présentés dans le deuxième article, ont été convergents avec ceux du réseau Sentinelles en termes d'intensité de l'épidémie de grippe, mais avec une anticipation de plus de 2 semaines. Toute anticipation d'épidémie permettant d'organiser les soins en amont et d'atténuer son intensité en suivant les préconisations des agences régionales de santé [83], les épidémies à venir devraient pouvoir être mieux anticipées et ainsi permettre de sauver des vies.

Enfin, l'étude sur la fin de vie, a montré que l'on pouvait générer une typologie de la fin de vie (voir la table 7 § 3-5-4-2), reflet souvent d'une typologie de début de séjour amorcée dans l'étude sur les vaccinations pour le sous-groupe des résidents vaccinés contre la grippe (voir la figure 14 § 3-5-3-2). Les travaux doivent se poursuivre, mais il semble qu'une période de 100 jours permette de collecter un volume suffisant de d'informations syndromiques pour pouvoir brosser un profil de santé des résidents. Cela ouvre la voie à la prédiction de la durée de séjour dès les premières semaines.

Enfin et surtout, la construction de la BBV n'a nécessité aucune ressource supplémentaire de la part du personnel soignant puisque toutes les données recueillies ont été extraites à partir du SI

de santé existant. Une fois l'outil d'extraction réalisé au moyen de la plateforme Pentaho® et un silo de stockage pour les quatre tables de données alloué, la BBV peut être enrichie chaque semaine des nouveaux résidents entrés la semaine précédente et des données syndromiques de la population des résidents présents et vivants dans les établissements utilisant le logiciel de gestion sur lequel s'appuie l'outil de requêtage. Cette manière de procéder permet donc de bâtir assez rapidement, une fois l'outil de requêtes construit, une cohorte de taille assez importante avec des moyens informatiques relativement réduits et des ressources humaines faibles.

4-3 Ce que la BBV n'est pas

Bien que la BBV ait prouvé son utilité et son efficacité en tant qu'outil de surveillance contre la grippe et la gastro-entérite aiguë, ou pour d'autres pathologies comme les chutes, ce n'est pas un SSS au sens habituel du terme. Contrairement aux quatre exemples exposés plus haut : SurSaUD (figure 5 § 2-2-3), Sentinelles, ESSENCE et ESPnet (figures 6 et 7 § 2-2-3), nos syndromes ne s'appuient pas sur des cas définis cliniquement et sont donc moins précis. Néanmoins, l'étude de la précision des syndromes grippe et gastro-entérite aiguë faite dans le deuxième article a montré la fiabilité du système. En effet, nous avons obtenu des précisions de 98% et 96% pour la définition de nos syndromes grippe et GEA durant les semaines pics des épidémies respectives de l'hiver 2016-2017, sans que l'algorithme construit sur les saisons antérieures ait subi de modifications pour la nouvelle saison.

Ce n'est donc pas une cohorte de personnes âgées au sens habituel du terme [77], notamment si on compare la BBV aux cohortes médicales françaises présentées table 2 § 2-3-2. Pour générer nos données nous n'avons procédé ni à des examens cliniques, ni réalisé des prélèvements biologiques, ni utilisé des auto-questionnaires [84], ni enfin étudié un effet d'exposition à des facteurs de risque [85] ou recherché des associations entre plusieurs pathologies [86].

Ce n'est pas non plus un panel bien que la BBV en présente certains aspects de ceux présentés table 3 § 2-3-3. Ainsi, à aucun moment, nous n'avons procédé à des entretiens périodiques sur

des sujets récurrents comme le statut socio-économique, les relations sociales ou l'activité professionnelle antérieure et recherché des facteurs de risque éventuellement associés [87].

4-4 Perspectives

L'ensemble de ces travaux ont montré l'étendue des sujets possibles à explorer en santé publique pour le sujet âgé et ce, sur une cohorte de 41 000 personnes sur une période de presque 7 années.

Bien que nous exploitions des données surtout textuelles, nos variables syndromiques possèdent un véritable versant quantitatif et structuré qui permet d'examiner des changements au cours du temps, pour différentes tranches d'âge, ou différents sous-groupes et ce sur toute la durée du séjour.

Mais nos variables syndromiques gardent un versant purement qualitatif à l'aide de leur description textuelle toujours disponible. Ces données textuelles ont permis de générer un thésaurus qui peut être enrichi à tout moment par de nouvelles variables et/ou de nouveaux concepts. Ainsi, nous venons d'ajouter 2 nouveaux syndromes *apathie* et *parkinson* dans le cadre d'un nouveau partenariat avec l'association FranceParkinson [88]. Le nouveau syndrome *parkinson* pourra améliorer les connaissances et permettre un accompagnement renforcé des résidents atteints de cette pathologie.

Nous allons également tenté de rapprocher nos données syndromiques *demence* avec d'autres données syndromiques dans le cadre du partenariat avec l'association FranceAlzheimer, alors qu'un partenariat a déjà permis de travailler sur les techniques non médicamenteuses chez les résidents souffrant de la maladie d'Alzheimer [89 - 90].

Les travaux amorcés sur un *syndrome de fragilité* pourront se poursuivre en travaillant avec les médecins gériatres en interne et en externe, de même pour les travaux de prédiction de la durée de séjour. Nous pourrons alors probablement déterminer de nouveaux leviers dans la prise en charge et ainsi progresser dans les soins apportés au quotidien aux résidents en institution chez Korian et ailleurs.

Il reste que la BBV s'appuyait sur le SI EasySoins® qui prend fin cette année pour être remplacé par un nouveau SI NetSoins® avec une indexation et une organisation des informations de santé sensiblement différentes. Néanmoins ce nouveau SI sera déployé sur l'ensemble des

EHPAD Korian soit deux fois plus d'établissements. Ce sera alors l'occasion de développer de nouveaux algorithmes syndromiques plus performants et d'enichir notre liste de syndromes [91 - 92].

Enfin nous tenterons de rapprocher le silo des prescriptions, à l'heure actuelle gérées par les applications Medissimo® et Robotics®. Aujourd'hui, seule Medissimo possède un SI avec les identités des résidents et permettrait de faire de la véritable médecine personnalisée. Cela ouvrirait également le champ de la recherche aux études cliniques avec traitements médicamenteux sur les sujets âgés, population où le recrutement de patients est souvent difficile.

Cependant, au vu de la nouvelle loi sur la protection des données sensibles GDPR qui entrera en vigueur le 25 mai 2018 [93], s'il devient possible de fusionner le silo de la BBV avec celui de Medissimo, cela nécessitera à la fois d'en informer chaque résident de manière nominative, comme les contrats d'hébergement le prévoient aujourd'hui, et de définir des nouveaux processus d'anonymisation, permettant à la fois de suivre les traitements de chaque résident et leur prise en charge tout en respectant leur anonymat.

De nouvelles réponses techniques devront donc être trouvées alors que de nombreux défis restent posés en termes d'interopérabilité entre les différentes applications EHR exploitant des données socio-économiques, comportementales, environnementales et de santé pour favoriser une médecine véritablement 'personnalisée' et un système de santé authentiquement 'apprenant' [94].

Références

1. INSEE : Nathalie Blanpain, Olivier Chardon, division Enquetes et etudes demographiques in Insee Premiere N¢X1320 - octobre 2010 - Projections de population a l'horizon 2060
2. Ministere des affaires sociales et de la sante : Vieillissement actif et solidarite entre les generations. 14 decembre 2011
3. Plan Solidarité grand âge http://www.cnsa.fr/parcours-de-vie/plans-de-sante-publique/plan-solidarite-grand-age
4. QUENTIN (B.). (2013). Ethique et vieillissement. GERONTOLOGIE ET SOCIETE
5. Arai H, Ouchi Y, Yokode M, Ito H, Uematsu H, Eto F, Oshima S, Ota K, Saito Y, Sasaki H, Tsubota K, Fukuyama H, Honda Y, Iguchi A, Toba K, Hosoi T, Kita T; Members of Subcommittee for Aging.Toward the realization of a better aged society: messages from gerontology and geriatrics.
6. Washko MM, Campbell M, Tilly J. Accelerating the translation of research into practice in long term services and supports: a critical need for federal infrastructure at the nexus of aging and disability.
7. Dépendance 4 cohortes - Rapport final INSERM Projet Dépendance 4 cohortes épidémiologiques Haute Normandie, Paquid, 3Cités et AMI - Novembre 2011 http://www.cnsa.fr/documentation/_projet_dependance_4_cohortes_cnsa_version_finale_nov2011_.pdf
8. Kristel Gilis, Direction technique Agirc-Arrco - La demographie et l'etat de sante des personnes agees de 80 ans et plus
9. Loi n° 2009-879 du 21 juillet 2009 portant reforme de l'hopital et relative aux patients, a la sante et aux territoires
10. Brazil K., Maitland J., Ploeg J. et al. Identifying Research Priorities in Long Term Care Homes JAMDA Vol. 13, Issue 1, Pages 84.e1-84.e4
11. Données épidémiologiques et sociologiques – Faculté de médecine de Toulouse http://www.medecine.ups-tlse.fr/dcem3/module05/54_poly_viellissement_2.pdf
12. Bouyer J. Epidemiologie, principes et methodes quantitatives, 2009
13. http://invs.santepubliquefrance.fr/Dossiers-thematiques/Environnement-et-sante/Syndromes-collectifs-inexpliques
14. Projet HygEHPAD http://mesurs.cnam.fr/spip.php?article31
15. Comment Passer d'un Usage Individuel du Dossier Résident à un Bénéfice Collectif ? workshop EPICLIN 9 2015 à Montpellier Mai 2015
16. New Methods to Evaluate Physiotherapy Care in Nursing Homes workshop Nursing Home Research à Toulouse Décembre 2015
17. Benaim C., Froger J., Compan B. et al. Évaluation de l'autonomie de la personne âgée. Annales de Réadaptation et de Médecine Physique Volume 48, Issue 6, July 2005, Pages 336-340. https://doi.org/10.1016/j.annrmp.2005.04.005
18. Delespierre T, Denormandie P, Bar-Hen A, Josseran L. Empirical Advances with Text Mining of Electronic Heath Records. Journal: BMC Medical Informatics and Decision Making DOI: 10.1186/s12911-017-0519-0
19. Institute of Medicine, authors. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academies Press; 2001. Jun 1, p. 15. http://books.nap.edu/books/0309072808/html/index.html.
20. Gunter, Tracy D; Terry, Nicolas P. "The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions". Journal of Medical Internet Research. 7 (1): e3. PMC 1550638. PMID 15829475. doi:10.2196/jmir.7.1.e3.
21. Apsden P., Corrigan JM, Wolcott J, Erickson SM, editors. Committee on Data Standards for Patient Safety, Board on Health Care Services, Institute of Medicine, authors. Patient Safety: Achieving a New Standard for Care. Washington, DC: The National Academies

Press; 2004. p. 4. http://www.nap.edu/catalog/10863.html.
22. Kausdal R., Kaushal R, Bates DW. Computerized Physician Order Entry (CPOE) with Clinical Decision Support Systems (CDSSs) In: Shojania KG, Duncan BW, McDonald KM, Wachter RM, editors. Making Health Care Safer: A Critical Analysis of Patient Safety Practices. Evidence Report/Technology Assessment, No. 43, Chap 6. AHRQ Publication No - 01-E058 (Prepared by the University of California at San Francisco - Stanford University Evidence-based Practice Centre) Rockville, MD: Agency for Healthcare Research and Quality; 2001. [2004 Dec 16]. http://www.ahrq.gov/clinic/ptsafety/
23. Le programme de médicalisation des systèmes d'information (PMSI) http://www.caducee.net/DossierSpecialises/systeme-information-sante/pmsi.asp
24. Logiciels d'Aide à la Prescription médicale certifiés selon le référentiel de la HAS : https://www.has-sante.fr/portail/jcms/c_672760/fr/logiciels-d-aide-a-la-prescription-pour-la-medecine-ambulatoire-certifies-selon-le-referentiel-de-la-has
25. What is a personal health record?". HealthIT.gov. Office of the National Coordinator for Health IT. https://www.healthit.gov/providers-professionals/faqs/what-personal-health-record
26. Le dossier médical partagé http://www.dmp.gouv.fr/
27. Lamberts M. Electronic Health Records The View From the Trenches. pp 9-22. https://www.cdc.gov/cdcgrandrounds/pdf/GREHRAllFINAL21Jul2011.pdf
28. Mostashari F. Public Health and Meaningful Use of Electronic Health Records Opportunities, Realities, and a Proposed Approach. Pp 50 – 67. https://www.cdc.gov/cdcgrandrounds/pdf/GREHRAllFINAL21Jul2011.pdf
29. Birkhead G.S., Klompas M., Shah N.R. Uses of Electronic Health Records for Public Health Surveillance to Advance Public Health. Annu. Rev. Public Health 2015. 36:345–59. doi: 10.1146/annurev-publhealth-031914-122747
30. Tomines A., Readhead H., Readhead A. et al. Applications of Electronic Health Information in Public Health: Uses, Opportunities & Barriers EGEMS (Wash DC). 2013; 1(2): 1019. doi: 10.13063/2327-9214.1019
31. Loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. Version consolidée au 28 octobre 2017 https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000886460
32. https://www.cnil.fr/fr/vos-demarches-en-ligne
33. Nsubuga P., White M.E., Thacker S.B. et al. Chapter 53. Public Health Surveillance: A Tool for Targeting and Monitoring Interventions Disease Control Priorities in Developing Countries 2nd edition. Jamison DT, Breman JG, Measham AR, et al., editors. Washington (DC): The International Bank for Reconstruction and Development / The World Bank; New York: Oxford University Press; 2006.
34. Astagneau P., Ancelle T. La surveillance épidémiologique. Surveillance syndromique (Chapitre 12), 2011, Ed. Lavoisier.
35. Josseran L, Fouillet A. La surveillance syndromique : bilan et perspective d'un concept prometteur. Rev Epidemiol Sante Publique (2013), http://dx.doi.org/10.1016/j.respe.2013.01.094
36. Josseran L., Fouillet A., Caillère N. et al. Assessment of a Syndromic Surveillance System Based on Morbidity Data: Results from the Oscour® Network during a Heat Wave. PLoS ONE 5(8): e11984. doi:10.1371/journal.pone.0011984
37. Fouillet A., Medina S., Medeiros H. et al. La Surveillance Syndromique en Europe: le projet triple-S. BEH 3-4 | 21 janvier 2014 | p 79, La surveillance syndromique en France en 2014.
38. http://invs.santepubliquefrance.fr/fr/Espace-professionnels/Surveillance-syndromique-SurSaUD-R
39. Josseran L. La Surveillance Syndromique à l'InVS. La veille sanitaire fondée sur l'enregistrement automatique des données métiers. InVS septembre 2010. http://esante.gouv.fr/sites/default/files/L.%20JOSSERAN%20-%20INVS%20-%20La%20surveillance%20syndromique.pdf.
40. https://websenti.u707.jussieu.fr/sentiweb/?page=presentation
41. Lombardo J.S. JOHNS HOPKINS APL TECHNICAL DIGEST, VOLUME 24, NUMBER 4 (2003)

42. Happel Lewis S.L., Hurt-Mullen K., Loschen W. et al. Active Biosurveillance in an urban metropolitan area. (Montgomery County area) November 18, 2003
43. ESSENCE Overview - The Evolution of ESSENCE pp 1-17 https://www.cdc.gov/nssp/documents/essence-training-presentation-phi-conference.pdf
44. Lazarus R., Klompas M., Campion F.X. et al. Electronic Support for Public Health: Validated Case Finding and Reporting for Notifiable Diseases Using Electronic Medical Data. J Am Med Inform Assoc. 2009;16:18 –24. DOI 10.1197/jamia.M2848
45. Klompas M., Klompas M, McVetta J, Lazarus R, Eggleston E, Haney G, et al. 2012. Integrating clinical practice and public health surveillance using electronic medical record systems. Am. J. Public Health 102(S3):S325–32
46. Lazarus R., Klompas M, Campion FX, McNabb SJ, Hou X, et al. 2009. Electronic support for public health: validated case finding and reporting for notifiable diseases using electronic medical data. J. Am. Med. Inf. Assoc. 16(1):18–24
47. Goldberg M., Zins M. Les études de cohorte: principes et méthodes. Apport des cohortes à la connaissance de la santé. adsp n°78 mars 2012 pp14-20
48. Van Ganse E., Belhassen M. SNIIRAM: Primary and Secondary Case Resource Use in France. January 23, 2015. https://fr.slideshare.net/RespiratoryEffectivenessGroup/sniiram-primary-and-secondary-care-resource-use-in-france (Accessed July 6, 2017)
49. Moulis G., Lapeyre-Mestre M., Palmaro A. et al. Review - French health insurance databases: What interest for medical research? _ Les bases de données de l'assurance maladie française : quel intérêt pour la recherche médicale ? La Revue de médecine interne 36 (2015) 411–417
50. Laurie H. Panel studies. Panel studies. Oxford bibliographies Online 2013. DOI: 10.1093/obo/9780199756384-0108
51. Le portail Epidémiologie France https://epidemiologie-france.aviesan.fr/
52. The US Health and Retirement Study http://hrsonline.isr.umich.edu/
53. The English Longitudinal Study of Ageing .http://www.natcen.ac.uk/taking-part/studies-in-field/elsa-50plus-health-and-life/
54. Banks J., Batty G.D., Nazroo J. et al. The dynamics of ageing.: Evidence from the English Longitudinal Study of Ageing 2002-15 (Wave 7) October 2016. http://www.elsa-project.ac.uk/publicationDetails/id/8696
55. The Survey of Health, Ageing and Retirement in Europe http://www.share-project.org/home0.htmlhttps://www.cdc.gov/obesity/data/surveillance.html
56. Gibbons C., Richards S., Valderas J.M. et al. Supervised Machine Learning Algorithms Can Classify Open-Text Feedback of Doctor Performance With Human-Level Accuracy J. Med. Internet Res 2017 vol 19 iss 3 e65 p.1 DOI:10.2196/jmir.6533
57. http://www.has-sante.fr/portail/upload/docs/application/pdf/2013-04/referentiel_concernant_levaluation_du_risque_de_chutes_chez_le_sujet_age_autonome_et_sa_prevention.pdf
58. http://www.has-sante.fr/portail/upload/docs/application/pdf/2009-06/chutes_personnes_agees_synthese.pdf
59. http://ageing.oxfordjournals.org/content/38/2/194.full
60. http://www.cdc.gov/homeandrecreationalsafety/falls/nursing.html
61. Fulop T., Pawebec G, Castle S. et al. Immunosenescence and Vaccination in Nursing Home Residents. Aging and Infectious Diseases 2009 ; 48 :443-8 DOI : 10.1086/596475
62. Scemama A. La vaccination des personnes âgées et les recommandations du calendrier vaccinal 2009 concernant le personnel et les résidents en EHPAD. Journée du CCLIN Paris Nord du 6 mai 2009.
63. Dorrington M. G., Bowdish D.M.E. Immunosenescence and novel vaccination strategies for the elderly. frontiers in IMMUNOLOGY 28 June 2013 DOI : 10.3389/fimmu.2013.00171
64. Loukov D., Naidoo A., Bowdish D.M.E. Immunosenescence: implications for vaccination programs in the elderly. Development and Therapy 6 August 2015

http://dx.doi.org/10.2147/VDT.S63888
65. Grippe Bulletin hebdomadaire Semaine 09 08/03/2017 Santé Publique France. http://invs.santepubliquefrance.fr/Dossiers-thematiques/Maladies-infectieuses/Maladies-a-prevention-vaccinale/Grippe/Grippe-generalites/Donnees-de-surveillance/Archives/Bulletin-epidemiologique-grippe-semaine-9.-Saison-2016-2017 accessed 30th March 2017
66. Falissard B. Déploiement d'une matrice de corrélation sur la sphère unité de R 3. Revue de statistique appliquée„ tome 43, n° 2 (1995), p. 35-48.
67. Husson J.F., Josse J., Pagès J. Principal component methods – hierarchical clustering – partitional clustering : why would we need to choose for visualizing data ? Technical report Agrocampus 2010.
68. https://cicelysaundersinternational.org/research/
69. Zheng L., Finucane A. M., Oxenham D. et al. How good is primary care at identifying patients who need palliative care? A mixed methods study. EUROPEAN JOURNAL OF PALLIATIVE CARE , 2013; 20 (5)
70. DOC d'accompagnement Korian TECHNIQUE 5-1 Le chariot Korian Cicely
71. Monod S., Rochat E., Büla C. et al. Désir de mort chez les personnes âgées : que savons-nous de cette réalité ? Forum Med Suisse 2013 ;13(51-52) ; 1063-1064
72. Fizzala A. Les durées de séjour en EHPAD – Les dossiers de la DREES mai 2017, n°15. http://drees.solidarites-sante.gouv.fr/IMG/pdf/15_dossiers_drees_final.pdf, accédé le 13 octobre 2017
73. Allen D. N., Golstein G., Cluster Analysisin Neuropsychological Research : Recent Applications, DOI 10.1007/978-1-4614-6744-1_1 Springer Science+Business Media New York 2013
74. https://cran.r-project.org/web/packages/fastcluster/index.html, accédé le 13 octobre 2017
75. Qualité de vie en Ehpad (volet 1) - De l'accueil de la personne à son accompagnement. Recommandations de Bonnes Pratiques Professionnelles ANESM Décembre 2010 http://www.anesm.sante.gouv.fr/IMG/pdf/reco_qualite_de_vie_ehpad_v1_anesm.pdf, accédé le 13 octobre 2017
76. https://www.cdcfoundation.org/what-public-health
77. Singh-Manoux A. Les cohortes au niveau international : histoire et perspective. adsp n°78 mars 2012 pp31-33
78. Vogenberg F. R., Isaacson Barash C., Pursel M. Personalized Medecine Part I : Evolution and Development into Theranostics. P&T. October 2010. Vol 35 N°10
79. Vogenberg F. R., Isaacson Barash C., Pursel M. Personalized Medecine Part II : Ethical, Legal, and Regulatory Issues. P&T. November 2010. Vol 35 N°10
80. Pancrazi M. P., Métais P. Prise en charge non médicamenteuse dans les démences sévères. Psychologie NeuroPsychiatrie Vieillissement 2005 ; vol. 3 (Suppl. 1) : S42-S50
81. Andreeva V., Dartinet-Chalmery V., Kloul A. "Snoezelzn" ou les effets de la stimulation multisensorielle sur les troubles du comportement chez les personnes âgées démentes à un stade avancé. NPG Volume 11, Issue 61. February 2011, pages 24-29 DOI : 10.1016/j.npg.2010.09.003
82. Wenisch E., Stocker A., Bourrellis C. et al. Méthode de prise en charge globale non médicamenteuse des patients déments institutionnalisés A global intervention program for institutionalized demented patients. Revue Neurologique Volume 161, Issue 3, March 2005, Pages 290-298 https://doi.org/10.1016/S0035-3787(05)85035-0
83. http://france3-regions.francetvinfo.fr/centre-val-de-loire/epidemie-grippe-ars-renouvelle-ses-recommandations-prevention-1170701.html (Accessed June 23, 2017°)
84. Goldbohm R. A., van den Brandt P. A., Brants H. A M. et al. Validation of a dietary questionnaire used in a large-scale prospective cohort study on diet and cancer. European Journal of Clinical Nutrition (1994) 48, 253 – 265.
85. Gustavsson P., Hogstedt C. A cohort study of swedish capacitor manufacturing workers exposed to polychlorinated biphenyls (PCBs) American Journal of Industrial Medecine 1997 DOI: 10.1002/(SICI)1097-0274(199709)32:3<234::AID-AJIM8>3.0.CO;2-X

86. Shah A. D., Langenberg C., Rapsomaniki E. et al. Type 2 diabetes and incidence of cardiovascular diseases: a cohort study in 1.9 million people. Lancet Diabetes Endocrinal 2015; 3-105-13 http://dx.doi.org/10.1016/S2213-8587 (14) 70219-0
87. Loftus C., Yost M., Sampson M. et al. Regional PM2.5 and asthma morbidity in an agricultural community: A panel study. http://doi.org/10.1016/j.enres.2014.10.030
88. https://www.silvereco.fr/lassociation-france-parkinson-et-korian-signent-un-partenariat/3190051
89. Gueyraud C., Anaut M., Denormandie P. et al. Jeu et Maladie d'Alzheimer. Le cadre ludique dans la prise en charge de la démence. ERES « Empan » 2016/2 n° 102 pages 116 à 122. https://www.cairn.info/revue-empan-2016-2-page-116.htm
90. Gueyraud C., Anaut M., Denormandie P. et al. Dementia and non-pharmacological therapy, the effectiveness of play. Soins Gerontol. 2017 May - Jun;22(125):27-31. doi: 10.1016/j.sger.2017.03.006.
91. Abbé A. Analyse de données textuelles d'un forum médical pour évaluer le ressenti exprimé par les internautes au sujet des antidépresseurs et des anxyolitiques. https://www.theses.fr/196877113
92. https://gargantext.org/
93. Conseil européen. « Le règlement général sur la protection des données », http://www.consilium.europa.eu/fr/policies/data-protection-reform/data-protection-regulation/ 11 avril 2016.
94. Evans R. S. Electronic Health Records : Then, Now, and in the Future. Yearb Med Inform. 2016 May 20; Suppl 1:S48-61. Doi: 10.15265/IYS-2016-s0006.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique

Titre : Du dossier résident informatisé à la recherche en santé publique : application des méthodes de surveillance en temps réel à des données médico-sociales de la personne âgée et exploration de données de cohorte pour la santé publique.

Mots clés : EES, requêtes SQL, santé publique, SSS, textmining, EHPAD

Résumé: La France connaît un vieillissement accru de sa population. Même si de nombreuses cohortes de personnes âgées existent déjà dans le monde et que la part de ces personnes vivant dans des structures d'hébergement collectif (EHPAD, cliniques de soins de suite) augmente, la connaissance de cette population reste lacunaire. Certains grands groupes privés de maisons de retraite s'équipent d'entrepôts de bases de données qui comprennent à la fois des données structurées décrivant les résidents et leurs traitements et pathologies, mais aussi des données textuelles au format libre. Saisies par le personnel soignant, celles-ci permettent d'analyser la prise en charge et les soins.

Le but de cette thèse a été de transformer ces données textuelles pour les intégrer dans une base de données d'un nouveau type, de constituer une cohorte de santé publique à partir des résidents Korian et d'organiser un système de surveillance grippe et gastro-entérites syndromique fonctionnel pour la personne âgée.

Title : From a nursing home electronic resident data warehouse to public health research: applying public health surveillance systems methods to a real time long term care database and building a resident cohort study.

Keywords: EHR, SQL requests, public health, SSS, textmining, nursing home

Abstract: French population is rapidly aging. Senior citizens ratio is increasing and our society needs to rethink its organization to know better this fast growing population group. Even if numerous cohorts of elderly people already exist worldly and as they live in growing numbers in nursing homes and outpatient treatment clinics, knowledge of this population is still missing. Today several health and medico-social structure groups such as Korian invest in big relational data bases enabling them to get real-time information about their residents. They contain at the same time structured data describing them as well as their treatments and pathologies, but also free-textual data detailing their daily care by the medical staff.

This thesis objective was to transform these textual data to integrate them into a new type of database, enabling a public health cohort inception from Korian residents and organize an influenza and gastro-enteritis operational surveillance system for elderly people.