Surveillance réactive de la mortalité fondée sur les causes médicales de décès en texte libre : application de méthodes de traitement automatique des langues

Yasmine Baghdadi

UNIVERSITÉ PARIS-EST

ÉCOLE DOCTORALE DE SANTE PUBLIQUE (ED570)

Thèse de doctorat

Santé publique-Epidémiologie

BAGHDADI Yasmine

SURVEILLANCE RÉACTIVE DE LA MORTALITÉ FONDÉE SUR LES CAUSES MÉDICALES DE DÉCÈS EN TEXTE LIBRE : APPLICATION DE MÉTHODES DE TRAITEMENT AUTOMATIQUE DES LANGUES.

Thèse dirigée par Dr GALLAY Anne

Co-encadrée par Dr FOUILLET Anne

Soutenue le 16 Octobre 2019

Jury :

Pr Philippe Quénel : Président

Dr Marie-Hélène Metzger : Rapporteur

Pr Jean-Baptiste Meynard : Rapporteur

Pr Marc Cuggia : Examinateur

Dr Anne Gallay : Directrice de thèse

Dr Anne Fouillet : Co-encadrante de thèse

Remerciements

Tout d'abord je tiens à remercier Anne Gallay et Anne Fouillet, Directrice et Co-encadrante de ma thèse.
Merci de m'avoir permis de travailler sur ce sujet riche et multidisciplinaire, de m'avoir guidée et conseillée, encouragée et poussée à me dépasser pendant ces 3 années, toujours avec bienveillance.

Je tiens à remercier particulièrement Anne F. Merci pour ton extrême patience, je sais que je ne t'ai pas rendu la tâche facile. Tu as toujours été disponible et ton sens de la pédagogie et de la précision m'ont fait devenir meilleure.

Je remercie le Dr Marie-Hélène Metzger, le Pr Jean-Baptiste Meynard, le Pr Marc Cuggia et le Pr Philippe Quénel de me faire l'honneur de participer à mon jury de thèse.

Je tiens à remercier le CépiDc et plus particulièrement Grégoire Rey et Aude Robert. Merci pour toute l'aide que vous m'avez apportée durant cette thèse, aussi bien matérielle que pour nos échanges constructifs. Je vous remercie également d'avoir participé à mes COPIL et de toujours avoir soulevé les bonnes questions.

Je remercie également Cyril Grouin et Pierre Zweigenbaum du CNRS-LIMSI de m'avoir apporté leur expertise en TAL.
Merci Cyril de m'avoir accueillie dans ton cours pour m'apprendre les bases du TAL. Je te remercie aussi pour ton suivi attentif, ta disponibilité et pour nos échanges toujours enrichissants. Merci Pierre d'avoir partagé tes connaissances et ton expertise avec l'équipe et de t'être rendu disponible pour ce projet.

Je remercie Yann Le Strat, directeur de la DATA. Merci Yann, pour ton regard critique, tes conseils avisés et ta disponibilité.
Je remercie également Céline Caserio-Schönemann directrice de l'unité ABISS. Merci Céline pour ton éclairage et ta vision générale sur ce projet.

Je remercie particulièrement Alix Bourrée pour son travail. Merci Alix d'avoir travaillé à mes côtés pendant 6 mois et de m'avoir permis d'approfondir mes connaissances en TAL, sans toi le résultat n'aurait pas été le même.

Je tiens également à remercier Alexandre Cornec. Merci Alexandre d'avoir « donné vie » à ces résultats à travers l'application.

Je remercie Jean-Claude Desenclos, directeur scientifique de Santé publique France de m'avoir fait confiance et pour m'avoir permis de faire partie de la première promotion des doctorants de Santé publique France.

Je remercie tous les membres de mon COPIL, Catherine Ha, Linda Lasbeur, Aurélie Névéol, Didier Che, Frank Golliot, Alain Le Tertre d'avoir apporté leurs regards critiques sur mes travaux et leurs conseils.

Je remercie également toute l'équipe SurSaUD, Cécile, Isabelle, Marie-Michèle et Jérôme de m'avoir accueilli et soutenu pendant ces trois ans.

Plus personnellement, je tiens à remercier Charlotte Castel. Merci Charlotte pour tout le soutien que tu m'as apporté pendant cette dernière année, nos discussions, nos samedis et bien sûr merci pour tes pâtisseries.

Merci à toute l'équipe des coureurs de SPF, Amélie, Clémence, Emilie, Maelle, Côme, Duc, Edouard, Minh-Canh, Minh-Taï, Mohamed, Paul-Henri. Merci pour toutes ces sorties qui m'ont permis de décompresser et de ne pas lâcher.

Je tiens aussi à exprimer toute mon affection à ma famille, merci pour votre soutien. Maman et Myriam merci pour vos relectures !

Martine, Jean-Pierre, je vous remercie chaleureusement pour votre présence et votre soutien. Merci Martine pour la relecture.

Je remercie également tous mes proches et particulièrement Ambre, Charlotte, Hélène, Larissa, Pauline, Adrien, Pierre mais aussi Alix, Camille, Clémence, Olivia, Stéphanie, Adrian, Jérôme, Mathieu, les Nicolas, Olivier, Tristan. Merci à tous d'avoir été là et de m'avoir soutenue.

Enfin, je tiens à remercier celui qui m'a soutenue sans faille, au quotidien, dans les bons moments comme dans les moins bons. Yann, tu m'as épaulée durant ces 3 années, tu m'as supportée et as su me remotiver pour aller au bout de ce projet. Je ne te remercierai jamais assez. Merci d'être dans ma vie.

Résumé

A partir des décès certifiés électroniquement de 2012 à 2016 en France, la thèse vise à mettre en œuvre et évaluer les performances de méthodes de traitement automatique des langues, pour classer les causes médicales de décès disponibles en texte libre dans des regroupements syndromiques (RS) pertinents pour la surveillance réactive de la mortalité à visée d'alerte et d'évaluation d'impact sanitaire.

Près de 100 RS répondant aux objectifs ont été définis. Deux méthodes de classification ont été développées : une méthode à base de règles linguistiques et une méthode par apprentissage supervisé (SVM). Deux modèles SVM ont été développés utilisant différentes combinaisons de caractéristiques. Le développement et l'évaluation des performances des méthodes se sont appuyés sur 4 500 certificats de décès annotés. L'évaluation a porté dans un premier temps sur 7 RS, puis a été étendue à 60 autres RS. Les évolutions mensuelles des RS attribués par les méthodes ont été comparées sur l'ensemble des décès de 2012 à 2016 (204 000 décès).

La méthode par règles et le modèle SVM incluant l'ensemble des caractéristiques ont obtenu des performances élevées (F-mesure≥0,95) pour la classification des causes dans 31 RS. L'évolution temporelle de ces RS obtenus par les deux méthodes était comparable. En moyenne, les causes de décès au sein d'un certificat sont classées dans 3,7 RS. Une méthode de pondération équilibrée des RS a été proposée pour prendre en compte ces causes multiples lors de l'analyse en routine de la mortalité pour la surveillance à visée d'alerte et d'évaluation d'impact.

Ces résultats permettent d'améliorer la surveillance réactive actuellement fondée sur des données d'état-civil.

Mots clés : Surveillance syndromique, Mortalité, Causes médicales de décès, Traitement automatique des langues, Apprentissage automatique

REAL-TIME MORTALITY SURVEILLANCE BASED ON FREE-TEXT MEDICAL CAUSES OF DEATH: APPLICATION OF NATURAL LANGUAGE PROCESSING (NLP) METHODS

Abstract

Based on electronic death certificates from 2012 to 2016 in France, this thesis aims to implement and evaluate the performance of natural language processing methods, to classify medical causes of death in free-text format into relevant mortality syndromic groups (MSG) for reactive mortality surveillance and health impact assessment.

Close to 100 MSGs meeting the objectives of the surveillance were defined. Two classification methods were developed: a rule-based method and a supervised machine learning method (SVM). Two SVM models were developed using different combinations of features. The development and evaluation of the performances of the methods were based on 4,500 annotated death certificates. The evaluation was initially based on 7 MSGs and was then extended to 60 other MSGs. The variations of the monthly number of MSGs assigned by the two methods were compared using the whole death certificates from 2012 to 2016 (204,000 deaths).

The rule-based method and the SVM model including all the features obtained high performances (F-mesure≥0.95) for the classification of causes into 31 MSGs. The monthly variations of those MSGs were comparable. In average, a death certificate contains 3.7 MSG. We proposed a balanced weighting method of the MSG to take these multiple causes into account in the routine analysis of mortality for alert and impact assessment.

These results complete and enrich the reactive surveillance currently based on administrative data.

Keywords: Syndromic surveillance, Mortality, Medical causes of death, Natural language processing, Machine learning

Intitulé et adresse du laboratoire où la thèse a été effectuée

Santé publique France, Direction Appui, Traitements et Analyses des données (DATA),

Unité Applications, Big data, Surveillance Syndromique (ABISS)

12, rue du Val d'Osne

94410 Saint-Maurice

Productions scientifiques issues du travail de thèse

Articles publiés ou acceptés pour publication

Baghdadi Y, Gallay A, Caserio-Schönemann C, Fouillet A. Evaluation of the French reactive mortality surveillance system supporting decision making. *Eur J Public Health* 2018:cky251-cky251

Baghdadi Y, Bourrée A, Robert A, Rey G, Gallay A, Zweigenbaum A, Grouin C, Fouillet A. Automatic classification of free-text medical causes from death certificates for reactive mortality surveillance in France. Int J Med Inform. 2019;131:103915.

Baghdadi Y, Bourree A, Robert A, Rey G, Gallay A, Zweigenbaum P, et al. A New Approach to Compare the Performance of Two Classification Methods of Causes of Death for Timely Surveillance in France. Stud Health Technol Inform. 2019;264:925-9.

Robert A, Baghdadi Y, Zweigenbaum P, Morgand C, Grouin C, Lavergne T, Névéol, Fouillet A, Rey G. Développement et application de méthodes de traitement automatique des langues sur les causes médicales de décès pour la santé publique. *Bulletin Epidémiologique Hebdomadaire (Sous presse, publication Octobre 2019)*

Communications orales

Baghdadi Y, Gallay A, Caserio-Schönemann C, Fouillet A. A reactive mortality surveillance system in France for alert and decision-making. *Rev Epidemiol Sante Publique*. 2018;66:S257 European Congress of Epidemiology, du 4 au 6 Juillet 2018, Lyon (France)

Baghdadi Y, Bourrée A, Robert A, Rey G, Gallay A, Zweigenbaum A, Grouin C, Fouillet A. Automatic classification of free-text medical causes of death into mortality syndromic groups for reactive mortality surveillance in France
Congrès international de surveillance syndromique ISDS Du 30 Janvier au 1er Février 2019, San Diego (USA)

Baghdadi Y, Gallay A, Caserio-Schönemann C, Thiam MM, Fouillet A. A strategy of analysis of free-text french e-death certificates using machine learning
Congrès international de surveillance syndromique ISDS Du 30 Janvier au 1er Février 2019, San Diego (USA)

Baghdadi Y, Bourrée A, Robert A, Rey G, Gallay A, Zweigenbaum A, Grouin C, Fouillet A. A new approach to compare performance of two classification methods of causes of death for timely

surveillance in France.

Congrès international MedInfo Du 25 au 30 Août 2019, Lyon (France)

Communications affichées

Baghdadi Y, Gallay A, Caserio-Schönemann C, Fouillet, A. Environmental or infectious events: A threat for vulnerable population

European Public Health Conference (EPHC), du 1er au 4 Novembre 2017, Stockholm (Suède)

Baghdadi Y, Gallay A, Caserio-Schönemann C, Thiam M-M, Fouillet A. Towards real-time mortality surveillance by medical causes of death: A strategy of analysis for alert. Rev Epidemiol Sante Publique. 2018;66:S402.

European Congress of Epidemiology, du 4 au 6 Juillet 2018, Lyon (France)

Table des matières

Liste des Tableaux

Liste des Figures

Liste des Annexes

Liste des Abréviations

ARS : Agence Régionale de Santé

ASTER : Système d'Analyse et de Surveillance épidémiologique en TEmps Réel

CépiDc : Centre d'épidémiologie sur les causes médicales de Décès

CIM : Classification Internationale des Maladies

CLEF : Conference and Labs of the Evaluation Forum

CPS : Carte de Professionnel de Santé

CS2A : Collaboration de Services pour la Surveillance et l'Alerte

DGS : Direction Générale de la Santé

ESSENCE : Electronic Surveillance System for the Early Notification of Community-based Epidemics

GIP : Groupement d'Intérêt Public

Insee : Institut National de la Statistique et des Etudes Economiques

Inserm : Institut national de la santé et de la recherche médical

InVS : Institut de Veille Sanitaire

OMS : Organisation Mondiale de la Santé

OSCOUR : Organisation de la Surveillance Coordonnée des URgences

RNIPP : Répertoire National d'Identification des Personnes Physiques

RODS : Real-time Outbreak and Disease Surveillance

RS : Regroupement Syndromique

SurSaUD : Surveillance Sanitaire des Urgences et des Décès

SVM : Support Vector Machine

TAL : Traitement Automatique des Langues

2SCE : Surveillance Spatiale des Epidémies

Introduction générale

La mortalité s'oppose à l'immortalité et représente la condition finie de l'être humain. Durant des années, l'homme a cherché à en connaître les causes et a commencé à mettre en œuvre les moyens d'y faire face. Ce n'est que tardivement, au XVI^{ème} siècle, avec John Graunt que le mot mortalité a renvoyé au phénomène d'étude statistique et de compréhension des causes de décès. Jusqu'alors, la mortalité était utilisée en démographie pour expliquer les évolutions des populations.

Nous reviendrons sur l'évolution de l'utilisation des informations liées aux décès et à leurs causes au cours de l'Histoire, puis nous présenterons le circuit actuel des données de mortalité à travers le système SurSaUD et nous décrirons brièvement ce qu'est le domaine du traitement automatique des langues.

I/ La mortalité et les causes de décès : Historique

1) L'état civil

En France avant le Concile de Trente, l'identification des personnes était fondée sur la reconnaissance physique et la perception des visages.

Ce sont les registres paroissiaux qui ont été les précurseurs de l'état civil, le clergé y inscrivant les baptêmes, mariages et sépultures. En effet, l'ordonnance de Villers-Cotterêts en 1539, bien que peu suivie, impose la tenue des registres paroissiaux ou de catholicité. Ce n'est qu'en 1667, que la tenue des registres en double exemplaire est rendue obligatoire et permet de réduire la perte des informations durant les périodes de guerre ou lors d'incendies fréquents à cette époque. Pendant plusieurs décennies, cette ordonnance est mal appliquée et les informations ne sont pas transmises à la justice royale (1).

L'Edit de Louis XIV de 1691 créé les greffiers d'état civil, qui ont la charge de la gestion des archives en recevant chaque année une copie de chaque registre paroissial. Pourtant c'est seulement avec l'ordonnance royale de 1736, qui complète celle de 1667, que l'obligation de tenue en double des registres se généralise à l'ensemble des paroisses. Ce texte indique non seulement l'obligation de signature, mais détaille aussi les informations qu'il convient d'enregistrer au moment des baptêmes, mariages ou sépulture. Il spécifie aussi l'obligation de préciser les ondoiements[1] pour les enfants mort-nés (2).

[1] Rite simplifié de baptême en cas de danger de mort

Cependant une partie de la population de confession religieuse autre que catholique (Protestants et Juifs) échappe à l'état civil. Il faut attendre 1664 pour que soit donnée officiellement aux pasteurs la mission de constater et enregistrer légalement les baptêmes et mariages et de tenir les registres en double exemplaire. Avec la révocation de l'édit de Nantes, en octobre 1685, les baptêmes et les mariages des réformés ne peuvent être inscrits légalement que sur les registres de catholicité. Quant aux décès, ils sont inscrits sur des registres en exemplaire unique et rédigés par des officiers de justice ou de police à partir de 1736.

L'assemblée législative de 1792 définit un nouveau mode de « constat de l'état civil des citoyens ». La tenue des registres est retirée aux curés, transmise aux maires et devient ainsi laïque. Elle décide la confection de tables annuelles et décennales de mortalité et confirme le dépôt en double exemplaire des registres aux greffes des tribunaux.

Aujourd'hui, les décès sont enregistrés par les bureaux d'état civil de la commune de décès. Si l'enregistrement des décès apparait rapidement dans l'histoire et devient systématique et obligatoire, il n'en est pas de même pour les causes ayant conduit à la mort.

2) Les causes de décès

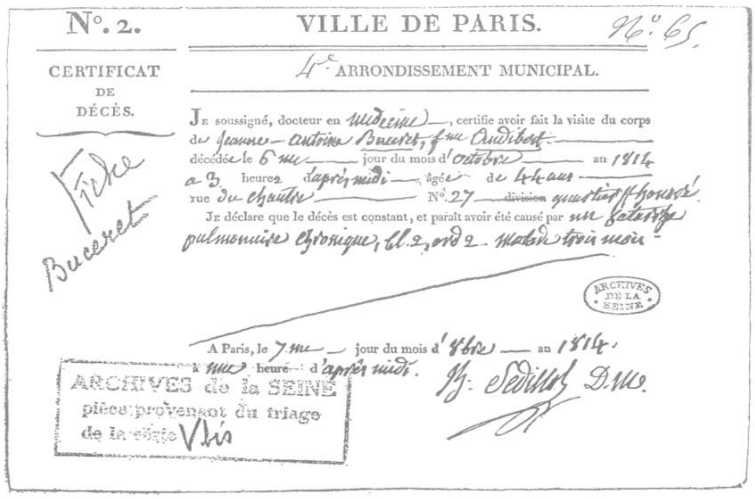
C'est durant les grandes épidémies du Moyen-âge qui ont ravagé l'Europe au milieu du XIVème siècle que commencent les décomptes réguliers des décès et de leurs causes dans les grandes villes (3).

A partir du XVIème siècle, on retrouve les premiers décomptes de mortalité. A Londres, les premiers bulletins de mortalité ou « bills of mortality » sont établis et contiennent des mentions des causes ayant entrainé le décès. A Paris, à la même époque Colbert ordonne que soit établi le relevé mensuel des décès par paroisse avec en marge les principales maladies observées durant cette période (4).

Ce n'est qu'une centaine d'années plus tard, en 1767, que Razoux inspiré par les travaux de Boissier de Lacroix de 1731, publie des tables nosologiques avec des tableaux chiffrés détaillants des causes de décès enregistrées à l'Hôtel Dieu de Nîmes durant 5 années. Puis en 1776, l'enquête de la Société Royale de Médecine auprès de nombreux médecins répartis en France, dont le but est de rassembler le plus grand nombre d'informations sur les épidémies observées, attire l'attention sur les statistiques des causes de décès.
A la suite de cette enquête, il sera demandé à certains curés de noter la cause du décès sur les actes de décès. Ces causes, fournies par l'entourage du défunt, restent souvent vagues.

Le 21 octobre 1802, une circulaire est adressée aux maires des arrondissements de Paris. Elle mentionne qu'il est nécessaire de faire « un travail sur les causes des maladies qui règnent le plus souvent à Paris » et met à disposition des médecins un modèle « d'état de décès ». Il doit contenir

les noms et adresses des personnes décédées, leur sexe, leur âge, les causes et le genre de leurs maladies. Dès lors, les médecins commencent à remplir des bulletins de causes de décès (Figure 1).



N°. 2. VILLE DE PARIS. 96.e 65.

CERTIFICAT DE DÉCÈS.

4.e ARRONDISSEMENT MUNICIPAL.

Je soussigné, docteur en Médecine, certifie avoir fait la visite du corps de Jeanne-Antoine Bucaret, femme Audibert, décédée le 6 me jour du mois d'octobre an 1814, à 3 heures après midi, âgée de 64 ans, rue du chantier N°. 27 division quartier d'Honoré. Je déclare que le décès est constant, et paraît avoir été causé par une Catarrhe pulmonaire chronique, et 2, ord 2. malade trois mois.

Fille Bucaret

ARCHIVES DE LA SEINE

A Paris, le 7 me jour du mois d'8bre an 1814, à une heure d'après midi. JJ. Sedillot D.M.

ARCHIVES de la SEINE pièce provenant du triage de la série Vbis

Figure 1 : Certificat de décès sous la Révolution et l'Empire (XIX^{ème} siècle) (Source : Biraben, Essai sur la statistique des causes de décès en France sous la Révolution et le Premier Empire)

Le conseil de salubrité créé la même année, prendra une orientation plus médicale et s'intéressera spécialement à la prévention des épidémies. En 1808, les premiers tableaux de mortalité sont envoyés au préfet de Paris puis la décision est prise de dresser un « tableau des maladies considérées comme causes de mort ». Il s'agit de la première nomenclature de causes de décès élaborée en France.

En 1886, une circulaire enjoint les maires des villes de plus de 10 000 habitants d'effectuer des relevés bimensuels des décès dus à certaines maladies épidémiques très précises. Ce relevé s'étendra à l'ensemble du territoire en 1906. Parallèlement, la première classification française des maladies voit le jour et contient 27 rubriques. Quelques années plus tard, en 1893, la 1ère version de la Classification Internationale des Maladies (CIM) est mise en place.

C'est en 1925 qu'un changement majeur intervient et que l'on se rapproche de la collecte des causes de décès telle que nous la connaissons aujourd'hui. Le recueil des causes de décès et l'élaboration des statistiques se font selon des principes et des méthodes systématisés dans le but de répondre à des préoccupations de santé publique. On voit apparaître les bulletins individuels de décès. Déjà utilisés pour la statistique d'état civil depuis 1907, ils ne servaient pas, à ce moment-là, à recueillir les causes de décès. Les maires en charge des tableaux des causes de décès, à cette époque, sont déchargés de cette tâche qui est confiée à la Statistique Générale de France.

Les causes de décès sont inscrites sur le bulletin d'état civil, elles résultent des déclarations faites à la mairie au moment où l'on établit l'acte de décès. Elles proviennent soit du médecin traitant, soit de la famille. Les bulletins sont ensuite regroupés par trimestre et envoyés à la préfecture qui les transmet à la Statistique Générale de France. Ainsi, l'analyse des causes devient centralisée, un numéro est attribué à chaque cause en fonction d'une liste de 38 rubriques, liste abrégée issue de la nomenclature internationale.

En 1937, le certificat médical de décès individuel et confidentiel est mis en service. Il est obligatoirement rempli par un médecin et vient s'adjoindre au bulletin de décès (état civil). Les médecins inspecteurs, en charge de détecter et surveiller les épidémies, sont les destinataires de ces certificats. Ces derniers, après avoir transcrit les causes sur un bulletin anonyme, les transmettent de façon trimestrielle à la Statistique Générale de France où est assuré le codage des informations démographiques et médicales. En 1955, un nouveau certificat de décès, conforme aux recommandations de l'Organisation Mondiale de la Santé (OMS) est mis en place.

Dès 1941, ce n'est plus la Statistique Générale de France qui a la charge du codage et de la statistique des causes de décès, mais l'Institut National de la Statistique et des Etudes Economiques (Insee) et ce jusqu'en 1968, date à laquelle la 8ème révision de la CIM est mise en application et marque la cession de la réalisation du chiffrement des causes de décès à l'Institut National de la Santé et de la Recherche Médicale (Inserm).

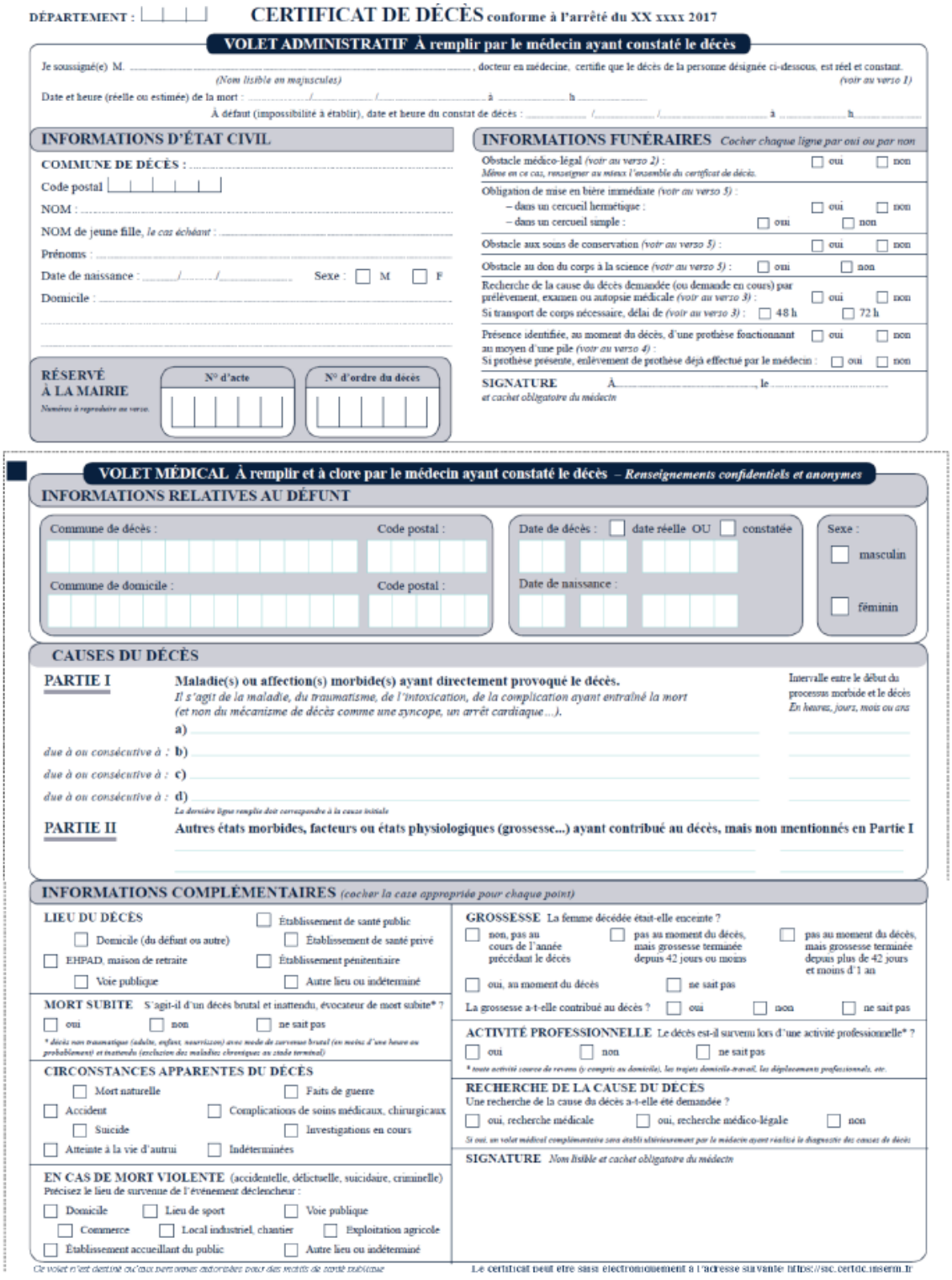
Dès lors, l'Insee sera notamment en charge du recensement des décès basé sur les données d'état civil, et l'Inserm, avec la création en 1968 du Centre d'épidémiologie sur les causes médicales de décès (CépiDc) sera en charge du codage et de la statistique nationale des causes de décès.

3) Les données de mortalité aujourd'hui

2.1-Le certificat de décès

Depuis le décret du 15 Avril 1919 (Article 8) en France, il est obligatoire de déclarer un décès à la mairie de la commune du décès dans les 24 heures suivant la constatation par le médecin afin d'obtenir le certificat de décès qui autorise la fermeture du cercueil. Le certificat est rédigé sur un modèle établi par le Ministère chargé de la santé et basé sur les recommandations de l'OMS. Il précise la ou les causes de décès, aux fins de transmission à l'Inserm (5).

Le certificat de décès est aujourd'hui conçu sur le modèle de certificat proposé par l'OMS et est constitué de 2 parties (Figure 2) :



DÉPARTEMENT : |_|_|_|

CERTIFICAT DE DÉCÈS conforme à l'arrêté du XX xxxx 2017

VOLET ADMINISTRATIF À remplir par le médecin ayant constaté le décès

Je soussigné(e) M. ..... , docteur en médecine, certifie que le décès de la personne désignée ci-dessous, est réel et constant. *(voir au verso 1)*

(Nom lisible en majuscules)

Date et heure (réelle ou estimée) de la mort : ..../..../.... à .... h ....

À défaut (impossibilité à établir), date et heure du constat de décès : ..../..../.... à .... h ....

INFORMATIONS D'ÉTAT CIVIL

COMMUNE DE DÉCÈS : .....

Code postal |_|_|_|_|_|

NOM : .....

NOM de jeune fille, *le cas échéant* : .....

Prénoms : .....

Date de naissance : ..../..../.... Sexe : ☐ M ☐ F

Domicile : .....

RÉSERVÉ À LA MAIRIE

Numéros à reproduire au verso.

N° d'acte

N° d'ordre du décès

INFORMATIONS FUNÉRAIRES *Cocher chaque ligne par oui ou par non*

Obstacle médico-légal *(voir au verso 2)* : ☐ oui ☐ non

Même en ce cas, renseigner au mieux l'ensemble du certificat de décès.

Obligation de mise en bière immédiate *(voir au verso 3)* :

– dans un cercueil hermétique : ☐ oui ☐ non

– dans un cercueil simple : ☐ oui ☐ non

Obstacle aux soins de conservation *(voir au verso 3)* : ☐ oui ☐ non

Obstacle au don du corps à la science *(voir au verso 3)* : ☐ oui ☐ non

Recherche de la cause du décès demandée (ou demande en cours) par prélèvement, examen ou autopsie médicale *(voir au verso 3)* : ☐ oui ☐ non

Si transport de corps nécessaire, délai de *(voir au verso 3)* : ☐ 48 h ☐ 72 h

Présence identifiée, au moment du décès, d'une prothèse fonctionnant au moyen d'une pile *(voir au verso 4)* : ☐ oui ☐ non

Si prothèse présente, enlèvement de prothèse déjà effectué par le médecin : ☐ oui ☐ non

SIGNATURE À ..... le .....

et cachet obligatoire du médecin

VOLET MÉDICAL À remplir et à clore par le médecin ayant constaté le décès – *Renseignements confidentiels et anonymes*

INFORMATIONS RELATIVES AU DÉFUNT

Commune de décès : Code postal :

Commune de domicile : Code postal :

Date de décès : ☐ date réelle OU ☐ constatée

Date de naissance :

Sexe : ☐ masculin ☐ féminin

CAUSES DU DÉCÈS

PARTIE I **Maladie(s) ou affection(s) morbide(s) ayant directement provoqué le décès.**

Il s'agit de la maladie, du traumatisme, de l'intoxication, de la complication ayant entraîné la mort (et non du mécanisme de décès comme une syncope, un arrêt cardiaque…).

Intervalle entre le début du processus morbide et le décès *En heures, jours, mois ou ans*

a) .....

due à ou consécutive à : b) .....

due à ou consécutive à : c) .....

due à ou consécutive à : d) .....

La dernière ligne remplie doit correspondre à la cause initiale

PARTIE II **Autres états morbides, facteurs ou états physiologiques (grossesse…) ayant contribué au décès, mais non mentionnés en Partie I**

INFORMATIONS COMPLÉMENTAIRES *(cocher la case appropriée pour chaque point)*

LIEU DU DÉCÈS ☐ Établissement de santé public ☐ Domicile (du défunt ou autre) ☐ Établissement de santé privé ☐ EHPAD, maison de retraite ☐ Établissement pénitentiaire ☐ Voie publique ☐ Autre lieu ou indéterminé

MORT SUBITE S'agit-il d'un décès brutal et inattendu, évocateur de mort subite* ? ☐ oui ☐ non ☐ ne sait pas

** décès non traumatique (adulte, enfant, nourrisson) avec mode de survenue brutal (en moins d'une heure en probablement) et inattendu (exclusion des maladies chroniques au stade terminal)*

CIRCONSTANCES APPARENTES DU DÉCÈS ☐ Mort naturelle ☐ Faits de guerre ☐ Accident ☐ Complications de soins médicaux, chirurgicaux ☐ Suicide ☐ Investigations en cours ☐ Atteinte à la vie d'autrui ☐ Indéterminées

EN CAS DE MORT VIOLENTE (accidentelle, délictuelle, suicidaire, criminelle) Précisez le lieu de survenue de l'événement déclencheur : ☐ Domicile ☐ Lieu de sport ☐ Voie publique ☐ Commerce ☐ Local industriel, chantier ☐ Exploitation agricole ☐ Établissement accueillant du public ☐ Autre lieu ou indéterminé

GROSSESSE La femme décédée était-elle enceinte ? ☐ non, pas au cours de l'année précédant le décès ☐ pas au moment du décès, mais grossesse terminée depuis 42 jours ou moins ☐ pas au moment du décès, mais grossesse terminée depuis plus de 42 jours et moins d'1 an ☐ oui, au moment du décès ☐ ne sait pas

La grossesse a-t-elle contribué au décès ? ☐ oui ☐ non ☐ ne sait pas

ACTIVITÉ PROFESSIONNELLE Le décès est-il survenu lors d'une activité professionnelle* ? ☐ oui ☐ non ☐ ne sait pas

** toute activité source de revenu (y compris au domicile), les trajets domicile-travail, les déplacements professionnels, etc.*

RECHERCHE DE LA CAUSE DU DÉCÈS Une recherche de la cause du décès a-t-elle été demandée ? ☐ oui, recherche médicale ☐ oui, recherche médico-légale ☐ non

Si oui, un volet médical complémentaire sera établi ultérieurement par le médecin ayant réalisé le diagnostic des causes du décès

SIGNATURE *Nom lisible et cachet obligatoire du médecin*

Ce volet n'est destiné qu'aux personnes autorisées pour des motifs de santé publique

Le certificat peut être saisi électroniquement à l'adresse suivante https://sic.certdc.inserm.fr

Figure 2 : Modèle du certificat de décès général en vigueur depuis 2018, France

- Un volet administratif contenant des données administratives du défunt, incluant son nom, sa date de naissance, son sexe, ainsi que le lieu du décès et les informations sur les conditions d'inhumation.
- Un volet médical, anonyme, contenant entre autres, des informations sociodémographiques (date de naissance et de décès, sexe…) et les causes médicales du décès. La séquence morbide ayant conduit au décès y est rapportée de la cause initiale à la cause immédiate. L'OMS définit la cause initiale comme « a) la maladie ou le traumatisme qui a déclenché l'évolution morbide conduisant directement au décès, ou b) les circonstances de l'accident ou de la violence qui ont entraîné le traumatisme mortel ». La cause immédiate est la cause finale ayant directement entrainé le décès.

  Les causes associées n'ayant pas directement conduit au décès sont également mentionnées dans ce volet.

  On retrouve également dans le volet médical des informations sur le lien du décès avec une grossesse, un accident de travail ou encore le type de lieu de décès.

Outre ce modèle général de certificat de décès, il existe un certificat de décès spécifique (certificat néonatal (Annexe 1)) pour les décès survenant dans les 28 premiers jours de la vie. Ce certificat construit sur le même modèle que le certificat général, contient des informations supplémentaires relatives aux parents du défunt, à l'accouchement, ainsi qu'une séquence morbide spécifique (6).

2.2-Le circuit des données issues du certificat de décès depuis 1968

2.1 Le circuit des certificats de décès papier

Dès que le médecin a complété le certificat de décès et a clos le volet médical, le certificat est transmis au bureau d'état civil de la mairie de la commune de décès. La mairie rédige alors deux documents :

- L'avis 7 bis qui comporte le nom de la personne décédée et les informations d'état civil. Il est transmis à l'Insee par voie électronique de façon automatique ou non automatique (selon les communes) pour la mise à jour du Répertoire National d'Identification des Personnes Physiques (RNIPP). Seules les données transmises de manière automatique sont ensuite transférées à Santé Publique France.
- L'avis 7 ou bulletin 7 (cf. Figure 3) qui contient les mêmes informations que l'avis 7 bis mais sans le nom du défunt. Il est envoyé à l'Agence Régionale de Santé (ARS).

Le volet médical du certificat de décès est envoyé à l'ARS avec l'avis ou bulletin 7. Après ouverture, le volet médical est envoyé au CépiDc qui se charge de sa numérisation, son codage et sa validation (ce codage des causes de décès fait l'objet d'une validation avant d'être mise à disposition). Ce processus prend du temps et la base de données complète d'une année N est disponible avec un délai de 18 à 24 mois (Figure 3).

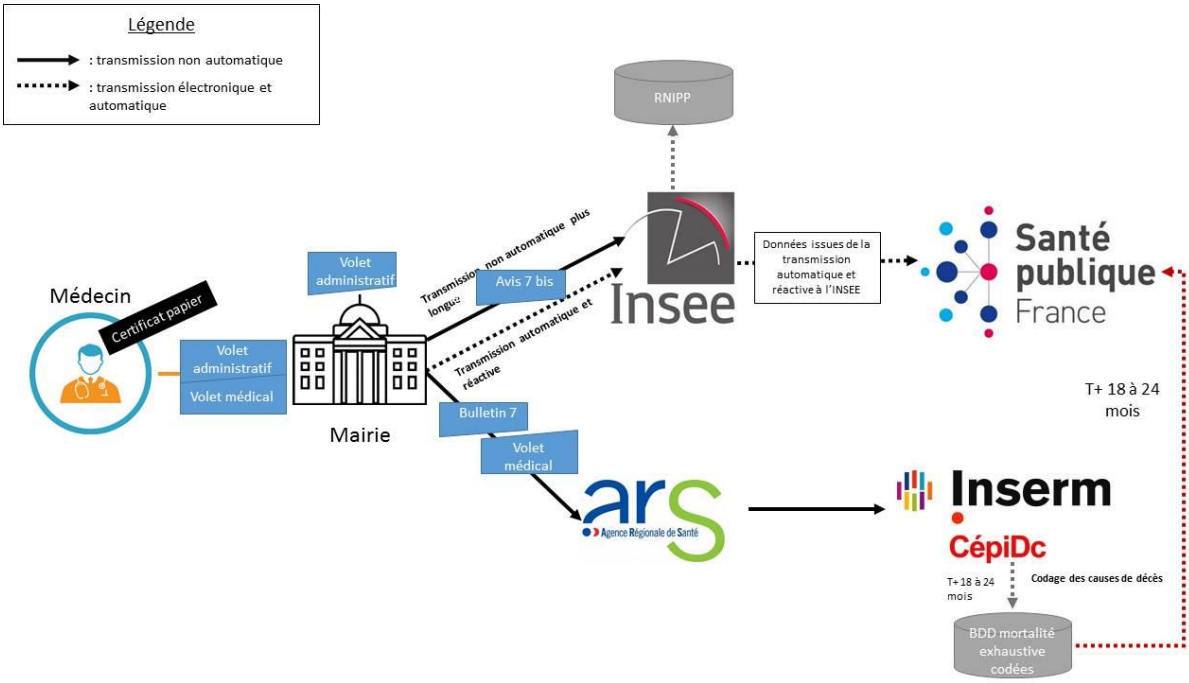


Figure 3 : Circuit de transmission des certificats de décès rédigés sur un formulaire papier, France, 2019

2.2 Circuit du certificat de décès électronique

Le certificat électronique est rempli (2 volets) par le médecin via l'application web CertDc (https://sic.certdc.inserm.fr/login.php). Une fois saisi, le volet administratif est imprimé et apporté au bureau d'état civil de la commune de décès, et suit le même circuit que le volet administratif papier.

En revanche, les informations contenues dans le volet médical sont directement envoyées par voie électronique au CépiDc et sont mises à disposition de Santé publique France. Les causes de décès sont alors disponibles sous forme brute (texte libre et non codées) dans les minutes qui suivent la validation du certificat par le médecin. Les étapes de codage et de validation sont toujours exécutées par le CépiDc et la base de données complète et codée de l'année N est toujours disponible dans les mêmes délais que pour le format papier (Figure 4).

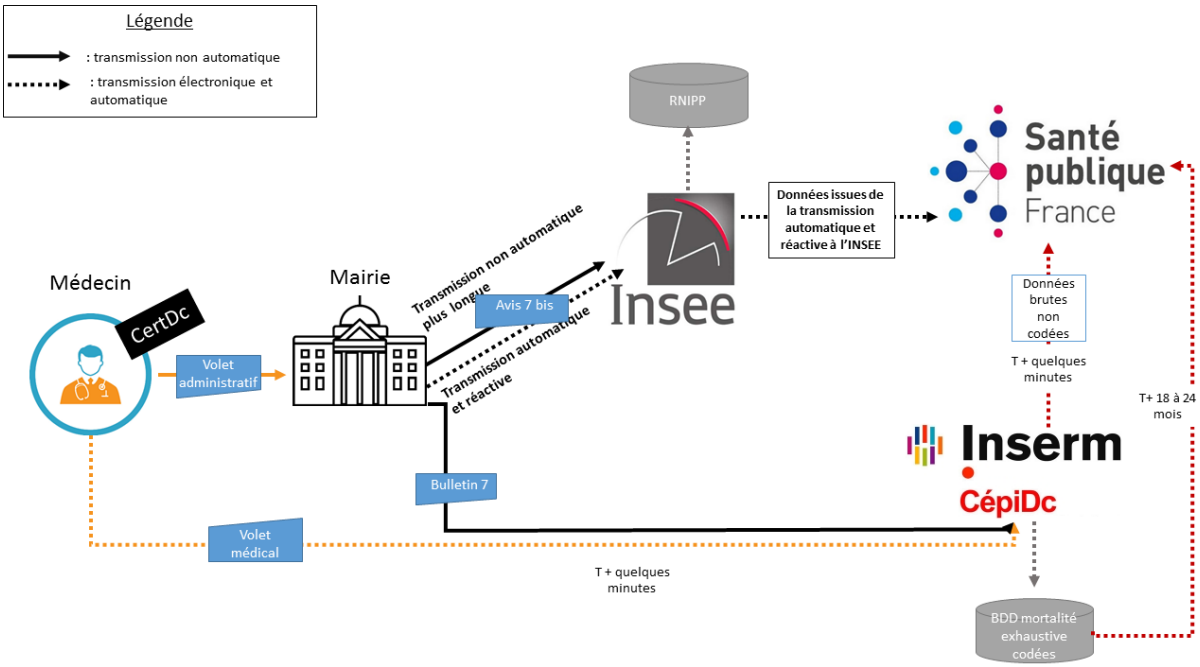


Figure 4 : Circuit de transmission du certificat de décès rédigés sur format électronique, France, 2019

2.3 La codification des causes de décès

Afin de produire les statistiques nationales sur la mortalité, l'Inserm-CépiDc doit dans un premier temps, coder les causes médicales de décès rédigées par les médecins sous forme de texte libre. La codification comporte deux étapes :

- L'attribution d'un code à chaque maladie, traumatisme ou cause externe de décès mentionné sur le certificat sur la base de la Classification Internationale des Maladies 10ème révision (CIM-10)
- La sélection et le codage de la cause initiale de décès.

Le codage assure la qualité et la standardisation des données et permet une comparaison internationale des statistiques produites. Des codeurs sont en charge de cette étape. Cependant des variations de codages liées aux codeurs ont conduit à la mise en place et l'utilisation à partir de 2011, en complément, d'un logiciel de codage automatique: Iris (7).
Le codage des causes et la sélection de la cause initiale sont régis par des règles édictées par l'OMS. Ces règles permettent de sélectionner la cause initiale tout en respectant le plus possible les informations rapportées par le médecin. Si le certificat de décès ne respecte pas le processus morbide, d'autres règles permettent d'identifier la cause qui correspondrait le mieux à la cause initiale et ainsi de reconstruire la séquence morbide.

II/ La surveillance de la mortalité en France

1) La surveillance de la mortalité avant 2003

La certification des décès permet la mise à jour du suivi administratif et démographique de la population. L'enregistrement des causes médicales de décès répond à un objectif de santé publique, par le biais (6)**:**

- de la réalisation d'études épidémiologiques permettant d'identifier et quantifier les causes, de décès et de les hiérarchiser,
- de la surveillance et de l'alerte sanitaire.

Outre l'utilisation de ces données pour la surveillance en routine de la mortalité sur laquelle nous reviendrons ensuite, le CépiDC fournit les statistiques nationales des causes de décès. Celles-ci permettent d'identifier les priorités de santé publique, notamment les priorités d'orientation des problématiques de recherche.
La Direction de la recherche, des études, de l'évaluation et des statistiques (DREES) et Santé publique France produisent périodiquement un rapport de l'état de santé de la population dans lequel ils

consacrent un chapitre aux principales causes de décès (8). Sont également produites par différents acteurs (Insee, Inserm, Institut national des études démographiques…) des études tendancielles de la mortalité liées aux grandes causes nationales de décès (Cancer, maladies cardio-vasculaires, alcool, tabac, …), ainsi que des atlas de la mortalité mettant en évidence des disparités géographiques des causes de mortalité et leur évolution au cours du temps (9).

A l'échelon international, les données de mortalité par causes de décès sont diffusées à Eurostat (La statistique Européenne) et à l'OMS. Elles permettent de positionner la France lors des différentes études comparatives internationales des systèmes de santé et de prévention.

2) La crise sanitaire de 2003

Bien que la première grande loi de 1902 instaure les principes de la surveillance sanitaire, ce n'est qu'avec les premières crises sanitaires majeures et mondiales du SIDA dans les années 1980 puis du sang contaminé, de la vache folle ou encore de l'amiante dans les années 1990, qu'émerge, en France, la nécessité de mettre en place un réel système de surveillance sanitaire. Le Réseau National de la Santé Publique (RNSP) est alors créé en 1992 sous forme d'un Groupement d'Intérêt Public (GIP)(10), et voit son statut modifié en 1998. Elle devient agence d'état sous la forme de l'Institut de Veille Sanitaire (InVS) (11).

Cependant, l'épisode de canicule de l'été 2003 va mettre en évidence des failles dans la surveillance sanitaire en France. En effet, bien qu'ayant pour mission « d'alerter les pouvoirs publics en cas de menace pour la santé publique », d'essayer de prévenir ces menaces et de « mener toutes actions nécessaires », l'InVS avait un fonctionnement tourné vers l'analyse *a posteriori* et non vers l'intervention *a priori* (12, 13). Le système de surveillance en place en 2003 n'a alors pas permis de donner l'alerte rapidement sur l'augmentation du nombre de décès dans les périodes de fortes chaleurs. Ce n'est donc que tardivement que l'impact de la chaleur d'intensité sans précédent a pu être appréhendé, grâce aux données de mortalité hospitalière de l'APHP (Assistance Publique des Hôpitaux de Paris) et aux données d'activité des pompes funèbres générales. L'ampleur des conséquences de cet événement, avec une surmortalité de près de 15000 décès mesurée avec d'autres sources de données quelques mois après, a conduit à la remise en question du système de surveillance en place à cette époque et notamment son manque de réactivité (13).

3) La surveillance réactive de la mortalité à travers le système de surveillance syndromique SurSaUD®

3.1 La surveillance syndromique

Le concept de surveillance syndromique est apparu pour la première fois au début des années 90 à la suite d'une épidémie de gastro-entérite qui avait touché plus de 400 000 personnes dans le Milwaukee. Il avait été montré que les ventes d'anti diarrhéiques hors prescription avaient plus que triplé, plusieurs semaines avant que l'épidémie n'ait été identifiée par les autorités (14). Cet événement a fait émerger les premières réflexions sur l'intérêt d'utiliser de telles données pour la détection rapide de phénomènes inattendus.

Mise en place concrètement aux Etats-Unis à la suite des attentats du World Trade Center en 2001 pour faire face à la menace bioterroriste, la surveillance syndromique avait pour objectif initial la détection rapide de phénomènes inattendus pouvant révéler la présence à grande échelle de l'impact sanitaire d'un agent biologique. L'utilisation et la définition d'un tel système a évolué au cours des 10 années suivantes. En 2004, les CDC (Centers for Disease Control and Prevention) américains définissent la surveillance syndromique comme « une approche dans laquelle les intervenants sont assistés par des procédures d'enregistrement automatique des données pour le suivi et l'analyse épidémiologique en temps réel ou proche du temps réel ; cela afin de détecter plus tôt qu'il n'aurait été possible de le faire, des événements habituels ou inhabituels sur la base des méthodes traditionnelles de surveillance »(15). En 2011, le projet européen Triple S, ayant pour but de faire un inventaire des systèmes européens de surveillance syndromique existants et de proposer des orientations scientifiques et techniques pour leur mise en place, a présenté une définition élargie de la surveillance syndromique. Elle précise les données susceptibles d'être utilisées en particulier, s'élargit aux systèmes vétérinaires et souligne sa complémentarité avec les surveillances spécifiques (16).

La surveillance syndromique est une surveillance populationnelle. Toutes les étapes (collecte, analyse, interprétation, diffusion) se font en temps réel ou proche du temps réel, d'où également l'utilisation de l'expression de « surveillance réactive » dans la suite de ce mémoire. L'objectif principal de ce type de surveillance est d'identifier précocement une variation anormale, attendue ou inhabituelle, à travers des sources de données réactives dans un but d'alerte. Pour cela, le dispositif se base sur la collecte de données existantes et sans sélection *a priori*. Le type de données collectées est multiple, il peut s'agir de signes cliniques, de diagnostics ou encore de proxy de l'état de santé constituant un diagnostic prévisionnel ou syndrome (diagnostic clinique non confirmé, absentéisme, ventes de médicaments, …).

La surveillance syndromique a également comme objectifs de suivre l'évolution des événements, qu'ils soient détectés par le système ou signalés par d'autres canaux, et de fournir une évaluation rapide de l'impact sanitaire de l'événement sur la santé de la population. Enfin, ce type de système contribue à la réassurance des décideurs, à travers la surveillance de risques potentiels lors de la survenue de situations environnementales (tempêtes, inondations, …), industrielles ou de grands rassemblements de population (événements sportifs, festivals, rassemblements politiques, ..) en fournissant de l'information sur l'impact ou l'absence d'impact de la situation en cours.

Le système RODS (real-time outbreak and diseases surveillance) (17) est le plus ancien (créé et développé à partir de 1999) et le plus abouti des systèmes de surveillance syndromique. Il est basé sur l'analyse et la catégorisation en syndromes des motifs de recours aux urgences exprimés par le patient et reportés en texte libre. Aujourd'hui ce système est utilisé dans de nombreuses villes et états américains mais aussi hors des Etats-Unis, au Canada (18), à Taïwan (19).

Egalement, le projet ESSENCE (Electronic Surveillance System for Early Notification of Community based Epidemics) (20) est à l'origine un programme militaire pour la surveillance sanitaire des troupes déployées à l'étranger. ESSENCE II est la version civile de ce programme. Il est basé sur des sources de données à la fois civiles et militaires et de format très hétérogène. Outre l'objectif de surveillance sanitaire, ce programme a aussi pour objectif de développer et évaluer de nouvelles techniques analytiques pour l'identification rapide de phénomènes anormaux de santé et d'évaluer la pertinence des indicateurs de santé utilisés.

Enfin, le dernier système nord-américain est le système de la ville de New-York mis en place au lendemain des attentats du 11 septembre 2001 et modifié quelques mois plus tard (21). Il était initialement basé sur un questionnaire simple d'une page, destiné à surveiller douze syndromes pour identifier les conséquences d'une éventuelle attaque bioterroriste et évaluer l'impact de l'effondrement des tours. Après quelques mois, la mise en place du système a été revue avec notamment une automatisation de la transmission et de l'analyse des données, et l'intégration de la surveillance quotidienne des ventes de certains médicaments.

En Europe, Public Health England (PHE) a développé plusieurs systèmes de surveillance syndromique basés sur différentes sources de données. Historiquement, en Angleterre, la surveillance syndromique reposait sur le NHS direct, une ligne d'aide téléphonique au travers de laquelle du personnel infirmier répondait directement à la population en les conseillant ou en les orientant vers une consultation médicale (22). Les infirmiers saisissaient dans un logiciel les informations des symptômes évoqués sous forme de codes. Aujourd'hui, plusieurs autres sources de données s'ajoutent à celle-ci : le nombre de consultations chez les omnipraticiens durant les heures et jours ouvrés de consultation pour les indicateurs cliniques connus tels que la grippe, l'asthme, la

pneumonie, la gastroentérite, etc. (GP in hours syndromic surveillance system), le nombre de consultations imprévues et d'appels aux omnipraticiens tous les jours fériés, en soirées ou pendant la nuit (GP out-of-hours syndromic surveillance system) et le nombre de consultations quotidiennes dans un réseau des services d'urgence (Emergency Department Syndromic Surveillance System) mis en place à l'occasion des Jeux Olympiques de Londres de 2012 (23).

En France, depuis 2004 le service de santé des armées utilise un système de surveillance syndromique : le Système d'Analyse et de Surveillance épidémiologique en TEmps Réel (ASTER) (24). Ce système a été construit pour répondre aux objectifs de maintien de la capacité opérationnelle des combattants qui dépend du maintien de leur santé. ASTER est un système de télé-épidémiologie basé sur la coexistence de 2 réseaux de communication :

- Une source de données : les *réseaux de déclaration*, réseaux « 2SE » (Surveillance spatiale des épidémies), qui assurent l'infrastructure humaine, matérielle et organisationnelle autorisant la déclaration et le suivi des cas par les médecins d'unités, en collaboration avec la direction locale du Service de santé.
- Un système d'analyse : un *réseau d'analyse*, dénommé CS2A (collaboration de services pour la surveillance et l'alerte), qui va chercher les données sur le réseau de déclaration pour procéder aux traitements automatiques requis dans le cadre de la surveillance pour les différents experts impliqués dans ce processus. (http://cybertim.timone.univ-mrs.fr/recherche/projets-recherche/ASTER)

Ce système est utilisé en routine depuis 2004 et a permis de détecter précocement plusieurs épidémies de dengue et de paludisme chez les militaires en Guyane (25).

Jusqu'en 2008, les systèmes de surveillance sont généralement fondés sur des données de morbidité. En Europe, un inventaire de l'ensemble des systèmes de surveillance syndromique a été conduit par le projet Triple-S et a identifié une trentaine de système de surveillance humaine (26).

Entre 2008 et 2011, le projet européen EuroMomo (European monitoring of excess mortality for public health action) (27) a été mené dans la perspective de construire et d'assurer une surveillance sanitaire réactive fondée sur la mortalité à l'échelle européenne. L'objectif principal était d'homogénéiser les modes d'enregistrement en routine des données de mortalité dans les pays européens et de mettre en œuvre un outil d'analyse statistique commun, dont les résultats seraient comparables d'un pays à l'autre. Aujourd'hui, ce système collecte chaque semaine les données de mortalité de plus de 20 pays, afin d'effectuer des analyses à l'échelle européenne mais aussi de comparer l'évolution hebdomadaire de la mortalité entre les pays.

Finalement, un système de surveillance syndromique en population générale est également en place en France depuis 2004 et permet de surveiller en routine et de façon réactive la santé de la population : le système de Surveillance Sanitaire des Urgences et des Décès (SurSaUD®)

3.2 Le système de surveillance syndromique SurSaUD®

Le système SurSaUD® a été mis en place en France suite aux conséquences exceptionnelles de la canicule de 2003 (28). Ce système a été construit dans l'objectif de détecter précocement la survenue d'événements habituels ou inhabituels, d'en suivre l'évolution et d'en mesurer l'impact sur la santé. Le système SurSaUD® est alimenté par 4 sources de données (Figure 5) :

- Deux sources de données de morbidité :
    - Les données de recours aux urgences issues du réseau OSCOUR (Organisation de la Surveillance Coordonnée des Urgences), incluant en 2019 plus de 730 structures d'urgences,
    - Les données des actes effectués par le réseau des associations SOS Médecins incluant 62 des 63 associations.
- Deux sources de mortalité :
    - Les données issues de l'état civil fournies par l'Insee,
    - Les données issues du volet médical des certificats de décès rédigés par voie électronique.

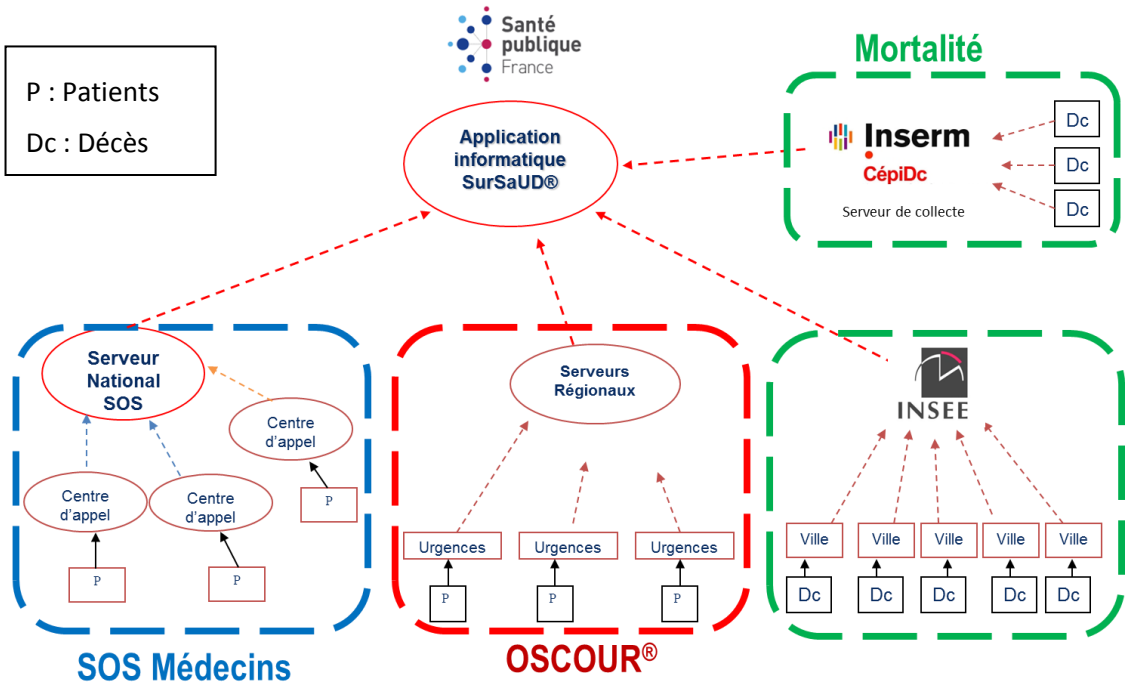


Figure 5 : Sources de données du système de surveillance syndromique SurSaUD, France, 2019

Les données des 4 sources sont collectées quotidiennement sous forme de données individuelles et de façon automatique. Cette collecte a été approuvée par la CNIL (Commission Nationale de l'Informatique et des Libertés).
Enregistrées au sein de l'application SurSaUD, ces données vont faire l'objet de contrôle automatisé de qualité avant d'être analysées.

L'analyse des données de morbidité est basée sur la construction de regroupements syndromiques (ensemble de diagnostics ou symptômes) définis pour répondre aux objectifs de surveillance sanitaire. Des méthodes statistiques permettent d'établir un seuil au-delà duquel le nombre observé de passages aux urgences ou actes médicaux est statistiquement supérieur au nombre attendu de passages ou actes médicaux. Ces seuils mettent en évidence des événements inhabituels et permettent d'alerter les décideurs (29).

L'analyse de ces données est faite à différents niveaux géographiques et est rapportée à l'aide de tableaux de bord quotidiens et de bulletins de surveillance hebdomadaires permettant de suivre l'évolution de ces indicateurs et de transmettre les informations aux décideurs.

3.3 La surveillance syndromique de la mortalité

Depuis 2004, la surveillance syndromique de la mortalité s'appuie essentiellement sur les données d'état civil issues du volet administratif du certificat de décès. Santé publique France reçoit quotidiennement les données transmises par l'Insee de façon informatisée et automatique. Ces données permettent d'effectuer une surveillance quantitative toutes causes, en routine et ainsi répondre aux objectifs de la surveillance syndromique : la détection d'évènements attendus ou inattendus à visée d'alerte et l'évaluation d'impact. Parallèlement et à la suite de la canicule de 2003, la certification électronique des décès a été mise en place en 2007.

La certification électronique permet de recevoir dans les minutes qui suivent la validation du certificat de décès par le médecin, les informations sur les causes médicales de décès. La disponibilité des causes de décès de façon réactive permet d'envisager en complément de l'analyse quantitative, une analyse qualitative de la mortalité en temps quasi réel.

L'analyse des causes de décès repose sur le traitement de données disponibles sous un format texte libre. En effet le médecin complète le volet médical en décrivant le processus morbide avec ces propres mots. L'exploitation des causes médicales de décès pour la surveillance réactive de la mortalité nécessite donc l'utilisation de méthodes de traitement automatique des langues.

III/ Le Traitement Automatique des Langues dans le domaine de la santé

Le Traitement Automatique des Langues (TAL) est une discipline à la frontière de la linguistique, de l'informatique et de l'intelligence artificielle. Elle concerne la conception de systèmes et techniques informatiques permettant de manipuler le langage humain dans tous ses aspects.

Deux grands domaines se distinguent, 1) l'extraction d'informations qui permet de retrouver des informations préalablement définies dans un texte et 2) la récupération d'informations qui permet à partir d'une requête de regrouper tous les documents en lien avec cette requête (ex : Google).

Avec l'utilisation de plus en plus fréquente de dossiers de santé électroniques, dans lesquels une quantité d'informations est généralement reportée en texte libre, l'extraction d'information est devenue un des éléments qui facilite l'utilisation de ces données pour la prise de décisions ou la surveillance sanitaire. Cette tâche fait référence à l'extraction automatique de concepts, d'entités et d'événements, ainsi que de leurs relations et attributs associés à partir de texte libre.

Dans le domaine biomédical, de multiples travaux ont utilisé des systèmes d'extraction d'information (Tableau 1). L'identification de pathologies tient une place importante dans les études utilisant ce type de méthode. On retrouve en tête les études sur l'identification des différents types de cancers (30, 31), suivi des maladies cardiovasculaires (32, 33) et du système digestif (34, 35). D'autres auteurs se sont attachés à identifier les syndromes cliniques et les concepts biomédicaux communs à partir de rapports de radiologie (36, 37), ou de résumés de dossier de sortie d'hospitalisation (38). D'autres ont cherché à identifier à partir de données d'urgence les circonstances de blessures (39) ou les accidents du travail dans un objectif de surveillance sanitaire (40). La surveillance des infections associées aux soins est un domaine où l'utilisation du TAL est aussi en forte expansion (41-43). Plus particulièrement, des travaux ce sont intéressés au développement d'un système de surveillance syndromique des infections associées aux soins dans un hôpital à partir des données du dossier médical des urgences(44).

Des études se sont intéressées à l'utilisation de ces méthodes pour la pharmacovigilance et la surveillance de la consommation de certains médicaments (45, 46). Ces méthodes sont aussi utilisées pour la surveillance de la mortalité pour pneumonie ou grippe (47, 48), par cancer (49, 50), en lien avec la consommation de drogues (51), avec une mention de diabète ou de VIH (48).

Une autre utilisation du TAL dans le domaine biomédical, concerne le codage automatique des causes de décès. « Le CLEF e-Health challenge », qui se déroule chaque année depuis 2013, vise à développer des méthodes de TAL pour l'extraction et la recherche d'information à partir de données médicales (52). Depuis 2015, une tâche plus spécifique vise à coder automatiquement des causes de

décès en code CIM-10. Ce défi consiste à mettre à disposition une base de données des certificats de décès incluant les causes médicales rédigées par un médecin en texte libre et les codes CIM-10 associés. Les équipes participantes doivent proposer la méthode la plus performante pour attribuer automatiquement un code CIM-10 à chaque cause (53-56).

Sur le plan méthodologique, plusieurs revues de la littérature, dont celle de Meystre en 2008 (57) et plus récemment celle de Wang en 2018 (58), ont recensé les différents types de méthodes utilisées pour l'extraction d'information dans le domaine biomédical. Les défis organisés pour le codage automatique des causes de décès sont aussi une source utile pour la recherche de méthodologies.

On peut diviser les approches d'extraction d'information en deux grandes catégories : l'apprentissage fondé sur des règles linguistiques et l'apprentissage machine.

La première famille de méthodes dites par règles linguistiques consiste à extraire l'information textuelle à partir de règles de décision définies a priori, telles que la recherche par mots clés, l'application d'expressions rationnelles, l'analyse syntaxique, etc. Les méthodes de classification par apprentissage automatique se divisent en deux branches :

- Les méthodes par apprentissage supervisé (59-63),
- Les méthodes par apprentissage non supervisé (64).

L'identification et la mise en œuvre des méthodes les plus performantes d'après la littérature nous permettront d'exploiter les causes de décès issues de la certification électronique pour la surveillance réactive de la mortalité dans un objectif de détection, d'alerte et d'évaluation d'impact.

Tableau 1: Etudes internationales utilisant des méthodes de TAL dans le domaine biomédical, revue de la littérature jusqu'en Juin 2019

	Domaine d'application	Objectif	Méthodes
Yang[38] (2009)	Challenge en TAL	Identifier automatiquement chez les patients le statut obèse et 15 comorbidités liées, à l'aide de leur résumé de décharge clinique	• Projection de dictionnaire • Méthode par règles • SVM
Flynn[36] (2010)	Recherche en pharmaco-épidémiologie	Augmenter le nombre de diagnostics spécifiques d'AVC en codant automatiquement SMR01 à partir des données du système informatisé d'information radiologique (CRIS)	Recherche de mots clés et détection de négation
McKenzie[40] (2010)	Surveillance accident du travail/ blessures liées au travail	Comparer les méthodes de recherches d'information textuelle afin d'identifier les blessures liées au travail aux urgences	• Recherche de mots clés • Recherche de mots clés à partir d'un dictionnaire • Logiciel Leximancer : étudie la fréquence et la cooccurrence des termes
Savova[32] (2010)	Recherche en génétique	Extraction du phénotype à partir des notes de radiologie pour identifier les maladies artérielles périphériques	Mayo's clinical Text Analysis and Knowledge Extraction System (cTAKES)(65) : ensemble d'outils de TAL pour traiter les textes cliniques et extraire l'information.
Xu[30] (2011)	Etude épidémiologique	Identification des patients avec un cancer colorectal à partir de l'ensemble du dossier patient électronique (compte-rendu radiologie, compte-rendu analyses…)	Identification de concepts liés au cancer (UMLS) : • A partir de l'outil MedLEE (approche à base de règles) • SVM (Support Vector Machine) (Apprentissage supervisée Identification d'un cas de cancer : • Développement de règles • Méthodes d'apprentissage : Foret aléatoire, Ripper, SVM, Régression logistique
Gerbier[44] (2011)	Surveillance syndromique des infections associées aux soins	Extraire et coder les informations retrouvées dans le dossier médical des urgences pour la surveillance syndromique des infections associées aux soins	• Outil UrgIndex et ECMT
Davis[47] (2012)	Surveillance Mortalité	Codage automatique des décès avec mention de pneumonie ou grippe	• MetaMap[66]
Iyer[46] (2013)	Pharmacovigilance	Détecter de nouvelles interactions médicamenteuses et l'identification de solutions de rechange aux associations médicamenteuses à risque.	Recherche de mots clés
Ludvigsson[34] (2013)	Systèmes d'aide à la décision clinique	Tester deux algorithmes de symptômes, de signes et de codes diagnostiques, afin de construire un modèle qui permettra d'identifier	Combinaison de recherche de mots clés

		les personnes à risque élevé de maladies cœliaques et qui bénéficieraient du dépistage de ces maladies	
Zuccon[37] (2013)	Systèmes d'aide à la décision clinique	Développer et évaluer des techniques d'apprentissage automatique qui permettent d'identifier les fractures des membres et d'autres anomalies à partir de rapports de radiologie	• Classifieur multinomial bayésien naïf • Classifieur SMO (Sequential minimal optimization): SVM • Classifieur Spegasos : variante du SVM
Roch[35] (2014)	Recherche sur la détection précoce et prévention du cancer du pancréas	Identifier automatiquement les patients présentant des kystes pancréatiques grâce au dossier patient électronique à l'aide de méthode de TAL.	Méthodes à base de règles et recherche de mots clés à l'aide d'expressions rationnelles
Branch-Elliman[42] (2015)	Surveillance	Evaluer si l'examen manuel des dossiers pouvait être complété ou remplacé par un algorithme enrichi de TAL pour l'identification d'infections des voies urinaires associées aux cathéters en temps réel	Recherche de mots clés ou phrase clés
Carrell[45] (2015)	Surveillance	Identifier l'usage problématique des opioïdes chez les patients recevant un traitement aux opioïdes à partir des dossiers cliniques	• Méthode par règles • Recherche d'expressions dans un dictionnaire • Expressions rationnelles
Chen[39] (2015)	Recherche sur les blessures et profils de causes de blessures	Faire correspondre des informations textuelles au code des blessures graves et au code externe	• Techniques de factorisation matricielle : - décomposition des valeurs singulières (SVD) - factorisation matricielle non négative (NNMF). • Arbre de décision, Modèle bayésien naïf, SVM, Réseau de neurones, KNN, Boosting
Koopman[48] (2015)	Surveillance Mortalité	Classer les certificats de décès selon quatre pathologies : Diabète, Grippe, Pneumonie, VIH	• Recherche de mots clés • SVM
Sada[31] (2016)	Recherche/Epidémiologie	Identification des cancers hépatocellulaires à partir du dossier patient électronique	• Conditionnal Random Field
Trinidad[51] (2016)	Surveillance mortalité	Identifier des drogues impliquées dans les décès	• Recherche de mots clés
Gundlapalli[43] (2017)	Systèmes d'aide à la décision clinique	Elaborer un algorithme de TAL afin d'extraire les concepts liés aux infections des voies urinaires associées aux cathéters à partir de rapports médicaux électroniques	• Recherche de mots à partir d'un lexique
Koopman[50] (2018)	Surveillance Mortalité	Identifier si un cancer particulier est la cause initiale d'un décès	Méthode combinant une méthode par règles et un SVM
Tvardik[41] (2018)	Surveillance	Détection des infections associées aux soins dans les dossiers médicaux	Outil SYNODOS

IV/ Problématique et objectifs de la thèse

Si, dans un délai très court, la disponibilité des causes médicales de décès issues des certificats électroniques constitue une avancée majeure pour la surveillance qualitative et réactive de la mortalité, l'analyse de ces données est rendue complexe par leur contenu. Les certificats de décès contiennent la description de la séquence morbide qui a conduit au décès, ainsi que les causes associées ou comorbidités qui ne sont pas directement liées au décès. Cette séquence peut être un enchaînement de pathologies, la description de symptômes ou syndromes, des traumatismes ou intoxications. Elle peut contenir des diagnostics ou encore des traitements. Des informations similaires peuvent être exprimées de manière différente selon le médecin certificateur. La diversité d'expressions d'une même cause de décès rend difficile le suivi de chacune d'elles pour la surveillance réactive. Il devient alors nécessaire de regrouper les expressions en groupes homogènes pour suivre les variations de la mortalité par cause. La première question porte sur **la meilleure façon d'organiser les causes en groupes homogènes, qui constitueront des indicateurs que nous suivrons au cours du temps pour répondre aux objectifs de surveillance syndromique**.

Par ailleurs, les causes sont collectées sous la forme de texte libre. On retrouve des formulations différentes d'une même cause, des abréviations, des acronymes, des fautes d'orthographe ou de frappe. On trouve aussi des informations temporelles qui précisent des antécédents, des actes médicaux antérieurs (prothèse, intervention chirurgicale), ou bien la survenue d'une pathologie ou d'un symptôme par rapport à un autre. L'exploitation de ces données nécessite dans un premier temps d'appliquer des prétraitements afin d'harmoniser les textes. La seconde question vise 1) **à identifier les méthodes de traitement automatique des langues les plus appropriées pour classer les causes de décès dans les groupes homogènes, 2) à mettre en œuvre ces méthodes et les évaluer à partir de nos données.**

Une fois les causes médicales classées dans les groupes homogènes, il reste à définir comment analyser et interpréter ces indicateurs pour la surveillance en routine. Le certificat de décès rempli par le médecin contient plusieurs mots qui rendent compte de la séquence morbide. Pour un même certificat, on retrouve différentes causes de décès (soit dans le processus morbide, soit dans les causes associées), qui pourront être classées dans différents groupes homogènes. La prise en compte de l'ensemble des informations du certificat de décès va conduire à un décompte des effectifs de l'ensemble des groupes qui sera supérieur au nombre

de décès total toutes causes. Ce constat pose la question **de la méthode de prise en compte des causes multiples de décès afin de 1) mesurer le bon volume de décès, et 2) prendre en compte l'ensemble des causes dont l'enchaînement ou la coexistence ont contribué au décès pour l'analyse d'impact.**

Hypothèse :

Notre hypothèse de recherche est que la mise en œuvre de méthodes de classification des causes médicales de décès sous forme de texte libre dans des groupes homogènes, nous permettra d'analyser en temps quasi réel la mortalité par cause et ainsi compléter et enrichir la surveillance réactive toutes causes.

Objectifs :

L'objectif de nos travaux était de mettre en œuvre le système de surveillance syndromique de la mortalité à partir des causes médicales de décès en texte libre, en appliquant des méthodes de traitement automatique des langues.

Notre démarche se décompose en 4 objectifs :

- Décrire le système de surveillance de la mortalité en place depuis 2004 et basé sur les données d'état civil,
- Définir les groupes homogènes (que l'on appellera « regroupements syndromiques » dans la suite de ces travaux) à utiliser pour la surveillance réactive par cause,
- Développer et évaluer les méthodes issues du traitement automatique des langues pour le classement des causes de décès dans ces groupes homogènes,
- Proposer une méthodologie de prise en compte des causes multiples de décès, afin de mesurer le volume correct de décès et de prendre en compte l'ensemble des causes contenues dans les certificats pour l'évaluation d'impact.

Chapitre I : Bilan de la surveillance en routine de la mortalité depuis 2011

Suite aux conséquences extraordinaires de la canicule de 2003 qui ont mis en évidence un défaut de détection d'un événement sanitaire majeur, un nouveau système de surveillance réactif, le système SurSaUD, a été mis en place en France. Ce système a pour objectif de détecter précocement et de suivre en temps réel l'évolution de la morbidité et la mortalité liée à des événements, ainsi que d'en évaluer l'impact quantitatif sur la population.

Depuis 2004, la surveillance en temps réel de la mortalité s'appuie sur les données d'état civil du certificat de décès. L'analyse en routine de ces données permet de produire des bulletins décrivant les variations de la mortalité toutes causes, aux niveaux national, régional et départemental. Ces bulletins sont transmis aux autorités sanitaires, afin de les aider dans leurs prises de décisions pour la gestion des événements potentiellement à l'origine d'une surmortalité.

Bien que ce système de surveillance soit en place depuis près de 15 ans, il n'a jamais été décrit selon ses caractéristiques, son fonctionnement et son utilité.

Les CDC américains ont produit des recommandations qui permettent d'évaluer les performances d'un système de surveillance de santé publique (67). Outre l'utilité du système, il convient aussi de décrire les attributs du système :

- La simplicité d'un système de surveillance se réfère à la fois à sa structure et à sa facilité d'utilisation. Un système souple sous-entend qu'il s'adapte facilement à l'évolution des demandes d'information, à l'évolution de définitions ou de technologies.
- La qualité des données est reflétée par la complétude et la validité des données collectées.
- Un système de surveillance est défini comme acceptable lorsque des personnes et des organismes souhaitent participer à ce système.
- La sensibilité du système de surveillance est définie par sa capacité à détecter une variabilité inhabituelle, y compris la capacité à surveiller l'évolution du nombre de cas au cours du temps.
- Un système de surveillance qui est représentatif décrit avec précision l'occurrence d'un événement lié à la santé au fil du temps et sa répartition dans la population par lieu et par personne.
- La rapidité reflète la vitesse entre les étapes d'un système de surveillance.
- La stabilité fait référence à la fiabilité (c'est-à-dire la capacité de recueillir, de gérer et de fournir des données de façon appropriée sans défaillance) et à la disponibilité (la capacité d'être opérationnel lorsqu'il est nécessaire) du système de surveillance.

Afin de décrire et évaluer le système de surveillance de mortalité français, nous allons dans ce chapitre, présenter les données utilisées (collecte, fiabilité, et qualité des données) pour cette surveillance, puis nous décrirons certains attributs (couverture, population enregistrée) du système de surveillance. Nous décrirons ensuite les performances de ce système (capacité à détecter les variations connues ou inhabituelles de la mortalité), ainsi que son utilité pour les décideurs. Enfin, nous discuterons des limites du système.

I/ Les données de mortalité transmises par l'Insee

L'Insee reçoit les données issues des volets administratifs des certificats de décès saisies par les bureaux d'état civil. Seule une partie des bureaux d'état civil envoie chaque nuit de façon automatique et réactive des données saisies dans la journée. Le reste transmet ses données de façon non automatique et dans un délai de temps plus long. La base de données complète est ainsi consolidée après plusieurs semaines.

Les données transmises à Santé publique France par l'Insee sont issues des bureaux d'état civil envoyant leurs données de façon réactive et automatique chaque nuit.

Plus précisément, vers minuit, chaque jour, l'Insee compile dans un unique fichier, l'ensemble des données saisies par les bureaux d'état civil, ainsi que les données sur les décès survenus dans ces communes dans les 30 jours précédents. La transmission du fichier à Santé publique France est automatisée et s'effectue entre minuit et 3h du matin, 6 jours par semaine. Les données ne sont pas transmises à l'agence dans la nuit du samedi au dimanche. Ce fichier est déposé automatiquement sur un serveur de l'agence.

Au démarrage de la transmission des données à l'agence par l'Insee, les données étaient issues d'un échantillon fixe de 147 communes qui enregistraient 50% de la mortalité en France. Ces communes étaient essentiellement des communes de grandes tailles ou des communes sur lesquelles étaient localisés des établissements de santé. En 2006, l'échantillon s'est ensuite élargi à 1042 communes qui enregistraient 68% de la mortalité nationale. En 2011, l'échantillon s'est étendu à 3062 communes. Depuis 2015, l'échantillon s'élargit automatiquement, intégrant les données de tous nouveaux bureaux d'état civil transmettant les données de façon automatique à l'Insee. A ce jour, l'agence reçoit les données de plus de 7300 communes (La France est découpée en environ 35 000 communes).

La description du système réalisée dans ce travail repose sur les données saisies dans l'échantillon de 3062 communes que l'on appellera dans la suite de ce chapitre « échantillon de

routine ». Cet échantillon permet d'avoir des données historiques depuis 2011 et permet donc de comparer les évolutions de la mortalité sur plusieurs années consécutives.

Ces données contiennent des informations démographiques et administratives : l'année de naissance, le sexe, la date de décès et la commune du décès. Ces variables ne contiennent aucune donnée manquante. Par ailleurs, on récolte également la date de transmission des certificats de décès permettant de mesurer le délai entre le décès et la date de transmission des données à l'agence.

En effet, compte tenu des délais légaux de déclaration d'un décès (24h hors week-end et jours fériés) et du délai de saisie des informations par les bureaux d'état civil, on observe un délai entre la survenue du décès et la réception des informations par l'agence. Ces délais de transmission peuvent être allongés ponctuellement. C'est notamment le cas lors de jours fériés, week-end allongés, ponts, vacances scolaires, très forte période épidémique. En moyenne, l'agence reçoit les informations de 50% des décès survenus un jour J dans un délai de 3 jours, 90% dans un délai de 7 jours et 95% dans un délai de 10 jours (Figure 6).

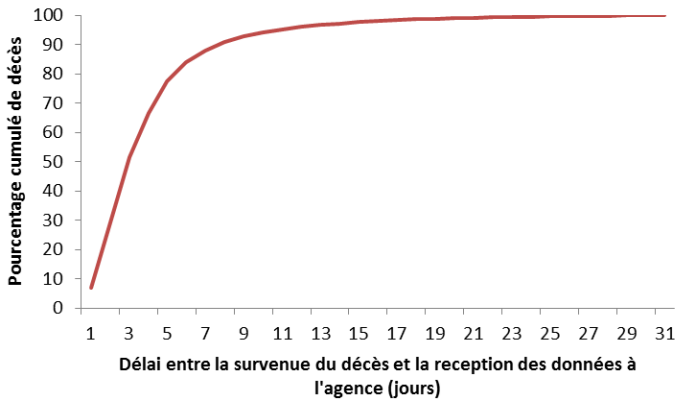


Figure 6 : Pourcentage cumulé de décès reçus en routine par Santé publique France selon le délai de transmission des données (en jours), France, 2017

Dans un délai inférieur à 7 jours, la part cumulée de décès disponibles à Santé publique France varie fortement selon le jour de survenue des décès, essentiellement du fait de l'ouverture des bureaux d'état civil uniquement les heures et jours ouvrés. Ainsi, pour des décès survenus en début de semaine, le délai de disponibilité des données à Santé publique France pour ces certificats sera plus court que pour des décès survenus en fin de semaine (Figure 7).

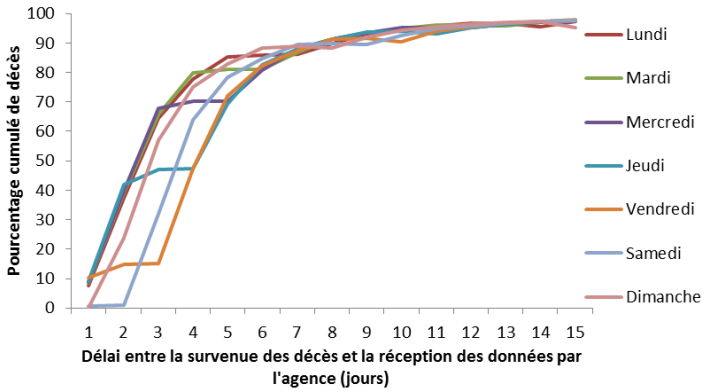


Figure 7 : Pourcentage cumulé de décès reçus à Santé publique France en fonction du délai de transmission (en jours) selon le jour de la semaine de survenue du décès, France, 2017

Ces caractéristiques de délais devront être considérées pour l'interprétation des variations de la mortalité et la communication auprès des décideurs.

II/ Méthode

1) Estimation de la couverture du système de surveillance

La couverture du système de surveillance de la mortalité est définie comme le nombre de décès enregistrés par le système (« Echantillon de routine ») sur une période donnée divisé par le nombre exhaustif de décès sur la même période.

Le nombre exhaustif de décès est fourni par le Centre d'épidémiologie sur les causes médicales de décès (Inserm-CépiDC) en charge de la base nationale exhaustive de mortalité.

$$Couverture\ du\ système = \frac{Nb\ décès\ INSEE_{échantillon\ routine}}{Nb\ décès\ exhaustif\ CépiDc}$$

En raison des délais de codage des causes de décès par le CépiDc, seules les données exhaustives de mortalité des années 2011, 2012 et 2013 étaient disponibles au moment de l'étude.

La couverture du système a été calculée entre 2011 et 2013 aux niveaux national, régional et départemental, ainsi que par classe d'âge : <=1 an, 2-14 ans, 15-44 ans, 45-64 ans, 65-84 ans, et 85 ans et plus.

Afin de mesurer la stabilité de la couverture au cours d'une année, entre 2011 et 2013, la couverture mensuelle du système a été calculée aux niveaux national et régional. La couverture

était considérée comme stable si la différence entre les couvertures mensuelles d'un mois à l'autre était inférieure à 2%.

En l'absence de la disponibilité des données exhaustives du CépiDc pour les années 2014 à 2016, nous ne pouvions pas évaluer la stabilité de la couverture annuelle et mensuelle sur cette période. Nous avons donc fait l'hypothèse que la couverture était stable sur ces années et qu'elle correspondait à la moyenne de la couverture annuelle et mensuelle pour les années 2011 à 2013.

Afin de vérifier notre hypothèse, nous avons comparé le nombre de décès Insee total issu de la base complète de l'Insee (« Données Insee ») au nombre de décès extrapolé par année et par mois. Le nombre de décès extrapolé a été calculé comme le nombre de décès enregistré par le système (« échantillon de routine ») divisé par la couverture annuelle moyenne calculée entre 2011 et 2013. Les nombres annuels et mensuels de décès extrapolés ont été calculés par an et par mois entre 2014 et 2016.

$$Nombre\ de\ décès\ extrapolé_{2014-2016} = \frac{Nb\ décès\ INSEE_{échantillon\ de\ routine\ (2014-2016)}}{Couverture\ moyenne\ entre\ 2011\ et\ 2013}$$

Cette différence entre le nombre extrapolé et le nombre total Insee a été exprimée en proportion par rapport au nombre de décès Insee total (« Données Insee »). Une très faible différence suggérait que la couverture entre 2014 et 2016 était stable et similaire à celle de 2011 à 2013. Nous avons défini des critères arbitraires pour décrire ces différences :

- Différence comprise entre [-1% et 1%] : couverture comparable,
- Différence supérieure à ±1% : couverture non comparable.

2) Description des décès enregistrés

La répartition des décès enregistrés par le système entre 2011 et 2016 a été décrite par sexe, puis par classe d'âge (0-14 ans, 15-64 ans, 65-84 ans, 85 ans et plus) aux niveaux national et régional et présentée en proportions de décès par rapport à l'ensemble des décès enregistrés par le système. L'âge moyen de décès a aussi été calculé aux niveaux national et régional et pour chaque année.

3) Détection des augmentations inhabituelles du nombre de décès et performance du système de surveillance

Un des principaux objectifs du système de surveillance de la mortalité est la détection précoce d'augmentation attendue ou inattendue de la mortalité. Cette détection s'appuie sur un modèle statistique qui a été développé et mis en place dans le cadre du projet Européen Euromomo. Ce projet, impliquant des acteurs nationaux et internationaux comme l'European Center for Disease Prevention and Control (ECDC) ou encore le Bureau régional de l'OMS pour l'Europe, a été mis en place dans l'objectif de développer et d'assurer une surveillance coordonnée de la mortalité dans tous les pays européens. Chaque pays participant contribue à l'évaluation d'impact des risques associés aux principales menaces pour la santé aux niveaux national et européen, en effectuant de manière réactive des analyses de variations hebdomadaires de la mortalité toutes causes pour différents groupes de population. Les informations produites sur l'impact d'un événement sur la mortalité sont compilées et partagées avec les autorités européennes pour améliorer les réponses sanitaires.

Plus spécifiquement, ce consortium a développé et mis à disposition des pays participants un modèle statistique conçu pour la détection précoce, la mesure et la comparaison en temps quasi réel des indicateurs de surmortalité toutes causes, dans différents groupes de population (68). Pour que le modèle réponde à la fois aux exigences liées à la réactivité d'un système de surveillance et au délai de transmission des données et donc de leur disponibilité, le choix du pas de temps de l'analyse s'est porté sur l'échelle hebdomadaire.

Le modèle implémenté sur les données françaises est basé sur la comparaison du nombre hebdomadaire de décès observés au nombre hebdomadaire de décès attendus. La modélisation du nombre hebdomadaire de décès attendus s'appuie sur plusieurs hypothèses : 1/ la mortalité hebdomadaire est comparable à une série temporelle poissonienne avec une tendance et parfois un cycle sinusoïdal sur une période de 1 an (69, 70), 2/ les mortalités hivernales et estivales sont modifiées par des facteurs tels que les infections virales en hiver (71, 72) ou les vagues de chaleur en été (73) entrainant des excès de mortalité d'amplitude variable, 3/le printemps et l'automne, à l'inverse, sont des périodes durant lesquelles des facteurs extérieurs sont moins susceptibles d'entrainer des excès de décès.

A partir de ces hypothèses, le nombre hebdomadaire de décès attendus a alors été estimé à partir d'un modèle linéaire généralisé de type régression de poisson corrigé pour la surdispersion et utilisant uniquement les données des mois de printemps et automne. Le

modèle est ajusté sur une période historique valide avec un minimum de 3 ans et un maximum de 5 ans de données. Il a été ajusté de façon différente selon des sous-groupes d'analyse :

- Prise en compte d'une tendance linéaire et une saisonnalité :
    - Au niveau national et pour les régions métropolitaines tous âges, et pour les 15-64 ans, 65-84 ans et les 85 ans et plus.
- Prise en compte d'une tendance linéaire mais pas de saisonnalité :
    - Au niveau national et pour les régions métropolitaines pour les 0-14 ans,
    - Pour les régions d'Outre-mer tous âges, 0-14 ans, 15-64 ans, 65-84 ans et 85 ans et plus.

Ce modèle fournit également le seuil au-delà duquel le nombre observé de décès est significativement supérieur au nombre attendu de décès.

Afin de décrire et comparer les variations hebdomadaires de la mortalité entre les régions et/ou les classes d'âges en France, nous avons également utilisé un indicateur complémentaire : le Z-score. En effet, l'écart entre les nombres de décès observés et attendus, correspondant à un excès de décès s'il est positif, ne peut être comparé entre des régions ou groupes d'âges de taille de population différente. Le Z-score est défini comme la différence entre le nombre de décès observés et le nombre de décès attendus, divisée par l'écart-type du nombre attendu. Les fluctuations hebdomadaires de la mortalité sont considérées habituelles quand le Z-score est compris entre -2 et 2.

Différents niveaux de Z-score ont été considérés afin de caractériser les variations de la mortalité (68):

- Z-score <2 : pas d'excès de mortalité significatif,
- Z-score compris entre [2-4[: faible excès de mortalité,
- Z-score compris entre [4-6[: excès de mortalité moyen,
- Z-score≥6 : excès de mortalité élevé.

Une alarme statistique (ou signal) a été déclenchée quand le Z-score observé pour une semaine donnée dépassait le seuil de 2.

Une alerte sanitaire a été considérée si une alarme statistique était concomitante avec un événement spécifique et identifié, ou si le Z-score dépassait 2 pendant au moins deux semaines consécutives, sans connaissance d'un évènement potentiellement à l'origine de cette hausse.

Les périodes d'excès de mortalité identifiées entre 2012 et 2016 ont été décrites et analysées.

Une caractéristique importante d'un système de surveillance dont l'objectif est la détection de variation habituelle ou inhabituelle de la mortalité est sa sensibilité, c'est-à-dire sa capacité à détecter une augmentation réelle, ici de mortalité (67). Pour évaluer les performances de détection du système, nous avons calculé la sensibilité et la spécificité du système aux niveaux national et régional pour les années 2012 et 2013. Pour cela, nous avons comparé sur la période 2012 à 2013 les nombres de semaines et les dates pour lesquelles des alarmes ont été détectées en utilisant les données exhaustives de l'Inserm-CépiDc, au nombre de semaines et aux dates pour lesquelles des alarmes ont été détectées en utilisant les données Insee reçues en routine (« Echantillon de routine »). Le même modèle a été appliqué aux deux jeux de données.

4) Utilité du système de surveillance

Selon les CDC, un système de surveillance est utile s'il contribue à la prévention et au contrôle des effets indésirables d'événements sur la santé, ou s'il permet de déterminer qu'un événement indésirable sur la santé, que l'on croyait sans danger, est important (67).

Afin d'illustrer l'utilité du système, nous avons identifié 3 exemples d'événements ayant conduit à un excès de mortalité. Nous avons décrit ces événements, ainsi que le processus de remontée des informations aux décideurs.

Ces travaux ont nécessité la collecte de données supplémentaires pour les 3 exemples suivants :

- L'étude des variations de la mortalité durant les hivers : Les périodes d'épidémies de grippe entre 2012 et 2016 en France (74-78), et le nombre hebdomadaire de diagnostics de grippe du réseau Sentinelles entre 2012 et 2016 (79),
- L'étude des variations de la mortalité durant les étés : Le nombre de vagues de chaleur et leurs périodes entre 2012 et 2016 (80),
- L'étude de variations inattendues de la mortalité : Le nombre hebdomadaire de cas de Chikungunya en Guadeloupe durant l'épidémie de 2014 (81).

L'évaluation des actions et mesures de gestion prises suite à la remontée des informations produites par le système n'a pas été effectuée dans ce travail.

L'ensemble des analyses a été effectué avec R 3.3.0.

III/ Résultats

1) La couverture du système

1.1 Au niveau national

Le système de surveillance de la mortalité enregistrait 77,6% de la mortalité totale en France en 2011, 77,2% en 2012 et 77,6% en 2013. La couverture était stable sur ces 3 années avec des variations de ± 0.5% (Figure 8).

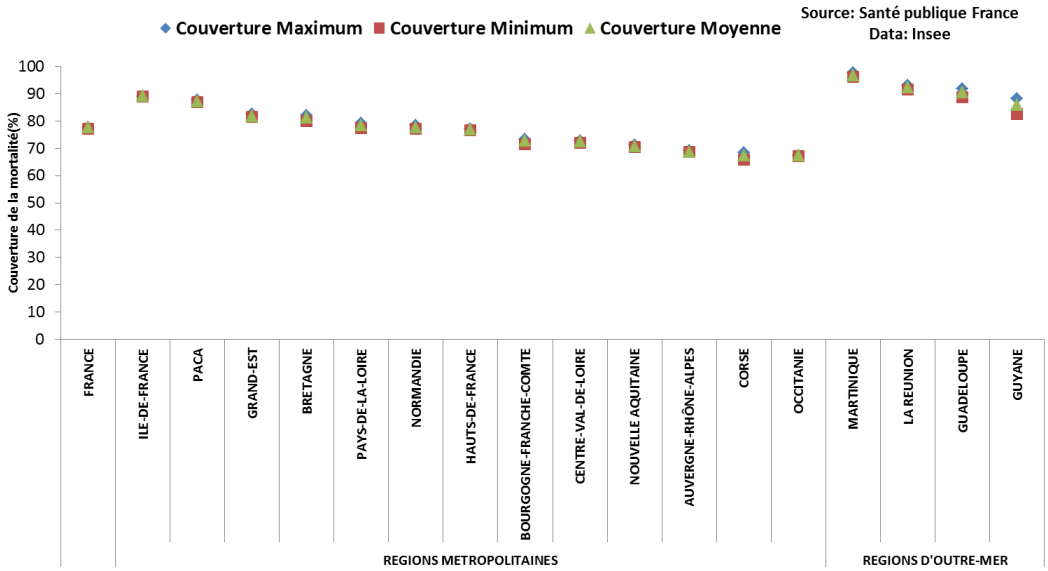


Figure 8 : Couverture moyenne, maximum et minimum du système de surveillance syndromique de la mortalité fondée sur les données d'état-civil, tous âges entre 2011 et 2013 en France aux niveaux national et régional

1.2 Au niveau régional

La couverture moyenne du système de 2011 à 2013 en région métropolitaine variait de 67% en région Occitanie à 89% en région Ile-de-France (Figure 8). Sept régions sur 13 avaient une couverture moyenne supérieure ou égale à la couverture moyenne au niveau national allant de 77,5% à 89% de couverture.

Cette couverture restait stable sur les 3 années pour chacune des régions, avec des variations allant jusqu'à ± 2% d'une année à l'autre. Le détail des couvertures est présenté en Annexe 2.

De même, pour les régions d'Outre-mer, les couvertures variaient de 86% pour la Guyane à 96% pour la Martinique (Figure 8) et étaient stables pour les 3 années avec des variations de moins de 2%, sauf pour la Guyane pour laquelle nous avons observé une variation de ±5,5% en 3 ans.

1.3 Au niveau départemental

La couverture de la mortalité par département en France métropolitaine variait de 42% pour la Lozère à 98% pour Paris. En Outre-mer, elle variait de 86% pour la Guyane à 96% pour la Martinique (Figure 9). Cette couverture était stable pour les 3 années avec des variations inférieures à 2% sur l'ensemble des années pour la majorité des départements. Les variations de la couverture sur les 3 années étaient supérieures à 2% pour 39 départements : 29 départements avaient des variations comprises entre 2% et 4 % et 10 départements avaient des variations de couvertures supérieures à 4% dont le Puy-de-Dôme, Savoie, Haute-Savoie, Moselle et la Seine-et-Marne pour lesquels on observait des variations allant de 5,5% (Puy-de-Dôme) à 13,7 % (Moselle). Le détail des couvertures par département est présenté en Annexe 2.

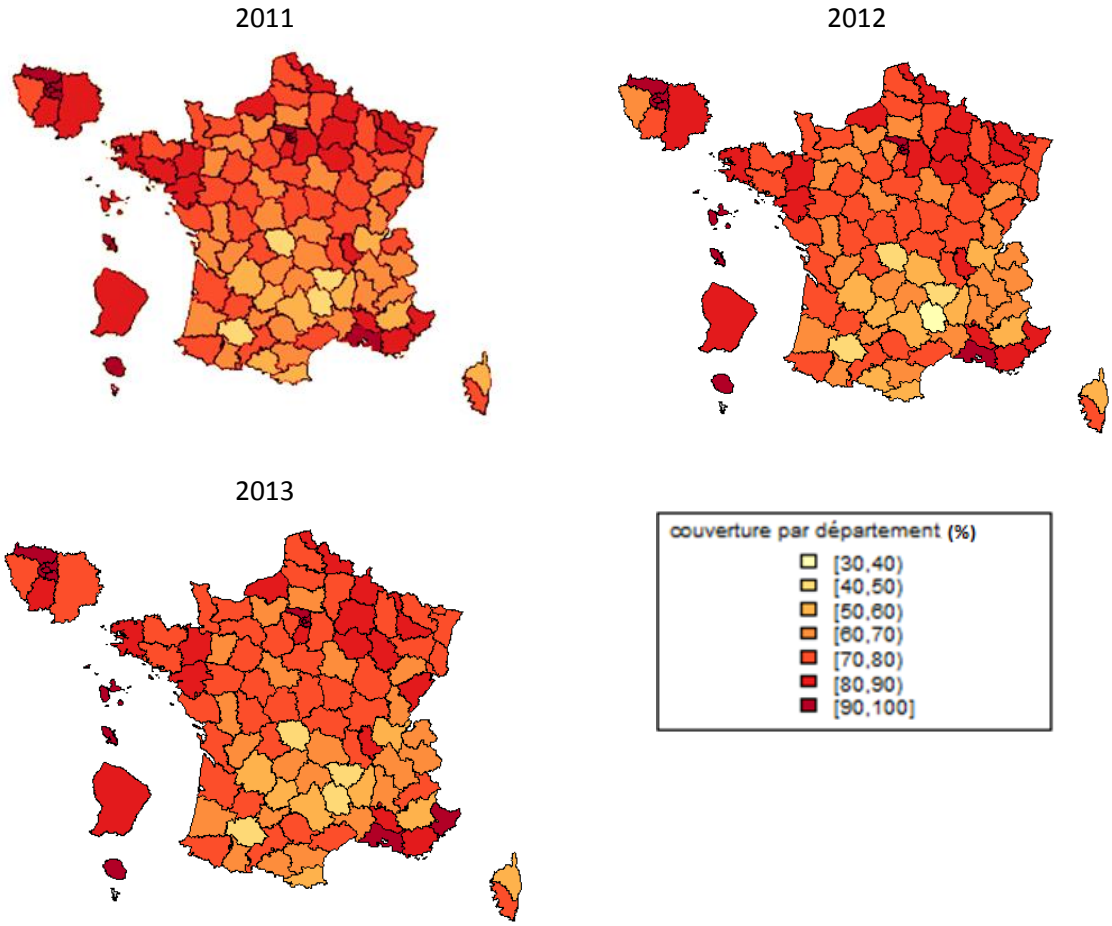


Figure 9 : Couverture du système de surveillance syndromique de la mortalité fondé sur les données d'état-civil entre 2011 et 2013 par département français

1.4 Couverture par classe d'âge

Au niveau national, la couverture de la mortalité variait en fonction des classes d'âge. Alors que la couverture moyenne était de 93% entre 2011 et 2013 chez les enfants ≤ 1 an, la couverture était comprise entre 70% et 75% chez les 15-44 ans et les personnes âgées de 85 ans et plus. Pour les 4 autres classes d'âge, la couverture était comprise entre 78% et 82% en moyenne sur les 3 années (Figure 10). On observait une stabilité de la couverture sur les 3 années pour chacune des classes d'âge.

L'ordre de couverture des classes d'âge observé au niveau national (les enfants ≤1 an avaient la couverture la plus élevée, les 15-44 ans et ≥85 ans avaient les couvertures les plus basses et les autres classes d'âge avaient des couvertures intermédiaires) l'était aussi pour les régions métropolitaines mais à des niveaux différents de ceux des nationaux. La couverture du système pour la surveillance des enfants âgés de 0 à 1 an était comprise entre 87% et 100% sur l'ensemble des régions métropolitaines, celles des personnes de 15-44 ans et des 85 ans et plus étaient comprises entre 58% et 87%, alors qu'elle était comprise entre 67% et 90% pour les autres classes d'âge. Ces couvertures étaient stables sur les 3 années sauf pour la classe d'âge 2-14 ans et pour la région Corse (Annexe 3).

Pour les régions d'Outre-mer, les couvertures par classe d'âge avaient un ordre différent de celui du niveau national et des régions métropolitaines. Les niveaux de couverture étaient plus homogènes, compris entre 85% et 90%, pour l'ensemble des classes d'âge sauf pour la Guyane (Figure 10).

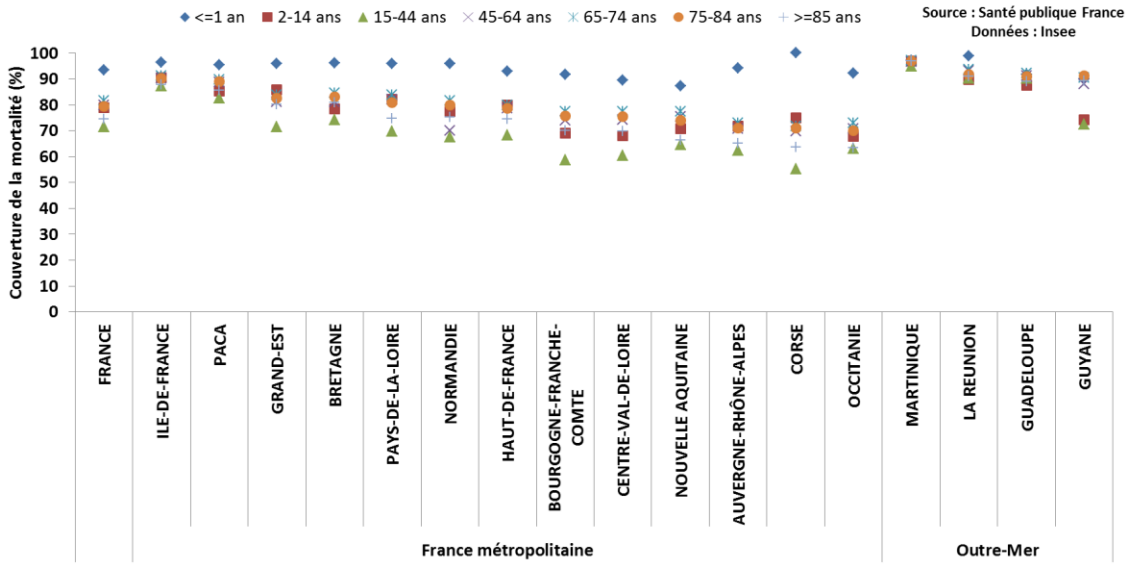


Figure 10 : Couverture du système de surveillance syndromique de la mortalité fondé sur les données d'état-civil par classe d'âge entre 2011 et 2013 en France et au niveau régional, moyenne

1.5 Stabilité temporelle de la couverture

La couverture mensuelle de la mortalité entre 2011 et 2013 était stable au niveau national, avec des variations inférieures à ± 1% par rapport à la couverture moyenne.

Au niveau régional, la couverture mensuelle était stable pour la majorité des régions métropolitaines avec des différences entre la couverture mensuelle et la couverture moyenne comprise entre]-2 et +2%[à l'exception de la Corse et du Centre-Val-de-Loire, pour lesquelles on observait des variations de couverture supérieure à ±2% par rapport à la couverture moyenne.

La couverture mensuelle dans les régions d'Outre-mer présentait des variations allant de ± 2% à ±13%.

On observait une très faible différence de ± 1% entre le nombre de décès mensuel au niveau national (« Données Insee ») et le nombre de décès extrapolés calculés pour chaque mois entre 2014 et 2016 (Figure 11), sauf pour les mois de Mars, Octobre et Novembre 2014 et le mois de Novembre 2015, pour lesquels les variations étaient comprises entre +1% et +2%.

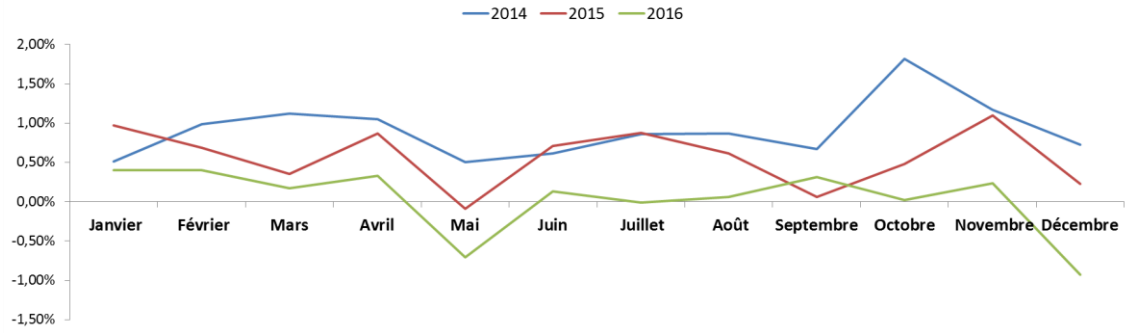


Figure 11 : Différence entre le nombre décès total estimé par l'Insee («Données Insee ») et le nombre de décès extrapolé par mois sur la période 2014-2016, exprimée en proportion, Tous âges, France

2) Description des décès enregistrés

Entre 2011 et 2016, 49,2% des décès enregistrés étaient des femmes. Plus de 81% des personnes décédées étaient âgées de 65 ans et plus. Moins de 1% étaient des enfants de moins de 15 ans (Tableau 2). Une augmentation de l'âge moyen de décès de 1,4 ans entre 2011 et 2016. Sur la période 2011-2016, l'âge moyen de décès dans les régions d'Outre-mer était plus faible que dans les régions métropolitaines et plus particulièrement en Guyane (58,8 ans vs 77,7 ans). Aussi, 2 à 11% des décès étaient des enfants de moins de 15 dans ces régions.

Tableau 2 : Proportion de décès enregistrés par le système de surveillance syndromique de la mortalité fondé sur les données d'état-civil entre 2011 et 2016 en France, par sexe, classes d'âge, âge moyen de décès au niveau national et régional.

	N	Sexe (%)	Classes d'âge (%)				
Variables socio démographiques/ année		**Femme**	**0-14 ans**	**15-64 ans**	**65-84 ans**	**>=85ans**	**Age moyen de décès (Années)**
FRANCE							
2011	424 802	48,6	0,9	19,2	40,8	39,2	76,9
2012	441 656	49,2	0,8	18,1	40,0	41,0	77,5
2013	442 540	49,1	0,9	17,8	39,6	41,8	77,6
2014	437 637	49,1	0,8	17,5	39,5	42,1	77,7
2015	462 939	49,4	0,8	16,6	38,9	43,8	78,3
2016	460 545	49,4	0,8	16,3	38,7	44,3	78,3
2011-2016	2 670 119	49,2	0,8	17,5	39,5	42,1	77,7
REGION (2011-2016)							
Ile-de-France	396 383	50,0	1,5	20,7	38,3	39,9	76,0
Auvergne-Rhône-Alpes	268 399	48,9	0,9	16,0	39,9	43,3	78,3
Nouvelle-Aquitaine	262 490	48,3	0,5	15,7	38,9	44,9	79,0
Provence-Alpes-Côte-D'Azur	256 629	49,9	0,6	15,1	38,8	45,6	79,1
Haut-de-France	248 753	49,1	0,7	21,1	41,4	36,7	76,1
Grand-Est	243 274	50,1	0,7	17,3	41,7	40,3	77,6
Occitanie	222 614	48,4	0,7	15,9	38,9	44,5	78,6
Bretagne	161 591	49,9	0,6	16,6	39,1	43,8	78,5
Pays-de-la-Loire	152 361	48,3	0,7	17,1	38,6	43,6	78,3
Normandie	147 892	49,2	0,6	18,1	39,8	41,6	77
Bourgogne-Franche-Comté	125 101	49,0	0,6	15,6	39,9	43,9	78,8
Centre-Val de Loire	108 413	48,3	0,6	15,3	38,8	45,3	79,1
Corse	12 071	48,6	0,2	15,4	42,3	42,1	78,8
La Réunion	25 721	45,3	3,2	28,9	42,3	25,6	69,9
Guadeloupe	16 613	46,2	1,9	23,7	39,8	34,5	73,9
Martinique	17 775	48,9	1,5	18,5	41,1	38,9	76,5
Guyane	4 039	42,5	10,7	39,7	32,0	17,6	58,8

3) Détection des augmentations inhabituelles du nombre de décès et performance du système de surveillance.

Sur la période 2012-2016, on dénombrait 180 semaines en excès de décès au total au niveau national et un total de 1179 alarmes sur l'ensemble des régions sur la même période (Z-score≥2) pour les cinq classes d'âge.

Parmi celles-ci, on observait 3 périodes d'excès de décès moyen à élevé pendant plusieurs semaines consécutives. Des périodes d'excès de décès faible et ponctuel ont aussi été observées avec uniquement une ou deux semaines consécutives avec une augmentation significative de la mortalité (Z-score ≥2).

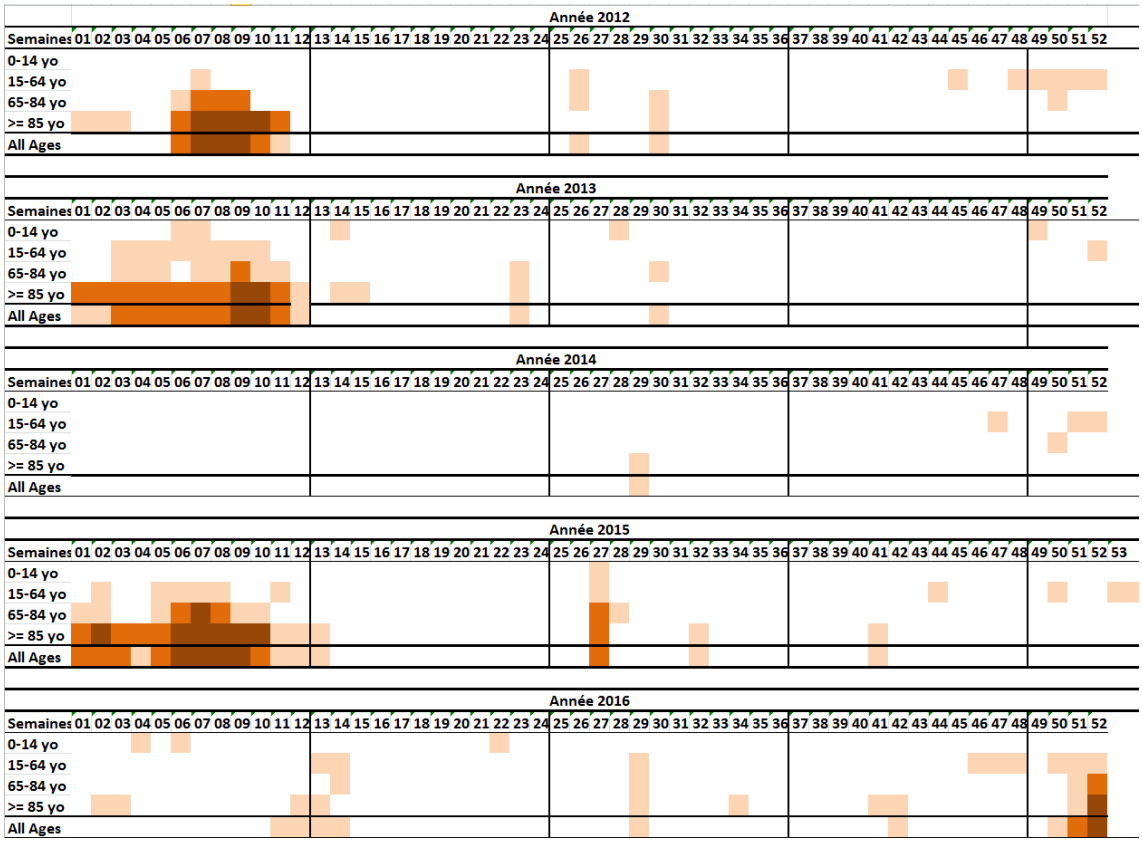
Les périodes d'excès de décès élevé ont été observées durant les hivers 2011-2012, 2012-2013 et 2014-2015 et ont duré respectivement 6, 12 et 13 semaines consécutives, avec un Z-score atteignant un pic supérieur à 6. Durant ces périodes, les personnes âgées de plus de 85 ans étaient les plus touchées même si on observait un excès de décès de faible niveau chez les 15-64 ans (Figure 12a).

Entre 2012 et 2016, durant les périodes estivales, 5 périodes d'excès de décès moyen de décès ont été observées avec un Z-score maximum compris entre 2 et 4. Durant ces périodes, les personnes âgées de plus de 85 ans étaient aussi les plus concernées (Figure 12a).

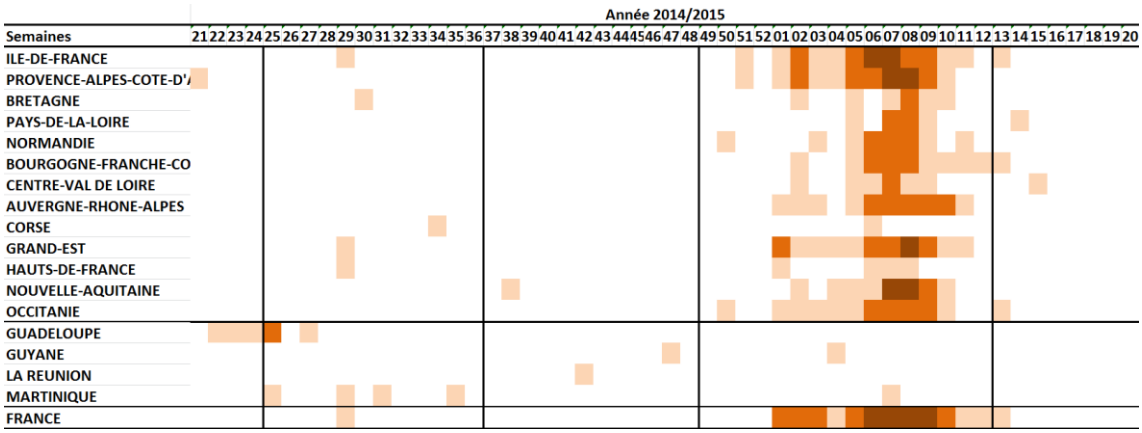
La figure 12b décrit le niveau de Z-score par semaine au niveau régional entre 2014 et 2015. On observait une augmentation du nombre de décès pour l'ensemble des régions métropolitaines (sauf la Corse) durant l'hiver, d'amplitudes et de durées différentes.

Dans les régions d'Outre-mer, sur les mêmes périodes, des excès de mortalité limités ont été observés sauf pour la Guadeloupe. Entre juin et juillet 2014, un excès de décès a été identifié dans cette région entre les semaines 22 à 25 et en semaine 27 avec un Z-score compris entre 4 et 6 en semaine 25 (Figure 12b).

(a)



(b)



Légende: Blanc: Pas d'excès de décès (Z-score <2); beige: Faible excès de décès (Z-Score compris entre [2; 4[ ); Orange: Excès de décès moyen (Z-Score compris entre [4-6[); Marron: Excès de décès élevé (Z-Score ≥6)

Figure 12 : Niveau de l'écart hebdomadaire standardisé de la mortalité par rapport au seuil (Z-score) en France, entre 2012 et 2016 par classe d'âge (a), sur la période 2014-15, tous âges confondus, dans les régions (b)

La sensibilité et la spécificité du système à détecter les semaines en excès de décès étaient respectivement de 0,96 et 0,99 au niveau national, tous âges. La sensibilité du système variait entre les classes d'âges au niveau national, allant de 0,65 pour les 15-44 ans à 0,86 pour les personnes âgées de 85 ans et plus. De même, la spécificité du système variait de 0,93 pour les 15-64 ans à 0,99 pour les 0-14 ans.

Au niveau régional, la sensibilité du système à détecter les semaines en excès de décès variait entre 0,42 et 1 sauf pour la Corse, alors que la spécificité était supérieure à 0,89 pour l'ensemble des régions (Tableau 3).

Tableau 3 : Sensibilité et spécificité du système à détecter les semaines où le Z-score dépassait 2 pour les différents groupes d'âge en France et par région pour tous les âges entre 2012 et 2013

	Nombre de semaines où le Z-score dépasse 2 avec les données exhaustives du CépiDc	Sensibilité	Specificité
France			
0-14 ans	4	0,75	0,99
15-64 ans	17	0,65	0,93
65-84 ans	19	0,74	0,94
≥85 ans	21	0,86	0,96
Tous âges	23	0,96	0,99
Régions			
Normandie	24	0,79	0,94
Grand-Est	23	0,83	0,95
Ile-de-France	22	0,77	0,94
Bourgogne-Franche-Comte	22	0,55	0,89
Nouvelle-Aquitaine	20	0,85	0,97
Auvergne-Rhône-Alpes	19	0,89	0,98
Provence-Alpes-Côte-D'azur	17	1,0	1,0
Haut-de-France	16	0,56	0,92
Pays-de-la-Loire	14	0,43	0,92
Bretagne	12	0,83	0,98
Centre-val de Loire	12	0,42	0,93
Corse	12	0,08	0,89
Occitanie	9	0,89	0,99
Guyane	4	0,5	0,98
La Réunion	3	0,67	0,99
Guadeloupe	0	-	1
Martinique	**0**	-	1

4) Utilité du système

Entre 2012 et 2016, les périodes hivernales d'excès de décès élevés étaient concomitantes avec les périodes d'épidémies de grippe (Figures 13a et 13b) et plus particulièrement chez les personnes âgées de 85 ans et plus (Figure 13a). Lors de la période hivernale d'excès de décès de 2015, on a pu observer une augmentation du nombre de décès avant le début de l'épidémie. Cette augmentation suggère que des facteurs autres que l'épidémie grippale ont entrainé une augmentation du nombre de décès en hiver.

Durant la période estivale, 3 des 5 périodes d'excès ponctuels observés entre 2012 et 2016 étaient concomitants avec des vagues de chaleur (Figure 13a et 13b).

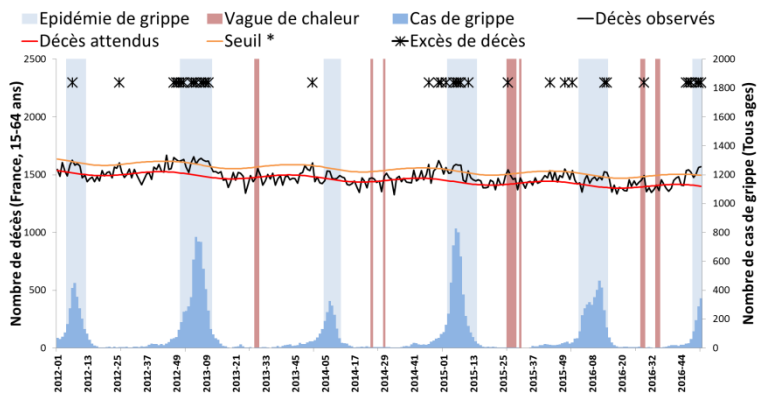
L'excès de décès survenus en Guadeloupe entre juin et juillet 2014 s'est produit durant la période d'épidémie de chikungunya qui a duré de la semaine 13 à 35 de 2014 (Figure 13c). Cet excès a touché toutes les classes d'âge et était concomitant avec le pic de l'épidémie en semaine 24 de 2014.

Nous avons aussi pu observer des variations de la mortalité qui étaient concomitantes avec des événements clairement identifiés tels que les attentats du 13 Novembre 2015 à Paris (Figure 13d). Cette augmentation soudaine et élevée de la mortalité était particulièrement notable pour les personnes âgées de 15 à 44 ans.

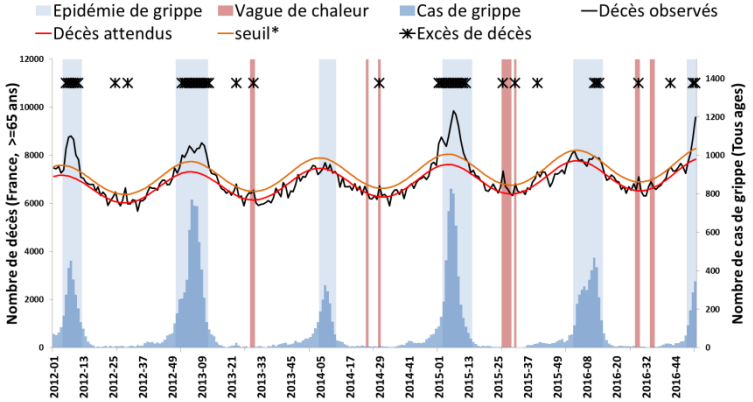
Chaque semaine, des bulletins décrivant les tendances de la mortalité et fournissant les chiffres sur l'augmentation du nombre de décès en cours sont produits. Du fait des délais dans la disponibilité des données, l'analyse des fluctuations d'une semaine calendaire complète est effectuée avec un délai minimum de deux semaines afin de disposer d'une complétude satisfaisante et d'une validité suffisante. L'évaluation quantitative de l'impact d'un événement sur la mortalité est réalisée à partir des données consolidées, soit dans un délai minimum de 3 semaines après la survenue de l'événement.
Ces bulletins font partie des supports de la réunion hebdomadaire de surveillance sanitaire qui a lieu chaque semaine à la Direction Général de la Santé (DGS). Ils sont aussi publiés chaque mercredi sur le site internet de Santé publique France.
Lorsqu'un excès de décès majeur est observé, une note décrivant le phénomène et fournissant des indications méthodologiques pour l'interprétation des données est rédigée et envoyée à la DGS. En cas de demande du Ministère en charge de la santé, les chiffres de la mortalité peuvent être fournis quotidiennement.

Cette étude a fait l'objet d'un article publié dans European journal of public health (82).

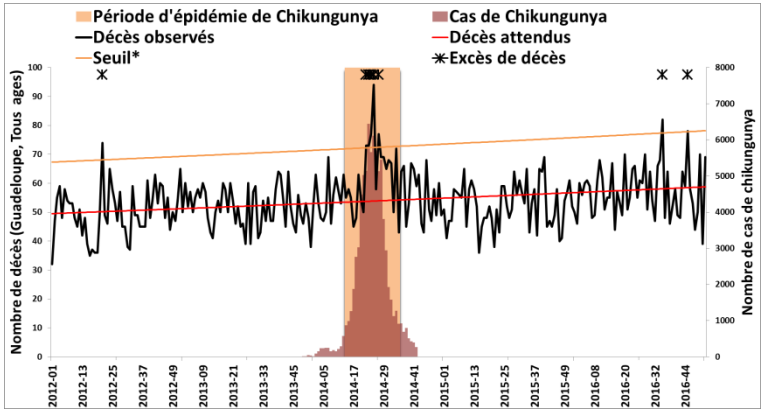
(a)



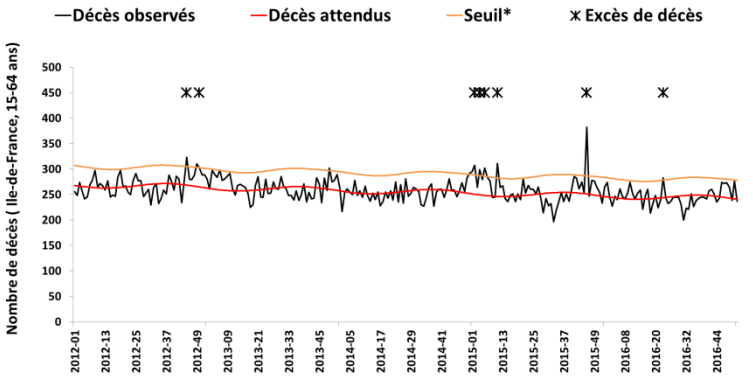
(b)



(c)



(d)



* Seuil au delà duquel le nombre observé de décès est significativement supérieur au nombre attendu

Figure 13 : Fluctuations hebdomadaires de la mortalité observée et attendue entre 2012 et 2016, (a) avec les épidémies de grippe et les vagues de chaleur en France, pour les 15-64 ans, (b) pour les >=65 ans, (c) avec le nombre hebdomadaire de cas de chikungunya en Guadeloupe tous âges, (d) en Ile-de France pour les 15-64 ans lors des attentats

IV/ Discussion

Si on se réfère aux attributs décrits par les CDC pour caractériser un bon système de surveillance, le système de surveillance de la mortalité est simple et acceptable puisque la structure est basée sur la collecte d'un seul type de données reçues quotidiennement et automatiquement par l'agence (transmises par un seul partenaire) et celles-ci ne contiennent aucune donnée manquante.

De plus, le système de surveillance de la mortalité enregistrait 77,5% de la mortalité totale entre 2011 et 2013 et couvrait l'ensemble du territoire ainsi que toutes les classes d'âge. Sur cette période, la couverture de la mortalité restait stable entre les années et au cours d'une année. Cette stabilité suggère que la couverture de la mortalité n'est pas influencée par les saisons et n'est pas modifiée par une variation de la mortalité.

Cependant, au niveau régional et de surcroit au niveau départemental, la couverture de la mortalité était plus hétérogène. La variabilité du niveau de couverture doit être prise en compte notamment pour l'extrapolation du nombre de décès lors de la mesure d'impact d'un événement sur la mortalité. L'extrapolation du nombre de décès doit être considérée avec prudence et doit être accompagnée d'éléments permettant son interprétation.

Le système a démontré sa capacité à détecter et à évaluer l'impact d'un événement sur la mortalité. Ces excès observés affectaient l'ensemble des classes d'âge et plus particulièrement les personnes âgées de 85 ans et plus.

Pour qu'un système d'alerte et de surveillance soit considéré comme efficace, il doit être capable de détecter un événement lorsqu'il se produit réellement, c'est-à-dire qu'il doit être sensible. Pour notre système, il s'agit de détecter un excès de décès (z-score ≥2). Au niveau national, 96% des excès de mortalité ont pu être repérés par le système sur la période d'étude. Cependant, au niveau régional les performances de détection étaient plus hétérogènes.
La sensibilité du système étant particulièrement liée au nombre hebdomadaire de décès plus le nombre hebdomadaire de décès était faible, plus la sensibilité du système était faible. De même, la sensibilité du système à détecter les semaines en excès de décès est dépendante de la couverture du système : plus la couverture du système était basse, plus la sensibilité était faible.

La surveillance en routine et l'analyse de ces données conduites au niveau national mais aussi régional s'appuient sur une méthodologie développée à l'échelle européenne. La production de bulletins hebdomadaires fournissant l'analyse et l'interprétation des variations de la mortalité aux niveaux national et régional permet de faciliter l'évaluation d'une situation sanitaire et

d'appuyer les décisions des autorités. Il est aussi utile à l'adaptation, au maintien et au renforcement des recommandations relatives aux mesures de contrôle et de prévention pour les futurs événements.

A cela s'ajoute une analyse basée sur la même méthodologie à l'échelle européenne, dans plus de 20 pays (83). Chaque semaine, les pays européens envoient leurs résultats issus du modèle pour qu'ils soient centralisés et analysés de façon groupée. Les résultats sont partagés à travers des bulletins transmis à l'ECDC et disponibles sur le site internet d'Euromomo (http://www.euromomo.eu/index.html). Des rapports et articles scientifiques sont aussi produits après chaque épisode d'épidémie hivernal.

Un système utile malgré certaines limites

La surveillance de la mortalité s'appuie sur un modèle statistique développé au niveau européen dont la construction a dû répondre au plus petit dénominateur commun à des fins de faisabilité et de comparabilité. Le système a été adapté au processus de collecte, gestion et analyse des données de l'ensemble des pays.

Une des limites du modèle est le choix de l'analyse à un pas de temps hebdomadaire. Cette échelle est appropriée pour mesurer les variations de la mortalité lors d'événements qui perdurent pendant plusieurs semaines telles que les épidémies hivernales. En revanche, elle est peu adaptée aux événements ponctuels et de courte durée comme par exemple certaines vagues de chaleur. En utilisant ce modèle pour l'évaluation d'impact d'un tel événement, le nombre de décès potentiellement lié à celui-ci sera alors moins précis. Une méthode plus adaptée pour l'évaluation des vagues de chaleur s'appuyant sur ces données a été mis en place (84).

Le choix d'un pas de temps quotidien pour le modèle Euromomo pourrait être une réponse à cette limite pour tout type d'événement.

Une autre limite du modèle est le calcul du nombre de décès attendu. Le modèle prend en compte la saisonnalité et la tendance observables à différents niveaux géographiques et pour les classes d'âges « Tous âges », « 0-14 ans », « 15 ans et plus ». Une évolution de ce modèle vise à inclure la structure d'âge de la population en offset pour améliorer l'estimation du nombre de décès attendu et prendre en compte les évolutions démographiques de la population, notamment celles de plus de 65 ans.

Le système de surveillance de la mortalité a permis d'identifier sur la période d'étude 3 périodes d'excès de décès élevé de plusieurs semaines consécutives et 5 périodes d'excès faible de 1 à 2

semaines consécutives entre 2012 et 2016. Ces périodes étaient majoritairement identifiées durant les périodes de grippe hivernales ou durant les vagues de chaleur estivales. Cependant, l'absence d'information sur les causes de décès ne nous a pas permis de déterminer la contribution exacte de ces événements à ces excès de décès.

Il est bien connu que le virus de la grippe contribue largement aux excès de mortalité hivernaux, et plus particulièrement lors d'épidémie à virus A (H3N2) (85, 86). La simultanéité d'autres événements hivernaux tels que la circulation d'autres virus respiratoires ou encore les vagues de froid intense, peuvent aussi contribuer à l'augmentation de la mortalité (87, 88). Santé publique France dispose aujourd'hui d'un modèle prenant en compte ces événements et permettant d'estimer la part attribuable de la grippe (de l'ordre de 70% de la mortalité totale) durant un excès de mortalité hivernale (89)

Le même constat peut être fait pour les vagues de chaleur. Elles sont connues pour toucher plus particulièrement les personnes vulnérables telles que les personnes âgées et les nourrissons et ce malgré la mise en place de mesures de prévention dès les premières augmentations de la température (90-94).

L'absence des causes de décès est plus particulièrement problématique lorsqu'un excès de décès survient en l'absence de tout événement clairement identifié pouvant affecter les variations de la mortalité, ou pour identifier la part de différents facteurs ou événements. Face à une telle situation, une investigation peut être menée avec l'aide de professionnels de santé lorsque la surmortalité concerne un groupe d'âge spécifique ou un périmètre géographique limité.

Dans le cas d'un excès de décès plus important, seules des hypothèses peuvent être formulées. Cela a notamment été le cas lors de l'excès de décès survenu durant l'épidémie de Chikungunuya en Guadeloupe en 2014. Le virus du Chikungunya est connu pour être transmis par les piqûres des moustiques *Aedes*. Les signes cliniques typiques de la maladie sont la fièvre et une arthralgie sévère (95). Cependant, cette infection n'est pas considérée comme létale, bien qu'une surmortalité ait été observée lors d'une précédente épidémie de Chikungunya à La Réunion en 2005-2006 (96, 97).

Si un lien causal n'a pas pu être établi entre cette épidémie et l'excès de décès, le système de surveillance a néanmoins permis de mettre en lumière, un potentiel effet inattendu de cette épidémie. Ceci a permis de compléter les connaissances sur cette arbovirose.

Enfin, en l'absence de causes de décès le système de surveillance est aussi capable d'identifier et de mesurer l'impact d'événements très spécifiques comme les attentats terroristes du 13 novembre 2015.

Ce système a donc permis de produire des informations pour les décideurs. L'utilité des mesures mises en place suite à ces remontées d'informations resterait à évaluer, en lien avec les différents acteurs concernés.

Vers une surveillance plus réactive et complète

Depuis 2015, les données de tout nouveau bureau d'état civil envoyées de façon automatique à l'Insee sont transférées à Santé publique France. En 2019, l'échantillon contient plus de 7300 bureaux d'état civil, ce qui permet d'augmenter la couverture de la mortalité sur le territoire. Cependant, cet échantillon pourra être utilisé en routine pour la surveillance quand un historique de données suffisant sera disponible. En effet, le modèle Euromomo qui estime le nombre attendu de décès requiert un historique de données de 3 à 5 années minimum.

En 2007, la mise en place de la certification électronique des décès a permis d'enrichir le système de surveillance de la mortalité avec une seconde source de données. La certification électronique des décès permet au médecin de déclarer un décès via une application web sécurisée. Le volet médical du certificat contenant les causes de décès sous forme de texte libre est alors transmis dans les minutes qui suivent à l'agence (98) (95% des certificats électroniques sont disponibles le lendemain du décès (99)). En plus d'augmenter la réactivité du système, une étude pilote montré la pertinence de l'utilisation de ces données pour répondre aux objectifs de la surveillance en routine de la mortalité (99). Cette nouvelle source de données pourra remédier au délai d'enregistrement des données du système actuel (90% de la mortalité reçue en 7 jours). Cependant, l'exploitation de ces données nécessite l'utilisation de méthodes appropriées de traitement automatique des langues en raison de leur format texte libre.

Dans le chapitre qui suit, nous présentons une description de cette seconde source de données.

Chapitre II : Une nouvelle source de données : La certification électronique des décès

I / Les données issues de la certification électronique des décès

1) Le volet médical du certificat électronique de décès : circuit et construction

Lors d'un décès, le médecin a la possibilité de remplir un certificat de décès au format électronique à travers l'application web « CertDc ». Cette application sécurisée gère à la fois les connexions CPS (Carte de professionnel de santé) et la connexion par identifiant ou mot de passe. Après la validation d'un certificat de décès électronique par le médecin, les données du volet médical sont envoyées sur le serveur de l'Inserm-CépiDc et sont immédiatement transmises à Santé publique France. Les médecins peuvent toutefois modifier ou compléter le certificat de décès dans les 96 heures qui suivent sa validation. Chaque modification entraine un renvoi complet des données du volet médical au CépiDc et à Santé publique France.
Le volet médical du certificat de décès est anonyme et contient (Figure 14) :

- Une partie contenant des informations sociodémographiques (date de naissance, date de décès, commune de décès…)
- Les causes de décès reportées dans 2 parties :
    - La partie I contient la séquence morbide qui a entrainé le décès. Cette séquence est indiquée de la cause initiale (champ (d)) à la cause immédiate (champ (a)). Au regard de chaque champ, le médecin dispose d'un champ dans lequel il peut indiquer l'intervalle de temps entre la survenue de la cause et le décès.
    - La partie II contient les autres états morbides ayant contribué au décès mais n'étant pas cités en partie I.
- Une partie « information complémentaire » qui permet de donner des précisions sur le type de lieu de décès, le statut grossesse pour les femmes ou bien indiquer si une autopsie est demandée par le médecin.

Dans la suite du mémoire, les champs seront énumérés comme les champs 1 à 4 pour les champs de la partie I (champs a à d) et champs 5 et 6 pour les deux champs de la partie II.

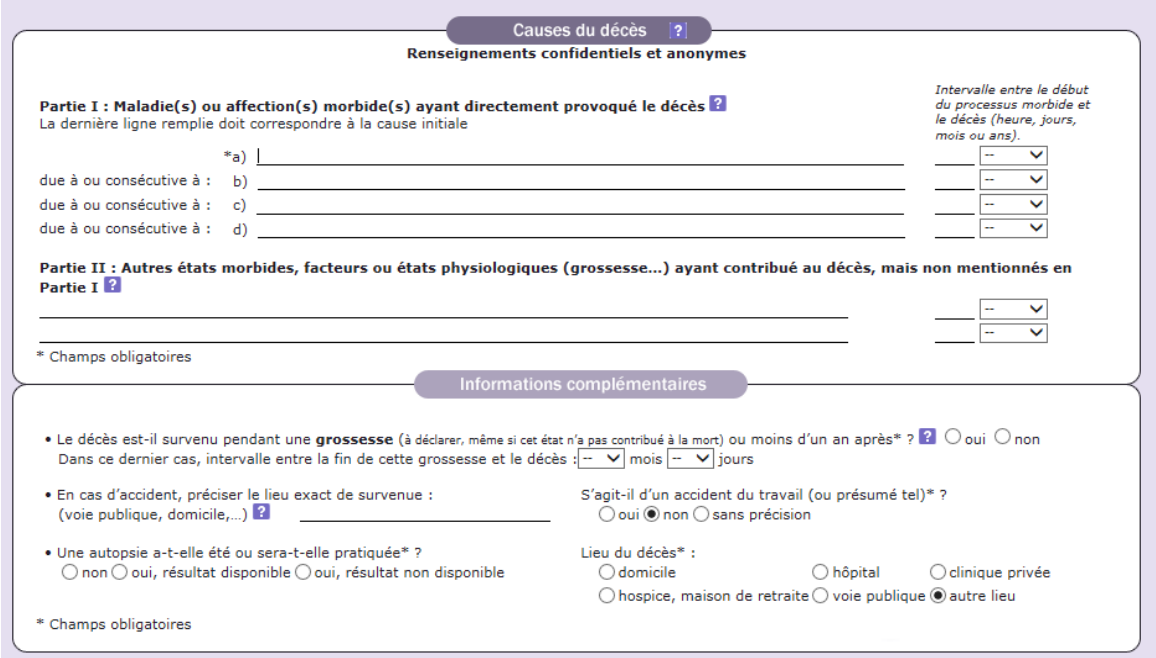


Figure 14 : Extrait du volet médical du certificat de décès général, 2017

Les causes médicales contenues dans le volet médical, ainsi que l'intervalle de temps, sont remplies sous forme de texte libre (pas d'adjonction d'une aide à la saisie ou d'un dictionnaire médical) par le médecin et reçu en l'état à Santé publique France.

Chaque champ du certificat de décès peut être renseigné avec plusieurs mots qui peuvent contenir des fautes d'orthographe, des abréviations, sigles. Ces termes expriment une ou plusieurs causes de décès correspondant à un symptôme (ex : « fièvre »), une pathologie spécifique (ex : « cirrhose »), un diagnostic (ex : « apparition de troubles de la déglutition nécessitant une trachéotomie »), un traumatisme (« traumatisme cranien avec hemorragie cerebrale et trauma de la face ») ou encore un traitement (« redicive eoa oesophage traite par radio chimiotherapie en 2009 »).

2) Le déploiement de la certification électronique en France depuis 2007

Le système de certification électronique est en place depuis 2007. En 2009, une note signée conjointement par le DGS et la directrice de l'hospitalisation et de l'organisation des soins a été adressée aux directeurs des ex-ARH (Agence régional de l'hospitalisation) afin de « […] participer à l'accélération du déploiement de ce dispositif dans le cadre d'actions contre la pandémie grippale. L'objectif est de déployer la certification électronique des décès dans 250 établissements du réseau « sentinelle » urgence de l'InVS avant octobre 2009 » (100).
Du 1er janvier 2007 au 30 juin 2011, 141 établissements avaient utilisé la certification électronique des décès au moins une fois et 4 à 5 % de la mortalité nationale était collectée par cette voie.

De 2011 à 2013, la part des décès certifiés électroniquement des décès a stagné autour de 5% de la mortalité totale (Figure 15).

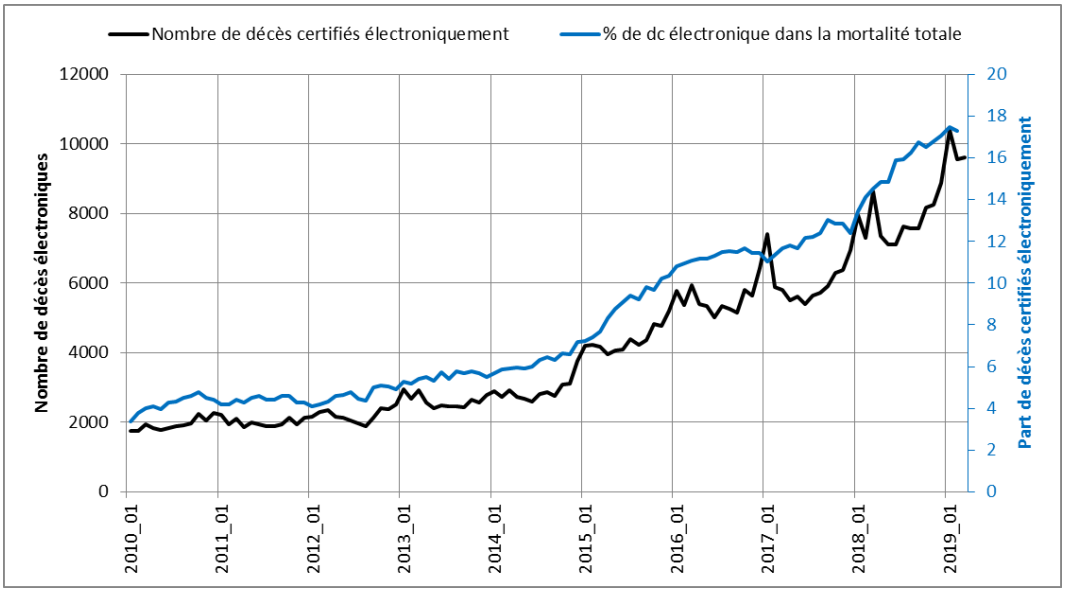


Figure 15 : Evolution mensuelle du nombre de décès certifiés par voie électronique et de la part des décès certifiés électroniquement parmi l'ensemble des décès entre 2010 et 2018 en France

Afin d'inciter les établissements à s'engager dans la certification électronique, en 2013, une instruction de la DGS (101) aux ARS (Agence Régionale de Santé) a été publiée. Elle fixait un objectif pour 2015, de 20% de la mortalité enregistrée par cette voie dans chaque région. Une seconde instruction suivra en 2016 (102), augmentant l'objectif à 40%. Ces instructions, prises

en compte sérieusement par certaines ARS, vont donner une impulsion à l'utilisation de la certification électronique. Même si les objectifs ne sont pas atteints dans toutes les régions, on note une augmentation progressive de la proportion de décès enregistrés par cette voie, passant au niveau national de 6,3% en 2013 à 15,6 % en 2018 (Figure 16).

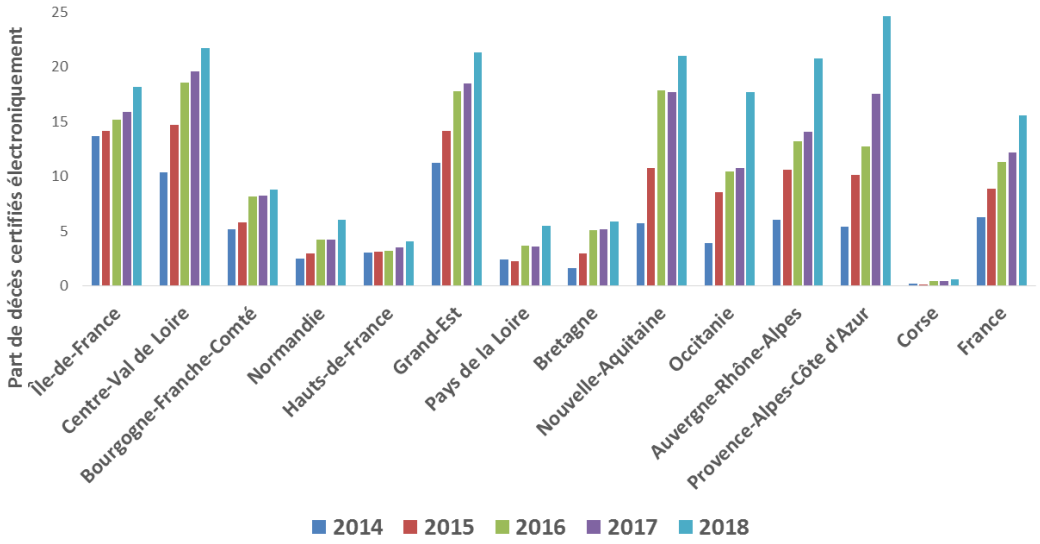


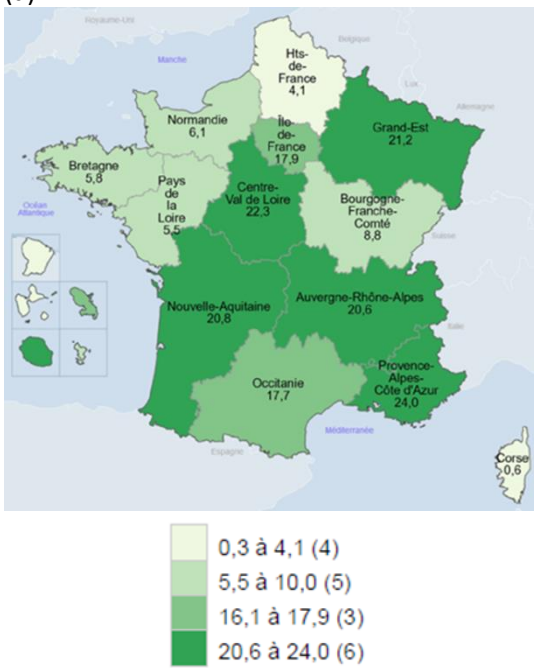
Figure 16 : Evolution annuelle de la proportion de décès certifiés par voie électronique entre 2014 et 2018 en France et par région métropolitaine

Ce taux variait, en 2018 au niveau régional sur l'ensemble du territoire de 0,6% en Corse à 24 % en Provence-Alpes-Côte d'Azur. On notait un taux de déploiement supérieur à 20% dans cinq régions : Auvergne-Rhône-Alpes (20,6%), Nouvelle-Aquitaine (20,8%), Grand-Est (21,2%), Centre-Val-de-Loire (22,3%), la Réunion/Mayotte (23,9%) (Figure 17a).

Au niveau départemental, un taux supérieur à 40% est noté pour 5 départements : les Hauts-de-Seine (52%), le Cher (47%), la Haute-Savoie (45%), l'Allier (43%) et les Alpes-Maritimes (43%). Toutefois, un tiers des départements enregistrait moins de 5% de la mortalité par voie électronique (Figure 17b).

Dans la suite de nos travaux, nous avons travaillé sur la période 2012-2016. De 5 à 12% de la mortalité nationale était enregistrée par cette voie sur cette période. L'ensemble des régions utilisait la certification électronique des décès.

(a)



(b)

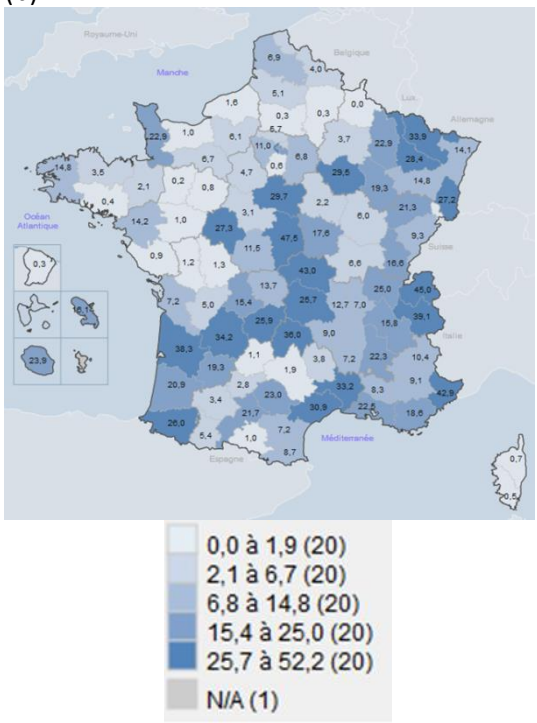


Figure 17 : Estimation de la part (%) de décès certifiés par voie électronique en 2018 parmi l'ensemble de la mortalité au niveau régional (a) et départemental (b)

II/ Description des décès certifiés électroniquement entre 2012 et 2016

Au total, 203 797 décès ont été certifiés par voie électronique entre 2012 et 2016.

1) Répartition des certificats de décès électroniques

La répartition des décès certifiés par voie électronique entre 2012 et 2016 est présentée dans le tableau 4. Les hommes constituaient 53 % des décès certifiés par voie électronique et 78% étaient âgés de 65 ans et plus. On observait une augmentation de l'âge moyen de décès de 1,1 ans entre 2012 et 2016 (Tableau 4).

Tableau 4 : Proportion de décès, par sexe, par classe d'âge et moyenne d'âge au niveau national, par années et pour la période 2012-2016 ; Certificats de décès électroniques 2012-2016, France

	N	Sexe (%)	Classes d'âge (%)				
Variables socio démographiques / année		**Femme**	**0-14 ans**	**15-64 ans**	**65-84 ans**	**≥85 ans**	**Age moyen de décès (Années)**
FRANCE							
2012	25 982	46,1	0,9	22,9	44,1	32,1	74,5
2013	30 317	45,6	1,0	22,5	43,7	32,9	74,6
2014	30 931	45,8	0,9	21,4	44,6	33,1	75,1
2015	52 519	45,8	0,6	19,7	43,4	43,3	76,2
2016	64 048	46,4	0,6	18,7	43,3	37,5	76,7
2012-2016	203 797	47,0	0,7	20,5	43,7	35,1	75,8

Au niveau régional, les certificats de décès enregistrés électroniquement concernaient majoritairement des hommes sauf pour la Bretagne. Plus de 70% des décès enregistrés dans chacune des régions étaient des personnes âgées de 65 ans ou plus, sauf pour la Corse, la Réunion et la Guadeloupe (Figure 18). Très peu de décès ont été enregistrés par voie électronique en Guyane et en Martinique sur cette période. Les résultats ne sont pas présentés pour ces deux régions.

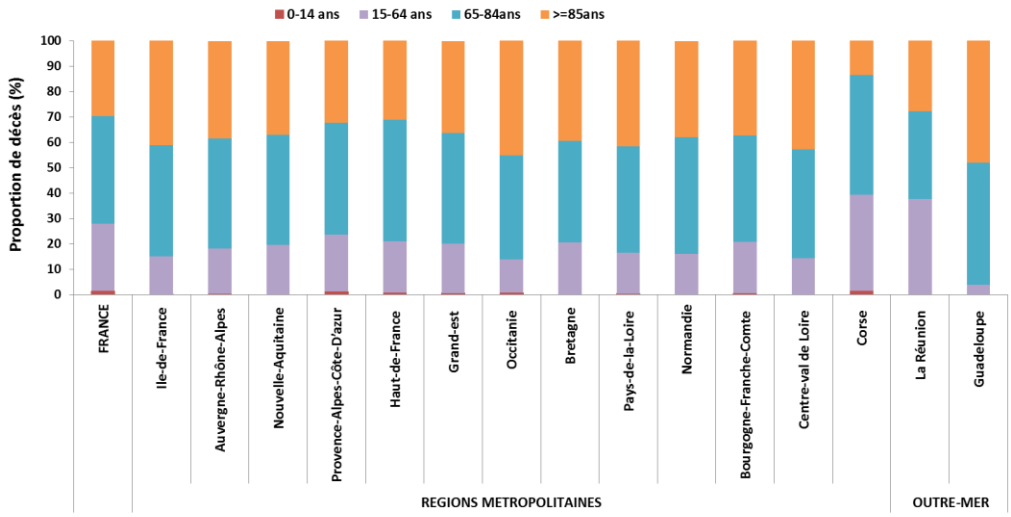


Figure 18 : Proportion de décès par classe d'âge au niveau national et régional, certificats de décès électroniques 2012-2016, France

Les certificats de décès étaient enregistrés majoritairement dans des établissements de santé publics (Figure 19). L'âge moyen de décès variait en fonction du type de lieu de décès de 87 ans en EHPAD à 47,7 ans pour les décès sur la voie publique.

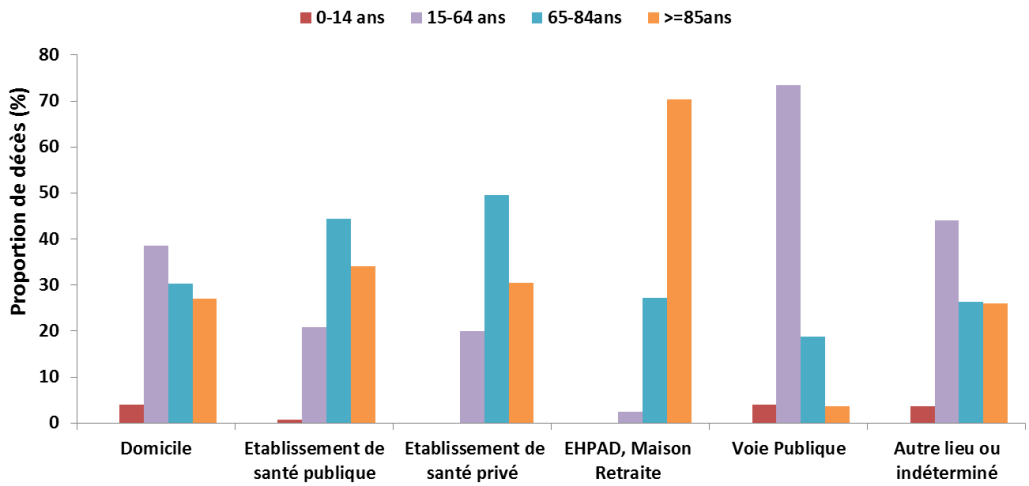


Figure 19 : Proportion de décès par classe d'âge selon le type de lieu de décès au niveau national ; certificats de décès électroniques 2012-2016, France

2) Répartition des causes dans les champs de certificats

Sur l'ensemble des 203 797 certificats, nous avons dénombré 591 509 champs contenant au moins une cause de décès. En moyenne, 2,9 champs par certificat de décès étaient remplis.

Parmi l'ensemble des certificats, 16% avaient les 4 champs de la partie I complétés (champs 1 à 4) et 35% avaient au moins le premier champ de la partie II rempli (champ 5) (Figure 20).

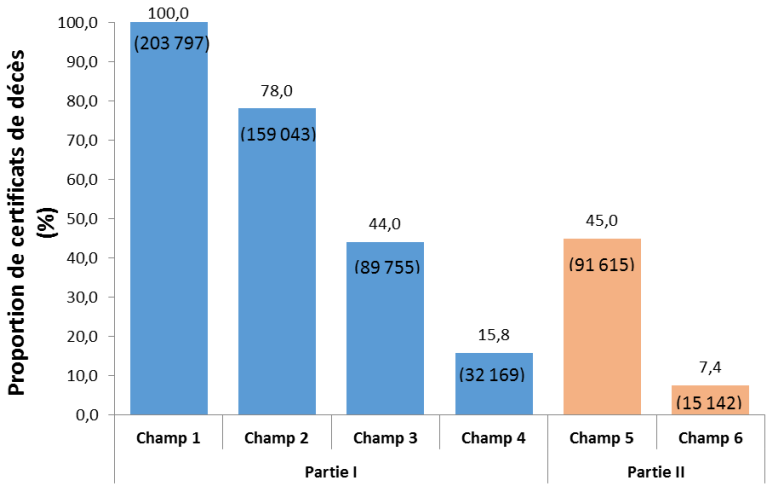


Figure 20 : Répartition des certificats électroniques de décès de 2012 à 2016 selon le champ de cause renseigné, France

3) Répartition du nombre moyen de mots par certificat et par champ

Les champs de certificats de décès contiennent à la fois des mots apportant des informations sémantiques et des mots vides (mots sans information sémantique). Par exemple dans la phrase « Syndrome de défaillance multiviscérale», « de » est un mot vide.

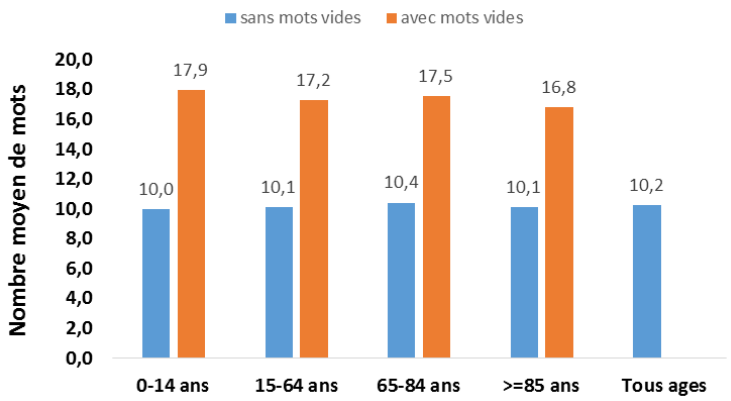


Figure 21 : Nombre moyen de mots par certificat de décès avec ou sans mots vides par classe d'âge et tous âges ; Certificats électroniques de décès de 2012-2016, France

En moyenne, on retrouvait 10,2 mots dans un certificat de décès. Le nombre de mots moyen par certificat variait selon les classes d'âge. Le nombre moyen de mots par champ était comparables pour chacune des classes d'âges (Figure 21).

Le nombre moyen de mots par champ variait de 5,5 à 5,7 mots dans la partie I. Le champ 5 (premier champ de la partie II) contenait en moyenne 7,7 mots (Figure 22).

On observait que les champs de la partie I (Champs 1 à 4) contenaient en moyenne entre 3,2 et 3,5 mots par champs sans prendre en compte les mots vides. Le champ 5 contenait en moyenne le plus grand nombre de mots (4,8 mots) (Figure 22).

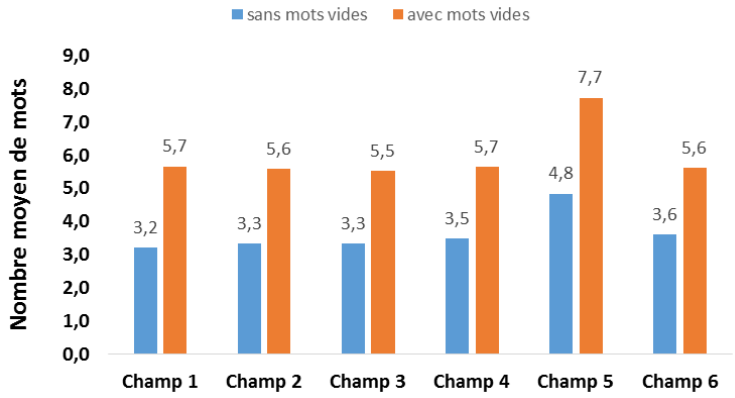


Figure 22 : Nombre moyen de mots par champ avec ou sans mots vides, tous âges, Certificats électroniques de décès de 2012-2016 France

En synthèse, la certification électronique reste limitée sur le territoire, avec une représentativité sociodémographique imparfaite de la mortalité totale et une répartition géographique hétérogène sur le territoire. Elle fournit toutefois des causes médicales de décès qui sont :

- de nature variée : diagnostics, symptômes, traumatismes, antécédents, actes médicaux, causes externes, états physiologiques, etc. ; que nous réunirons sous l'appellation « entité nosologique » dans la suite du document,
- d'antériorité variable : pouvant aller de quelques minutes avant le décès à plusieurs années,
- l'expression d'un processus d'évolution de l'état de santé, d'une survenue progressive ou brutale d'une pathologie/symptôme, de son aggravation/décompensation, etc.
- multiples au sein d'un même certificat.

La première étape pour l'exploitation de ces causes pour la surveillance à visée d'alerte consiste à définir les regroupements syndromiques à suivre pour répondre aux objectifs de détection et d'évaluation d'impact.

Chapitre III : Définition et construction des regroupements syndromiques

Les regroupements syndromiques (RS) de mortalité ont été définis comme étant la réunion d'expressions de causes de décès (entités nosologiques) ayant un sens pour la surveillance réactive de la mortalité à visée d'alerte et d'évaluation d'impact.

I/ La démarche de définition des regroupements syndromiques

1) Exploration des dictionnaires

Afin de définir la liste des regroupements syndromiques pour la détection d'événements connus ou inattendus et l'évaluation d'impact, nous avons, dans un premier temps, exploré différentes classifications, dictionnaires ou thesauri dans le domaine biomédical. Nous avons répertorié 4 classifications d'intérêt :

- **La CIM-10 : la classification internationale des maladies 10ème révision** est une classification codant les maladies, signes, symptômes, circonstances sociales et causes externes de maladies ou de blessures. C'est une classification largement utilisée pour le codage diagnostic à l'hôpital en France et le codage des causes des certificats de décès. Elle est découpée en 21 chapitres reprenant les pathologies par grands appareils, les tumeurs, des symptômes non classés par appareil, les causes externes et les facteurs de recours aux soins. Elle contient plus de 55 000 codes uniques. Cette classification existe dans différentes langues et fournit un langage commun grâce auquel les professionnels de santé peuvent échanger des informations partout dans le monde.
- **La SNOMED : Systematized Nomenclature of Medecine : Clinical Terms**, est un système international de terminologie clinique. Il comprend une liste de termes cliniques contrôlés, détaillés et validés médicalement avec leurs synonymes. Cette terminologie est utilisée dans les soins directs aux patients pour documenter les plaintes physiques et psychologiques, les symptômes, les circonstances, les maladies, les interventions chirurgicales, les diagnostics, les résultats et les décisions thérapeutiques.

  La SNOMED comporte plus de 350 000 concepts cliniques, répartis en 19 groupes hiérarchiques (les médicaments, la recherche biomédicale, la chimie des molécules, les diagnostics médicaux, etc.). Chaque concept a une définition clinique et est référencé par un identifiant : le SNOMED concept ID. Les concepts sont inter-reliés, une relation spécifie un lien entre deux concepts (https://www.health.belgium.be/fr/terminologie-et-systemes-de-codes-snomed-ct).

  La version française de cette classification n'est pas encore achevée, seuls 16 groupes hiérarchiques sont disponibles.

- **Le MeSH : Medical Subject Headings,** est le thésaurus de référence dans le domaine biomédical. La NLM (National Library of Medecine) qui l'a construit en 1954 le met à jour chaque année. La traduction française de ce thésaurus a été réalisé par l'Inserm pour la première fois en 1986, depuis il le met à jour chaque année.

  Le MeSH est organisé en 16 catégories thématiques allant de l'anatomie aux lieux géographiques. Chacune de ces catégories est organisée en arborescence de descripteurs (MeSH headings), qui peut comprendre jusqu'à onze niveaux hiérarchiques. Les descripteurs sont constitués de concepts, les concepts sont eux-mêmes constitués de termes et un terme est un mot ou un ensemble de mots exprimant une notion (https://www.nlm.nih.gov/mesh/concept_structure.html).

  La catégorie thématique « Maladie » est organisée en 26 descripteurs reprenant les appareils et systèmes mais aussi les troubles professionnels, liés à l'environnement, les états, signes et symptômes pathologiques. On retrouve aussi une thématique psychiatrie et psychose.
- **L'UMLS : Unified Medical Language System**, a été construit spécifiquement pour le traitement et l'exploitation informatique du langage de la santé et de la biomédecine. C'est le regroupement d'un grand nombre de dictionnaires et classifications disponibles en plusieurs langues. Il est constitué de 3 entités, le Metathesaurus, le semantic network et le specialist lexicon.
  - Le Metathesaurus regroupe tout le vocabulaire, codes, thesauri existants. Soixante-deux pour cent du metathesaurus est constitué à partir de sources en anglais. Il est organisé en concepts. Chaque concept possède un code unique.
  - Le semantic network est l'entité qui permet de regrouper chaque concept du metathesaurus par groupes sémantiques. Chacun des concepts du metathesaurus peut appartenir à plusieurs groupes sémantiques.
  - Le specialist lexicon est un ensemble de programmes informatiques conçus pour aider au traitement du langage.

Cette étape d'exploration de ces divers thesauri, classifications ou dictionnaires avait pour but de relever les différents types de classifications existants, afin d'identifier celle à partir de laquelle nous pourrions recréer une arborescence qui répondrait à nos objectifs. Cependant, la complexité des arborescences, notamment du MeSH et de l'UMLS, ne nous semblait pas adaptée à la classification des causes de décès. La CIM-10, qui est une classification largement utilisée et dont l'arborescence décrit les appareils, semblait plus appropriée.

Afin d'établir la liste et la définition des regroupements syndromiques utiles pour répondre aux objectifs de détection à visée d'alerte et d'évaluation d'impact, nous avons choisi de baser nos définitions sur la CIM-10. L'arborescence que nous avons définie est proche mais pas strictement identique à celle de la CIM-10.

2) Définition des regroupements syndromiques : découpage de la CIM-10

Outre la CIM-10, la définition des regroupements syndromiques s'est appuyée sur une autre ressource appartenant à la catégorie des dictionnaires : le dictionnaire du CépiDc. Il répertorie l'ensemble des expressions d'une cause de décès qui ont été retrouvées dans un certificat de décès au moins une fois depuis le début des années 2000 et le code CIM-10 qui lui a (ont) été associé(s). C'est une ressource très riche puisqu'elle contient une grande variété d'expressions de causes de décès : plus de 150 000 expressions.

Les regroupements syndromiques ont alors été définis comme étant des listes de codes CIM-10 et d'expressions correspondant à ces codes, chaque code CIM-10 et chaque expression appartenant à un unique regroupement syndromique.

Afin d'élaborer la liste de codes CIM-10 appartenant à un regroupement syndromique, nous avons passé en revue les 21 chapitres de la CIM-10.

A partir des groupes de chaque chapitre, nous avons construit nos propres regroupements. Un regroupement syndromique peut être :

- La fusion de plusieurs groupes d'un même chapitre ou de chapitre différents,
- Le rassemblement de codes CIM-10 appartenant à différents groupes, d'un même chapitre ou de chapitres différents.

Le principe général était de regrouper les codes CIM-10 constituant un ensemble homogène de causes pour répondre aux objectifs de surveillance de la mortalité.

Pour la définition de certains regroupements syndromiques, nous nous sommes basés sur des définitions existantes dans la littérature. Par exemple, le regroupement syndromique « Maladies à Déclaration Obligatoire » contient l'ensemble des maladies à déclaration obligatoire listées par le Ministère de la santé.
(https://www.legifrance.gouv.fr/affichCode.do?idSectionTA=LEGISCTA000006190444&cidTexte=LEGITEXT000006072665).

D’autres définitions de regroupements syndromiques, comme « Suicide » ou « Noyade » se sont appuyées sur celles utilisées dans des études produites par l’agence.

Plus généralement, nous nous sommes également inspirés de l’expérience et des définitions utilisées pour la surveillance syndromique de la morbidité, menée à partir des deux sources de données du système SurSaUD® (structures d’urgences et associations SOS Médecins).

Au total, l’ensemble des regroupements syndromiques disjoints permettront ainsi de décomposer la mortalité totale et de quantifier la contribution de chacun lors de la mesure de l’impact d’un événement sur la population. A partir de l’ensemble des regroupements syndromiques définis, nous avons déterminé une liste de regroupements syndromiques qui sera spécifiquement suivi en routine pour la détection à visée d’alerte. Etaient retenus dans la liste des regroupements syndromiques suivis en routine pour la surveillance, les regroupements syndromiques répondant aux caractéristiques suivantes :

- Regroupements syndromiques de pathologies spécifiques exprimant une relation "directe" avec un événement connu ou attendu, par exemple les regroupements syndromiques « Grippe », « Bronchiolite », « Gastro-entérite/Diarrhée », « Méningite Virale », « MDO infectieuses », etc.
- Regroupements de pathologies ou signes cliniques peu spécifiques mais caractérisant un état de santé atypique et/ou alarmant en termes de sévérité ou de dégradation rapide : « Fièvre d’origine inconnu », « Sepsis », « Douleurs », « Nausées et vomissements ».

3) Validation des définitions

L’étape de validation des définitions des regroupements syndromiques s’est appuyée sur des experts des thématiques ou des pathologies, pour lesquelles nous souhaitons valider *a priori* des définitions.

Le principe est de 1) proposer la (les) définitions des regroupements syndromiques d’une thématique spécifique au groupe d’experts, 2) discuter la liste des codes CIM-10 définissant chaque regroupement syndromique, 3) aboutir à une définition convenant à l’ensemble des experts et à l’équipe.

II/ Liste des regroupements syndromiques définis

Quatre-vingt-dix-huit regroupements syndromiques ont été définis et sont répartis dans 20 thématiques différentes (Tableau 5).

Soixante-trois regroupements syndromiques ont été spécifiquement définis pour répondre aux objectifs de détection précoce à visée d'alerte. Les trente-cinq autres regroupements syndromiques, s'ils n'ont pas vocation à être suivis en routine dans un objectif de détection précoce, permettront de prendre à compte, lors de l'évaluation rapide de l'impact d'un événement, la part des pathologies chroniques, comorbidités, antécédents, … dans la mortalité totale. Cela permettra d'identifier ou décrire les causes indirectes, comorbidités ou les facteurs de risques associés à l'événement et ainsi contribuer à déterminer des profils de décès (par exemple des personnes âgées souffrant de plusieurs pathologies chroniques ou des personnes plus jeunes sans aucune présence de comorbidité). Ces regroupements ont d'autant plus d'importance en situation d'émergence ou d'événements inattendus.

Les définitions des regroupements syndromiques de la thématique ***« cardio et cérébrovasculaire »*** ont été revues et validées. Cette validation a été menée avec deux experts épidémiologistes de l'équipe des maladies cardio-vasculaires de la direction des maladies non transmissibles et des traumatismes de Santé publique France. Les définitions des regroupements syndromiques de cette thématique ont été présentées aux experts. Sur cette base, ils nous ont proposé des modifications qui s'appuyaient sur leur expertise des maladies cardio-vasculaires mais aussi sur des définitions de regroupements de pathologies publiées dans plusieurs de leurs études (103-105)

III/ Discussion

La mise en œuvre opérationnelle de la surveillance syndromique de la mortalité par cause nécessite dans un premier temps de définir les regroupements syndromiques à suivre en routine et d'être en mesure de décomposer la mortalité totale afin d'identifier les causes contributives à une variation habituelle ou inhabituelle de la mortalité.

A ce jour, la validation des définitions des regroupements syndromiques est une étape toujours en cours. Même si la validation des premiers regroupements syndromiques a été effectuée avec l'aide d'épidémiologistes internes à l'agence, il conviendra d'étendre le panel d'experts à des

cliniciens et épidémiologistes extérieurs pour la validation des autres regroupements.
La méthode Delphi pourrait être une méthode envisagée pour la validation (106) des définitions d'un ou plusieurs regroupements syndromiques. Elle vise à recueillir, par l'entremise d'un questionnaire ouvert, l'avis justifié d'un panel d'experts (épidémiologistes et cliniciens) dans différents domaines. La procédure, basée sur la rétroaction, évite la confrontation des experts. Les résultats d'un premier questionnaire sont communiqués à chaque expert et sont accompagnés d'une synthèse des tendances générales et particulières, des avis et des justifications. Dès lors, chacun est invité à réagir et à répondre à un deuxième questionnaire élaboré en fonction des premiers avis recueillis, l'étape se répète et ce jusqu'à l'obtention d'une convergence aussi forte que possible des réponses.

Une autre méthode de validation des regroupements syndromiques serait de croiser les données de sources complémentaires. Elle s'appuierait par exemple sur la comparaison des évolutions temporelles des regroupements syndromiques construits à partir des certificats de décès à celles des regroupements syndromiques construits à partir d'une autre source de données prise en référence. Cette méthode a été explorée pour valider deux regroupements syndromiques de la thématique « Maladies respiratoires » en utilisant une source de données de morbidité du système SurSaUD (Données de recours aux urgences). Sa principale difficulté réside dans l'identification de source de données pertinente à considérer en référence.

La validation de l'ensemble des définitions des regroupements syndromiques n'a pu être réalisée dans le cadre de ces travaux et constitue une perspective.

Afin de poursuivre la démarche de mise en œuvre de la surveillance syndromique de la mortalité, nous avons identifié des méthodes issues du traitement automatique des langues pour la classification automatique des causes de décès dans les regroupements syndromiques définis dans ce chapitre. Ces méthodes ont ensuite été mises en œuvre et évaluées sur les causes de décès issues de la certification électronique entre 2012 et 2016.

Tableau 5: Liste des regroupements syndromiques (RS)° établie en 2018

Thématiques /Regroupements syndromiques	RS définis pour l'alerte	RS non définis pour l'alerte	RS pour l'évaluation d'impact
1-CARDIO et CEREBROVASCULAIRE			
1-Cardio et circulatoire chronique		**x**	**x**
1-Cardio Infectieux	**x**		**x**
1-Décompensation/insuffisance cardiaque aigue	**x**		**x**
1-Embolie pulmonaire	**x**		**x**
1-Maladies cérebrovasculaires chroniques		**x**	**x**
1-Mort subite cardiaque	**x**		**x**
1-Non informatif		**x**	**x**
1-Oedeme cérébral		**x**	**x**
1-Rupture d'anévrisme non cérébral	**x**		**x**
1-Syndrome coronarien aigue	**x**		**x**
1-AVC/Hémorragie cérébrale	**x**		**x**
1-Troubles du rythme de la conduction / fibrillation	**x**		**x**
2-CANCERS/TUMEURS			
2-Cancers/Tumeurs		x	x
3-MALADIES RESPIRATOIRES			
3-Asphyxie et anomalies de la respiration	**x**		**x**
3-Asthme	**x**		**x**
3-Bronchiolite		**x**	**x**
3-Dyspnée	**x**		**x**
3-Grippe	**x**		**x**
3-Infections respiratoires aigües basses	**x**		**x**
3-Insuffisance respiratoire aiguë	**x**		**x**
3-Maladies respiratoires aigües	**x**		**x**
3-Maladies respiratoires chroniques		**x**	**x**
3-ORL		**x**	**x**
4-SYMPTOMES			
4-Autres chocs	**x**		**x**
4-Chocs cardio et hypovolémique	**x**		**x**
4-Coma	**x**		**x**
4-Douleur	**x**		**x**
4-Fièvre d'origine inconnue	**x**		**x**
4-Grabataire/ Sénile		**x**	**x**
4-Malaise et fatigue	**x**		**x**
4-Mort subite	**x**		**x**
4-Nausées et vomissement	**x**		**x**
4-Résultats biologiques anormaux		**x**	**x**
4-Symptômes du trouble de l'alimentation		**x**	**x**
4-Symptômes généraux	**x**		**x**
4-Troubles cognitifs et de l'humeur	**x**		**x**

Thématiques /Regroupements syndromiques	RS définis pour l'alerte	RS non définis pour l'alerte	RS pour l'évaluation d'impact
5-MALADIES INFECTIEUSES			
5-Arbovirose / Fièvres virales moustique	x		x
5-Gastro-entérite - Diarrhée	x		x
5-Hépatites virales	x		x
5-Infections bactériennes	x		x
5-Infections intestinales	x		x
5-Infections parasitaires		x	x
5-Infections Sexuellement Transmissibles	x		x
5-Infections virales	x		x
5-Maladies à Déclaration Obligatoire infectieuses	x		x
5-Méningite virale	x		x
5-Resistance	x		x
5-Sepsis	x		x
5-Syndrome du choc toxique	x		x
6-MALADIES ENDOCRINIENNES ET NUTRITIONNELLES			
6-Deshydratation	x		x
6-Maladies aigües endocriniennes	x		x
6-Anomalies métaboliques		x	x
6-Maladies chroniques endocriniennes		x	x
7- MALADIES DU SYSTÈME NERVEUX			
7-Epilepsie	x		x
7-Maladies chroniques du système nerveux		x	x
7-Maladies inflammatoires du système nerveux central	x		x
7-Tableaux neuropsychologiques aigus	x		x
8-FACTEURS ET MOTIFS DE RECOURS AUX SOINS			
8-Antécédants		x	x
8-Autres facteurs et recours aux soins		x	x
8-Examens et soins		x	x
9-TRAUMA ET EMPOISONNEMENT			
9-Brûlure, corrosion	x		x
9-Complications		x	x
9-Effet de causes externes	x		x
9-Effet toxique lié à d'autres substances	x		x
9-Effet toxique lié à des substances médicinales	x		x
9-Effet toxique lié à l'alcool	x		x
9-Lesions traumatiques et présence de corps étrangers	x		x
9-Séquelles		x	x

Thématiques /Regroupements syndromiques	RS définis pour l'alerte	RS non définis pour l'alerte	RS pour l'évaluation d'impact
10-TROUBLES MENTAUX ET DU COMPORTEMENT			
10-Troubles mentaux liés à des substances	x		x
10-Autres troubles mentaux		x	x
10-Troubles mentaux aigus	x		x
10-Troubles mentaux chroniques		x	x
11-MALADIE DE LA PEAU ET DU TISSU CELLULAIRE SOUS CUTANE			
11-Maladies aigües de la peau et tissu	x		x
11-Maladies chroniques de la peau et tissu		x	x
12-MALADIE DU SANG ET DES ORGANES HEMATOPOIETIQUES			
12-Maladies aigües du sang	x		x
12-Maladies chroniques du sang		x	x
12-Purpura et affections hémorragiques	x		x
13-MALADIES DE LA GROSSESSE, NEONATALES ET CONGENITALES			
13-Maladies aigues de la grossesse	x		x
13-Maladies congénitales		x	x
13-Maladies de la grossesse		x	x
13-Maladies Néonatales		x	x
14-MALADIES DE L'ŒIL ET SES ANNEXES			
14-Maladies de l'œil		x	x
15- MALADIES DE L'APPAREIL DIGESTIF			
15-Maladies aigües de l'appareil digestif	x		x
15-Maladies chroniques de l'appareil digestif		x	x
16-CAUSES MAL DEFINIES			
16-Autres causes de mortalité mal définies et non précisées	x		x
17-CAUSES EXTERNES			
17-Accidents et autres causes externes		x	x
17-Accident de transport		x	x
17-Agression		x	x
17-Chaleur	x		x
17-Evénements environnementaux	x		x
17-Froid	x		x
17-Noyade accidentelle	x		x
17-Suicides	x		x
18-MALADIES GENITO-URINAIRES			
18-Insuffisance rénale aiguë	x		x
18-Maladies aigües de l'appareil génito-urinaire	x		x
18-Maladies chroniques de l'appareil Génito-urinaire		x	x
19-MALADIES DE L'OREILLE ET DE L'APOPHYSE MASTOIDE			
19-Maladies de l'oreille		x	x
20- MALADIES DU SYSTÈME OSTEO ARTICULAIRE, MUSCLES ET TISSU CONJONCTIF			
20-Maladies du système ostéo-articulaire		x	x

Chapitre IV: Mise en œuvre et évaluation de la classification des causes de décès dans les regroupements syndromiques

Dans ce chapitre nous nous intéresserons à la mise en œuvre et à l'évaluation de méthodes de traitement automatique des langues pour la classification des causes de décès dans des regroupements syndromiques. Un éventail de méthodes disponibles peut être envisagé pour cette tâche. Dans un premier temps, nous avons recherché et sélectionné à travers la littérature les méthodes les plus appropriées pour répondre à notre objectif, puis nous les avons mises en œuvre et avons évalué leurs performances.

I/ Sélection des méthodes pour la classification des causes de décès

Le tableau 6 présente une synthèse des études considérées dans la littérature pour l'exploration des méthodes utilisées.

Dans une étude parue en 2017, Mujtaba et al. ont testé différentes méthodes pour identifier les causes de décès liés à des accidents à partir des rapports d'autopsie (107). Ils ont cherché à identifier les causes de décès liées à des accidents correspondants à 9 codes CIM-10 différents. Cette classification avait pour objectif d'aider les pathologistes à déterminer les causes de décès par autopsie. Plusieurs méthodes d'apprentissage ont été testées en utilisant l'outil Weka : un SVM (Support Vector Machine), un modèle bayésien naïf, un modèle du k-plus proche voisin, un modèle basé sur un arbre de décision et un modèle basé sur les forêts aléatoires. Les forêts aléatoires et les arbres de décision, paramétrés à l'aide d'une sélection de caractéristiques par des experts, ont donné les meilleurs résultats (F-mesure= 0,90).

Dans leur étude, Butt et al. ont cherché à identifier les décès liés à des cancers (108). Pour cela à partir d'un échantillon de 5000 certificats de décès, ils ont testé différentes méthodes d'apprentissage automatique regroupées dans l'outil Weka. Le SVM obtenait les meilleurs résultats avec une F-mesure de 0,98, suivi des arbres de décision dont la F-mesure était de 0,92. Deux autres méthodes (un modèle bayésien naïf et le boosting d'algorithme (AdaBoost)) ont obtenu des performances inférieures aux deux autres modèles.

Koopman et al. ont utilisé des méthodes d'apprentissage automatique et plus spécifiquement des classifieurs SVM, pour identifier les décès liés à des cancers (109). A partir des textes issus des certificats de décès, des vecteurs de caractéristiques binaires ont été construits et ont été utilisés pour entrainer deux modèles SVM : 1) un classifieur binaire pour identifier la présence de cancer dans le certificat de décès, 2) un ensemble de classifieurs pour identifier le type de cancer en utilisant la CIM-10. Leur système s'est avéré très efficace pour identifier les cancers en cause initiale (F-mesure=0,94). Il s'est aussi avéré efficace pour déterminer les types de cancers fréquents (F-mesure= 0,7). Cependant les cancers rares pour lesquels ils disposaient de peu de

données pour l'entrainement étaient difficiles à classer avec précision (F-mesure = 0,12).
L'efficacité des méthodes de traitement automatique des langues pour l'identification de cas de pneumonie ou de grippe dans les certificats de décès a aussi été démontrée par Davis et al (47). Dans leur étude, les auteurs ont développé le codage automatique des causes de décès en « Concept Unique Identifier » de l'UMLS à partir de l'outil MetaMap (66). Ils ont ensuite cherché à identifier les décès avec des mentions de grippe et pneumonie à partir des définitions des CDC. Les performances de codage de l'outil MetaMap, puis l'identification des certificats avec mention de grippe et pneumonie ont été comparées à une classification des certificats basée sur une recherche de mots clés. Le codage avec l'outil MetaMap suivi de l'identification des certificats avec mention de grippe et pneumonie obtenait une F-mesure de 0,99. L'identification avec une recherche par mots clés obtenait une F-mesure de 0,96.

Dans une autre étude, Koopman et al. ont utilisé une méthode d'apprentissage automatique et une méthode par règles afin de classer les certificats de décès selon quatre maladies : le diabète, le VIH, la grippe et la pneumonie (48). La méthode à base de règles s'appuyait sur un ensemble de règles et de combinaison de mots clés, tandis que la méthode par apprentissage était à nouveau un SVM, utilisant des vecteurs de caractéristiques binaires. Un modèle distinct a été construit pour chacune des quatre pathologies. Les résultats ont montré que la classification des certificats dans l'une des 4 classes était très précise à la fois pour la méthode à base de règles (F-mesure=0,95) et pour les SVM (F-mesure=0,94). La classification la plus fine en codes CIM-10 a démontré des résultats variables, avec des classifications moins précises pour les blocs de codes avec peu de données d'entrainement.

Les méthodes de TAL ont également été utilisées sur les causes médicales de décès exprimées en texte libre dans des objectifs de codage automatique en codes CIM-10.
Depuis 2016, le CLEF e-Health challenge (Conference and Labs of the Evaluation Forum) proposait une tâche de codage automatique des causes de décès. Ces challenges sont basés sur des données réelles. En 2016, un corpus en Français avait été proposé. Le challenge 2017 était basé sur deux corpus : l'un français l'autre américain, et celui de 2018 proposait trois corpus de langues différentes : un corpus français, un corpus hongrois et un corpus italien. Les corpus comprenaient les données renseignées en texte libre sur chaque ligne des certificats électroniques et le code CIM-10 obtenu après codage manuel des données. Ces corpus étaient considérés comme la référence avec laquelle l'outil développé était évalué. Les corpus étaient décomposés en jeux d'apprentissage et de développement.

Les équipes recevaient ensuite les corpus de test sans les codes CIM-10 et s'engageaient alors à retourner les résultats fournis par leur(s) outil(s). L'évaluation était effectuée sur ces jeux en comparaison avec les codes CIM-10 obtenu par codage manuel.
Ces trois éditions avaient respectivement attiré 5 équipes en 2016, 11 équipes en 2017 et 14 équipes en 2018.
En 2016, le système le plus performant avait utilisé des dictionnaires construits à partir d'autres tâches. Il obtenait une F-mesure égale à 0,85 (54). Parallèlement au challenge mais en s'appuyant sur ses données, Zweigenbaum et al. ont présenté des méthodes hybrides pour le codage CIM-10 des certificats de décès (55), combinant la recherche de correspondances avec le dictionnaire et l'apprentissage machine supervisé. Le meilleur modèle hybride correspondait à la combinaison des résultats produits par les méthodes basées sur les dictionnaires et les méthodes basées sur l'apprentissage (SVM), dépassant légèrement le meilleur système du challenge 2016, avec une F-mesure de 0,86.

Durant l'édition 2017, les outils développés par les équipes faisaient appel à des techniques très diverses (53). La plupart utilisait des ressources terminologiques, dont les dictionnaires fournis par les organisateurs. Des outils préexistants, ainsi que des structures de données et des méthodes d'apprentissage comme les SVM ou les réseaux de neurones, ont été combinés de façon variée.
Les meilleurs résultats sur le corpus américain étaient obtenus à l'aide de méthodes utilisant des réseaux de neurones, avec une F-mesure de 0,85 sur l'ensemble des causes (110).
Sur le corpus français, les meilleurs résultats étaient obtenus par une méthode à base de règles utilisant des ressources terminologiques propres et un correcteur orthographique (F= 0,80) sur l'ensemble des causes (111).
L'équipe organisatrice du LIMSI obtenait, avec une méthode hybride faisant appel à des règles et à une classification par SVM, des résultats plus élevés ou proches sur le corpus américain (F-mesure=0,84) sur l'ensemble des causes (56).

En 2018, la même équipe a obtenu les meilleurs résultats sur les trois corpus (112). Elle avait développé un réseau de neurones qui cartographiait les extraits de textes en entrée et les codes CIM-10 en sortie. Ils avaient obtenu une F-mesure de 0,84 pour le corpus français, de 0,96 pour le corpus hongrois et de 0,95 pour le corpus italien.

Plus récemment, Duarte et al. ont développé des méthodes d'apprentissage profond (réseau de neurones) pour classer les causes de décès en codes CIM-10 (113). Leur méthode combine des plongements de mots (word embedding), des unités récurrentes et un mécanisme d'attention. Les évaluations ont été menées pour plusieurs niveaux de codage : pour le code correspondant

au chapitre de la CIM-10, pour le code correspondant au groupe et pour le code complet. La méthode obtenait des performances (F-mesure) de 0,63, 0,40 et 0,27 pour chaque niveau de codage respectif.

En synthèse, parmi l'ensemble de ces études, les méthodes qui obtenaient le plus fréquemment les meilleures performances étaient les SVM ainsi que les méthodes à bases de règles et les réseaux de neurones. Sur la base de cette revue des méthodologies utilisées dans la littérature, nous avons choisi de mettre en œuvre et évaluer une méthode à base de règles linguistiques et la méthode SVM pour classer les causes de décès dans les regroupements syndromiques.

Tableau 6 : Tableau récapitulatif des méthodes de classification des informations biomédicales identifiées dans la littérature et leurs performances pour des tâches proches de la classification des causes de décès dans les regroupements syndromiques, Recherche de la littérature jusqu'en Juin 2019

Auteur (Année)	Domaine application	Objectif	Méthodes	Performances
Davis[47] (2012)	Surveillance	Identifier les mentions de grippe et de pneumonie dans les certificats de décès	• Utilisation de l'outil MetaMap : Il décompose le texte saisi en mots ou en phrases, en les catégorisant en termes standardisés, puis en faisant correspondre les termes aux concepts de l'UMLS. Pour chaque terme apparié, MetaMap le classe dans un type sémantique puis retourne l'identifiant unique du concept (CUI) • Utilisation d'une méthode basée sur la recherche de mots clés	• L'outil MétaMap obtenait une F-mesure de 0,99 • La méthode basée sur la recherche de mots clés obtenait une F-mesure = 0,96
Butt[108] (2013)	Epidémiologie	Identification automatique des certificats de décès contenant une mention de cancer	SVM, modèle bayésien naïf, un arbre de décision, et boosting d'algorithme	• SVM : F-mesure=0,98 • Bayésien naïf : F-mesure=0,95 • Arbre de décision : F-mesure =0,97 • Boosting d'algorithme :F-mesure= 0,90
Koopman[109](2015)	Statistique de mortalité	• Identifier les certificats de décès avec des mentions de cancer • Identifier le type de cancer en attribuant le code CIM-10 correspondant	Deux modèles SVM ont été développés : • Un classifieur binaire (Cancer/non cancer) • Un classifieur multi-label pour chaque type de cancer	Classifieur SVM binaire : F-mesure=0,94 Classifieur SVM multi-label : • F-mesures comprises entre 0,2 et 0,85 pour les cancers les plus fréquents • F-mesures comprises entre 0,05 et 0,65 pour les cancers les plus rares
Koopman[48](2015)	Surveillance	Classer les certificats de décès pour diabète, grippe, pneumonie et VIH	• Un classifieur SVM binaire unique a été construit pour identifier chacune des 4 pathologies • Un classifieur SVM binaire unique a été construit pour l'attribution de chaque code CIM-10 Une méthode par règles basée sur la recherche de mots clés pour identifier les 4 pathologies	• Performance des classifieurs SVM pour identifier les quatre pathologies : F-mesure= 0,89 ; 0,97, 0,99, 0,93 pour la grippe, pneumonie, diabète, VIH respectivement • Performances de la méthode par règles : F-mesure= 0,92 ; 0,97 ; 0,97 ; 0,89 pour la grippe, pneumonie, diabète, VIH respectivement • Performances des classifieurs pour l'attribution des codes CIM-10 : F-mesure comprise entre 0,44 et 0,97 selon les codes CIM-10
Névéol[54] (2016)	Codage automatique	• Coder automatiquement les causes de décès en code	Construction de deux terminologies à partir des ressources proposées et utilisation d'un taggeur pour indexer les	• F-mesure = 0,85

	des causes de décès	CIM-10	certificats de décès et générer les codes.	
Zweigenbaum[55] (2016)	Codage automatique des causes de décès	Coder automatiquement les causes de décès en code CIM-10	Combinaison d'une projection de dictionnaire et d'un modèle SVM	F-mesure=0,86
Cabot[111] (2017)	Codage automatique des causes de décès	Coder automatiquement les causes de décès en code CIM-10 (Corpus Français)	Utilisation de l'outil CIM-IND, conçu pour reconnaitre les entités nommées. Il s'appuie sur un dictionnaire et méthodes TAL (correcteur orthographique)	F-mesure=0,80
Miftahutdinov[110] (2017)	Codage automatique des causes de décès	Coder automatiquement les causes de décès en code CIM-10 (Corpus américain)	Réseau de neurones	F-mesure=0,85
Mujtaba[107] (2017)	Surveillance	Classer les rapports d'autopsie dans neuf causes de décès	Cinq modèles différents contenus dans l'outil Weka ont été testés : SVM, Forêt aléatoire, bayésien naïf, k-plus proche voisin, arbres de décision	Les forêts aléatoires et les arbres de décision, paramétrés à l'aide d'une sélection de caractéristiques par des experts, obtenaient les meilleurs résultats : F-mesure= 0,90
Zweigenbaum[56] (2017)	Codage automatique des causes de décès	Coder automatiquement les causes de décès en code CIM-10 (Corpus américain)	Amélioration de la méthode développée précédemment (Zweigenbaum (2016)) : • Ajout de caractéristiques dans le dictionnaire, • Utilisation d'un classifieur multi label, • Ajout de l'âge de décès en caractéristiques d'apprentissage du SVM • Ajout des résultats de la projection de dictionnaire en caractéristiques d'apprentissage	F-mesure =0,84
Atutxa[112] (2018)	Codage automatique des causes de décès	Coder automatiquement les causes de décès en code CIM-10 (Corpus français, hongrois, italien)	Réseau de neurones	Corpus français : F-mesure=0,84 Corpus hongrois : F-mesure =0,96 Corpus italien : F-mesure= 0,95
Duarte[113] (2018)	Codage automatique des causes de décès	Coder automatiquement les causes de décès en code CIM-10	• Réseau de neurones : apprentissage profond. • Combinaison de plongements de mots, des unités récurrentes et un mécanisme d'attention	Code correspondant au chapitre : F-mesure= 0,62 Code correspondant au groupe de codes : F-mesure=0,40 Code complet : F-mesure=0,27

II/ Préparation des données

1) Matériel et ressources

1.1 Les données

Les volets médicaux des certificats de décès électroniques reçus entre 2012 et 2016 et pour lesquels nous avions à la fois les informations démographiques (date de naissance, sexe, type de lieu de décès) et les causes de décès exprimées par le médecin sous forme de texte libre, ont été sélectionnés (Tableau 7). Ces certificats étaient disponibles pré-anonymisés (absence du nom et du prénom). Nous disposions aussi des causes de décès codées en CIM-10 par le CépiDc, reçues dans un délai de plusieurs mois après leur validation.

Les certificats de décès néonataux n'ont pas été pris en compte dans cette étude. Au total, l'étude a porté sur 203 797 décès.

Tableau 7 : Exemple de causes médicales de décès d'un homme de 41 ans dont le décès est survenu en 2012

Identifiant	Date de décès	Age	Sexe	Champ /rang	Causes de décès
22203632	**13.09.2012**	**41**	**M**	**1-0**	**IDM***
22203632	**13.09.2012**	**41**	**M**	**2-0**	**cardiopathie ischémique**
22203632	**13.09.2012**	**41**	**M**	**3-0**	**insuffisance resp aigüe**
22203632	**13.09.2012**	**41**	**M**	**5-0**	**Surpoids /diabète évoluée malaidie parquinson**

*Infarctus du myocarde

1.2 Le dictionnaire

Le CépiDc a mis à disposition de Santé publique France le dictionnaire répertoriant l'ensemble des expressions des causes de décès retrouvées au moins une fois dans un certificat de décès et le ou les codes CIM-10 qui leur ont été attribués depuis le début des années 2000. Ce dictionnaire est mis à jour régulièrement. En 2016, il contenait plus de 150 000 entrées. Initialement conçu pour faciliter et homogénéiser le codage des causes de décès, il offre une grande variété d'expressions de causes.

Une même cause de décès pouvait être associée à plusieurs codes CIM-10 différents, en fonction du contexte du décès définis par les autres causes de décès rapportées dans le certificat de décès. C'est par exemple le cas de l'expression « intoxication au ciprolon » (Tableau 8). Cette expression était associée dans le dictionnaire au code X44 : « Intoxication accidentelle par des médicaments et substances biologiques et exposition à ces produits, autres et sans précision », au code X64 : « Auto-intoxication par des médicaments et substances biologiques et exposition à ces produits, autres et sans précision » et au code Y14 : « Intoxication par des médicaments et substances biologiques, autres et sans précision et exposition à ces produits, intention non déterminée ». Dans notre objectif

de classification, nous cherchons à classer les causes de décès indépendamment les unes des autres, c'est-à-dire sans prendre en compte le contexte. Nous avons donc choisi pour chaque expression associée à plusieurs codes CIM-10 (30 000 entrées) dans le dictionnaire, de conserver uniquement l'entrée la moins dépendante du contexte pour finalement obtenir un dictionnaire avec une entrée unique par cause (environ 120 000 entrées). Ainsi pour l'expression « Intoxication au Ciprolon », seule l'expression associée au code Y14 a été retenue (Tableau 8).

Dans une seconde étape, nous avons associé à chaque expression contenue dans le dictionnaire, le regroupement syndromique auquel il appartenait. Dans la suite des travaux, le dictionnaire du CépiDc modifié pour répondre à nos objectifs sera mentionné comme : le « dictionnaire adapté du CépiDc ».

Tableau 8 : Extrait du dictionnaire adapté du CépiDc après ajout des regroupements syndromiques associés aux expressions de causes de décès, 2019.

Libellé causes	Code CIM-10	Regroupement syndromique	Thématique	Code conservé[a]
défaillance multiviscérale	R688	Symptômes généraux	Symptômes	X
défaillance multi-viscérale	R688	Symptômes généraux	Symptômes	X
syndrome défaillance respiratoire aigüe	J960	Insuffisance respiratoire aiguë	Respiratoire	X
pneumopathie infectieuse	J189	IRAB	Respiratoire	X
bronchopneumopathie infectieuse	J180	IRAB	Respiratoire	X
embolies pulmonaires	I269	Embolie pulmonaire	Cardio et cérébrovasculaire	X
intoxication ciprolon	X44	Effets toxiques liés à des substances médicinales	Trauma et empoisonnement	
intoxication ciprolon	X64	Suicides	Causes externes	
intoxication ciprolon	Y14	Effets toxiques liés à des substances médicinales	Trauma et empoisonnement	X
lésions neurologiques	G98	Maladies chroniques du SN	Système nerveux	X
lésions neurologiques	P918	Tableaux neuropsycho aigüs	Système nerveux	
lésions neurologiques	P968	Maladies Néonatales	Maladies de la grossesse, néonatales et congénitales	
oedème aigu poumon	I501	Décompensation/insuffisance cardiaque aigue	Cardio et cérébrovasculaire	
oedème aigu poumon	J81	Décompensation/insuffisance cardiaque aigue	Cardio et cérébrovasculaire	X
apnée sommeil	G473	Maladies chroniques du SN	Système nerveux	X
apnée sommeil	P283	Maladies Néonatales	Maladies de la grossesse, néonatales et congénitales	
apnées obstructives sommeil	G473	Maladies chroniques du SN	Système nerveux	X
cardiopathie ischémique	I259	Cardio et circulatoire chronique	Cardio et cérébrovasculaire	X

[a] Code conservé dans le dictionnaire pour la classification

1.3 Sélection des regroupements syndromiques pour l'évaluation des méthodes

Afin d'évaluer les méthodes de classification des causes de décès dans les RS, nous avons choisi de sélectionner un nombre restreint de RS. Le choix des RS pour l'évaluation s'est appuyé sur :

- La finalité du RS, c'est-à-dire l'objectif pour lequel il sera suivi. Nous avons défini une liste de de RS qui seront spécifiquement suivis pour l'alerte (cf. Chapitre III).
- Le caractère spécifique ou non spécifique du RS, c'est-à-dire si le RS est défini pour suivre une pathologie bien définie (ex : Grippe, Asthme…) ou pour un ensemble plus large ou non spécifique (ex : Maladies chroniques, fièvre d'origine inconnue).
- Le caractère sensible ou non aux erreurs de classification, c'est-à-dire des RS qui ont des expressions de causes proches (ex : pneumopathie, pneumopathie d'inhalation). La proximité des définitions peut entrainer des difficultés pour les méthodes à classer les entités nosologiques dans l'un ou l'autre regroupement.

Pour que l'évaluation des méthodes sur l'échantillon de RS soit représentative de tous les types de RS définis, nous avons choisi sept RS qui possédaient au moins une des caractéristiques citées.

- Le RS **« Grippe » (G)** est un RS suivi pour l'alerte et est très spécifique.
- Le RS **« Insuffisance respiratoire aigüe » (IRA)** est un RS suivi pour l'alerte, sensible aux erreurs de classification notamment avec les RS « Asphyxie et Anomalie de la respiration » et « Maladies respiratoires aigües »
- Le RS **« Infections respiratoires aigües basses » (IRAB)** est un RS suivi pour l'alerte, et peu spécifique.
- Le RS **« Asphyxie et anomalie de la respiration » (AAR)** est un RS suivi pour l'alerte, peu spécifique et sujet aux erreurs de classification.
- Le RS **« Sepsis »** est un RS suivi pour l'alerte et spécifique.
- Le RS **« Maladies chroniques endocriniennes » (MCE)** est un RS non suivi pour l'alerte et peu spécifique qui contient un grand nombre de pathologies.
- Le RS **« Maladies chroniques de l'appareil digestif » (MCAD)** est un RS non suivi pour l'alerte et peu spécifique, qui contient un grand nombre de pathologies.

La définition de ces regroupements syndromiques, ainsi que des expressions des causes de décès sont présentées en annexe 4.

2) Découpage du corpus de certificats de décès électroniques de 2012 à 2016

2.1 Description du corpus

Les causes de décès n'ayant jamais été classées dans les regroupements syndromiques, à partir du texte libre, nous avons construit les RS à partir des causes de décès codées en CIM-10. Ces regroupements syndromiques fondés sur les codes constituent un proxi des regroupements syndromiques, permettant de nous donner un ordre d'idée de ce que nous cherchons à obtenir. Ils ne sont toutefois pas l'exact reflet des regroupements syndromiques construits à partir du texte libre, notamment pour les causes codées en fonction du contexte.

La description des corpus a été effectué à partir des causes de décès sous forme de codes CIM-10. Les RS étant définis par une liste de codes CIM-10, nous avons attribué à chaque code le RS auquel il appartenait.
Pour chaque RS, la proportion de certificats de décès contenant au moins un code appartenant aux RS d'intérêt parmi l'ensemble des certificats a été calculée. La proportion de chaque RS parmi l'ensemble des RS a aussi été calculée.

Parmi les 203 797 certificats de décès, 16,3% des certificats contenaient au moins une cause codée appartenant au RS « Maladies chroniques endocriniennes » et 16,5% des certificats contenaient au moins une cause codée appartenant au RS « Maladies chroniques de l'appareil digestif » sur l'ensemble de la période 2012-2016 (Tableau 9). La proportion de certificats contenant au moins une cause codée de chaque RS était stable entre 2012 et 2016.

On retrouvait 4,3 % du RS « Maladies chroniques endocriniennes » et 3,7% de RS « Maladies chroniques de l'appareil digestif » parmi l'ensemble des RS des certificats de 2012 à 2016. La proportion de chaque RS était stable entre 2012 et 2016.

Tableau 9 : Proportion de certificats de décès où sont retrouvés au moins une fois les RS par année et pour la période 2012-2016, et proportion de RS parmi l'ensemble des RS, certificats électroniques, France.

	Proportion de certificats de décès avec au moins une fois le RS						Proportion de RS parmi l'ensemble des RS attribués					
	2012	2013	2014	2015	2016	2012-2016	2012	2013	2014	2015	2016	2012-2016
Nb total certificats/RS attribués	**25 982**	**30 317**	**30 391**	**52 519**	**60 048**	**203 797**	**85 486**	**98 388**	**104 654**	**180 643**	**216 283**	**685 454**
Asphyxie et Anomalies de la respiration	0,1	0,2	0,2	0,4	0,3	0,2	0,8	0,8	1,0	1,0	1,0	0,9
Grippe	8,7	8,3	8,1	9	8,7	8,5	0,1	0,1	0,1	0,1	0,1	0,1
Insuffisance respiratoire aiguë	12,5	11,3	11,2	12,1	11,8	11,8	2,6	2,5	2,4	2,6	2,6	2,5
IRAB*	2,9	2,7	3,4	3,6	3,4	3,2	3,7	3,4	3,3	3,4	3,4	3,5
Sepsis	11,4	12,4	19,5	13,2	12,5	13,8	3,3	3,4	3,4	3,5	3,4	3,4
Maladies chroniques endocriniennes	17,2	16,1	16,4	16,2	15,8	16,3	4,5	4,3	4,2	4,1	4,1	4,3
Maladies chroniques de l'appareil digestif	17,3	16,8	17,3	15,7	15,3	16,5	3,9	3,8	3,9	3,5	3,5	3,7

*Infections respiratoires aigües basses

Aide à la lecture : En 2012,

- 25 982 certificats ont été rédigés sous forme électronique,
- 12,5% des certificats contenaient au moins une fois un RS « Insuffisance respiratoire aigüe »,
- 85 486 RS ont été attribués à partir des causes codées contenus dans les 25 982 certificats,
- Parmi ces 85 486 RS, 2,6% était des RS « Insuffisance respiratoire aigüe ».

2.2 Construction des corpus d'entrainement, de développement, d'évaluation interne et d'évaluation finale

Afin de constituer les échantillons qui seront utilisés pour l'évaluation de la classification des causes de décès dans les RS, nous avons construit des sous-échantillons de certificats de décès à partir de l'ensemble de certificats électroniques de 2012 à 2016.

Les ensembles d'entrainement et de développement ont été constitués à partir d'un tirage au sort aléatoire des certificats des décès survenus entre 2012 et 2014, selon la répartition suivante :

- Ensemble d'entrainement : 80% des certificats entre 2012 et 2014
- Ensemble de développement : 20% des certificats entre 2012 et 2014

L'ensemble d'évaluation interne a été constitué à partir des certificats des décès survenus en 2015 et l'ensemble d'évaluation finale à partir des certificats de décès de 2016 (Figure 23).

Ce découpage par année a été choisi car il permettait de simuler l'utilisation en routine, avec de nouvelles données contenant de potentielles nouvelles pathologies.

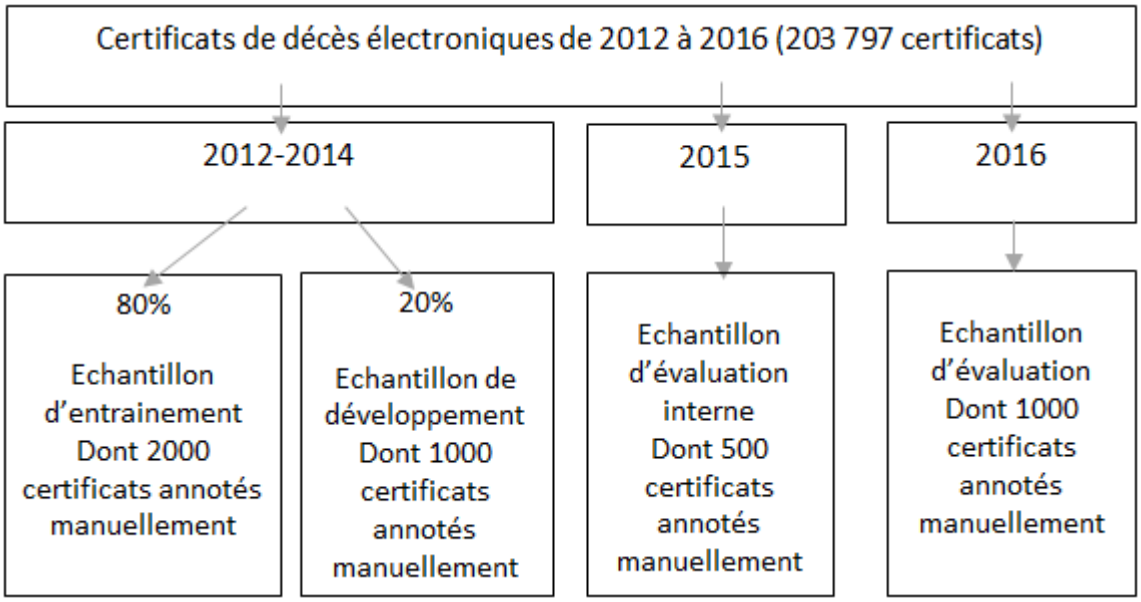


Figure 23 : Schéma de la construction des échantillons de certificats pour l'évaluation des méthodes de classification, certificats électroniques de 2012 à 2016

Parmi ces quatre ensembles de certificats de décès, quatre échantillons de certificats ont été tirés au sort aléatoirement et sans remise pour constituer les échantillons annotés (Figure 23):

- Echantillon d'entrainement : 2000 certificats tirés au sort au sein de l'ensemble d'entrainement,
- Echantillon de développement : 1000 certificats tirés au sort au sein de l'ensemble de développement,
- Echantillon d'évaluation interne : 500 certificats tirés au sort au sein de l'ensemble d'évaluation interne,
- Echantillon d'évaluation finale : 1000 certificats tirés au sort au sein de l'ensemble d'évaluation finale.

Afin de vérifier si les échantillons tirés au sort pour notre évaluation étaient représentatifs des ensembles à partir desquels ils ont été constitués, nous avons comparé les proportions des 7 RS (construits en utilisant les données codées en CIM-10) sélectionnés pour l'évaluation dans les échantillons aux proportions obtenues dans les ensembles complets de certificats (Tableau 10). Les proportions ont été comparées à l'aide d'un chi-2. Aucune différence statistiquement significative n'a été retrouvée

Tableau 10 : Proportion des 7 regroupements syndromiques parmi l'ensemble des regroupements syndromiques dans les échantillons d'entrainement, développement, évaluation interne et évaluation finale et dans les ensembles de certificats de décès de2012 à 2016

	Grippe	Insuffisance respiratoire aigue	Infections respiratoires aigües basses	Asphyxie/ anomalie de la respiration	Sepsis	Maladies chroniques endocriniennes	Maladies chroniques de l'appareil digestif
Echantillon d'entrainement	**0,12**	**2,30**	**3,40**	**1,04**	**3,21**	**4,97**	**3,88**
Echantillon de développement	**0,10**	**2,02**	**2,66**	**0,69**	**3,14**	**3,71**	**3,66**
Années 2012-2014	0,10	2,52	3,48	0,87	3,36	4,36	3,87
Echantillon d'évaluation interne	**0,10**	**2,21**	**2,40**	**1,03**	**2,84**	**4,17**	**3,97**
Année 2015	0,13	2,61	3,45	1,00	3,48	4,08	3,50
Echantillon d'évaluation finale	**0,12**	**2,59**	**3.56**	**1,00**	**3,53**	**3,90**	**3,47**
Année 2016	0,10	2,57	3,44	0,98	3,41	4,10	3,48

3) Annotation manuelle

Les 4 500 certificats de décès qui constituent les quatre échantillons ont été annotés manuellement et parallèlement par deux annotateurs : tous deux étaient des épidémiologistes de langue maternelle française. Ils avaient travaillé sur la définition des regroupements syndromiques.

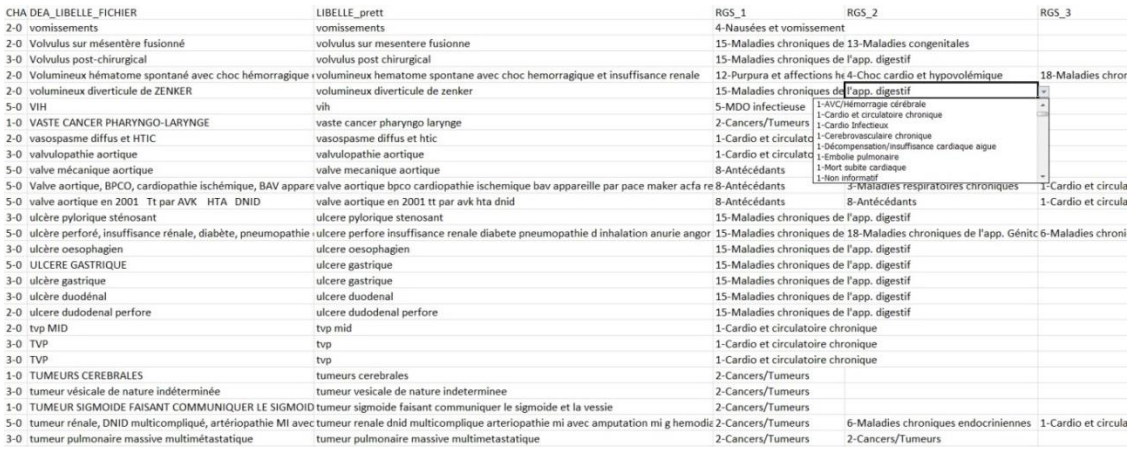
Quatre fichiers, correspondant à chaque échantillon de certificat, ont été construits. Les informations sociodémographiques des décès ont été retirées pour l'annotation ainsi que l'identifiant du certificat. Chaque certificat a été découpé en champs. Ainsi, les annotateurs disposaient d'un fichier contenant les champs des certificats de décès triés par ordre alphabétique.
Pour chaque champ, les annotateurs pouvaient choisir dans un menu déroulant parmi la centaine de RS définis et pouvaient associer autant de RS qu'ils souhaitaient à chaque champ (Figure 24).

Des règles d'annotation ont été définies en amont :

- Chaque champ d'un certificat était lu et annoté indépendamment des autres,
- Les annotateurs indiquaient pour chaque champ du certificat les RS qu'ils associaient aux causes de décès dans l'ordre de lecture du champ,
- Les annotateurs effectuaient leur travail d'annotation indépendamment l'un de l'autre.

Afin de faciliter l'annotation, les deux annotateurs disposaient de la définition détaillée des RS et du dictionnaire adapté du CépiDc, qui contient toutes les expressions des causes médicales de décès depuis le début des années 2000 et les regroupements syndromiques associés.

Les annotations des deux annotateurs ont ensuite été comparées. Les incohérences ont été discutées entre les annotateurs. Les définitions des RS et le dictionnaire du CépiDc ont été utilisés pour résoudre les conflits. Le taux d'accord (proportion d'annotations identiques) était de 0,90 dans l'échantillon d'évaluation finale. Les échantillons annotés finaux ont constitué les références pour l'évaluation.



CHA	DEA_LIBELLE_FICHIER	LIBELLE_prett	RGS_1	RGS_2	RGS_3
2-0	vomissements	vomissements	4-Nausées et vomissement		
2-0	Volvulus sur mésentère fusionné	volvulus sur mesentere fusionne	15-Maladies chroniques de	13-Maladies congenitales	
3-0	Volvulus post-chirurgical	volvulus post chirurgical	15-Maladies chroniques de	l'app. digestif	
2-0	Volumineux hématome spontané avec choc hémorragique	volumineux hematome spontane avec choc hemorragique et insuffisance renale	12-Purpura et affections hé	4-Choc cardio et hypovolémique	18-Maladies chror
2-0	volumineux diverticule de ZENKER	volumineux diverticule de zenker	15-Maladies chroniques de	l'app. digestif	
5-0	VIH	vih	5-MDO infectieuse		
1-0	VASTE CANCER PHARYNGO-LARYNGE	vaste cancer pharyngo larynge	2-Cancers/Tumeurs		
2-0	vasospasme diffus et HTIC	vasospasme diffus et htic	1-Cardio et circulato		
3-0	valvulopathie aortique	valvulopathie aortique	1-Cardio et circulato		
5-0	valve mécanique aortique	valve mecanique aortique	8-Antécédants		
5-0	Valve aortique, BPCO, cardiopathie ischémique, BAV appare	valve aortique bpco cardiopathie ischemique bav appareille par pace maker acfa re	8-Antécédants	3-Maladies respiratoires chroniques	1-Cardio et circula
5-0	valve aortique en 2001 Tt par AVK HTA DNID	valve aortique en 2001 tt par avk hta dnid	8-Antécédants	8-Antécédants	1-Cardio et circula
3-0	ulcère pylorique sténosant	ulcere pylorique stenosant	15-Maladies chroniques de	l'app. digestif	
5-0	ulcère perforé, insuffisance rénale, diabète, pneumopathie	ulcere perfore insuffisance renale diabete pneumopathie d inhalation anurie angor	15-Maladies chroniques de	18-Maladies chroniques de l'app. Génitc	6-Maladies chroni
3-0	ulcère oesophagien	ulcere oesophagien	15-Maladies chroniques de	l'app. digestif	
5-0	ULCERE GASTRIQUE	ulcere gastrique	15-Maladies chroniques de	l'app. digestif	
3-0	ulcère gastrique	ulcere gastrique	15-Maladies chroniques de	l'app. digestif	
3-0	ulcère duodénal	ulcere duodenal	15-Maladies chroniques de	l'app. digestif	
2-0	ulcere dudodenal perfore	ulcere dudodenal perfore	15-Maladies chroniques de	l'app. digestif	
2-0	tvp MID	tvp mid	1-Cardio et circulatoire chronique		
3-0	TVP	tvp	1-Cardio et circulatoire chronique		
3-0	TVP	tvp	1-Cardio et circulatoire chronique		
1-0	TUMEURS CEREBRALES	tumeurs cerebrales	2-Cancers/Tumeurs		
3-0	tumeur vésicale de nature indéterminée	tumeur vesicale de nature indeterminee	2-Cancers/Tumeurs		
1-0	TUMEUR SIGMOIDE FAISANT COMMUNIQUER LE SIGMOID	tumeur sigmoide faisant communiquer le sigmoide et la vessie	2-Cancers/Tumeurs		
5-0	tumeur rénale, DNID multicompliqué, artériopathie MI avec	tumeur renale dnid multicomplique arteriopathie mi avec amputation mi g hemodia	2-Cancers/Tumeurs	6-Maladies chroniques endocriniennes	1-Cardio et circula
3-0	tumeur pulmonaire massive multimétastatique	tumeur pulmonaire massive multimetastatique	2-Cancers/Tumeurs	2-Cancers/Tumeurs	

Figure 24 : Extrait d'un fichier contenant les champs de certificats à annoter par les annotateurs et les RS que les annotateurs ont attribué à chaque champ

III/ Classification des causes de décès dans les regroupements syndromiques

1) Prétraitement des données

Le prétraitement des données est une étape fondamentale de la classification des causes de décès. Ce prétraitement se décompose en 6 étapes qui sont illustrées par un exemple dans la Figure 25 :

1/ Négliger la casse : Les textes écrits en majuscules ont été transformés en minuscules.

2/ Supprimer les diacritiques : Les diacritiques (accents) ont été supprimées de tous les textes.

3/ Normaliser les mots composées : Cette étape consiste à dé-concaténer les mots composés afin d'obtenir 2 tokens distincts pour la phase de « tokenisation» (Découpage des textes en tokens).

Pour cela, un dictionnaire spécifique a été construit avec en clé tous les mots avec trait d'union et les même mots attachés sans le trait d'union trouvés dans le dictionnaire adapté du CépiDc, et en valeur, le mot final avec un espace au lieu du trait d'union.

Ex : cardio-vasculaire → cardio vasculaire

cardiovasculaire → cardio vasculaire

4/ Remplacer la ponctuation par des espaces : Tous les signes de ponctuation, à l'exception de la virgule, l'apostrophe et le point-virgule, ont été remplacés par un espace.

5/ Supprimer les mots vides : Un mot vide est un mot n'ayant pas de sens dans un texte. Nous avons établis la liste de mots vides suivante : «un», «une», «des», «le», «la», «l», «les», «au», «aux», «de», «du», «d», «en», «l'», «d'», «il», «alors», «dont». Le mot «a» était aussi considéré comme un mot vide sauf s'il apparaissait dans les expressions suivantes : «hepatite a», «grippe a », «viral a », «ig a», «chlid a», «type a», «stade a».

6/supprimer les multiples espaces dans les expressions et au début et à la fin : Les multi-espaces présents dans les textes ont été supprimés, ainsi que les espaces en début et fin de champ. Cela concerne notamment les espaces apparus lors des étapes précédentes.

Le prétraitement permet d'obtenir un ensemble plus restreint et homogène de termes qui seront pertinents à classer dans les RS. Il est effectué à la fois sur les échantillons de certificats de décès et sur le dictionnaire adapté du CépiDc.

Etapes de prétraitement	Exemple extrait des certificats de décès électroniques
Avant le prétraitement	**“**rupture d’anévrysme cérébral**”** **“** Démence vasculaire, malaidie de Parkinson, INSUFFISANCE RENALE CHRONIQUE **”**
Négliger la casse	**“**rupture d’anévrysme cérébral**”** **“** démence vasculaire, malaidie de parkinson, insuffisance renale chronique**”**
Supprimer les diacritiques	**“**rupture d’anevrysme cerebral**”** **“** demence vasculaire, malaidie de parkinson, insuffisance renale chronique**”**
Normaliser les mots composés (remplacement du trait d’union par un espace)	*cardio-vasculaire -> cardio vasculaire* *cardiovasculaire -> cardio vasculaire*
Remplacer la ponctuation par un espace (sauf la virgule, l’apostrophe, le point-virgule)	**“**rupture d’ anevrysme cerebral **“** demence vasculaire, malaidie de parkinson, insuffisance renale chronique**”**
Supprimer les mots vides (de, d, du, le, les, la…)	**“**rupture anevrysme cerebral**”** **“** demence vasculaire, malaidie parkinson, insuffisance renale chronique**”**
Supprimer les multiples espaces et/ou les espaces en début et fin de champs	**“**rupture anevrysme cerebral**”** **“**demence vasculaire, malaidie parkinson, insuffisance renale chronique**”**

Figure 25 : Exemple d’application de prétraitements sur les causes médicales de décès mis en œuvre en amont de leur classification dans les regroupements syndromiques

2) Méthode par règles

Cette méthode s'appuie sur le dictionnaire adapté du CépiDc. Le principe est d'attribuer un RS à une entité nosologique contenue dans un champ de certificat quand celle-ci correspond parfaitement à une expression contenue dans le dictionnaire adapté du CépiDc. Afin d'attribuer un RS à la majorité des entités nosologiques des certificats, nous avons mis en œuvre quatre étapes de traitement. A l'issue de chacune d'elles, on recherche si les entités nosologiques traitées correspondent à une expression du dictionnaire. La figure 27 résume l'ensemble des étapes de la méthode par règles.

2.1/ Règles de standardisation

La première étape de cette méthode est une normalisation des termes et expressions grâce à des règles de standardisation fournies par l'Inserm-CépiDc. Ces règles (plus de 900) ont été créées à l'origine pour faciliter le codage des causes de décès. Elles ont été modifiées pour répondre à nos objectifs de classification. Ces règles sont constituées d'expressions rationnelles qui permettent de supprimer des termes, d'en remplacer certains par leur synonyme ou encore de transformer une abréviation par son expression complète (Figure 26). Elles doivent être utilisées dans un ordre précis afin d'obtenir une normalisation des textes satisfaisante.

mainkey	rank	filterin	filterout
400acronymes	517	\bd\.?t\.?2\.?\b[.]?	diabete type 2
400acronymes	518	\bd\.?t\.?1\.?\b[.]?	diabete type 1
600correction	107	\binfractus\b	infarctus
600correction	108	\binsufisance\b	insuffisance
400synonymes	101	\babdominal\b	abdomen
400synonymes	102	\babdominale\b	abdomen
400synonymes	103	\babdominales\b	abdomen
400synonymes	104	\babdominaux\b	abdomen
400synonymes	105	\bavci\b	avc ischemique
500synonymes	106	\bfoie\b	hepatique

Figure 26 : Exemple de règles de standardisation, pour la normalisation des termes et expressions

2.2/ Les séparateurs

Cette étape utilise des repères de surface pour segmenter les entités nosologiques distinctement. Ces repères sont des signes de ponctuation («,», «;»), ou des mots («avec», «après», «sur», «puis», «pour», «en rapport avec», «en raison de» et «suite à», ou bien des signes (« + », « / ») qui délimitent la frontière entre deux entités nosologiques dans un champ de certificat.

Deux conjonctions de coordination («et», «ou») ont été identifiées comme susceptibles de mettre en facteur des entités (ex : « Hepatite A et B »). Les entités nosologiques incluant ces conjonctions sont développées en deux entités distinctes (ex : « Hepatite A », « Hepatite B »).

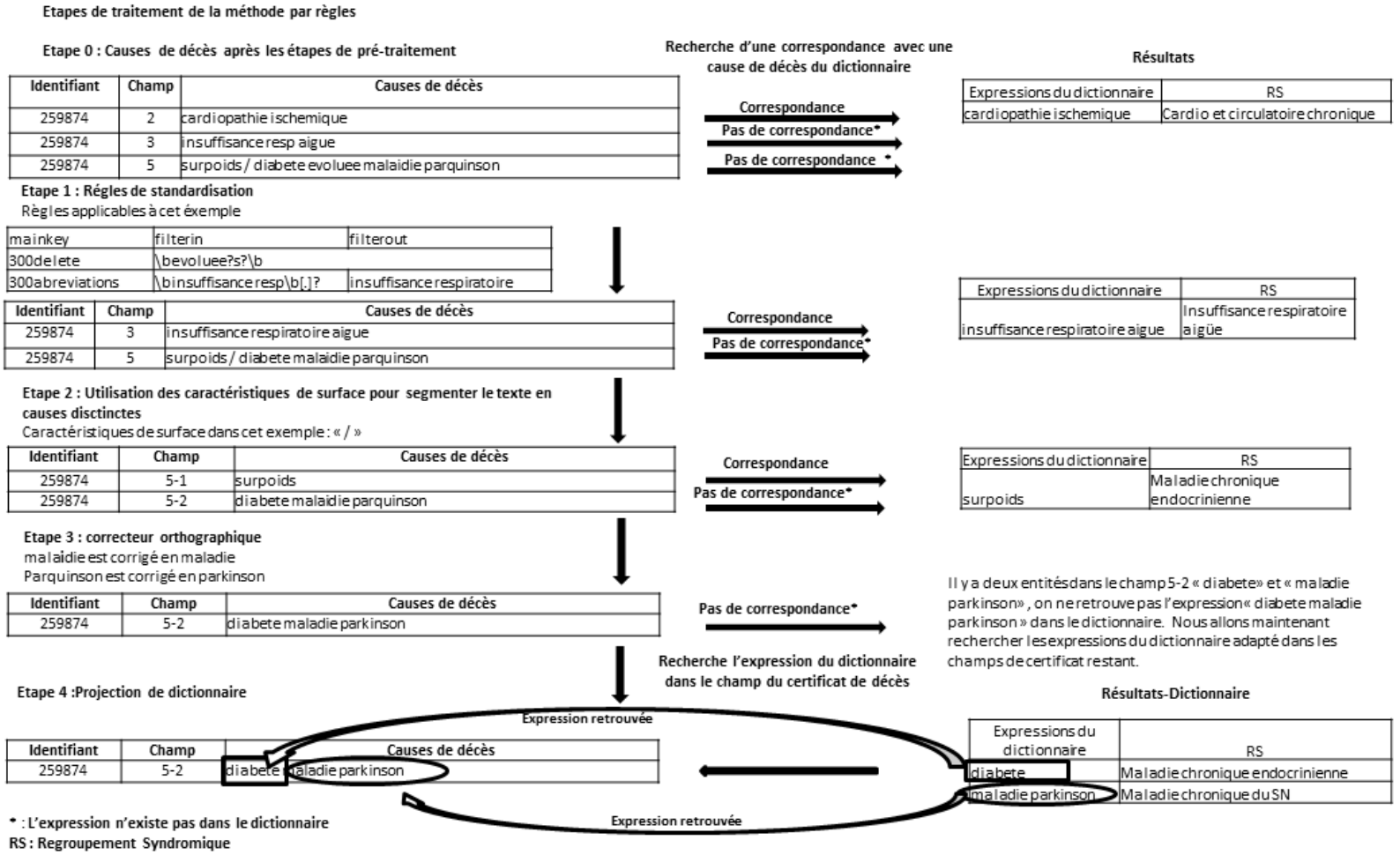


Figure 27 : Synthèse et illustration des 4 étapes de traitement de la méthode par règles pour la classification des causes de décès dans les regroupements syndromiques

Le mot «sous» a aussi été utilisé comme un délimiteur dans des expressions telles que «ACFA sous anticoagulant». Cette entité nosologique est segmentée en deux entités : le mot avant le mot «sous» constitue une première entité nosologique et le mot «sous» est gardé avec la seconde partie pour constituer une seconde entité («ACFA», «sous anticoagulant»).

2.3/ Correcteur orthographique

Afin de corriger les mots mal orthographiés présents dans les champs des certificats de décès, nous avons construit un correcteur orthographique.

2.3.1 Définition des mots mal orthographiés

Nous avons défini un mot mal orthographié comme étant un mot absent du vocabulaire défini par l'ensemble des mots retrouvés dans le dictionnaire adapté du CépiDC. Nous avons fait l'hypothèse que si un mot n'appartenait pas à ce vocabulaire (mot « hors vocabulaire »), celui-ci était mal orthographié.

2.3.2 Démarche

Le correcteur orthographique a été implémenté en utilisant une méthode de recherche d'information. Cette méthode contient deux étapes: la détection de mots mal orthographiés et la correction de ces mots.

- Méthode :

Cette méthode consiste à représenter tous les mots présents dans le vocabulaire dans un espace vectoriel à N dimensions. Si un mot est hors vocabulaire, nous effectuons une transformation vectorielle de ce mot, et nous lui attribuons comme correction le mot du vocabulaire le plus proche en termes de distance mathématique. C'est un algorithme de recherche du plus proche voisin (60) (Figure 28).

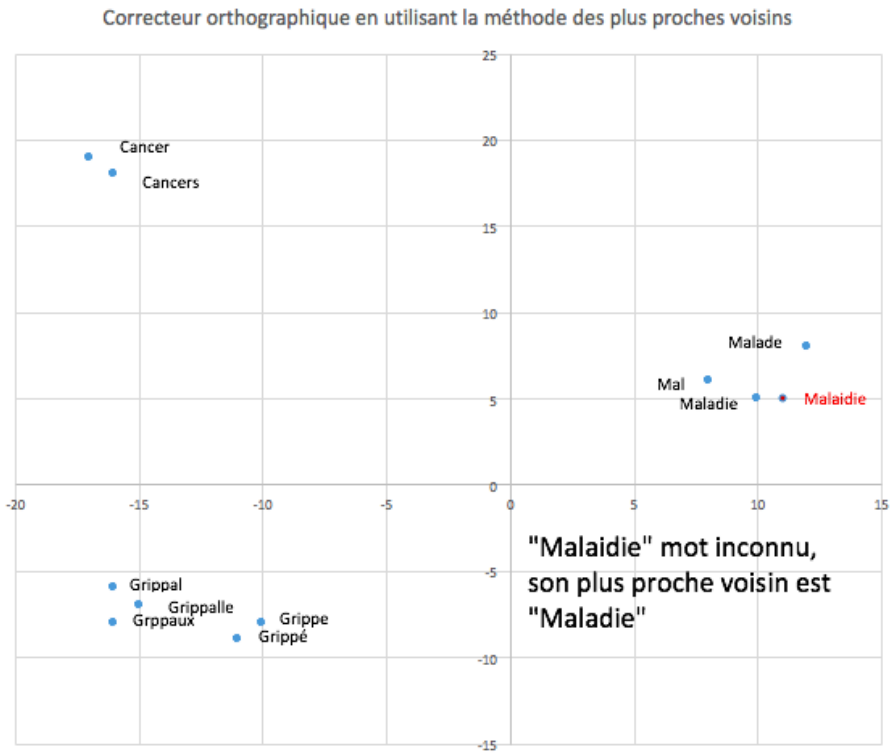


Figure 28 : Schéma du correcteur orthographique utilisant la méthode des plus proches voisins

- Caractéristiques d'apprentissage :

L'ensemble d'apprentissage pour l'implémentation du correcteur orthographique était l'ensemble des mots du vocabulaire.

Pour chaque mot, les unigrammes et bigrammes de caractères ont été construits. Chaque unigramme et bigramme de caractères était considéré comme une dimension du vecteur d'apprentissage. Pour chaque mot, on indiquait dans le vecteur le nombre de fois où l'unigramme et bigramme de caractères apparaissaient.

- Evaluation du modèle

Le modèle a été évalué sur 610 mots qui ont été annotés selon 5 catégories :

- ✓ Catégorie 1 : Le mot était mal orthographié et il a bien été corrigé (N=406),
- ✓ Catégorie 2 : Le mot était mal orthographié et il a mal été corrigé par le modèle (N=12),
- ✓ Catégorie 3 : Le mot était correctement orthographié et a été corrigé par le modèle par erreur (N=77),
- ✓ Catégorie 4 : Le mot était correctement orthographié mais a été remplacé par un mot de la même famille (apparenté à une lemmatisation) (N=55),
- ✓ Catégorie 5 : La compréhension du mot était ambiguë, la correction effectuée n'est pas évaluable (N=60).

Pour cette cinquième catégorie, les mots hors-vocabulaire que l'annotateur ne sait pas évaluer (60 en tout) étant très ambigus, ils sont considérés pour l'évaluation de ce modèle comme des mots mal orthographiés et mal corrigés par le modèle. Ils viennent s'ajouter aux 12 mots de la catégorie 2.

Deux tiers (478) des mots identifiés par le correcteur orthographique comme mal orthographiés, l'étaient bien. Quatre-vingt-cinq pourcent des mots étaient bien corrigés lorsqu'ils étaient détectés comme mal orthographiés

2.4/ Projection de dictionnaire

L'étape de projection de dictionnaire a été ajoutée aux 3 précédentes étapes suite à l'analyse des performances sur l'échantillon de développement.

Cette étape consiste à rechercher si l'une des expressions du dictionnaire adapté du CépiDc était contenue dans sa totalité dans chacun des champs du certificat. Plusieurs expressions du dictionnaire pouvaient être retrouvées dans un champ de certificat et plusieurs RS pouvaient ainsi être attribués à ce champ.

Cette étape était découpée en plusieurs phases :

1/ Le dictionnaire adapté était trié de l'expression dont le libellé était le plus long au libellé le plus court.

2/ Le dictionnaire trié était parcouru en cherchant une correspondance entre l'expression du dictionnaire parcouru et, tout ou partie d'un champ de certificat.

- Si une correspondance était trouvée avec la totalité du champ, on attribuait le regroupement syndromique correspondant. On passait ensuite au champ suivant, en recommençant à parcourir le dictionnaire depuis le début.
- Si la correspondance ne concernait pas la totalité du champ, on continuait de parcourir le dictionnaire sur la partie du champ restante afin de trouver une correspondance. Et ainsi de suite jusqu'à ce que tout le champ ait été parcouru.

3) Méthode par apprentissage supervisé

L'apprentissage supervisé est une technique d'apprentissage automatique où l'on cherche à produire des règles à partir d'un ensemble d'apprentissage (ou base de données d'apprentissage) contenant des exemples.

Si on considère un ensemble d'apprentissage constitué de couples $(x_n, y_n)_{1 \leq n \leq N}$ avec $x_n \in X$ et $y_n \in Y$ et X : l'ensemble des objets et Y : l'ensemble des étiquettes, l'objectif de l'apprentissage supervisé est de prédire les valeurs de y_n associées à chaque $x_n \in X$. On parle de fonction de prédiction (Figure 29)

Ici, l'ensemble des objets X est l'ensemble des champs de certificats contenant les entités nosologiques et Y est l'ensemble fini des 98 regroupements syndromiques. Lorsque l'ensemble Y est un ensemble fini, on parle d'une tâche de classification. On attribue à chaque objet (champ de certificat), une ou plusieurs étiquettes. La fonction de prédiction porte alors le nom de classifieur.

Parmi l'ensemble des méthodes supervisées employées pour la classification automatique de textes, nous avons choisi de mettre en place une Machine à vecteur de support (Support Vector Machine, SVM). Les SVM sont une généralisation des classifieurs linéaires.

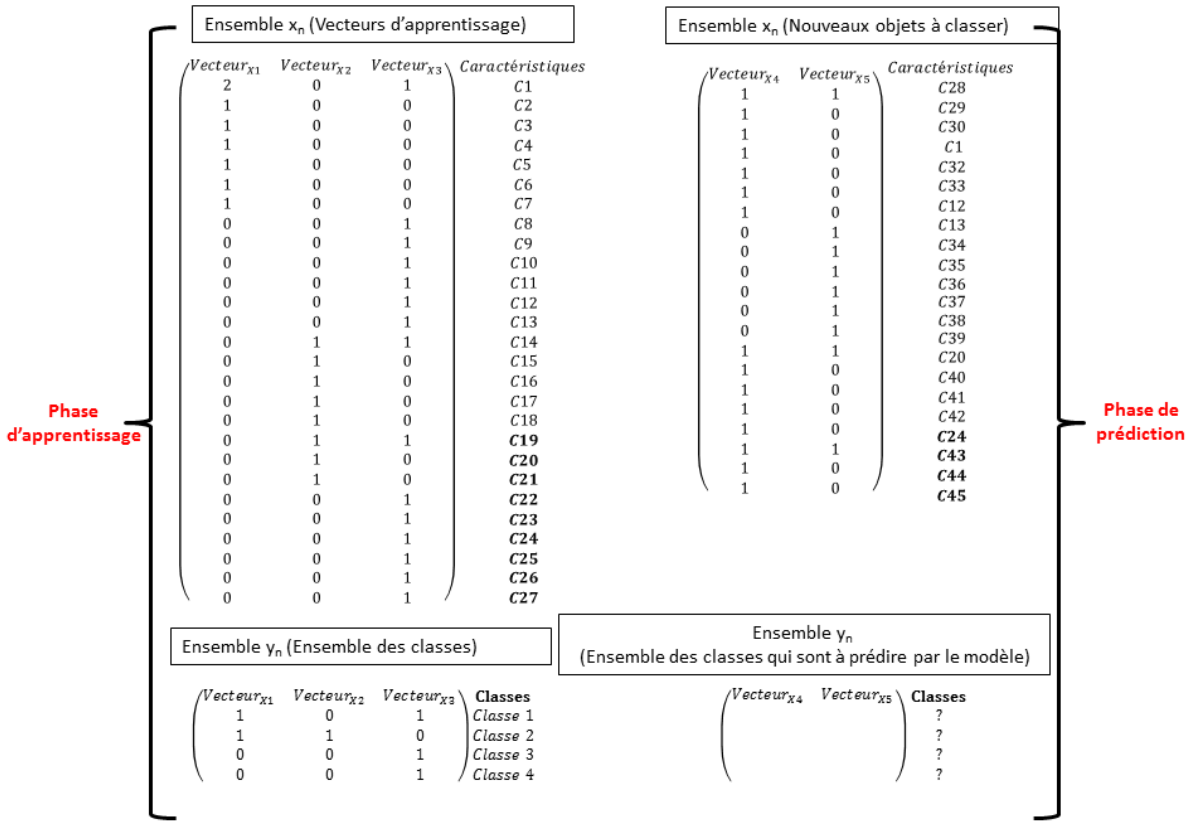


Figure 29: Schéma de la phase d'apprentissage et de la phase de prédiction d'un SVM

3.1 Les classifieurs linéaires

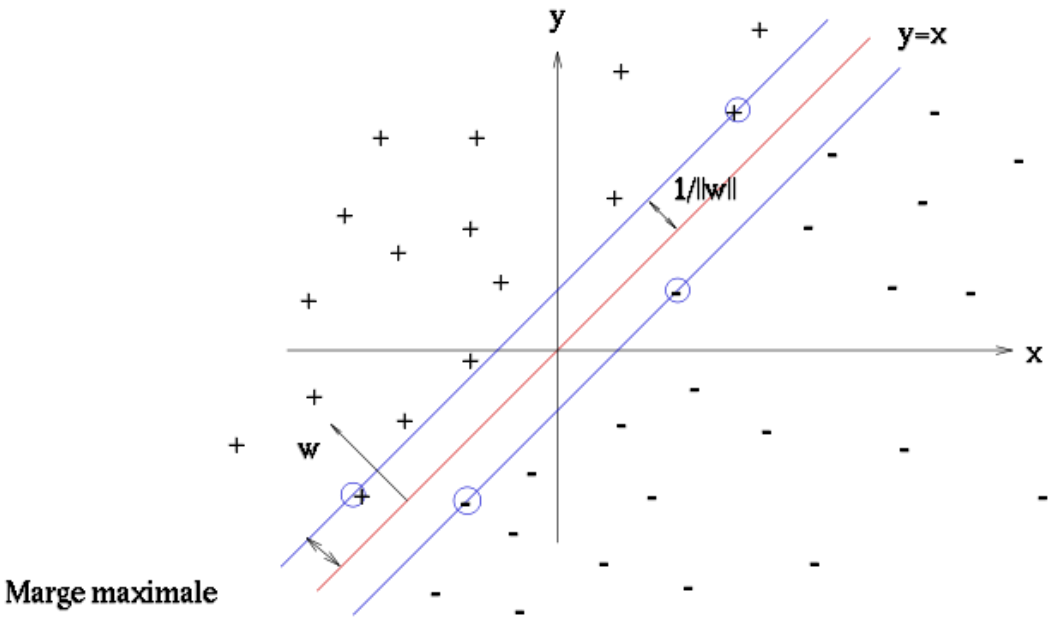
Les classifieurs linéaires classent dans des groupes (des classes) les échantillons qui ont des propriétés similaires, mesurées sur des observations. Un classifieur linéaire est un type particulier de classifieur, qui calcule la décision par combinaison linéaire des échantillons.

Géométriquement, on cherche l'équation d'un hyperplan qui sépare les données en deux espaces correspondant aux classes.

3.2 Les SVM

Les SVM reposent sur deux notions : la marge maximale et la fonction noyau.

- La marge maximale permet de trouver un unique plan séparateur. L'hypothèse initiale est que les données sont linéairement séparables. Il existe donc une infinité d'hyperplans séparateurs. On va alors chercher à obtenir l'unique hyperplan ayant la marge maximale, c'est-à-dire ayant les distances les plus grandes entre l'hyperplan et les échantillons (Figure 30).



La marge est la distance entre l'hyperplan (en rouge) et les objets les plus proches (entourés en bleu). Ces derniers sont appelés vecteurs supports

Figure 30 : Hyperplan optimal (en rouge) avec la marge maximale dans le cas de la méthode SVM

- La fonction noyau cherche à rendre des données linéairement séparables si elles ne le sont pas initialement. Les données sont généralement projetées dans un espace de dimension supérieure dans lequel on pourra trouver un hyperplan séparateur.

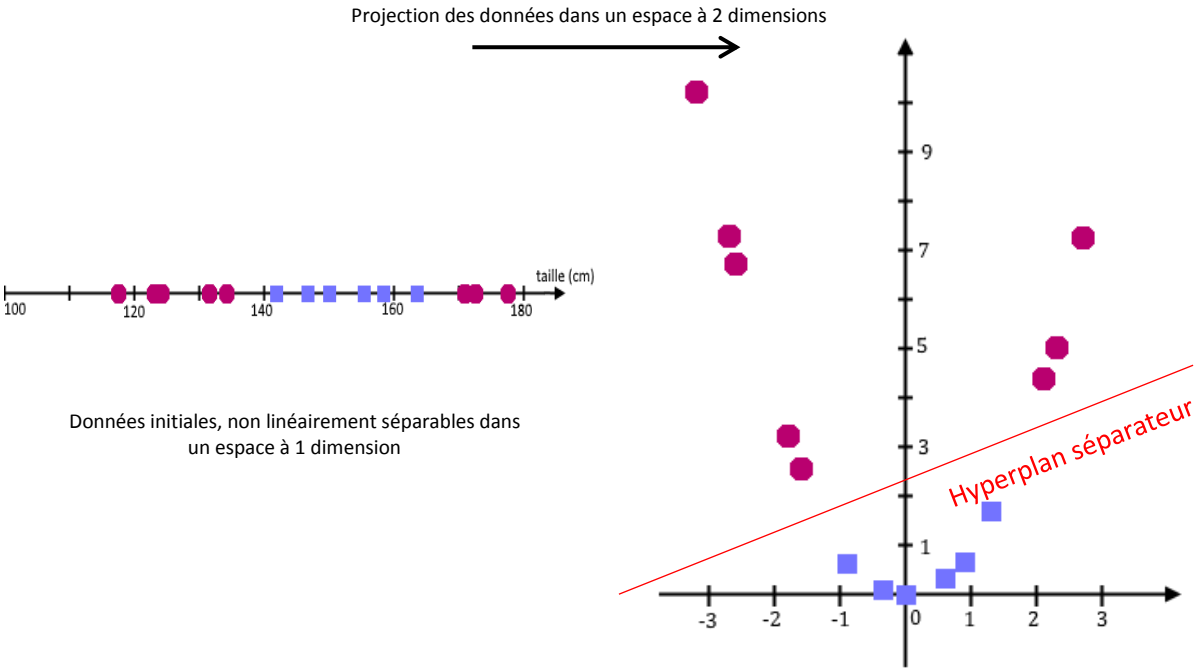


Figure 31 : Illustration de la fonction noyau, méthode SVM

Sur la Figure 31, à gauche est représentée en violet la taille (en cm) des enfants entre 12 et 16 ans et en bleu celles des autres personnes. Les données ne sont pas linéairement séparables. Lorsqu'on projette les données dans un espace à deux dimensions grâce à la fonction $\boldsymbol{\varphi \mapsto X}\left(\left(\frac{x-150}{10}\right), \left(\frac{x-150}{10}\right)^2\right)$, on observe que les données peuvent être séparées par la droite rouge (hyperplan).

3.3 Classification en classes multiples et en étiquettes multiples

Nous avons défini près de 100 RS dans lesquels les entités nosologiques doivent être classées.

Un champ de certificat de décès pouvant contenir plusieurs entités nosologiques, nous pouvons donc lui attribuer plusieurs classes. Nous devons alors utiliser un classifieur multi-étiquettes.

La stratégie la plus couramment utilisée pour traiter ce type de classification est nommée « un contre tous » (« one versus all » strategy) (Figure 32). Elle consiste à ajuster un classifieur par classe. Pour chaque classifieur, une classe est opposée à toutes les autres.

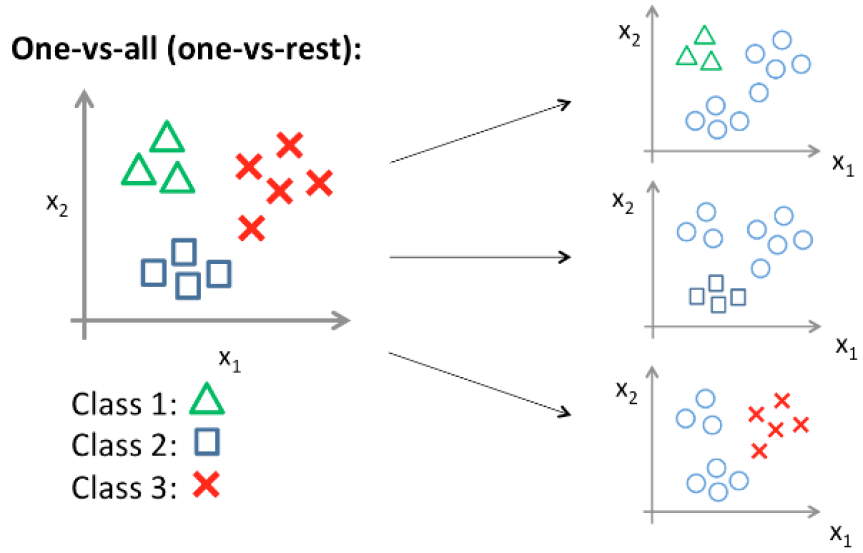


Figure 32: Schématisation de la stratégie « un contre tous » dans la méthode SVM

3.4 Caractéristiques d'apprentissage

Les caractéristiques d'apprentissage sont les éléments qui permettent de décrire les objets à classer (les entités nosologiques) et de les distinguer les uns des autres. Cette phase de description des objets en utilisant différentes caractéristiques est la phase de transformation vectorielle. Elle est effectuée sur l'échantillon d'apprentissage de 2000 certificats (cf phase d'apprentissage figure 27).

L'utilisation de caractéristiques d'apprentissage variées permet de tester des modèles utilisant des combinaisons différentes de caractéristiques d'apprentissage. Deux types de caractéristiques ont été testés : les caractéristiques de surface et les RS attribués par la méthode par règles.

3.4.1 Unigrammes et bigrammes de mots

Les premières caractéristiques identifiées sont les unigrammes et bigrammes de mots. Les unigrammes de mots sont tous les mots uniques rencontrés au moins une fois dans l'ensemble d'apprentissage et les bigrammes sont toutes les suites de deux mots consécutifs uniques rencontrées dans ce même ensemble d'apprentissage. Par exemple pour l'entité nosologique « cancer du sein droit », après l'étape de prétraitement elle devient : « cancer sein droit ». Les unigrammes de mots sont : [cancer], [sein], [droit] et les bigrammes de mots sont : [cancer, sein], [sein, droit].

Pour la phase d'apprentissage, afin d'établir la liste des unigrammes et bigrammes de mots, l'ensemble des données d'apprentissage a été découpé en mots. Pour cela, nous avons défini les délimiteurs d'un mot comme étant les espaces ou les caractères alphanumériques. Cette étape est

appelée la « tokénisation ». La liste est alors constituée de l'ensemble des unigrammes et bigrammes de mots uniques contenus dans l'ensemble des champs de l'échantillon d'apprentissage.
Pour chaque champ des certificats de décès (vecteur d'apprentissage), chaque unigramme et bigramme de mots de la liste est considéré comme une dimension du vecteur d'apprentissage. Pour chaque champ du certificat, on indique dans le vecteur d'apprentissage, le nombre de fois où l'unigramme et bigramme de mots de la liste apparait dans le champ.

Un exemple de vecteur de caractéristiques en utilisant les unigrammes et bigrammes de mots est présenté en Annexe 5.

3.4.2 Trigramme de caractères

La seconde caractéristique identifiée est le trigramme de caractères, c'est-à-dire toutes les suites de trois caractères consécutifs uniques retrouvées dans l'ensemble d'apprentissage. Par exemple pour l'entité nosologique « cancer », l'ensemble des trigrammes de caractères sera : [c,a,n], [a,n,c], [n,c,e], [c,e,r].

Cette caractéristique est moins sensible aux mots mal orthographiés que les unigrammes et bigrammes de mots. En effet, les termes « maladie » et « malaidie » vont avoir plusieurs trigrammes de caractères en commun et seront proches dans le plan, alors que ce seront deux unigrammes de mots différents donc deux dimensions vectorielles différentes. Un exemple de vecteur de caractéristiques en utilisant les trigrammes de caractères est présenté en Annexe 6.

3.4.3 Caractéristique fondées sur les résultats de la méthode par règles

La troisième caractéristique d'apprentissage que nous avons envisagée est l'utilisation des RS attribués par la méthode par règles. Le résultat de l'attribution des RS par la méthode par règles pour chaque entité nosologique est ajouté au vecteur d'apprentissage. Un exemple de vecteur d'apprentissage utilisant les RS attribués par la méthode par règles est présenté en Annexe 7.

3.5 Combinaison des caractéristiques d'apprentissage

Plusieurs combinaisons des caractéristiques d'apprentissage ont été envisagées pour être testées et déterminer laquelle permettrait d'obtenir les meilleures performances de classification :

- Unigramme et bigramme de mots,
- Trigramme de caractères,
- Union des unigrammes et bigrammes de mots et trigrammes de caractères,

- Union des unigrammes et bigrammes de mots, des trigrammes de caractères et des RS attribués par la méthode par règles.

Un modèle a été testé pour chaque combinaison.

3.6 Mise en œuvre de la méthode

Nous avons construit des modèles SVM avec noyau linéaire pour classer chaque champ de chaque certificat de décès, en utilisant les données d'apprentissage pour configurer les paramètres. Les réglages par défaut des hyper paramètres ont été sélectionnés (paramètre C égal à 1). Puisqu'il est possible d'associer 0, 1 ou plusieurs RS à chaque champ d'un certificat de décès, nous avons effectué une classification multi-étiquettes en utilisant la stratégie un contre tous (one vs rest).

IV/ Evaluation des performances des méthodes sur sept regroupements syndromiques

La méthode par règles permet d'attribuer des RS à l'ensemble des entités nosologiques d'un champ de certificat. Un même RS peut donc être attribué plusieurs fois pour un même champ de certificat. En revanche, la méthode SVM fournit pour chaque champ de certificat une liste de RS sans tenir compte de l'ordre, ni de la fréquence d'apparition dans le champ.
Lors de l'annotation, , un RS était attribué à chaque entité nosologique dans l'ordre de lecture du champ. Un même RS pouvait alors être attribué plusieurs fois pour un même champ de certificat.

Pour évaluer les performances de classification de la méthode par règles et de la méthode SVM lors des phases de développement et d'évaluation interne, nous avons considéré :

- Pour la méthode par règles, le nombre de RS attribués à un champ de certificat décès en prenant en compte leur ordre d'attribution,
- Pour la méthode SVM, le nombre de RS unique prédits pour un champ de certificat quel que soit l'ordre.

Lors de la phase d'évaluation finale, afin de comparer les performances des deux méthodes, nous avons appliqué le même critère d'évaluation pour les deux méthodes : le nombre de RS unique attribués à un champ de certificat de décès quel que soit l'ordre.

Les performances ont été mesurées en comparant le classement effectué par chaque méthode au classement effectué par les annotateurs lors de la phase d'annotation (référence).

1) Les mesures de performance

Pour chaque regroupement syndromique i, nous avons construit un tableau croisé (Tableau 11).

Tableau 11 : Tableau de contingence

RS i	**Référence (Annotation manuelle) 1 (RS i attribué)**	**Référence (Annotation manuelle) 0 (RS i non attribué)**
Système (Méthode de classification) 1 (RS i attribué)	VP	FP
Système (Méthode de classification) 0 (RS i non attribué)	FN	VN

VP= Vrai Positif, VN= Vrai Négatif, FP= Faux Positif, FN= Faux Négatif

En gardant l'exemple de la ***Grippe,*** le nombre de vrais positifs correspond au nombre de RS ***Grippe*** dans l'échantillon d'évaluation attribués à la fois par la référence (l'annotation manuelle) et par la méthode évaluée. Le nombre de faux positifs est le nombre RS ***Grippe*** attribués par la méthode mais pas par la référence. Le nombre de faux négatifs est le nombre de RS ***Grippe*** attribués par la référence mais non attribués par la méthode évaluée. Le nombre de vrais négatifs est le nombre de RS ***Grippe*** non attribué par la référence et par la méthode évaluée.

Trois mesures sont traditionnellement utilisées pour mesurer les performances de classification des méthodes : la Précision, le Rappel, et la F-mesure.

Ces mesures sont calculées pour chaque RS i à évaluer :

$$Précision_i\ (ou\ Valeur\ Prédictive\ Positive) = \frac{nb\ de\ champs\ de\ certificats\ de\ décès\ correctement\ attribués\ au\ RS\ i\ (VP)}{nb\ de\ champs\ de\ certificats\ de\ décès\ attribués\ par\ la\ méthode\ au\ RS\ i\ (VP+FP)}$$

$$Rappel_i(ou\ Sensibilité) = \frac{nb\ de\ champs\ de\ certificats\ de\ décès\ correctement\ attribué\ au\ RS\ i\ (VP)}{nb\ de\ champs\ de\ certificats\ de\ décès\ attribués\ par\ la\ référence\ au\ RS\ i\ (VP+FN)}$$

$$F-mesure_i = 2.\frac{précision_i\ .\ rappel_i}{précision_i + rappel_i}$$

Puis une mesure globale pour la méthode, regroupant les n RS évalués, peut être calculée comme suit :

$$Précision = \frac{\sum_{i=1}^{n} précision_i}{n}$$

$$Rappel = \frac{\sum_{i=1}^{n} rappel_i}{n}$$

$$F-mesure = \frac{\sum_{i=1}^{n} F-mesure_i}{n}$$

Une précision élevée signifie que les regroupements syndromiques prédits par la méthode sont corrects. Un rappel élevé signifie que les regroupements syndromiques à identifier l'ont été

correctement par le système. La F-mesure est une moyenne pondérée des deux mesures d'évaluation précédentes.

Nous avons considéré que les niveaux de performances élevés correspondaient à un rappel et une précision supérieure à 0,95.

La comparaison des performances des méthodes entre elles sur l'échantillon d'évaluation final de 1000 certificats de 2016 a été effectuée à l'aide d'un Z-test. Les différences de performances ont été calculées pour les méthodes deux à deux.
L'ensemble des méthodes a été programmé par Alix Bourrée (Stagiaire de M2 Linguistique Informatique, Université Paris Diderot) en langage python version 3.0 et avec l'aide de la librairie Scikit learn pour le SVM.

2) Performances des méthodes

2.1 Echantillon de développement de 1000 certificats de décès électroniques de 2012 à 2014

2.1.1 La méthode par règles

Chacune des trois premières étapes de la méthode par règles, prise séparément, puis sous la forme de combinaisons a été évaluée sur l'échantillon de développement de 1000 certificats.

Le regroupement syndromique ***Maladies chroniques endocriniennes*** (MCE) a été pris comme exemple pour commenter l'évolution des résultats. Les performances pour les sept RS sont présentées le tableau 12.

- **Sans traitement :**

L'évaluation après le prétraitement, c'est-à-dire sans aucune étape de la méthode par règles, indiquait une précision de 0,98, un rappel de 0,32 et une F-mesure de 0,48.

- **Application des règles de standardisation uniquement (Etape 1) :**

L'utilisation des règles de standardisation permettait d'obtenir une précision de 0,98, un rappel de 0,33 et une F-mesure de 0,50.

- **Prise en compte des séparateurs uniquement (Etape 2) :**

La segmentation des textes en utilisant les séparateurs permettait d'obtenir une précision de 0,99, un rappel de 0,72 et la F-mesure était de 0,83. Ces performances mettent en lumière l'apport majeur des séparateurs pour la classification des causes dans les RS.

- **Application du correcteur orthographique uniquement (Etape 3) :**

La correction des mots mal orthographiés permettait d'obtenir une précision de 0,98 et un rappel de 0,33 la F-mesure obtenue était de 0,50.

- **Combinaison des règles de standardisation + des séparateurs (Etapes 1 et 2) :**

La combinaison des séparateurs et des règles de standardisation permettait d'obtenir une précision de 0,99 et un rappel de 0,75.

- **Combinaison des séparateurs + correcteur orthographique (Etapes 2 et 3) :**

La combinaison des séparateurs et du correcteur orthographique permettait d'obtenir des performances similaires à la combinaison des règles de standardisation et des séparateurs.

- **Combinaison des règles de standardisation + correcteur orthographique (Etapes 1 et 3) :**

En revanche, la combinaison des règles de standardisation et du correcteur orthographique aboutit à un rappel plus faible (0,35) et une précision de 0,98.

- **Combinaison des règles de standardisation + séparateurs + correcteur orthographique (Etapes 1 à 3)**

La combinaison des 3 étapes permettait d'obtenir une précision de 0,99 et un rappel de 0,81. La F-mesure était de 0,89.

Les évolutions des performances pour chaque étape et les combinaisons d'étapes décrites pour le regroupement « Maladies chroniques endocriniennes » étaient également observées pour les 6 autres regroupements syndromiques (Tableau 12). Pour ces 6 autres regroupements, le rappel était toutefois nettement supérieur avant l'application des étapes de la méthode par règles (rappel sans traitement compris entre 0,44 et 1,00, vs. 0,32 pour le regroupement MCE).

Cette première analyse montrait l'importance de la segmentation des textes à l'aide des séparateurs pour identifier les différentes entités nosologiques sur un même champ de certificat. Les règles de standardisation et le correcteur orthographique permettaient également d'augmenter les performances de façon modérée.

La progression des performances après la segmentation des textes à l'aide des séparateurs était liée à la présence de plusieurs entités nosologiques séparées par des séparateurs dans un champ et cela pour un grand nombre de champs.

A partir de la combinaison des 3 étapes, la F-mesure était supérieure à 0,90 à l'exception des RS « Maladies chroniques endocriniennes » et « Maladies chroniques de l'appareil digestif », F-mesure

étant égale à 0,89 et 0,82 respectivement. L'analyse d'erreurs a montré que les champs pour lesquels il manquait des prédictions étaient des champs dans lesquels les entités nosologiques n'étaient pas délimitées par des séparateurs. Ce constat nous a mené à réfléchir à l'ajout de la quatrième étape de traitement pour augmenter les performances de la méthode pour la classification de ces entités nosologiques.

Tableau 12 : Evaluation des performances de classification de la méthode par règles selon différentes combinaisons d'étapes à partir de l'échantillon de développement de 1000 certificats électroniques de 2012 à 2014, France

	Ensemble des 7 RS	Grippe	IRA[1]	IRAB[2]	AAR[3]	Sepsis	MCE[4]	MCAD[5]
Nb dans l'échantillon	648	4	79	104	27	123	156	155
Sans traitement :								
Précision	0,99	1,00	1,00	1,00	1,00	1,00	0,98	0,99
Rappel	0,67	1,00	0,85	0,63	0,78	0,68	0,32	0,45
F-mesure	0,78	1,00	0,92	0,78	0,87	0,81	0,48	0,61
Règles de standardisation								
Précision	0,99	1,00	1,00	1,00	1,00	1,00	0,98	0,98
Rappel	0,69	1,00	0,87	0,71	0,78	0,70	0,33	0,47
F-mesure	0,80	1,00	0,93	0,83	0,87	0,82	0,50	0,63
Séparateurs								
Précision	0,99	1,00	1,00	0,99	1,00	1,00	0,99	0,99
Rappel	0,79	1,00	0,87	0,72	0,85	0,77	0,72	0,63
F-mesure	0,88	1,00	0,93	0,83	0,92	0,87	0,83	0,77
Correcteur orthographique								
Précision	0,99	1,00	1,00	1,00	1,00	1,00	0,98	0,99
Rappel	0,69	1,00	0,89	0,64	0,78	0,70	0,33	0,47
F-mesure	0,79	1,00	0,93	0,78	0,87	0,82	0,50	0,64
Règles de standardisation + séparateurs								
Précision	0,99	1,00	1,00	0,99	1,00	1,00	0,99	0,99
Rappel	0,82	1,00	0,90	0,82	0,85	0,79	0,75	0,67
F-mesure	0,90	1,00	0,95	0,89	0,92	0,88	0,85	0,80
Séparateurs + correcteur orthographique								
Précision	0,99	1,00	1,00	0,99	1,00	1,00	0,99	0,99
Rappel	0,75	1,00	0,91	0,75	0,44	0.7,6	0,78	0,64
F-mesure	0,84	1,00	0,95	0,85	0,62	0,86	0,87	0,78
Règles de standardisation + correcteur orthographique								
Précision	0,99	1,00	1,00	1,00	1,00	1,00	0,98	0,99
Rappel	0,71	1,00	0,90	0,72	0,78	0,72	0,35	0,50
F-mesure	0,81	1,00	0,95	0,84	0,87	0,84	0,52	0,66
Règles de standardisation + Séparateurs + correcteur orthographique								
Précision	0,99	1,00	1,00	0,99	1,00	1,00	0,99	0,99
Rappel	0,85	1,00	0,94	0,86	0,85	0,81	0,81	0,70
F-mesure	0,92	1,00	0,97	0,92	0,92	0,90	0,89	0,82

[1]Insuffisance respiratoire aigue
[2]Infections respiratoires aigües basses
[3]Asphyxie/ Anomalie de la respiration
[4]Maladies chroniques endocriniennes
[5]Maladies chroniques de l'appareil digestif

2.1.2 Méthode SVM

Différentes combinaisons de caractéristiques ont été évaluées pour la méthode SVM sur l'échantillon de développement de 1000 certificats.

Un modèle a été construit pour chaque combinaison de caractéristiques. L'apprentissage de chacun des modèles a été effectué sur l'échantillon d'entrainement de 2000 certificats. Les performances des modèles utilisant les différentes caractéristiques sont présentées dans tableau 13.

- **Les unigrammes et bigrammes de mots :**

Pour ce premier modèle, les unigrammes et bigrammes de mots permettaient d'obtenir des précisions élevées : 1,00 pour les regroupements « Grippe », « Insuffisance respiratoire aigüe » (IRA), « Asphyxie et anomalie de respiration « (AAR) et « Maladies chroniques endocriniennes » (MCE) et entre 0,94 et 0,98 pour les trois autres regroupements syndromiques. Le rappel était supérieur à 0,91 pour l'ensemble des regroupements syndromiques, sauf pour le RS « Maladies chroniques de l'appareil digestif » pour lequel le rappel était de 0,73.

- **Les trigrammes de caractères :**

Le modèle s'appuyant uniquement sur les trigrammes de caractères obtenait un rappel moyen de 0,92 et une précision moyenne de 0,93 pour l'ensemble des RS.

- **Combinaison des unigrammes et bigrammes de mots et des trigrammes de caractères :**

Le troisième modèle basé sur l'union des caractéristiques des deux modèles précédents aboutissait à des F-mesures supérieures à 0,93 pour l'ensemble des RS, sauf pour le regroupement « Maladies chroniques de l'appareil digestif » pour lequel la F-mesure était de 0,81.

- **Combinaison des unigrammes et bigrammes de mots et des trigrammes de caractères + les regroupements syndromiques attribués par la méthode par règles :**

L'ajout des regroupements syndromiques attribués par la méthode par règles au modèle combinant les caractéristiques de surface permettait d'obtenir des F-mesures comprises entre 0,97 et 1,00 sauf pour le RS « Maladies chroniques de l'appareil digestif » pour lequel la F-mesure était de 0,89.

Tableau 13 : Evaluation des performances de classification de la méthode par apprentissage automatique selon différentes combinaisons de caractéristiques à partir de l'échantillon de développement de 1000 certificats électroniques de 2012 à 2014, France

	Ensemble des 7 RS	Grippe	IRA[1]	IRAB[2]	AAR[3]	Sepsis	MCE[4]	MCAD[5]
Nb de RS dans l'échantillon	625	4	79	104	27	123	145	143
Unigrammes et Bigrammes de mots								
Précision	0,98	1,00	1,00	0,95	1,00	0,98	1,00	0,94
Rappel	0,91	1,00	0,91	0,91	0,93	0,95	0,92	0,73
F-mesure	0,94	1,00	0,95	0,93	0,96	0,97	0,96	0,82
Trigrammes de caractères								
Précision	0,93	1,00	0,96	0,91	0,89	0,97	0,98	0,80
Rappel	0,92	1,00	0,96	0,92	0,93	0,95	0,92	0,79
F-mesure	0,93	1,00	0,96	0,92	0,91	0,96	0,95	0,80
Unigrammes et Bigrammes de mots + Trigrammes de caractères								
Précision	0,96	1,00	0,99	0,93	0,96	0,99	0,99	0,85
Rappel	0,93	1,00	0,94	0,92	0,96	0,96	0,94	0,78
F-mesure	0,94	1,00	0,97	0,93	0,96	0,97	0,96	0,81
Unigrammes et Bigrammes de mots + Trigrammes de caractères + RS [6]								
Précision	0,98	1,00	0,99	0,97	1,00	99,16	0,99	0,93
Rappel	0,94	1,00	0,92	0,92	0,96	0,96	0,94	0,85
F-mesure	0,96	1,00	0,95	0,95	0,98	0,97	0,97	0,89

[1]Insuffisance respiratoire aigue
[2]Infections respiratoires aigües basses
[3]Asphyxie/ Anomalie de la respiration
[4]Maladies chroniques endocriniennes
[5]Maladies chroniques de l'appareil digestif
[6] Unigrammes et bigrammes de mots et trigrammes de caractères et regroupements syndromiques attribués par la méthode par règles.

Au total, un vecteur de caractéristiques ne prenant en compte que les unigrammes et bigrammes de mots permettait d'obtenir une précision élevée (0,98) et supérieure au rappel (0,94). Ceci indiquait que l'algorithme produisait peu d'erreurs de classification. En revanche, la présence de mots mal orthographiés et de dérivations des mots, qui n'étaient pas pris en compte par les unigrammes et les bigrammes de mots ne permettaient pas à l'algorithme d'associer un RS à une entité nosologique.

L'utilisation des trigrammes de caractères permettait d'obtenir un rappel généralement plus élevé qu'avec l'utilisation des unigrammes et des bigrammes de mots. Cette caractéristique de surface est en effet plus adaptée pour prendre en compte des termes avec et sans fautes d'orthographe car ils possèdent un grand nombre de trigrammes de caractères en commun.

L'utilisation, à la fois des unigrammes et bigrammes de mots, et des trigrammes de caractères permettent de prendre en compte ces différentes situations.

Enfin, la combinaison des caractéristiques de surface et des RS attribués par la méthode par règles faisait augmenter aussi bien la précision que le rappel par rapport aux modèles avec uniquement des caractéristiques de surface.

2.2 Echantillon d'évaluation interne de 500 certificats de décès électroniques de 2015

2.2.1 Méthode par règles

Les combinaisons des différentes étapes testées sur l'échantillon de développement ont été évaluées en ajoutant l'étape de projection de dictionnaire sur l'échantillon d'évaluation interne de 500 certificats de l'année 2015. Cette évaluation a permis de mesurer l'influence de cette quatrième étape de projection de dictionnaire sur les performances de la méthode.

Les combinaisons suivantes ont été testées :

- uniquement les séparateurs,
- les tables de standardisation et les séparateurs,
- le correcteur orthographique et les séparateurs,
- les tables de standardisation, les séparateurs et le correcteur orthographique
- les tables de standardisation, les séparateurs, le correcteur orthographique et la projection de dictionnaire

A l'image des résultats observés sur l'échantillon de développement, on observait une augmentation des performances de la méthode par règles à chacune des étapes, pour finalement obtenir une F-mesure comprise entre 0,94 et 1,00 pour chacun des RS avec la combinaison des quatre étapes (Tableau 14).

Le RS « Maladies chroniques endocriniennes » a été pris comme exemple pour commenter l'évolution des résultats.

- **Séparateurs (Etape 2):**

L'évaluation avec uniquement les séparateurs indiquait une précision de 1,00, un rappel de 0,67 et une F-mesure de 0,80.

- **Règles de standardisation + séparateurs (Etapes 1 et 2) :**

L'ajout des règles de standardisation faisait augmentait le rappel de 3 points et la précision restait de 1.

- **Séparateurs + correcteur orthographique (Etapes 2 et 3) :**

La combinaison des séparateurs et du correcteur orthographique n'entrainait pas d'évolution des performances pour ce RS par rapport à l'utilisation uniquement des séparateurs.

- **Règles de standardisation + séparateurs + correcteur orthographique (Etapes 1 à 3) :**

L'ajout du correcteur orthographique à la combinaison contenant les règles de standardisation et les séparateurs n'entrainait pas non plus de d'évolution des performances pour ce RS par rapport à la combinaison règles de standardisation et séparateurs.

- **Règles de standardisation + séparateurs + correcteur orthographique+ projection de dictionnaire (Etapes 1 à 4) :**

En revanche, l'ajout de la projection du dictionnaire à la combinaison précédente entrainait une nette augmentation des performances avec une précision de 1,00 et un rappel de 0,95.

Les évolutions des performances décrites pour le RS MCE étaient également observées pour les 6 autres RS. Toutefois, on peut noter que l'ajout du correcteur orthographique, qui n'entrainait pas d'évolution des performances pour le RS MCE, entraine une évolution notable des performances pour le RS AAR. Ces résultats s'expliquent par la présence de fautes d'orthographe dans 8 champs sur le mot «insuffisance », qui ont été corrigées lors de l'ajout du correcteur orthographique.
Aussi, la projection de dictionnaire entrainait une augmentation importante du rappel de plus de 10 points par rapport au rappel lors de la combinaison des trois étapes pour l'ensemble des RS à l'exception du RS IRA.

Tableau 14 : Evaluation des performances de classification la méthode par règles pour différentes combinaison d'étapes en utilisant l'échantillon d'évaluation interne de 500 certificats électroniques de 2015, France

	Ensemble des 7 RS	Grippe	IRA[1]	IRAB[2]	AAR[3]	Sepsis	MCE[4]	MCAD[5]
Nb dans l'échantillon	341	2	45	49	21	58	85	81
Séparateurs								
Précision	0,98	1,00	1,00	0,97	1,00	0,97	1,00	0,98
Rappel	0,74	1,00	0,78	0,75	0,71	0,68	0,67	0,59
F-mesure	0,84	1,00	0,87	0,85	0,83	0,80	0,80	0,73
Règles de standardisation + séparateurs								
Précision	0,98	1,00	1,00	0,95	1,00	0,97	1,00	0,98
Rappel	0,78	1,00	0,82	0,85	0,76	0,74	0,70	0,62
F-mesure	0,87	1,00	0,90	0,90	0,86	0,84	0,82	0,76
Séparateurs + correcteur orthographique								
Précision	0,99	1,00	1,00	0,97	1,00	0,97	1,00	0,98
Rappel	0,78	1,00	0,93	0,77	0,71	0,72	0,67	0,64
F-mesure	0,86	1,00	0,96	0,86	0,83	0,83	0,80	0,77
Règles standardisation + Séparateurs + correcteur orthographique								
Précision	0,98	1,00	1,00	0,95	1,00	0,97	1,00	0,98
Rappel	0,83	1,00	0,95	0,87	0,80	0,77	0,70	0,69
F-mesure	0,89	1,00	0,97	0,91	0,89	0,86	0,82	0,81
Règles de standardisation + Séparateurs + correcteur orthographique + projection de dictionnaire								
Précision	0,99	1,00	1,00	0,94	1,00	0,98	1,00	0,98
Rappel	0,96	1,00	0,97	0,96	1,00	0,94	0,95	0,90
F-mesure	0,97	1,00	0,98	0,95	1,00	0,96	0,97	0,94

[1]Insuffisance respiratoire aigue
[2]Infections respiratoires aigües basses
[3]Asphyxie/ Anomalie de la respiration
[4]Maladies chroniques endocriniennes
[5]Maladies chroniques de l'appareil digestif

2.2.2 Méthode SVM

Suite à l'expérimentation des combinaisons de caractéristiques avec la méthode SVM sur l'échantillon de développement de 1000 certificats, nous avons choisi de ne retenir que les deux modèles SVM suivant :

- **Modèle SVM 1** basé sur la combinaison de caractéristiques de surface : Unigrammes et bigrammes de mots + trigrammes de caractères
- **Modèle SVM 2** basé sur la combinaison de caractéristiques de surface (Unigrammes et bigrammes de mots + trigrammes de caractères) et des RS attribués par la méthode par règles.

Ces modèles ont été entrainés à partir d'un échantillon de 3000 certificats combinant l'échantillon d'entrainement (2000 certificats) et l'échantillon de développement (1000 certificats). L'évaluation a été effectuée sur l'échantillon d'évaluation interne de 500 certificats.

Le modèle SVM 1 obtenait des performances comprises entre 0,91 et 1,00 en termes de F-mesure pour les 7 RS. La précision était supérieure à 0,94 pour l'ensemble des RS sauf pour IRAB (P=0,89) (Tableau 15). Ce modèle permettait d'obtenir un rappel compris entre 0,93 et 1,00, sauf pour le regroupement MCAD pour lequel le rappel était de 0,89.

Le modèle SVM 2 obtenait des F-mesures et des rappels supérieurs à 0,95 pour l'ensemble des RS sauf pour MCAD (F-mesure = 0,93) (Tableau 15). La précision du modèle était supérieure à 0,95 pour l'ensemble des RS sauf pour IRAB. La F-mesure moyenne du modèle SVM2 pour l'ensemble des 7 RS était de 0,96, elle était de 0,96 pour le modèle SVM 1.

Tableau 15 : Performances de classification des deux modèles SVM en utilisant l'échantillon d'évaluation interne de 500 certificats électroniques de 2015, France

	Ensemble des 7 RS	**Grippe**	**IRA[1]**	**IRAB[2]**	**AAR[3]**	**Sepsis**	**MCE[4]**	**MCAD[5]**
Nb de RS dans l'échantillon	316	2	45	49	20	55	73	72
Modèle SVM 1:								
Précision	0,96	1,00	0,98	0,89	0,95	0,96	0,99	0,94
Rappel	0,96	1,00	1,00	0,96	0,95	0,96	0,93	0,89
F-mesure	0,96	1,00	0,99	0,92	0,95	0,96	0,96	0,91
Modèle SVM 2								
Précision	0,97	1,00	1,00	0,94	0,95	0,96	1,00	0,96
Rappel	0,96	1,00	0,97	0,96	0,95	0,96	0,97	0,92
F-mesure	0,96	1,00	0,98	0,95	0,95	0,96	0,98	0,93

[1] **Insuffisance respiratoire aigue**
[2] **Infections respiratoires aigües basses**
[3] **Asphyxie/ Anomalie de la respiration**
[4] Maladie chroniques endocriniennes
[5] Maladies chroniques de l'appareil digestif

L'analyse des erreurs a montré que l'ajout des RS attribués par la méthode par règles en caractéristiques pouvait entrainer de la confusion et des erreurs de classification pour certains RS. En effet, le(s) RS ajouté(s) dans les caractéristiques peuvent ne concerner qu'une seule entité nosologique contenue dans un champ qui en contient plusieurs. Or cette cause de décès n'a pas été nécessairement associée de façon unique à ce RS dans un autre champ (qui ne contenait que cette entité), permettant un apprentissage précis de cette association. Dès lors, l'ajout des regroupements syndromiques d'une cause de décès au sein d'un champ contenant plusieurs causes de décès peut générer de la confusion, et notamment pour les regroupements syndromiques ayant déjà un rappel plus faible dans la méthode par règles. Ainsi, quand le système ne trouve pas de solution avec la méthode par règles, le modèle SVM 2 a une probabilité plus faible de trouver une bonne prédiction avec la méthode par apprentissage, intégrant ces résultats en caractéristiques.

2.3 Echantillon d'évaluation finale de 1000 certificats électroniques de 2016

2.3.1 Evaluation des performances des méthodes

L'évaluation finale de la méthode par règles et de la méthode SVM a été effectuée sur un échantillon de 1000 certificats de décès électroniques de 2016. Les modèles SVM ont été entrainés à partir d'un échantillon d'apprentissage de 3500 certificats (combinaison de l'échantillon d'entrainement, de développement et d'évaluation interne).

Les performances de la méthode par règles ont été évaluées en prenant en compte les 4 étapes de traitement.

La méthode à base de règles obtenait une F-mesure supérieure ou égale à 0,96 pour tous les RS. La précision était supérieure au rappel pour l'ensemble des RS (Tableau 16).

La F-mesure du modèle SVM 1 était supérieure ou égale à 0,95 pour 5 des 7 RS (grippe, IRA, IRAB, sepsis et MCE) (Tableau 16). Leur précision et leur rappel étaient supérieurs ou égaux à 0,95, sauf pour le RS MCE (rappel était égal à 0,94). Les F-mesures du modèle SVM 1 pour les deux autres RS (AAR et MCAD) étaient respectivement de 0,91 et. 0,88. La précision était supérieure au rappel, sauf pour le RS MCE, pour lequel le rappel était supérieur à la précision (Tableau 16).

Le modèle SVM 2 a montré des performances élevées, la F-mesure était supérieure ou égale à 0,98 pour l'ensemble des RS sauf pour les RS MCAD (F-mesure égale à 0,94) et AAR (F-mesure égale à 0,95). La précision et le rappel pour ces deux RS étaient respectivement de 0,94 et 0,94 pour le RS MCAD et de 1,00 et 0,91 pour le RS AAR.

En synthèse, les performances de la méthode par règles et du modèle SVM 2 étaient élevées, avec des F-mesures supérieures à 0,94. Le modèle SVM 1 obtenait des performances légèrement

inférieures aux deux autres méthodes, avec des F-mesures variant de 0,91 à 1,00 sauf pour les RS MCAD (0,88). Toutefois, aucune différence statistiquement significative n'a été observée entre les performances de la méthode par règles et celles du modèle SVM1 ou entre le modèle SVM 1 et le modèle SVM 2 ou la méthode par règles et le modèle SVM 2.

Tableau 16 : Evaluation des performances de la méthode par règles et des modèles SVM 1 et SVM 2 sur l'échantillon d'évaluation finale de 1000 certificats électroniques de 2016, France

	Ensemble des 7 RS	**Grippe**	**IRA[1]**	**IRAB[2]**	**AAR[3]**	**Sepsis**	**MCE[4]**	**MCAD[5]**
Nombre de RS dans l'échantillon	626	3	86	115	36	119	134	133
Méthode par règles (4 étapes)								
Précision	0,99	1,00	1,00	0,98	1,00	0,99	1,00	0,98
Rappel	0,97	1,00	0,97	0,95	0,97	0,95	0,99	0,95
F-mesure	0,98	1,00	0,98	0,96	0,98	0,97	0,99	0,96
Modèle SVM 1								
Précision	0,96	1,00	0,98	0,96	0,96	0,98	1,00	0,86
Rappel	0,94	1,00	0,96	0,95	0,86	0,96	0,94	0,90
F-mesure	0,95	1,00	0,97	0,95	0,91	0,97	0,97	0,88
Modèle SVM 2								
Précision	0,99	1,00	1,00	0,99	1,00	0,98	1,00	0,94
Rappel	0,97	1,00	0,96	0,99	0,91	0,98	0,97	0,94
F-mesure	0,98	1,00	0,98	0,99	0,95	0,98	0,98	0,94

[1]Insuffisance respiratoire aigue
[2]Infections respiratoires aigües basses
[3]Asphyxie/ Anomalie de la respiration
[4]Maladie chroniques endocriniennes
[5]Maladies chroniques de l'appareil digestif

2.3.2 Analyse d'erreurs

Pour mieux comprendre les performances des méthodes de classification, nous avons décrit plus en détail les erreurs de classification sur l'échantillon d'évaluation finale de 1000 certificats. Les cent dix-sept champs de certificats de décès qui contenaient au moins une entité mal classée ont été passés en revue : vingt-quatre champs concernaient la méthode par règles, soixante-quatre concernaient le modèle SVM 1 et vingt-neuf concernaient le modèle SVM 2. Parmi ces champs, certains ont été mal classés par plus d'une méthode.

Deux grandes catégories d'erreurs ont été identifiées : les erreurs de classification réelles (96,6%) et les erreurs liées à la référence (3,4%). Ces erreurs n'étaient pas spécifiques d'un RS. Parmi les erreurs de classification, nous avons observé (Tableau 17) :

- **Variations de mots ou fautes d'orthographe (15,4%)** : Variantes lexicales d'une entité nosologique ou variations orthographiques ayant entraîné une classification erronée ou l'absence de classification. Par exemple, « detreses respiratoire aigue » contenait une faute d'orthographe et l'expression « detresse respiratoire aigue » n'a pas pu être identifiée.
- **Combinaison de mots (19,7%)** : Des mots qui auraient dû être pris séparément mais qui ont été combinés pour constituer une seule entité nosologique avec un sens différent de celui que l'on voulait leur donner. Par exemple, « fausse route alimentaire asphyxiante » contenait deux entités nosologiques (« fausse route alimentaire », « asphyxiante ») qui auraient dû être classées dans deux RS différents « Autres causes externes de décès » et « Asphyxie et anomalie de la respiration ».
- **Confusion de classes (32,5%)** : L'entité a été détectée par le modèle, mais la mauvaise classe (RS) lui a été attribuée. Cette erreur était principalement due à la proximité de la définition de deux RS. Par exemple, « volvulus sigmoide » a été classé parmi les « Maladies aigües de l'appareil digestif » alors qu'il aurait dû être classé parmi les « Maladies chroniques de l'appareil digestif ».
- **Absence de l'entité dans le dictionnaire (10,3%)** : Ce type d'erreur est spécifique à la méthode par règles car elle repose sur le dictionnaire adapté du CépiDc. Si une entité nosologique était absente du dictionnaire, la méthode par règles ne pouvait pas affecter de RS à cette cause. Par exemple, « ischemie duodéno pancréatique » ou « sepsis point depart pulmonaire » étaient des entités nosologiques qui n'existaient pas dans le dictionnaire.
- **Absence de l'entité dans l'échantillon d'entrainement (18,8%)** : Cette erreur est spécifique aux modèles SVM. L'échantillon d'apprentissage était constitué de 3 500 certificats de décès annotés. Le nombre de certificats était insuffisant pour couvrir les multiples façons de rédiger

une cause de décès. Par exemple, il manquait « surpoids » dans l'échantillon d'apprentissage. Le classifieur n'a attribué aucun RS à cette entité.

- **Erreurs d'annotation (3,4 %)** : Il y avait des entités pour lesquelles l'annotation manuelle (référence) était incorrecte. Par exemple, « Bronchopneumopathie sur inhalation » aurait dû être classée par l'annotateur dans « Maladies respiratoires chroniques » mais a été classée dans "Infections respiratoires aiguës basses". Même si les méthodes ont correctement attribué le RS à l'entité nosologique, il n'en demeure pas moins qu'il s'agit d'une erreur de classification, en comparaison à la référence.

Tableau 17 : Répartition des erreurs de classification en nombre et en proportion selon les différentes catégories pour les 3 modèles – Echantillon d'évaluation finale de 1 000 certificats électroniques de 2016, France

	L'ensemble des modèles		**Méthode par règles**		**Modèle SVM 1**		**Modèle SVM 2**	
Catégories	N	%	N	%	N	%	N	%
Erreurs de classification	**113**	**96,6**	**22**	**18,0**	**63**	**53,8**	**28**	**23,9**
Variations de mots ou fautes d'orthographe	18	15,4	6	5,1	7	6,0	5	4,2
Combinaison de mots	23	19,7	0	0,0	18	15,5	5	4,2
Confusion de classes	38	32,5	4	3,4	23	19,7	11	9,4
Absence de l'entité dans le dictionnaire	12	10,3	12	10,3	0	0,0	0	0,0
Absence de l'entité dans l'échantillon d'apprentissage	22	18,8	0	0,0	15	12,8	7	6,0
Erreurs liées à la référence								
Erreurs d'annotation	4	3,4	2	1,7	1	0,9	1	0,8

Nous avons abouti à la mise en œuvre de 2 méthodes de classification que nous avons perfectionnées et évaluées pour la classification des causes de décès dans sept RS. L'ensemble de cette étude a fait l'objet d'un article accepté dans *l'International Journal of Medical Informatics*. Cependant, près de cent RS ont été définis pour nos objectifs de surveillance (Cf Chapitre III). Lors de chaque étape de construction et d'évaluation des méthodes, même si les résultats et l'analyse d'erreurs se concentraient sur les sept RS, les performances ont été calculées pour l'ensemble des RS. Nous avons donc cherché à décrire les performances de classification des méthodes sur d'autres RS, sur l'échantillon d'évaluation finale.

V/ Description des performances de la méthode par règles et du modèle SVM 2 pour la classification des causes de décès dans 60 regroupements syndromiques.

Les deux méthodes qui ont obtenu les meilleures performances pour la classification des causes de décès dans les 7 RS ont été retenues dans la suite de ce travail, c'est-à-dire la méthode par règles et le modèle SVM 2.

Dans un premier temps, nous avons décrit les performances des méthodes de classification des causes de décès dans les RS qui avaient été retrouvés (sur la base des causes codées en CIM-10) au moins trois fois et dans des proportions similaires dans les échantillons d'entrainement, de développement, d'évaluation interne et d'évaluation finale. Soixante RS répondaient à ces critères parmi les 93 RS non évalués dans l'étape précédente. Ces RS appartenaient à dix-huit thématiques différentes (Annexe 8).

Dans un deuxième temps, nous avons exploré comment les performances obtenues sur l'échantillon d'évaluation finale se reflétaient sur l'ensemble des certificats électroniques entre 2012 et 2016. Pour cela, nous avons utilisé la méthode par règles et le modèle SVM 2 pour classer l'ensemble des causes de décès certifiés par voie électronique entre 2012 et 2016 dans les 60 RS évalués ainsi que les 7 évalués précédemment.

1) Méthodes

1.1 Mesures d'évaluation des performances des deux modèles

Les mesures de Rappel, Précision et F-mesure ont été utilisées pour rapporter les performances des méthodes.

Afin de simplifier la description des résultats, nous avons défini trois groupes de RS en fonction des performances des méthodes obtenues sur l'échantillon d'évaluation finale :

- Groupe 1 : Les regroupements syndromiques pour lesquels les deux méthodes obtenaient des performances similaires. Nous avons défini des performances similaires comme les trois mesures d'évaluation (Rappel, Précision, F-mesure) appartenant simultanément au même niveau de performance. Les niveaux de performance de chaque mesure d'évaluation ont été définis selon quatre catégories : ≥0,95, [0,90-0,95 [, [0,85-0,90 [et<0,85.
- Groupe 2 : Les autres regroupements syndromiques pour lesquels des performances similaires étaient obtenues soit pour la méthode par règles, soit pour le modèle SVM 2.

- Groupe 3 : Les autres regroupements syndromiques pour lesquels les méthodes obtenaient des performances hétérogènes. Nous avons défini des performances hétérogènes comme la précision et le rappel appartenant à différents niveaux de performance.

Les niveaux de performance ont été déterminés en fonction des objectifs du système de surveillance de la mortalité. Un système de surveillance mis en œuvre pour la détection d'événements habituels ou inhabituels doit être fondé sur des méthodes avec à la fois un bon rappel et une bonne précision.

1.2 Comparaison de l'évolution du nombre mensuel de décès entre 2012 et 2016 en utilisant les 2 méthodes de classification

Nous avons classé l'ensemble des causes de décès issues des certificats électroniques entre 2012 et 2016 à l'aide de la méthode par règles (MPR) et du modèle SVM 2 afin de vérifier comment les performances mesurées sur l'échantillon d'évaluation finale se traduisaient sur l'ensemble des données. Le modèle SVM 2 a été entrainé à partir de l'ensemble des échantillons annotés (4500 certificats de décès). Toutes les causes de décès contenues dans les certificats ont été classées selon les deux méthodes.

Nous avons ensuite dénombré le nombre de certificats de décès par RS et par mois pour les deux méthodes (nb_E-certificat$_{MPR,mois}$ et nb_E-certificat$_{SVM\ 2,mois}$ pour chaque RS).

La comparaison des nombres mensuels de RS obtenus par les deux méthodes a été effectuée sur la base de :

1/ une analyse visuelle de l'évolution mensuelle de chaque regroupement syndromique. Nous avons défini 3 profils : a/ les RS pour lesquels les évolutions mensuelles en utilisant la méthode par règles et le modèle SVM 2 se recouvraient ; b/ les RS pour lesquels on observait une petite différence entre les deux courbes; c/ les RS pour lesquels on observait une grande différence entre les 2 courbes.

2/ la différence calculée (Δ_i) entre les nombres mensuels de certificats de décès entre les deux méthodes, exprimée en proportion pour le RS$_i$:

$$\Delta_i = \frac{\frac{\sum_{m1}^{m60}\left(\text{nb E-certificat}_{MPR,i,m}\text{-nb E-certificat}_{SVM\ 2,i,m}\right)}{60}}{\frac{nbEcertificat_{RPM,i} + nb\ Ecertificat_{SVM\ 2,i}}{2}}$$

où *m* correspondait à chaque mois de 2012 à 2016 (*m* = 60 mois au total) et *i* se rapportait à chaque RS analysé.

Nous avons défini des seuils de différence :

- Faible: Δ_i variant de -5% à 5%;
- Intermédiaire: Δ_i variant de [-10% à -5%] ou de [5% à 10%];
- Elevée: Δ_i inférieure à -10% ou supérieure +10%.

3/ le coefficient de corrélation calculé entre de nb_E-certificat$_{MPR,i}$ et nb_E-certificat$_{SVM\ 2,i}$.

Cette comparaison a d'abord été effectuée pour les sept RS évalués initialement, puis pour les seuls RS dont les performances (3 mesures) de l'une ou l'autre méthode étaient supérieures à 0,95.

1.3 Comparaison de l'évolution mensuelle des regroupements syndromiques à une source externe de données

Pour les regroupements syndromiques tels que "Grippe" ou "Infections respiratoires aiguës basses" qui devraient avoir un profil temporel similaire à celui des épidémies infectieuses hivernales (114), nous avons comparé visuellement les profils du nombre mensuel de certificats électroniques de décès pour ces 2 RS et le nombre mensuel de passages aux urgences pour ces pathologies. Pour ce faire, nous avons collecté dans le système SurSaUD le nombre mensuel de passages aux urgences avec un diagnostic clinique de "grippe" et d'"infections respiratoires aiguës basses" entre 2012 à 2016 (28).

2) Résultats

2.1 Description des performances de classification des deux méthodes dans 60 regroupements syndromiques à partir de l'échantillon d'évaluation finale de 1000 certificats électroniques de 2016, France

Parmi les 60 RS retenus, 22 appartenaient au groupe 1, 18 appartenaient au groupe 2 et 20 au groupe 3.

2.1.1 Regroupements syndromiques avec des performances similaires pour les deux méthodes (Groupe 1)

Parmi les 22 RS du groupe 1, les performances (3 mesures) des méthodes étaient simultanément supérieures à 0,95 pour 19 RS (Tableau 18).

Les RS appartenaient à neuf thématiques différentes (Figure 33) : 1-Maladies cardiaques et circulatoires (4 RS/11 RS évalués dans cette thématique), 2-Cancers/Tumeurs (1/1), 3-Maladies respiratoires (3/4), 4-Symptômes généraux (3/9), 5-Maladies infectieuses (1/6), 6-Maladies endocriniennes et nutritionnelles (1/2), 7-Maladies du système nerveux (4/4), 10-Maladies mentales et comportementales (1/3), et 15-Maladie du système digestif (1/1).

Deux RS ont obtenu des performances inférieures à 0,85 (« 5-Infections bactériennes », « 16-Autres causes de décès indéfinies et non spécifiées »).

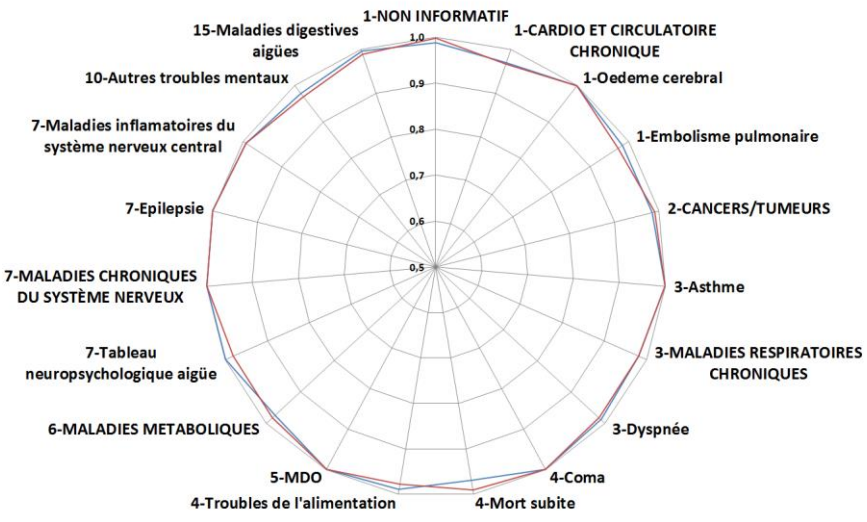


Figure 33 : F-mesure de la méthode par règles (bleu) et du modèle SVM 2 (rouge) pour les 19 regroupements syndromiques du groupe 1 avec un niveau de performance (3 mesures) supérieur à 0,95- échantillon d'évaluation finale de 1000 certificats électroniques de 2016, France

Tableau 18 : Distribution des regroupements syndromiques appartenant aux groupes 1 et 2 par niveau de performances, Echantillon d'évaluation finale, 1000 certificats, 2016-France

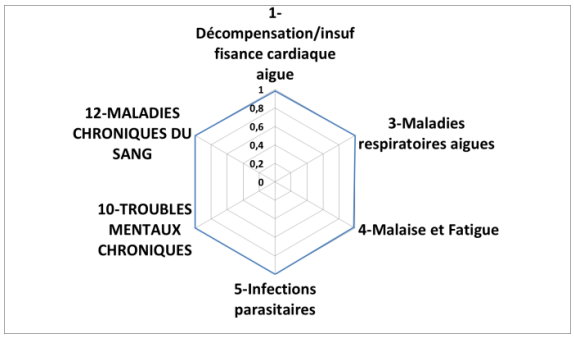
	Groupe 1	Groupe 2	
Niveau de performance*	2 méthodes	Méthode par règles	Modèle SVM 2
≥0,95	19	6	6
[0,90-0,95[	0	1	2
[0,85-0,90[	1	0	1
<0,85	2	3	1

*Rappel, précision et F-mesure appartenant simultanément au même niveau de performance

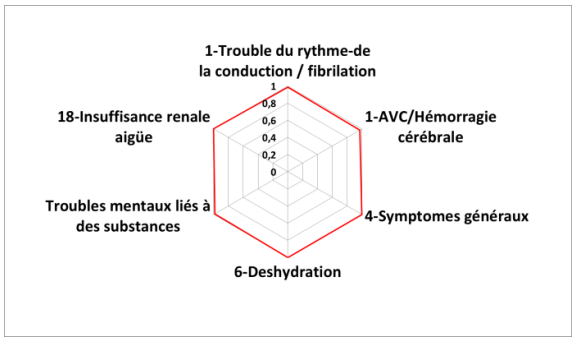
2.1.2 Regroupements syndromiques avec des performances similaires pour une des deux méthodes (Groupe 2)

Dix-huit regroupements syndromiques appartenaient au groupe 2. Des performances similaires (3 mesures) étaient observées uniquement avec la méthode par règles pour huit regroupements syndromiques. Des performances similaires étaient également observées uniquement avec le modèle SVM 2 pour huit autres RS. Des performances similaires appartenant à des niveaux différents ont été fournies par les deux méthodes pour 2 regroupements syndromiques (Tableau 18).

Les performances de la méthode par règles ou du modèle SVM 2 étaient supérieures à 0,95 pour la majorité des RS de ce groupe (12/18) (Figure 34). Les performances de l'une ou l'autre des méthodes étaient comprises entre [0,90 et 0,95[pour trois RS, tandis que des performances inférieures à 0,85 étaient mesurées pour quatre RS (Tableau 17).



Méthode par règles



Modèle SVM 2

Figure 34 : F-mesure de la méthode par règles (gauche) et du modèle SVM 2 (droite) pour les regroupements syndromiques avec un niveau de performance (3 mesures) supérieur à 0,95 pour l'une des deux méthodes (groupe 2), Echantillon d'évaluation finale de 1000 certificats. France, 2016

2.1.3 Regroupements syndromiques avec des performances hétérogènes (Groupe 3)

Le groupe 3 contenait vingt RS avec des niveaux hétérogènes de précision et de rappel entre les deux méthodes. Alors que la méthode par règles obtenait un rappel plus élevé que le modèle SVM 2 pour plus de la moitié des RS (11/20), la méthode par règles obtenait une précision soit supérieure soit inférieure à celle du modèle SVM 2 pour six RS et sept RS respectivement (Tableau 19). La précision des deux méthodes était similaire pour sept RS et le rappel était similaire pour trois RS.

Tableau 19 : Nombre de regroupements syndromiques du groupe 3 avec des niveaux de performance obtenus par la méthode par règles, inférieurs, supérieurs ou égaux aux niveaux de performance obtenus par le modèle SVM 2, Echantillon d'évaluation finale de 1000 certificats électroniques de 2016, France

	Précision	**Rappel**
Niveau de performance$_{MPR}$>Niveau de performance$_{SVM\ 2}$	6	11
Niveau de performance$_{MPR}$ <Niveau de performance$_{SVM\ 2}$	7	6
Niveau de performance$_{MPR}$ = Niveau de performance$_{SVM\ 2}$	7	3

2.2 Comparaison du nombre mensuel de certificats de décès classés dans les regroupements syndromiques par les deux méthodes

2.2.1 Les sept regroupements syndromiques évalués précédemment

Pour cinq des sept RS évalués précédemment, l'évolution du nombre mensuel de certificats de décès classés par la méthode par règles coïncidait avec l'évolution du nombre mensuel de certificats de décès classés par le modèle SVM 2 (Figure 35). On observait une faible différence entre les évolutions mensuelles pour les deux autres RS. Ils avaient tous de faibles différences Δ_i. (Tableau 20).

Tableau 20 : Nombre de regroupements syndromiques selon le niveau de différence Δ_i et l'analyse visuelle du nombre mensuel de certificat de décès classés par les 2 méthodes. France 2012-2016

		Analyse visuelle		
	Δ_i**	Dynamique similaire	Faible différence entre les dynamiques	Différence élevée entre les dynamiques
7 RS	Faible	5	2	0
19 RS du Groupe 1*	Faible	7	6	1
	Intermediaire	0	1	1
	Elevée	0	0	3
12 RS du groupe 2*	Faible	5	3	0
	Intermediaire	0	2	0
	Elevée	0	0	2

*RS avec des niveaux de performance supérieurs à 0,95.
** Différence Faible: Δ_i variant de -5% à 5%; Intermédiaire: Δ_i variant de [-10% à -5%] ou de [5% à 10%];
Elevée: Δ_i inférieure à -10% ou supérieure à+10%

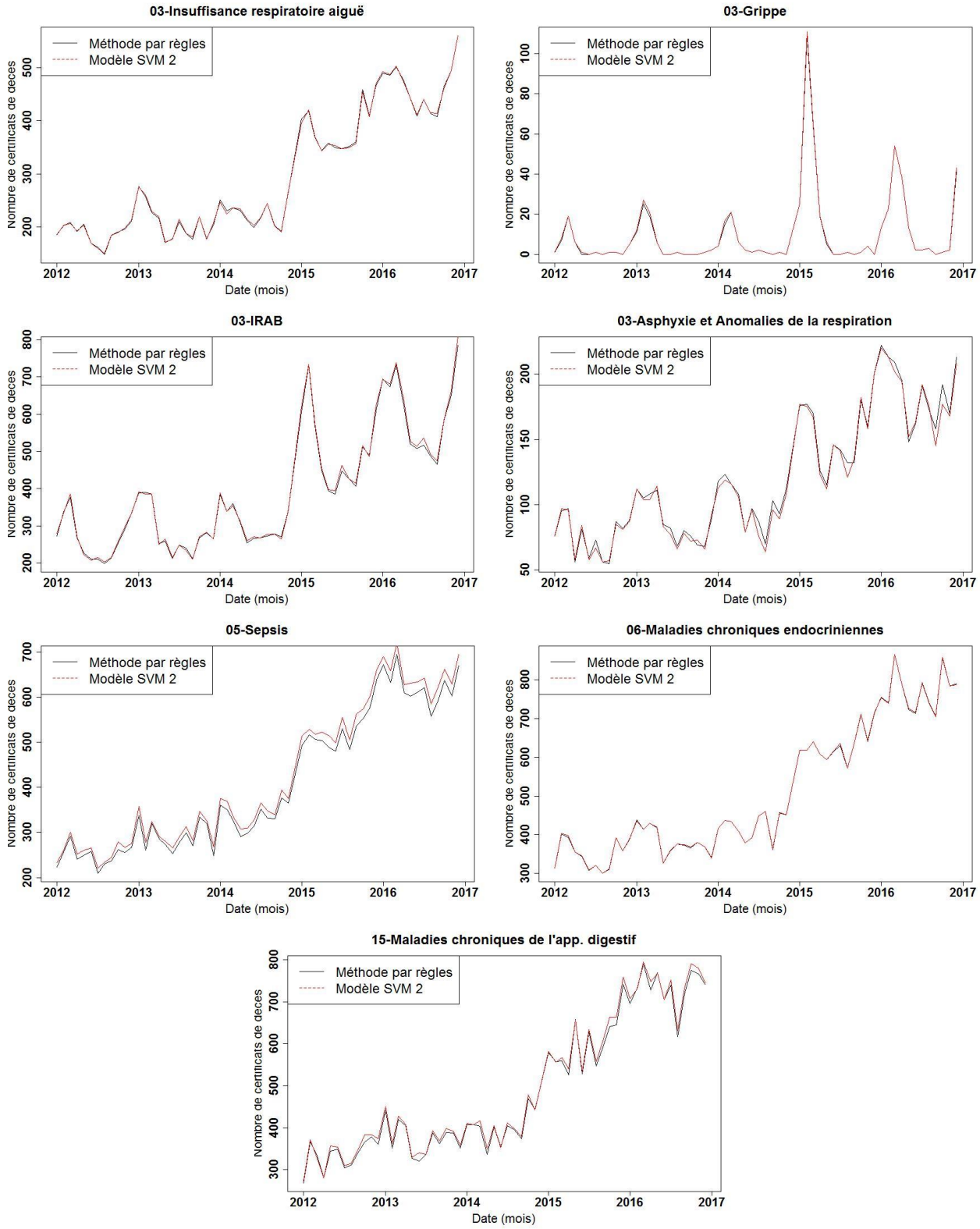


Figure 35 : Dynamique mensuelle de 2012 à 2016 des 7 regroupements syndromiques pour lesquels les performances de la méthode par règles et du modèle SVM 2 ont été initialement évaluées, ensemble des décès certifiés électroniquement entre 2012 et 2016, France

2.2.2 Regroupements syndromiques du Groupe 1 avec des performances supérieures à 0,95

Comme pour les RS évalués initialement, des dynamiques similaires ont également été observées pour 7/19 RS du Groupe 1 avec des performances supérieures à 0,95 dans l'échantillon d'évaluation finale (Tableau 20). La différence Δ_i de ces RS variait de -5,0 % à 5,0 % (Tableau 20) et leur coefficient de corrélation était très proche de 1.

Une faible différence visuelle entre les variations mensuelles des certificats de décès a également été observée pour 7/19 autres RS ayant des performances supérieures à 0,95 (Annexe 9). Les différences Δ_i variaient de -5,0 % à 5,0 %, à l'exception du RS " 7-tableau neuropsychologiques aigus " pour lequel la différence Δ_i était de 5,9 % (Annexe 9). Les coefficients de corrélation de ces RS étaient supérieurs à 0,99.

Une différence élevée entre les évolutions des nombres mensuels de décès utilisant les deux méthodes a été observée pour cinq autres RS qui atteignaient des performances supérieures à 0,95 pour les deux méthodes dans l'échantillon d'évaluation finale. La différence Δ_i de ces RS était pour la plupart inférieure à -10,0 % ou supérieure à 10,0 % (Tableau 20).

2.2.3 Regroupements syndromiques du Groupe 2 avec des performances supérieures à 0,95

Des évolutions superposées des nombres mensuels de décès fournis par les deux méthodes de classification étaient observées pour 5/12 RS du groupe 2 (Annexe 10) avec des performances de l'une ou l'autre des méthodes supérieures à 0,95 (Tableau 20). La différence Δ_I pour ces RS variait de -3,9% à 1,7% et les coefficients de corrélation étaient égaux à 1.

Cinq RS avaient également une faible différence entre les évolutions des nombres mensuels de décès des deux méthodes (tableau 20). La différence Δ_i était faible pour 3 RS avec un coefficient de corrélation supérieur à 0,99.

Une différence élevée des évolutions était observée pour deux RS ("10-Troubles mentaux chroniques" et "5-infections parasitaires"). Les différences Δ_i étaient de 12,5% et 24,3% et les coefficients de corrélation étaient de 0,98 et 0,96 respectivement.

2.3 Comparaison de l'évolution mensuelle des regroupements syndromiques à une source de données externe.

Les évolutions mensuelles du nombre de certificats de décès incluant un RS "Grippe" étaient très proches pour les deux méthodes. Comparativement à l'évolution du nombre mensuel de passages aux urgences pour grippe, la saisonnalité et les pics étaient concomitants, même si les amplitudes étaient différentes (Figure 36). Ces observations étaient également notées pour le RS "Infections respiratoires aiguës basses" (Figure 36).

Ces travaux, on fait l'objet d'un article accepté pour le congrès MedInfo.

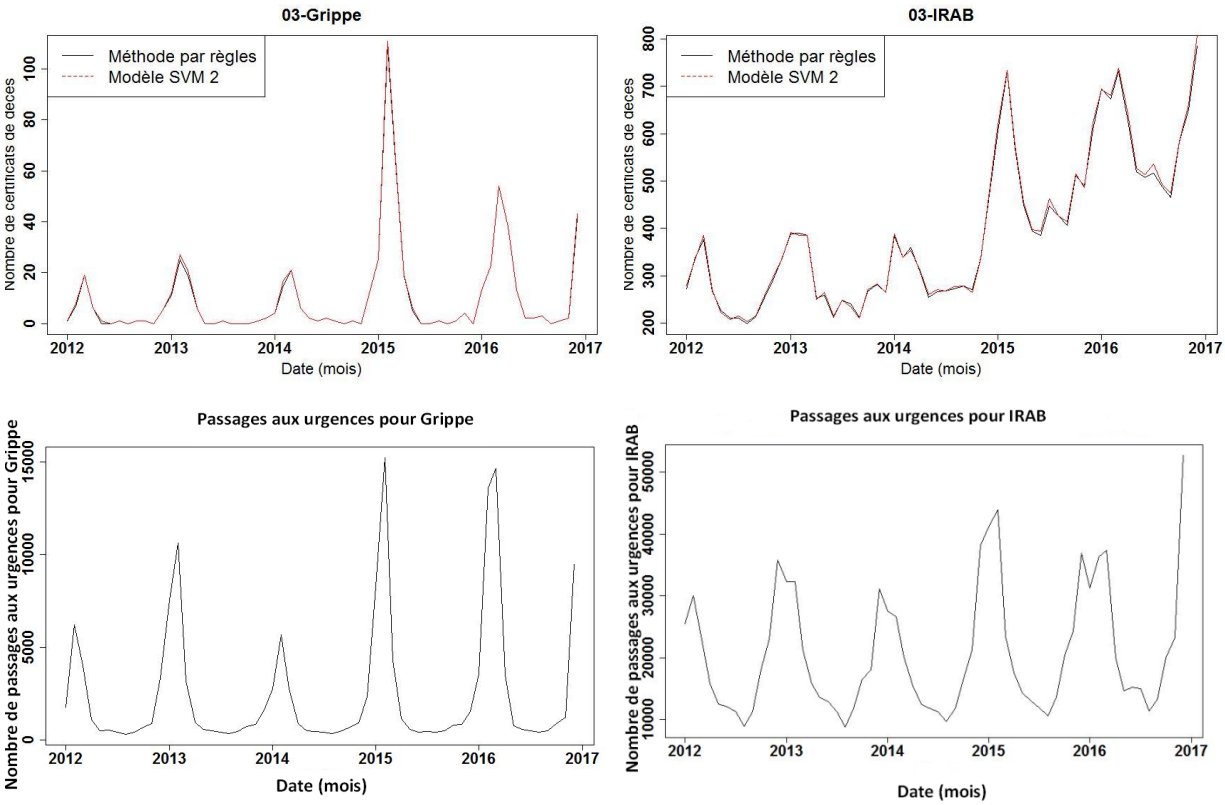


Figure 36 : Evolution du nombre mensuel de certificats de décès classés selon la méthode par règles et le modèle SVM 2 pour "Grippe" et "Infections respiratoires aiguës basses" (partie supérieure) et évolution du nombre mensuel de passages aux urgences pour les mêmes pathologies (Partie basse). France, 2012-2016 (Source des données : système SurSaUD (28))

VI/ Discussion

Les résultats présentés dans ce chapitre ont démontré que les performances de la méthode par règles et du modèle SVM 2 pour classer les causes de médicales de décès dans les sept regroupements syndromiques dédiés à la surveillance syndromique étaient élevées, avec des F-mesures supérieures à 0,94. La précision était supérieure au rappel pour les deux méthodes, témoignant de la fréquence supérieure des faux négatifs par rapport aux faux positifs. Le fait que la précision soit plus élevée que le rappel est lié à des champs des certificats de décès qui contenaient plusieurs entités nosologiques, dont certaines n'ont pas pu être détectées par le système. Le modèle SVM 1 obtenait des performances légèrement inférieures à celles des deux autres méthodes, la F-mesure variait de 0,91 à 1, sauf pour le RS « Maladies chroniques de l'appareil digestif » (0,88). Toutefois, aucune différence statistiquement significative n'a été observée entre les performances des méthodes prises deux à deux. La précision était également supérieure au rappel pour le modèle SVM 1.

Ces résultats concordent avec ceux de Koopman (48), Muscatello (115) et Shah (116). Cependant, il est important de noter que ces études portaient sur l'identification des pathologies spécifiques (pneumonie, grippe, VIH et diabète) alors que nos analyses concernaient des RS qui incluaient un plus large éventail de maladies (« Infections respiratoires aigües basses », « Maladies chroniques endocriniennes », « Maladies chroniques de l'appareil digestif »).

Les résultats ont mis en évidence que les méthodes obtenaient des performances de classification supérieures à 0,95 pour plus de la moitié des 60 regroupements syndromiques (19 RS du Groupe 1, 6 et 6 RS du Groupe 2) évalués sur l'échantillon d'évaluation finale de 1000 certificats de 2016.

Lorsque les méthodes de classification ont été appliquées à l'ensemble des décès survenus entre 2012 et 2016, nous avons observé que 13/19 RS du Groupe 1 et 8/12 RS du Groupe 2, ainsi que les 7 RS initialement évalués présentaient de faibles différences entre les nombres mensuels de certificats de décès obtenus par les deux méthodes, avec un coefficient de corrélation proche de 1.

Ces résultats suggèrent que la méthode par règles et le modèle SVM 2 initialement mis en œuvre, adaptés et évalués pour classer les causes médicales de décès dans sept RS, conviennent pour la classification des causes dans un plus grand nombre de RS. Cela est d'autant plus intéressant que parmi les 19 RS du groupe 1 et 18 du groupe 2, 21 RS étaient définis pour être suivis en routine dans un objectif d'alerte. Parmi ces 21 RS, 17 RS ont obtenus de bonnes performances et des corrélations élevées entre les deux méthodes.

Parmi les autres RS des groupes 1 et 2 avec pour lesquels les méthodes obtenaient des performances supérieures à 0,95, la majorité était associée à des différences intermédiaires ou élevées entre le

nombre mensuel de décès obtenus en utilisant les deux méthodes. La définition de ces RS contenait des maladies rares et/ou une grande variété de maladies. L'échantillon d'apprentissage comprenait un nombre insuffisant de certificats de décès contenant ces entités nosologiques pour saisir la variété de ces regroupements.

Pour la surveillance réactive de la mortalité en temps réel, la précision et le rappel sont importants. En effet, pour répondre à l'objectif de détection d'événements, un rappel élevé est nécessaire pour détecter la totalité des événements et une précision élevée est nécessaire pour limiter les fausses alarmes. La mesure d'impact est d'autant plus précise que le rappel et la précision sont élevés.

Choix des modèles pour la routine

Lors du choix d'un modèle, il serait simpliste de ne considérer que les performances globales comme seul critère. Dans le cas de la surveillance quotidienne de la mortalité, la flexibilité du modèle permettant la mise à jour en termes de nouvelles règles pour la méthode à base de règles, ou en termes de nouvelles données pour la phase d'apprentissage supervisé, est important. Comme les règles sont définies manuellement, mettre à jour les règles, revient à modifier le programme informatique. Les règles sont assez simples sur le plan du calcul et utilisent peu de ressources. Les modèles supervisés nécessitent une quantité suffisante de données d'entraînement annotées pour fonctionner correctement. Les méthodes d'apprentissage machine nécessitent l'extraction des caractéristiques des certificats de décès afin d'entrainer les modèles ; cette étape peut être coûteuse en calcul pour des ensembles de données plus importants (48, 50). Enfin, ces deux méthodes semblent complémentaires, ce qui favorise l'utilisation d'un modèle combiné.

En outre, l'utilisation de ces méthodes pour la surveillance quotidienne de la mortalité réactive implique également que le temps de fonctionnement de ces méthodes soit court. Nous avons mesuré les temps de fonctionnement des deux méthodes en simulant leur utilisation au quotidien : nous avons entrainé les modèles avec l'ensemble des certificats de décès annotés (4 500 certificats de décès) et prédit les RS d'un échantillon de 2000 certificats de décès approchant le nombre quotidien de décès observés en France (1500 décès par jours environ)
Pour la phase d'apprentissage, 2 minutes étaient nécessaires à l'entrainement du modèle SVM 1, tandis qu'il fallait 18 minutes pour le modèle SVM 2. Pour la classification des causes de décès de l'échantillon de 2000 certificats, le modèle SVM 1 prenait 30 secondes, tandis que le modèle SVM2 prenait 4 minutes 50 secondes pour prédire les RS. Enfin, la méthode fondée sur des règles prenait 4 minutes 40 secondes pour affecter les RS à l'échantillon de 2000 certificats. Ces temps d'exécution ont été mesurés dans les mêmes conditions que dans le reste de l'étude (128 Go de RAM, 32 cœurs)

et sont adaptés à une utilisation quotidienne, quelle que soit le modèle choisi. Mais on peut s'attendre à ce que le temps d'exécution soit plus long pour les modèles SVM avec un plus grand échantillon de certificats de décès pour l'apprentissage. Ces temps d'exécution restent acceptables pour la surveillance à visée d'alerte.

Perspectives

Nous nous sommes concentrés dans un premier temps sur l'évaluation des performances de classification des méthodes pour 7 RS parmi la centaine définis pour la surveillance réactive de la mortalité (117). Dans un second temps, nous avons décrit les performances des méthodes pour 60 autres RS. En dehors des performances élevées pour l'une ou l'autre des méthodes pour plus de la moitié d'entre eux, d'autres obtenaient des performances hétérogènes (groupe 3) plus complexes à analyser. Comme la mise en place d'un système de surveillance performant exige à la fois un rappel et une précision élevés, nous ne pouvons recommander l'analyse des RS du groupe 3, quelle que soit la méthode de classification. Le travail d'amélioration des performances des méthodes de classification pour ces RS doit être poursuivi pour permettre leur suivi en routine.

L'analyse des erreurs a montré que les erreurs n'étaient pas spécifiques des RS ciblés et qu'elles pouvaient se produire avec n'importe quel RS. Les erreurs restantes pourraient être dues à la complexité des langues humaines et de leur utilisation par les locuteurs, et à la difficulté technique de saisir cette complexité. Ces constats sont à rapprocher de la taille de l'échantillon d'apprentissage utilisé lors de l'évaluation finale qui était constitué de 3 500 certificats de décès annotés. Le nombre de certificats de décès pourrait être trop faible pour représenter avec précision les diverses expressions des causes de décès. L'ajout de certificats de décès annotés supplémentaires dans cet échantillon pourrait permettre de futures études visant à améliorer les résultats des modèles SVM.

Nous avons axé notre étude sur le développement d'une méthode à base de règles et d'une méthode d'apprentissage machine parce qu'elles répondaient à notre objectif et respectaient les exigences majeures : des méthodes faciles à utiliser pour une utilisation quotidienne, des méthodes facilement compréhensibles pour les épidémiologistes et des méthodes qui peuvent être régulièrement et rapidement actualisées. En outre, ces méthodes étaient connues pour leur haute performance. Des études plus récentes ont exploré la performance des méthodes d'apprentissage profond pour coder automatiquement les causes de décès dans la CIM-10 (49, 113). Les auteurs ont constaté que le modèle SVM surpassait certains des réseaux de neurones les plus simples, mais qu'un réseau neuronal combiné surpassait le modèle SVM (113). Ces méthodes pourraient constituer une

perspective intéressante pour comparer et améliorer notre performance en matière de classification et simplifier l'ingénierie manuelle des fonctions nécessaires au SVM.

Conclusion

Pour conclure, l'approche consistant à comparer les dynamiques temporelles des regroupements syndromiques, présentée dans ce chapitre, va au-delà de l'évaluation traditionnelle fondée sur un échantillon de test. Elle est surtout intéressante avec un grand nombre de catégories à évaluer ou lorsque les ressources annotées sont très limitées et ne saisissent pas la grande variété d'entités à classer. L'analyse de l'évolution temporelle du nombre de certificats de décès pour chaque regroupement syndromique (y compris la comparaison avec une source de données externe) permet de confirmer que ces regroupements syndromiques reflètent fidèlement la situation sanitaire ciblée que nous souhaitons suivre. Cette approche permet également de vérifier si les erreurs de classification de l'une des deux méthodes sont réparties de manière homogène dans le temps ou sont observées sur des périodes spécifiques. En effet, l'analyse en routine des causes médicales de décès en texte libre est essentielle pour améliorer le système actuel de surveillance de la mortalité, actuellement basé uniquement sur les informations administratives des bureaux d'état civil (82). Un tel système permettrait de fournir des informations spécifiques aux autorités de santé publique sur les causes de décès lors de crises de sanitaires (par exemple, les épidémies de dengue à La Réunion depuis septembre 2018) ou lors de périodes de surmortalité (par exemple, les épidémies d'hiver), ce qui permettrait aux décideurs d'adapter les messages de prévention et de mettre en place des mesures adaptées.

Pour ce faire, les chiffres communiqués lors de la mesure de l'impact d'un événement doivent refléter l'événement en cours. Or, un certificat de décès contient en moyenne 3,5 causes de décès qui peuvent appartenir à plusieurs regroupements syndromiques. Il est nécessaire que la somme des regroupements syndromiques soit égale au nombre de décès.

Afin de communiquer des informations interprétables pour les décideurs, nous devons appliquer une méthodologie d'analyse adaptée à ces causes multiples de décès.

Chapitre V : Prise en compte des causes multiples :

pondération des causes de décès

I/ Analyse en causes multiples de décès

Comme nous l'avons vu dans le chapitre II, un certificat de décès contient en moyenne 10 mots. Ces mots expriment des causes de décès appartenant à la séquence morbide ou peuvent exprimer des causes associées ou des comorbidités.

Traditionnellement, lors du processus de codage de décès mené par le CépiDc, la cause initiale est déterminée pour chaque décès. La cause initiale est définie comme « a) la maladie ou le traumatisme qui a déclenché l'évolution morbide conduisant directement au décès, ou b) les circonstances de l'accident ou de la violence qui ont entraîné le traumatisme mortel ». L'établissement de cette cause repose sur des règles spécifiques de codage définis par l'OMS. Or, dans la plupart des cas un décès est le résultat d'un processus de plusieurs causes et parfois de comorbidités.
Il a été démontré que l'analyse de la mortalité considérant uniquement la cause initiale conduisait à une sous-estimation de la contribution de certaines causes de décès telles que le diabète, des pathologies dermatologiques et hématologiques ou encore de maladies rénales (118, 119). En effet l'analyse sur la seule cause initiale ne permet pas d'intégrer la possibilité qu'un décès soit attribuable à plusieurs causes simultanément. L'analyse en prenant en compte les causes multiples permet de décrire de façon plus précise la présence conjointe ou consécutive de conditions contributives.

Dans notre cas, les causes sont classées directement dans les regroupements syndromiques et la cause initiale n'est pas disponible à l'étape d'analyse pour la surveillance réactive. Chaque décès étant susceptible de contribuer à un ou plusieurs regroupements syndromiques, deux options sont envisagées : 1) établir des règles pour choisir le regroupement le plus contributif du décès, 2) prendre en compte l'ensemble des regroupements attribués au certificat afin de considérer les différents facteurs contributifs du décès.
Pour l'objectif de détection, la prise en compte de l'ensemble des regroupements syndromiques permettra de ne pas passer à côté de l'augmentation d'un regroupement syndromique peu spécifique, voire symptomatique qui serait le reflet d'un événement inconnu. De plus, un événement habituel ou inhabituel peut avoir des conséquences différentes selon le type de population. Mesurer son impact sur la mortalité nécessite de prendre en compte l'ensemble des causes afin de distinguer un profil de décès d'une personne sur un terrain avec des comorbidités, d'un profil d'une autre personne sans risque apparent de comorbidités ou conditions contribuables au décès.

Cependant, la prise en compte de l'ensemble des regroupements syndromiques entraine un problème d'ordre statistique lors de la mesure d'impact. Afin d'obtenir une somme des décès pour les différents regroupements syndromiques égale au nombre total de décès toutes causes, il

convient de pondérer les causes du certificat afin que la somme des contributions des regroupements syndromiques soit égale à un.

Des pré-requis auxquels la stratégie devra répondre ont été établis :

- La pondération devra être applicable de la même manière quel que soit le type de décès (enfant ou adulte, décès à domicile ou sur la voie publique).
- La pondération devra être identique tout au long de l'année. En effet, la pondération ne devait pas être ajustée selon l'événement en cours afin que les variations de la mortalité soient comparables au cours du temps.

Nous avons choisi d'appliquer une pondération équilibrée à l'ensemble des regroupements syndromiques d'un certificat. L'ensemble des regroupements syndromiques d'un même certificat de décès aura un poids égal à :

$$\frac{1}{Nombre\ de\ RS\ du\ certificat}.$$

Ainsi, la contribution de chaque décès à un regroupement syndromique va dépendre du nombre de regroupements syndromiques attribués à ce décès.

La description de la répartition des regroupements syndromiques dans les certificats de décès de 2012 à 2016, pour différentes population, nous aidera à comprendre l'influence d'une telle pondération.

Compte tenu des limites des performances des méthodes pour la classification des causes de décès dans certains RS et l'absence d'évaluation pour un tiers des regroupements syndromiques définis, les résultats de cette description sont à interpréter avec prudence.

II/ Matériels et Méthodes

1) Les données

Afin de limiter les erreurs de classification liées à la taille de l'échantillon d'apprentissage, ce travail a été effectué sur les causes de décès des certificats électroniques des décès survenus entre 2012 et 2016 ont été classées par la méthode par règles.

2) Répartition moyenne des regroupements syndromiques par certificat

Pour chaque certificat de décès, le nombre de regroupements syndromiques uniques par champ a été comptabilisé. La répartition moyenne des 63 regroupements syndromiques définis pour l'alerte et des 35 regroupements syndromiques non définis pour l'alerte a été calculée :

- Par partie de certificat : Partie I et II,
- Par champ de certificat,
- Par classe d'âge : <1 an, 1-14 ans, 15-44 ans, 45-64 ans, 65-84 ans, ≥85 ans,
- Par type de lieu de décès : Domicile, Etablissement de santé public, Etablissement de santé privé, EHPAD/maison de retraite, Voie publique, Autre lieu indéterminé.

Pour chaque certificat, la répartition des RS est calculée comme la proportion de RS parmi le nombre total de RS du certificat.

Le nombre moyen de regroupements syndromiques par certificat a aussi été calculé pour les mêmes variables.

3) Fréquence des regroupements syndromiques

Le nombre de chaque RS parmi l'ensemble des RS attribués a été décrit par classe d'âges, par type de lieu de décès, puis par champ de certificat (champs 1 à 6) pour l'ensemble des RS et par thématique sur l'ensemble de la période 2012-2016.

Pour mémoire, près de 100 RS ont été définis et sont répartis dans 20 thématiques différentes (Chapitre III).

4) Variation mensuelle de la répartition moyenne des regroupements syndromiques

Les variations mensuelles de la proportion de RS définis pour l'alerte et non définis pour l'alerte ont été décrites pour l'ensemble de la période 2012-2016 dans ce travail.

III/ Résultats

1) Répartition des regroupements syndromiques par certificat

Selon le champ de certificat :

Les certificats des décès survenus entre 2012 et 2016, contenaient en moyenne 3,7 regroupements syndromiques (Tableau 21). Au sein de chaque certificat, 39% des RS attribués en moyenne étaient des RS définis pour l'alerte et 61% étaient des RS non définis pour l'alerte (Tableau 21). En partie I, en moyenne, 46% des RS étaient définis pour l'alerte, alors qu'en partie II, cette proportion était de 26%.

Tableau 21 : Nombre moyen de regroupements syndromiques par certificat et répartition moyenne des regroupements syndromiques définis pour l'alerte ou non définis pour l'alerte par partie de certificat ; Certificats électroniques de 2012 à 2016 classés par la méthode par règles, France

	Répartition moyenne des RS définis pour l'alerte par certificat (%)	Répartition moyenne des RS non définis pour l'alerte par certificat (%)	Nombre moyen de RS par certificat (N)
Certificat entier	39	61	3,7
Partie I	46	54	2,7
Partie II	26	74	1,0

Le nombre moyen de RS par champ était le plus élevé pour le champ 5 (1,9 RS en moyenne). Il était de 1,2 RS pour chacun des 4 champs de la partie I du certificat (Figure 37).

La proportion des RS définis pour l'alerte était supérieure à la proportion de RS non définis pour l'alerte uniquement pour le champ 1 (Figure 37).

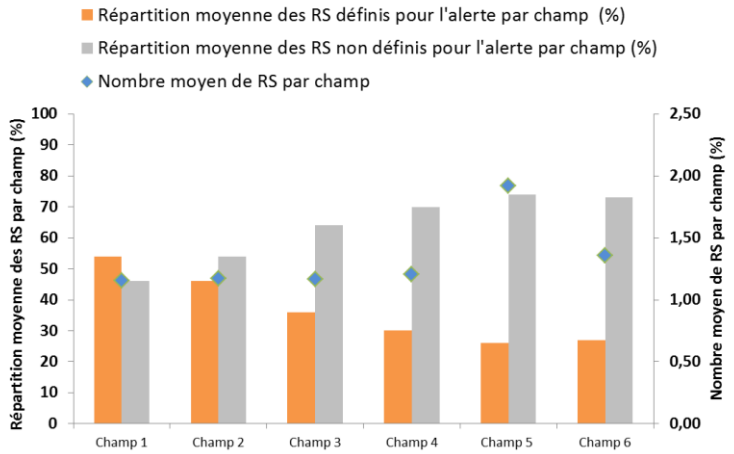


Figure 37 : Nombre moyen de regroupements syndromiques par champ et répartition moyenne de regroupements syndromiques définis pour l'alerte et non définis pour l'alerte par champ ; Certificats électroniques de 2012 à 2016 classés par la méthode par règles, France

<u>Selon la classe d'âge :</u>

Le nombre moyen de RS par certificat augmentait avec l'âge passant de 3,4 RS pour les moins de 1 an à 3,8 RS pour les personnes âgées de 85 ans et plus (Figure 38).

Les RS définis pour l'alerte étaient dans les mêmes proportions que les RS non définis pour l'alerte pour les classes d'âges jeunes (<1an, 1-14 ans et 15-44 ans) (Figure 39). Pour les classes d'âges plus

âgées (45-64 ans et 65-84 ans), les certificats comportaient en moyenne plus de 60% de RS non définis pour l'alerte alors qu'ils en contenaient 54% chez les personnes âgées de 85 ans plus.

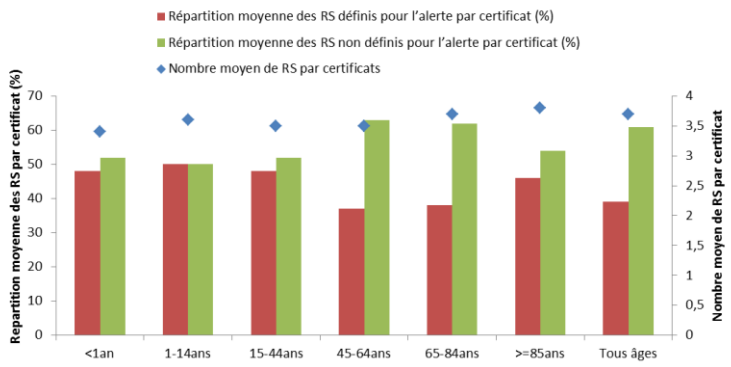


Figure 38 : Nombre moyen de regroupements syndromiques par certificat et répartition moyenne de regroupements syndromiques définis pour l'alerte et non définis pour l'alerte par certificat et par classe d'âge ; Certificats électroniques de 2012 à 2016 classés par la méthode par règles, France

<u>Selon le type de lieu de décès</u>

Le nombre moyen de RS par certificat variait selon le type de lieu de décès. Il était supérieur à 3,5 pour les décès survenant dans des établissements de santé publics, les EHPAD ou pour les autres lieux indéterminés (Figure 39). Le nombre moyen de RS était le plus faible pour les décès survenus dans les établissements de santé privés (3,2).

En dehors des décès ayant lieu à domicile, pour lesquels la répartition des RS définis pour l'alerte et non définis pour l'alerte étaient équivalentes, on retrouvait en moyenne une répartition des RS en faveur des RS non définis pour l'alerte et plus particulièrement pour les décès ayant eu lieu dans des établissements de santé privés (Figure 39). Pour les décès sur la voie publique, les RS définis pour l'alerte étaient de 57 % en moyenne.

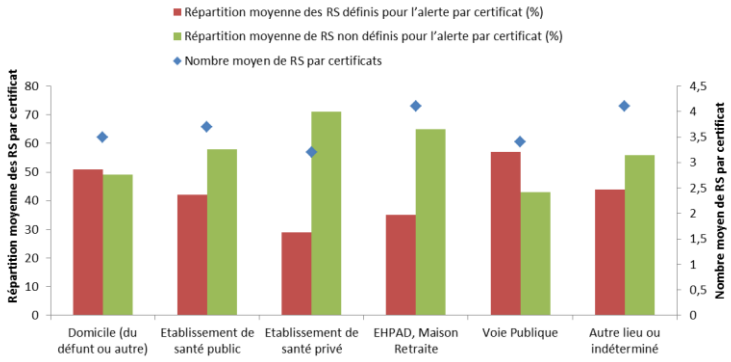


Figure 39 : Nombre moyen de regroupements syndromiques par certificat et répartition moyenne des RS définis pour l'alerte et non définis pour l'alerte par certificat et par type de lieu de décès ; Certificats électroniques de 2012 à 2016 classés par la méthode par règles, France

2) Fréquence des regroupements syndromiques et des thématiques dans les certificats électroniques reçus entre 2012 et 2016

Selon la classe d'âge

Pour l'ensemble des classes d'âge, la thématique « Cardio et Cérébrovasculaires » (de 16,0% à 31,6%) était la plus fréquente ou la deuxième plus fréquente et la thématique « Symptômes » (de 8,9% à 11,8%) était la troisième ou la quatrième plus fréquente selon les classes d'âge (Tableau 22). Chez les enfants de 1 à 14 ans, les thématiques les plus fréquentes après « Cardio et Cérébrovasculaires » étaient « Maladies respiratoires » (13,4%) et « Maladies du système nerveux central » (11,6%). Pour les adultes de 15 à 84 ans, la thématique « Cancers/Tumeurs » était retrouvée fréquemment en première ou en deuxième position (de 14,6% à 24,6%) selon les classes d'âge.

La distribution des RS les plus fréquents au sein des thématiques principalement retrouvées pour toutes les classes d'âge est présentée en Figure 40.

Pour la thématique « Cardio et Cérébrovasculaires », à part le RS « non informatif » le plus fréquemment retrouvé pour toutes les classes d'âge jusqu'à 64 ans, le RS « Cardio et circulatoire chronique » était le plus fréquent (Figure 40).

Pour la thématique « Maladies respiratoires », le RS « Maladies respiratoires chroniques » était le plus fréquent pour l'ensemble des classes d'âges (de 4,9 à 3,1%), suivi par « Asphyxie et anomalie de la respiration » (de 1,7 à 3,3%) pour les classes d'âge les plus jeunes jusqu'à 44 ans. Le RS « Infections respiratoires aigües basses » était retrouvé en deuxième position pour les classes d'âge les plus âgées (de 2,6 à 3,9%).

Pour la thématique « Symptômes », le RS le plus fréquent chez les moins de 1 an était « Mort subite » (4,4%). Pour l'ensemble des autres classes d'âge, les RS « Symptômes généraux » (1,7% à 3,5%) et « Coma » (de 1% à 3,9%) étaient les plus fréquents.

Le RS « Tableau neuropsychologiques aigus » était le plus fréquent parmi les RS de la thématique « Maladies du système nerveux central » pour les enfants jusqu'à 14 ans. Les « Maladies chroniques du système nerveux » étaient les plus fréquentes pour les autres classes d'âge.

Tableau 22 : Fréquence des thématiques par classe d'âge ; Certificats électroniques des décès survenus entre 2012 et 2016 et classés par la méthode par règles, France

Thématiques	<1 an (%)	1-14 ans (%)	15-44 ans (%)	45-64 ans (%)	65-84 ans (%)	>=85 ans (%)
1-Cardio et Cérébrovasculaires	19,8	18,6	16	17,1	24,0	31,6
2-Cancers/tumeurs	1,0	6,6	14,6	24,6	18,3	8,3
3-Maladies respiratoires	9,7	13,4	8,0	9,4	11,4	12,2
4-Symptômes	11,8	10,6	11,3	9,3	8,9	10
5-Maladies infectieuses	6,1	5,4	4,7	5,2	4,5	3,8
6-Maladies endocriniennes	4,1	3,8	2,2	3,4	5,0	6,0
7-Maladies du système nerveux central	8,7	11,6	5,8	3,6	3,7	3,7
8-Autres facteurs et motifs de recours aux soins	3,4	4,0	3,9	4,7	5,3	4,1
9-Trauma et empoisonnement	3,8	5,13	10	4,2	2,5	2,2
10-Troubles mentaux et du comportement	0,1	0,37	2,8	2,2	2,3	3,9
11-Maladies de la peau et du tissu cellulaire sous cutané	0,2	0,0	0,4	0,4	0,5	0,8
12-Maladies du sang et des organes hématopoïétiques	2,2	3,2	1,9	1,5	1,3	1,3
13-Maladies de la grossesse, néonatales et congénitales	20	4,8	1,6	0,7	0,4	0,2
14-Maladies de l'œil et de ses annexes	0	0,07	0,0	0,0	0,1	0,1
15-Maladies de l'appareil digestif	4,2	2,3	5,1	8,1	5,5	4,2
16-Causes mal définies	1,2	1,3	1,6	1,4	1,1	1,0
17-Causes externes	1,7	6,8	7,3	1,6	1	1,3
18-Maladies de l'appareil génito urinaire	0,8	0,7	0,8	1,5	3,	4,10
19-Maladies de l'oreille	0,0	0,0	0,0	0,0	0,0	0,0
20-Maladies du système ostéo-articulaire	1,0	1,2	2,0	1,1	1,2	1,3

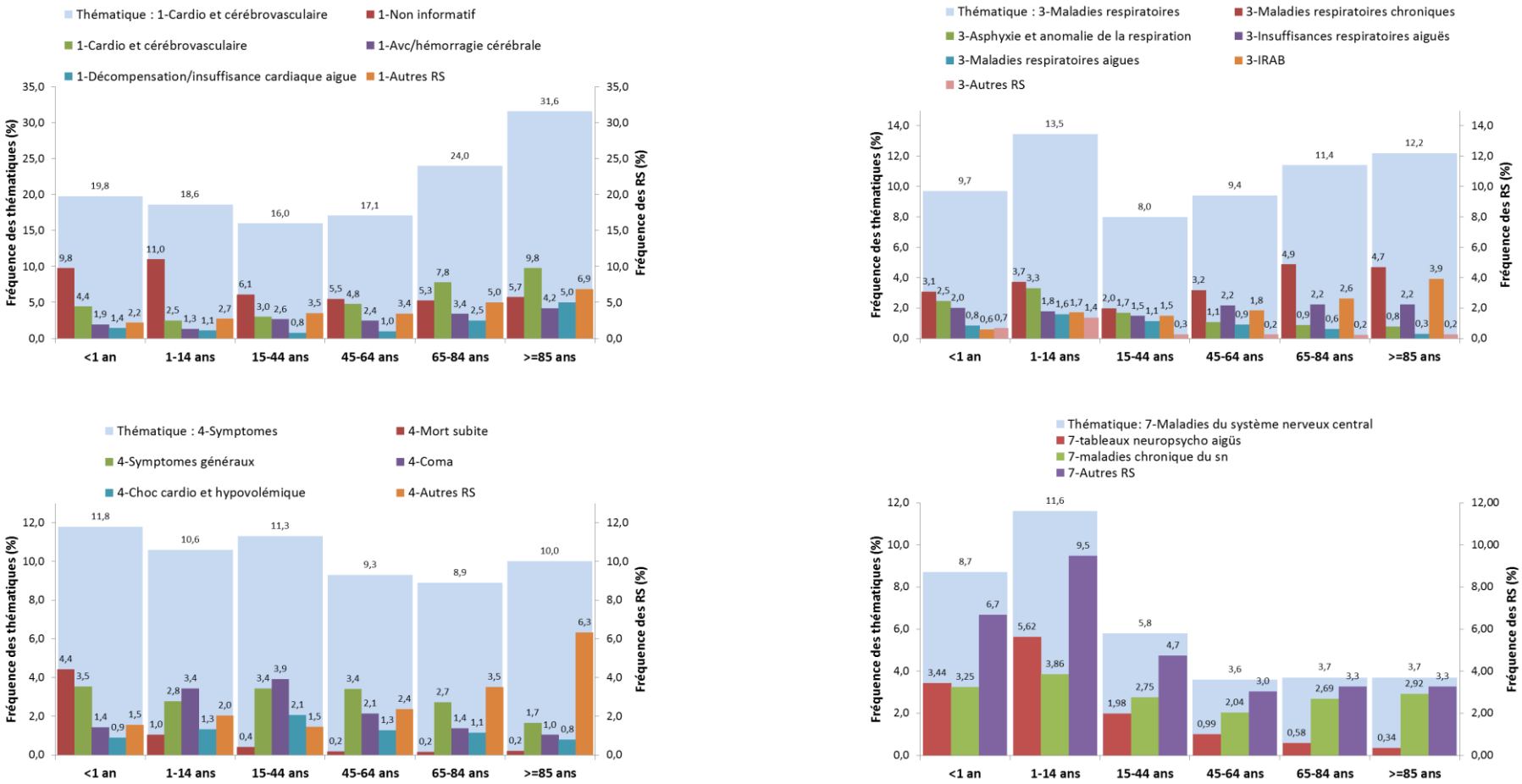


Figure 40 : Fréquence des regroupements syndromiques et des thématiques « Cardio et Cérébrovasculaires », « Maladies respiratoires », « Symptômes » et « Maladies du système nerveux central », par classe d'âge, Certificats électroniques de 2012 et 2016 et classés par la méthode par règles, France

Selon le type de lieu de décès

Les thématiques les plus fréquemment attribuées étaient différentes selon le type de lieu de décès (Tableau 23). La thématique « Cardio et Cérébrovasculaires » était la plus fréquente pour tous les types de lieu de décès sauf pour les décès de la voie publique, pour lesquels les thématiques les plus fréquentes étaient « Causes externes » (31,0%) et « Trauma et empoisonnement » (27,3%) et pour les établissement privés pour lesquels la thématique « Cancers/tumeurs » était la plus fréquente (28,7%).

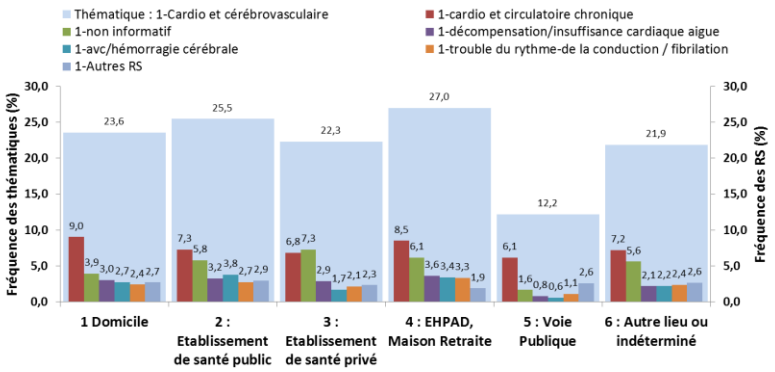
A domicile, en dehors de la thématique « Cardio et Cérébrovasculaires », les thématiques les plus fréquentes étaient : « Causes externes », « Trauma et empoisonnement » et « Symptômes ». Dans les établissements publics et privés, les thématiques « Cancers/Tumeurs » (15,3 % et 28,7% respectivement), « Symptômes » (9,3 % et 9,5% respectivement) et « Maladies respiratoires » (11,8% et 7,9% respectivement) étaient les plus fréquentes. Dans les EHPAD, la thématique « Symptômes » (13%) ainsi que la thématique « Troubles mentaux et du comportement » (9,3%) étaient majoritaires. Pour les autres lieux indéterminés on retrouvait fréquemment les « Cancers/Tumeurs » (10,4%) et les « Causes externes » (10,2%).

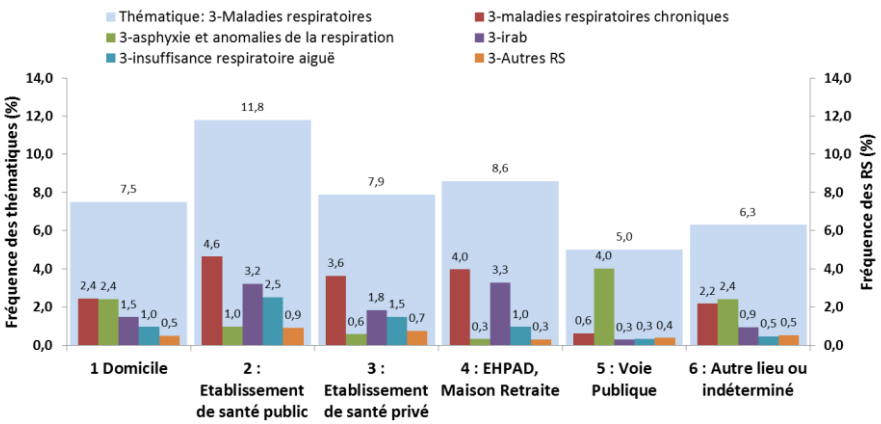
Les distributions des RS les plus fréquents pour les thématiques principalement retrouvées pour les différents types de lieu de décès sont présentées en Figure 41.

Le RS « Cardio et circulatoire chronique » était le plus fréquent de la thématique « Cardio et Cérébrovasculaires » pour l'ensemble des types de lieu de décès, suivi par le RS « Non informatif ». Pour la thématique « Maladies respiratoires » dans les établissements publics, privés et les EHPAD le RS « Maladies respiratoires chroniques » était le plus fréquent, alors que pour les autres lieux de décès le RS « Asphyxie et anomalie de la respiration » était le plus fréquent. Le RS « Infections respiratoires aigües basses » était le troisième plus fréquent pour tous les types de lieu de décès.

Tableau 23 : Fréquence des thématiques par type de lieu de décès; Certificats électroniques reçus entre 2012 et 2016 et classés par la méthode par règles, France

Thématiques	Domicile	Etablissement de santé public	Etablissement de santé privé	EHPAD, Maison retraite	Voie Publique	Autre lieu ou indéterminé
1-Cardio et Cérébrovasculaires	23,6	25,5	22,3	27,0	12,2	21,9
2-Cancers/tumeurs	6,5	15,3	28,7	5,8	1,3	10,4
3-Maladies respiratoires	7,5	11,8	7,9	8,6	5,0	6,3
4-Symptomes	8,9	9,3	9,5	13,0	4,2	9,2
5-Maladies infectieuses	1,7	4,7	3,0	1,8	0,4	1,3
6-Maladies endocriniennes	4,6	5,0	3,2	7,2	1,3	5,2
7-Maladies du système nerveux central	3,3	3,7	2,2	7,7	0,8	4,0
8-Autres facteurs et motifs de recours aux soins	2,9	4,6	7,0	4,5	1,2	5,5
9-Trauma et empoisonnement	9,0	2,9	1,8	1,8	27,3	8,7
10-Troubles mentaux et du comportement	6,1	2,6	1,6	9,3	4,5	3,7
11-Maladies de la peau et du tissu cellulaire sous cutané	0,4	0,6	0,5	1,0	0,0	0,6
12-Maladies du sang et des organes hématopoïétiques	1,1	1,4	1,0	0,8	1,2	1,6
13-Maladies de la grossesse, néonatales et congénitales	0,4	0,5	0,3	0,3	0,2	0,5
14-Maladies de l'œil et de ses annexes	0,0	0,1	0,1	0,2	0,0	0,3
15-Maladies de l'appareil digestif	3,5	5,7	4,7	3,1	1,2	2,8
16-Causes mal définies	6,8	0,8	2,1	2,8	2,7	3,3
17-Causes externes	9,7	1,2	0,6	1,5	31,0	10,2
18-Maladies de l'appareil génito urinaire	1,6	3,2	2,9	2,4	0,1	1,7
19-Maladies de l'oreille	0,0	0,0	0,0	0,0	0,0	0,0
20-Maladies du système ostéo-articulaire	2,5	1,2	0,8	1,4	5,0	2,7





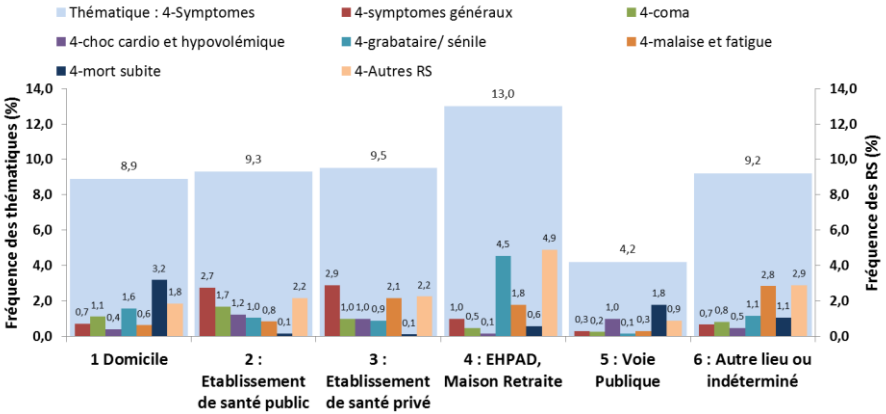


Figure 41 : Fréquence des regroupements syndromiques au sein des thématiques « Cardio et Cérébrovasculaires », « Maladies respiratoires » et « Symptômes » par type de lieu de décès; Certificats électroniques des décès de 2012 et 2016 et classés par la méthode par règles, France

Selon le champ de certificat :

Dans la partie I du certificat, les thématiques les plus fréquentes étaient « Cardio et Cérébrovasculaires » (de 20% à 28,3%) et « Cancers/Tumeurs » (de 21,6% à 17,3%) en première ou deuxième position selon les champs. Dans le champ 1, pour lequel la deuxième et troisième thématique les plus fréquentes étaient « Symptômes » (18,4%) et « Maladies respiratoires » (15,31%) (Tableau 24). La troisième position était obtenue pour la thématique « Maladies respiratoires » (15,2 % et 9,9% respectivement) dans les champs 2 et 3 et pour les thématiques « Autres facteurs de recours aux soins » (7, %) dans le champ 4.

Dans les champs de la partie II, les thématiques les plus fréquentes étaient « Cardio et Cérébrovasculaires» et « Maladies endocriniennes » en première et deuxième positions respectivement. Dans le champ 5, la troisième thématique la plus fréquente était « Troubles mentaux et du comportement », alors que dans le champ 6 la thématique « Autres facteurs et motifs de recours aux soins » était la troisième plus fréquente.

Pour la thématique « Cardio et Cérébrovasculaires », le RS « Cardio et circulatoire chronique » était le plus fréquents pour l'ensemble des champs, sauf pour le champ pour lequel le RS « Non informatif » était le plus fréquent (Figure 42). Le RS « AVC/hémorragie cérébrale » était le deuxième plus fréquent pour l'ensemble des champs de la partie I, alors que pour les champs de la partie II, le RS « Trouble du rythme et de la conduction » était le deuxième plus fréquent.

Pour la thématique « Maladies respiratoires », le RS le plus fréquents pour l'ensemble des champs à l'exception du champ 1 était «Maladies respiratoires chroniques ». Pour le champ 1, le RS le plus fréquent était « Insuffisance respiratoire aigüe ».

Les RS les plus fréquents pour la thématique « Symptômes » étaient « Symptômes généraux » pour les champs 1 et 2 et « Coma » pour les champs 3 à 6.

Pour la thématique « Maladies endocriniennes », le RS « Maladies chroniques endocriniennes » était le plus fréquent sur tous les champs sauf le champ 1.

Tableau 24 : Fréquence des thématiques par champ de certificats ; certificats électroniques des décès survenus entre 2012 et 2016 et classés par la méthode par règles, France

		Partie I			Partie II	
Thématiques	**Champ 1**	**Champ 2**	**Champ 3**	**Champ 4**	**Champ 5**	**Champ 6**
1-Cardio et Cérébrovasculaires	28,3	22,1	20,7	20,0	29,2	26,1
2-Cancers/tumeurs	11,4	17,3	20,8	21,6	8,1	7,5
3-Maladies respiratoires	15,3	15,2	10,0	6,9	5,9	5,8
4-Symptomes	18,4	7,9	5,0	4,4	5,5	5,8
5-Maladies infectieuses	5,9	6,1	4,3	3,6	1,8	2,2
6-Maladies endocriniennes	1,6	2,7	3,2	3,9	11,9	13,2
7-Maladies du système nerveux central	3,1	3,6	3,7	4,1	5,3	4,6
8-Autres facteurs et motifs de recours aux soins	2,3	3,7	5,8	7,2	6,8	7,6
9-Trauma et empoisonnement	1,7	3,2	4,4	5,1	3,1	3,6
10-Troubles mentaux et du comportement	0,5	1,4	2,6	4,2	7,5	6,8
11-Maladies de la peau et du tissu cellulaire sous cutané	0,3	0,7	0,7	0,7	0,9	1,1
12-Maladies du sang et des organes hématopoïétiques	0,9	1,8	1,8	1,4	1,5	1,9
13-Maladies de la grossesse, néonatales et congénitales	0,6	0,3	0,5	0,7	0,5	0,5
14-Maladies de l'œil et de ses annexes	0,0	0,0	0,0	0,0	0,3	0,3
15-Maladies de l'appareil digestif	4,2	7,2	8,3	7,5	3,5	3,7
16-Causes mal définies	2,5	0,8	0,5	0,4	0,3	0,3
17-Causes externes	0,5	2,0	3,2	3,8	0,6	0,8
18-Maladies de l'appareil génito-urinaire	1,9	2,8	3,1	2,8	5,3	5,9
19-Maladies de l'oreille	0,0	0,0	0,0	0,0	0,1	0,1
20-Maladies du système ostéo-articulaire	0,6	1,1	1,6	1,8	1,9	2,2

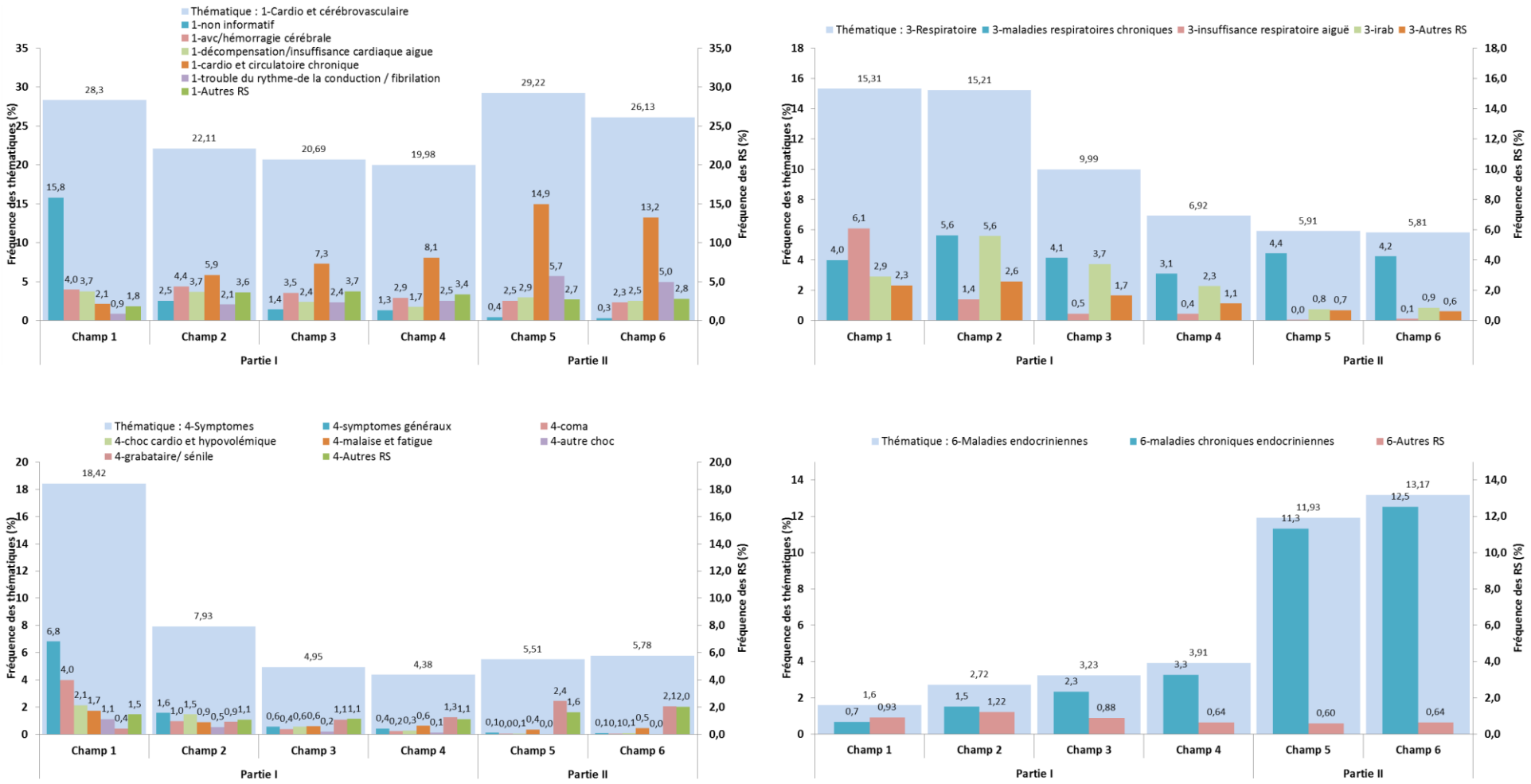


Figure 42 : Fréquence des regroupements syndromiques et des thématiques « Cardio et Cérébrovasculaires », « Maladies respiratoires », « Symptômes », « Maladies endocriniennes »par champ de certificats; Certificats électroniques des décès de entre 2012 et 2016 et classés par la méthode par règles, France

3) Variation mensuelle de la répartition moyenne des regroupements syndromiques par certificat

On observait une légère évolution saisonnière de la répartition moyenne des RS définis pour l'alerte au sein d'un certificat (Figure 43). Ainsi, chaque certificat contient en moyenne 40% de RS définis pour l'alerte pendant les mois estivaux contre 42% sur les mois hivernaux.

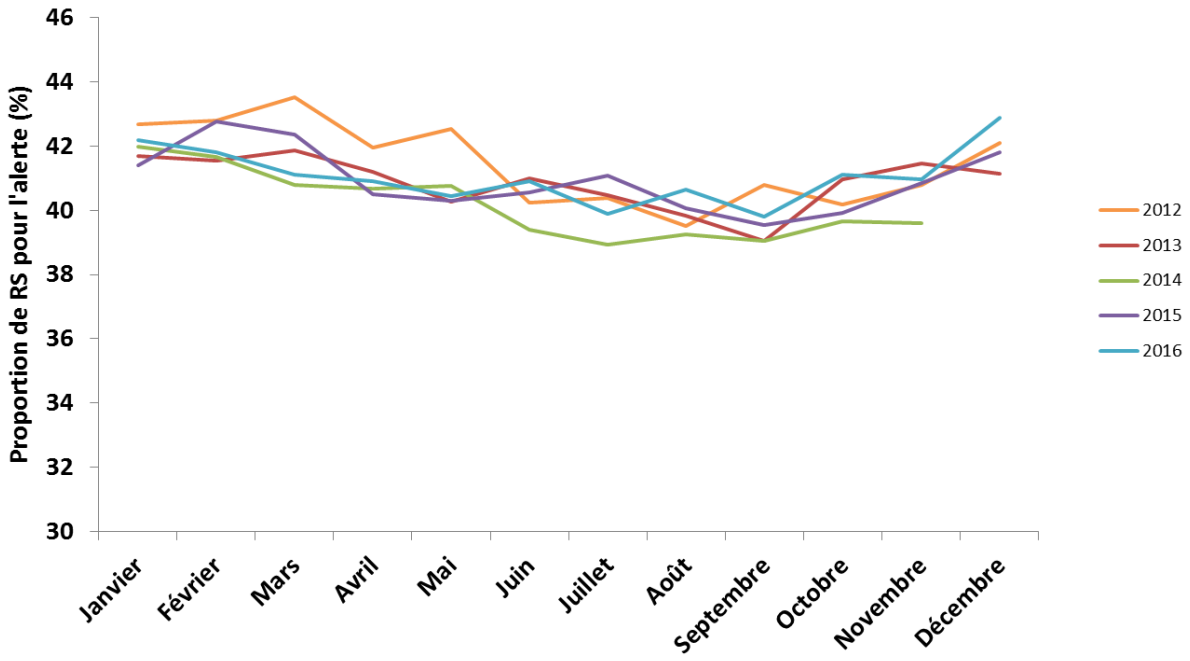


Figure 43 : Variation mensuelle de la répartition moyenne des regroupements syndromiques définis pour l'alerte au sein d'un certificat, Certificats électroniques de 2012 à 2016 classés par la méthode par règles, tous âges, France

IV/ Discussion

La description des regroupements syndromiques attribués par la méthode par règles pour les certificats électroniques des décès survenus entre 2012 et 2016 a montré que dans un certificat de décès, en moyenne 39% des RS étaient définis pour l'alerte : 46% dans la partie I et 26% dans la partie II.

Plus généralement, la répartition des RS était différente selon les classes d'âge et le type de lieu de décès. La répartition par champ de certificat indiquait que les RS définis pour l'alerte étaient retrouvés sur l'ensemble des champs.
La description de la fréquence des RS et des thématiques a montré que selon la classe d'âge et le type de lieu de décès, les thématiques et les RS les plus fréquents étaient variables. La description par champ de certificat a permis d'identifier des différences de thématiques et de RS entre la partie I et la partie II du certificat.
Enfin, les variations mensuelles des RS définis pour l'alerte mettaient en évidence une légère variation saisonnière.

L'analyse de la mortalité par regroupement syndromique en utilisant une pondération équilibrée des RS au sein de chaque certificat permettra de conserver l'ensemble de ces répartitions, alors qu'une autre stratégie de pondération donnant plus de poids aux RS définis pour l'alerte ou bien à la partie I du certificat par exemple entrainera une modification de ces répartitions.

Une pondération équilibrée présente les avantages suivant :

- Etre une pondération simple qui ne fait pas d'hypothèses *a priori* sur l'attribuabilité d'un type de RS dans le décès
- Respecter la séquence morbide telle qu'elle a été inscrite par le médecin,
- Prendre en compte tous les RS dans la contribution au décès.

Une des limites de cette pondération est que l'ensemble des causes sont mises sur le même plan : un antécédent ou une comorbidité aura le même poids qu'une pathologie aigüe. Elles n'ont pourtant pas forcément la même contribution au décès.

Des travaux ont été réalisés sur la pondération des causes de décès codées dans l'objectif d'établir des statistiques (120). Parmi les stratégies proposées, on trouvait une stratégie de pondération équilibrée sur l'ensemble des causes d'un certificat, et une stratégie qui proposait d'attribuer un poids plus important à la cause initiale et de répartir le reste du poids entre les causes se trouvant dans la partie II du certificat. Les autres causes de décès contenues dans la séquence morbide n'étaient alors pas comptabilisées comme ayant une part attribuable pour ce décès.

Ces stratégies ont été mises en œuvre sur les causes de décès codées et ne sont pas toutes transposables sur les causes de décès en texte libre.

Bien que notre choix se soit porté sur une stratégie de pondération équilibrée, d'autres stratégies peuvent être envisagées :

- Appliquer un poids plus important sur un type de RS, par exemple les RS spécifiquement définis pour l'alerte sanitaire ou sur une liste de RS définis préalablement comme étant les RS d'intérêt pour la surveillance de la mortalité,
- Appliquer un poids plus important aux RS qui constituent la séquence morbide (partie I du certificat).

Des travaux complémentaires pour discuter de la pertinence de ces différentes stratégies de pondération constituent une perspective à ce travail.

Conclusion générale et perspectives

Suite à la canicule de 2003, dont les conséquences sans précédent avaient révélé l'inefficacité du système de surveillance en place à détecter de manière précoce l'augmentation de la mortalité et à alerter les pouvoirs publics, un nouveau système de surveillance syndromique, le système SurSaUD, a été mis en place. Il s'appuie sur deux sources de données de morbidité et deux sources de données de mortalité dont la certification électronique des décès mise en place en 2007.

La surveillance syndromique de la mortalité a pour objectif de détecter précocement les variations habituelles ou inhabituelles de la mortalité et de mesurer l'impact d'un événement sur la santé de la population. La certification électronique des décès permet à Santé publique France de recevoir les données issues du volet médical dans les minutes qui suivent la validation d'un certificat. Cette seconde source, complémentaire de la source basée sur les données administratives, permet de disposer en temps réel des causes médicales de décès sous forme de texte libre.

Nos travaux visaient à mettre en œuvre le système de surveillance syndromique de la mortalité à partir des causes médicales de décès sous forme de texte libre, afin d'analyser en temps quasi réel la mortalité par cause et ainsi compléter et enrichir la surveillance réactive toutes causes.

Dans ce contexte, les objectifs de ces travaux étaient tout d'abord de décrire la surveillance actuelle de la mortalité s'appuyant sur les données administratives des certificats de décès, afin d'en souligner les avantages et limites. Ensuite, l'objectif était de définir les RS à suivre pour répondre aux objectifs de surveillance, puis de mettre en œuvre et évaluer des méthodes des traitements automatiques des langues pour classer les causes de décès en texte libre dans les RS. Enfin, le dernier objectif portait sur la méthodologie de prise en compte des causes multiples de décès pour l'analyse et l'interprétation des regroupements syndromiques.

En synthèse, nos travaux ont permis d'aboutir aux quatre points suivants :

- La description du système de surveillance de la mortalité à partir des informations administratives du certificat de décès a montré que ce système en place depuis 2004, est un système utile et efficace qui permet de fournir des éléments aux décideurs pour aider à évaluer une situation et identifier les mesures de gestion. La principale limite du système est qu'il s'appuie uniquement sur les données administratives du défunt et ne contient pas d'informations sur les causes médicales de décès, ce qui limite l'interprétation des variations observées, notamment lorsqu'elles ne sont pas concomitantes avec un événement identifié.

- La disponibilité des causes de décès en texte libre, grâce à la certification électronique, permet d'envisager une surveillance syndromique de la mortalité par cause. L'exploitation de telles données pour la surveillance a conduit à définir la liste de de près de cent regroupements syndromiques permettant de regrouper les causes de décès en groupes homogènes. Leurs définitions sont basées sur une liste de codes CIM-10 et complétées par une liste d'expressions d'une grande variété issues du dictionnaire adapté du CépiDc.

- Une fois les regroupements syndromiques définis, nous avons mis en œuvre et évalué deux méthodes issues du traitement automatique des langues permettant la classification automatique des causes de décès en texte libre dans les regroupements syndromiques. Ces deux méthodes étaient une méthode à base de règles linguistiques et une méthode par apprentissage automatique supervisé : un Support Vector Machine (SVM). L'évaluation des performances de ces méthodes a été effectuée dans un premier temps pour la classification des causes de décès dans un nombre restreint de sept regroupements syndromiques. Elle a montré que la méthode par règles et le deuxième modèle SVM (SVM 2) obtenaient des performances élevées avec des F-mesure supérieures à 0,95 pour l'ensemble des sept regroupements syndromiques. Le modèle SVM 1 obtenait une F-mesure comprise entre [0,90 et 0,95[pour deux regroupements syndromiques et supérieure à 0,95 pour les cinq autres regroupements syndromiques.

  A la suite de cette évaluation, nous avons étendu la description des performances de la méthode par règles et du modèle SVM 2 pour soixante autres regroupements syndromiques. Les deux méthodes obtenaient des performances supérieures à 0,95 pour vingt-quatre d'entre eux.

  Au final, les deux méthodes de classification des causes de décès ont été évaluées pour deux tiers des regroupements syndromiques (67) parmi la centaine de regroupements syndromiques définis pour la surveillance de la mortalité. Un tiers des regroupements syndromiques (31) n'a pas été évalué. Parmi les soixante-sept regroupements syndromiques évalués, les méthodes mises en œuvre obtenaient de très bonnes performances pour la classification des causes de décès dans trente-et-un regroupements syndromiques, soit près de la moitié des regroupements syndromiques évalués. C'est donc avec confiance que nous pourrons analyser les variations de la mortalité pour ces trente-et-un regroupements syndromiques.

  Le nombre restreint de certificats de décès inclus dans les échantillons utilisés pour l'évaluation des performances des méthodes explique, au moins en partie, les performances plus faibles obtenus pour les trente-six autres regroupements syndromiques évalués. En

effet, ces échantillons de 4500 certificats tirés aléatoirement, n'incluaient pas forcément toutes les variétés d'expressions de causes de décès, que ce soit pour les regroupements syndromiques couvrant une large palette de pathologies ou inversement pour les regroupements de pathologies rares (« Arboviroses/ Fièvres hémorragiques », « Bronchiolite », « Méningite virale »).

La taille limitée de nos échantillons est également à l'origine de l'exclusion des 31 regroupements syndromiques qui n'ont pas pu être évalués dans ce travail. Nos échantillons n'étaient en effet pas suffisamment représentatifs de la distribution de ces trente-trois regroupements syndromiques par rapport à celle observée dans l'ensemble de la mortalité.

Enfin, ces travaux ont permis de confirmer les performances élevées de ces méthodes pour la classification de données textuelles, performances déjà soulignées dans la littérature pour des tâches proches de la nôtre.

- Le nombre multiple de regroupements syndromiques attribué à chaque certificat de décès (3,7 regroupements syndromiques en moyenne) conduit à un décompte des effectifs de décès par regroupement syndromique supérieurs à celui du nombre total de décès. Nous avons proposé une méthode de pondération équilibrée des regroupements syndromiques au sein de chaque certificat, pour permettre que la somme des effectifs de décès associés à chaque regroupement syndromique soit égale au nombre total de décès.

Ces travaux ont fait l'objet d'une discussion détaillée à l'issue de chaque chapitre de ce mémoire. Nous pouvons toutefois souligner les principaux points suivants :

- Une liste de près de cent regroupements syndromiques a été proposée pour répondre aux objectifs de détection d'événements attendus ou inhabituels à visée d'alerte et d'évaluation d'impact. La définition de ces regroupements syndromiques est encore en cours de validation et doit se poursuive avec des cliniciens et experts de chaque thématique. De plus, cette liste de regroupements syndromiques est une proposition d'organisation pour répondre à nos objectifs, permettant de classer l'ensemble des pathologies, symptômes, traumatismes, actes médicaux ou antécédents. Elle doit rester flexible et évolutive pour être capable de s'adapter rapidement en cas d'émergence ou de besoin d'analyse plus ciblée sur certaines causes médicales de décès
- Nos travaux ont permis d'évaluer les performances de deux méthodes pour classer les causes médicales de décès dans soixante-sept regroupements syndromiques et d'obtenir de

très bonnes performances pour classer les causes dans trente-et-un d'entre eux. Ce travail doit se poursuivre pour permettre d'analyser avec confiance la mortalité pour l'ensemble des regroupements. Cela nécessite en premier lieu l'enrichissement des échantillons annotés de certificats, afin d'évaluer et améliorer les performances des méthodes sur les regroupements syndromiques restants.

Par ailleurs, des travaux internationaux récents ont mis en évidence les performances d'autres méthodes de classification d'informations textuelles, telles que les méthodes d'apprentissage profond ou des méthodes combinées. Il serait intéressant d'explorer l'apport de ces méthodes pour répondre à nos objectifs. Dans tous les cas, il est important que les méthodes choisies de classification des causes médicales de décès dans les regroupements syndromiques soient flexibles et puissent s'adapter rapidement, pour être capable d'évoluer en cas de situations nouvelles, telle que la survenue d'une émergence par exemple.

- Nous avons proposé d'analyser la mortalité en utilisant une pondération équilibrée des regroupements syndromiques attribués à chaque certificat, permettant ainsi de tenir compte dans nos analyses de l'ensemble des informations mentionnées par les médecins. L'analyse et l'interprétation de la décomposition de la mortalité totale par regroupement syndromique en utilisant cette pondération reste à mettre en place pour la surveillance en routine de la mortalité et l'évaluation d'impact sanitaire.

Perspectives

A l'issue de ces travaux, la première perspective découle directement des points discutés précédemment. En particulier, elle porte sur :

- la validation de la définition des regroupements syndromiques avec les experts et cliniciens,

- l'enrichissement des échantillons de certificats annotés pour améliorer les performances des regroupements syndromiques évalués dans ce travail et évaluer les performances des trente-et-un regroupements syndromiques non encore évalués.

Plus généralement, deux principales perspectives restent à mener pour permettre la mise en œuvre opérationnelle de la surveillance réactive à visée d'alerte et d'évaluation d'impact à partir des données la certification électronique des décès :

- la première perspective porte sur l'amélioration de la couverture de la mortalité enregistrée par voie électronique et la capacité à prendre en compte cette montée en charge dans l'analyse des regroupements syndromiques,

- la seconde vise à construire les outils de restitution et de visualisation des indicateurs pour les épidémiologistes qui seront en charge de mener cette surveillance réactive de la mortalité.

Déploiement de la certification électronique et prise en compte de ce déploiement pour l'analyse

Début 2019, la certification électronique des décès enregistre un peu plus de 15 % de la mortalité totale, avec une répartition hétérogène sur le territoire et une représentativité sociodémographique imparfaite. Afin d'accroitre cette couverture, des actions nationales et régionales sont mises en place progressivement par les ARS pour inciter les médecins à certifier les décès électroniquement. Parmi ces actions, une expérimentation visant à dématérialiser l'ensemble du certificat de décès (volet administratif et volet médical) a été menée en 2018 dans six villes françaises (Antibes, Aurillac, Créteil, La Rochelle, Montluçon, Villejuif) (121). La généralisation de cette dématérialisation complète du certificat permettra de lever un frein majeur au déploiement (l'impression du volet administratif après validation du certificat électronique de décès), notamment pour les décès à domicile et sur la voie publique.

Le déploiement progressif mais irrégulier de l'utilisation de la certification électronique par les établissements et les médecins généralistes ont une influence sur les variations temporelles de la mortalité et rendent délicat leur interprétation. En effet, il est difficile de distinguer l'évolution liée à

un événement sanitaire (ou à l'évolution saisonnière habituelle de la mortalité) de celle liée à la montée en charge du système (démarrage / interruption d'un nouvel établissement, d'un nouveau service hospitalier ou même d'un nouveau médecin). Ces deux phénomènes peuvent avoir un impact plus marqué sur certains regroupements syndromiques, mais pas forcément sur l'ensemble de la mortalité. Ainsi, en cas de survenue d'un événement sanitaire, certaines causes peuvent être exprimées par les médecins plus souvent qu'en dehors de cet événement. De même, le démarrage/interruption d'un établissement ou d'un service spécialisé (gériatrie, pédiatrique, réanimation) peut influencer plus spécifiquement certains regroupements syndromiques ou certaines sous-populations (enfants, personnes âgées par exemple).

La prise en compte de la montée en charge de la certification est une perspective de nos travaux. Plusieurs stratégies d'analyse sont possibles pour tenter de s'affranchir au moins en partie de ces effets :

1) L'analyse temporelle peut être envisagée sur un sous-ensemble d'établissements ayant tous démarré la certification électronique à partir d'une date donnée ("travail à établissement constant"). Cette stratégie reste toutefois une réponse partielle, puisque :
    - un établissement peut déployer la certification électronique progressivement dans son établissement (service par service) ou simultanément dans tout l'établissement,
    - elle ne peut pas s'appliquer au déploiement progressif des médecins libéraux dans une même commune, les médecins libéraux n'étant pas identifiés individuellement.
2) La surveillance de la mortalité par regroupement syndromique peut s'appuyer sur l'analyse de la proportion de décès de chaque regroupement syndromique, parmi l'ensemble des décès certifiés électroniquement. Cette analyse permet de s'affranchir grossièrement d'une variation liée à la montée en charge, mais ne permet pas de prendre en compte l'introduction d'un établissement/service très spécialisé.
3) L'analyse combinant les deux stratégies (analyse de la proportion de décès inclus dans un regroupement parmi l'ensemble des décès à échantillon constant) limitera l'influence de la montée en charge de la certification électronique dans l'interprétation des évolutions.

La mise en œuvre de ces stratégies de prise en compte de la montée en charge est nécessaire avant la mise en place en routine de la surveillance de la mortalité par cause.

Développement d'outils de restitution et visualisation pour les épidémiologistes

L'analyse réactive et en routine de la mortalité issue de la certification électronique nécessite pour les épidémiologistes de disposer d'un outil de restitution et de visualisation des indicateurs, automatiquement actualisé après la réception et le traitement des données à Santé publique France.

Ce travail est mené depuis février 2019 par Alexandre Cornec dans le cadre de son stage de troisième année d'école d'ingénieur en Systèmes d'Information. Cet outil en cours de développement qui prendra la forme d'une application développée sous R-Shiny, permettra dans une première version :

1/ de suivre la progression du déploiement de la certification électronique, à travers un tableau de consultation des établissements certificateurs en fonction de leurs caractéristiques et date de démarrage, ou à travers des cartes géographiques (Figure 44),

2/ d'analyser sous la forme de graphiques ou tableaux, les variations temporelles (quotidiennes, hebdomadaires, mensuelles) des regroupements syndromiques selon différents axes : classe d'âge, sexe, zone géographique (national, région, département), que l'utilisateur peut choisir à travers des filtres adaptés,

3/ de comparer ces variations entre différentes catégories (regroupements syndromiques, classes d'âge, zones géographiques) ou d'une année sur l'autre (Figure 45),

4/ d'analyser la répartition des décès certifiés électroniquement parmi l'ensemble des décès, en fonction de ces différentes catégories.

L'ensemble des effectifs présentés dans cette application reposera sur les causes médicales de décès classées selon les deux méthodes de classification décrites dans ce document (méthodes par règles et modèle SVM 2) et sur l'application de la méthode de pondération équilibrée proposée dans le chapitre V. Cette application, une fois finalisée et automatisée, devra être accompagnée d'une formation pour les utilisateurs.

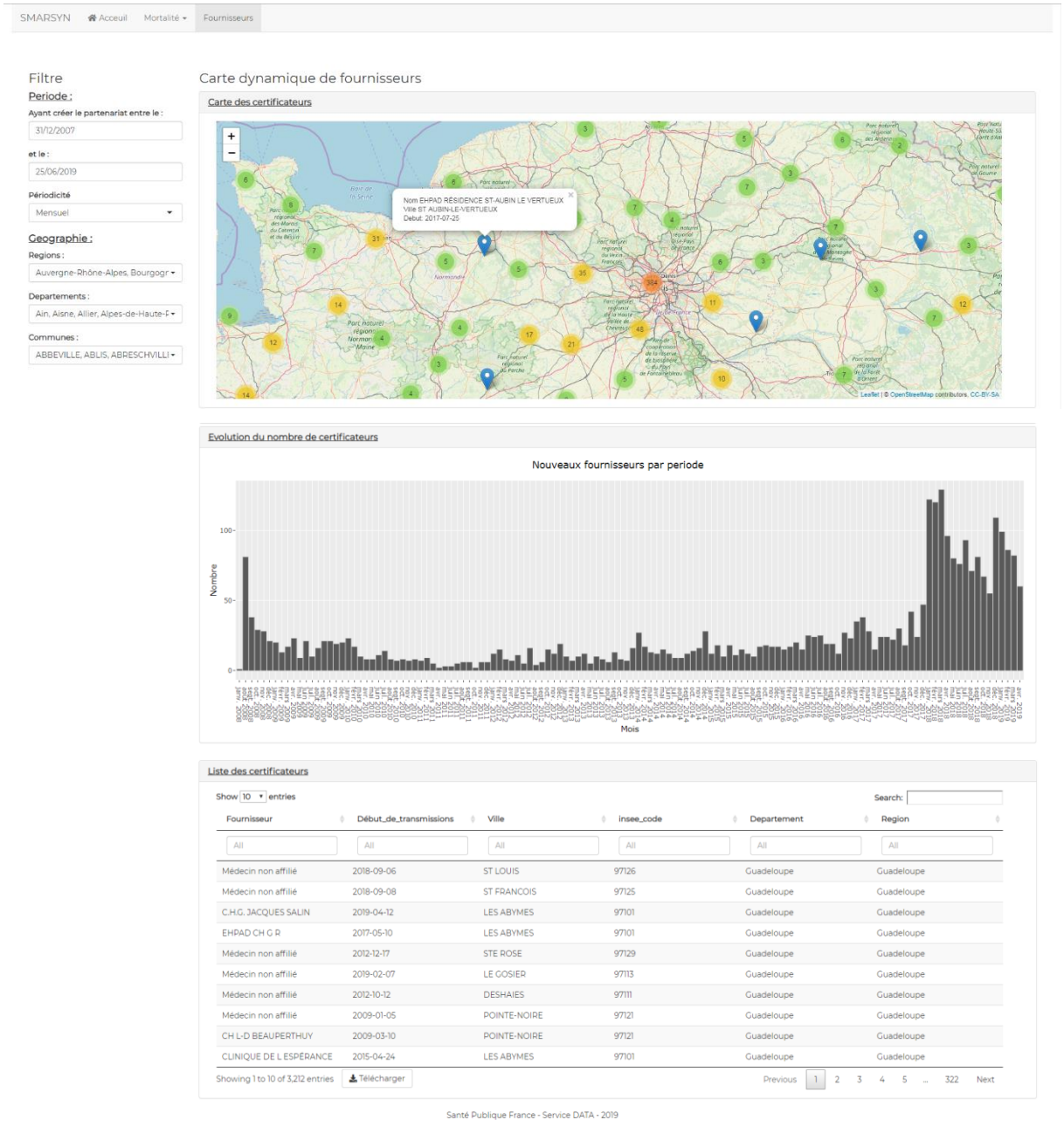


Figure 44: Exemple d'écran pour le suivi du déploiement de la certification électronique des décès

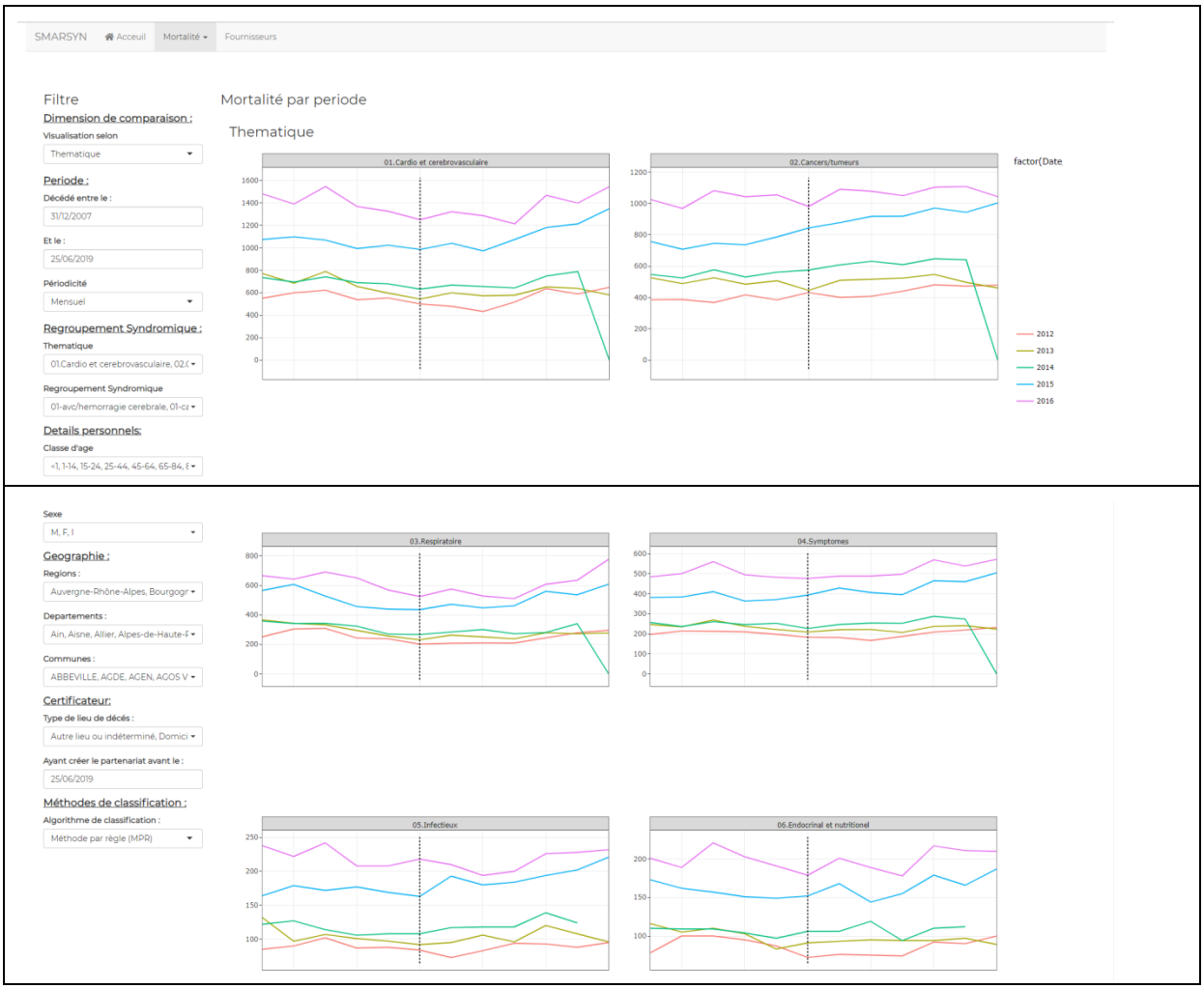


Figure 45 : Exemple d'écran pour la comparaison de l'évolution mensuelle des effectifs de décès pour différents regroupements syndromiques entre les années 2012 à 2016, pour une sélection de classes d'âges – France

Bibliographie

1. Archives départementales de Tarn-et-Garonne. Les registres paroissiaux et l'état-civil. 2014.

2. Noiriel G. L'identification des citoyens. Naissance de l'état civil républicain. Genèses Sciences sociales et histoire. 1993:3-28.

3. Bouvier-Colle MHV, J. ; Hatton, F. Mortalité et causes de décès en France. In: INSERM E, editor. 77-95. Paris: Edition INSERM. doin.; 1990.

4. Biraben J-N. Essai sur la statistique des causes de décès en France sous la Révolution et le Premier Empire. Annales de Démographie Historique. 1973:59-70.

5. Article L2223-42 Code général des collectivités territoriales, (2016).

6. Rey G. Les données des certificats de décès en France : processus de production et principaux types d'analyse. La Revue de Médecine Interne. 2016;37(10):685-93.

7. Iris Institute. Iris [17/06/2019]. Available from: https://www.dimdi.de/dynamic/en/classifications/iris-institute/#about-iris.

8. DREES, ABM, AFDPHE, CNR, DARES, DGS, et al. L'état de santé de la population en France. Paris: DREES, Santé publique France, 2017.

9. Rican S, Jougla E, Roudier Daval C, Gancel S, Gourdon G, Salem G, et al. Atlas de la mortalité par cancer en France métropolitaine : évolution 1970-2004. Institut national de la santé et de la recherche médicale(INSERM), 2008 2008-12. Report No.

10. Jouan. M. Réseau national de santé publique. Actualité et dossier en santé publique. 1995;13:10.

11. Houssin D, Coquin Y. Le dispositif français de sécurité sanitaire. Bull Epidémiol Hebd Hors série novembre. 2008.

12. Ledrans M, Isnard H. Impact sanitaire de la vague de chaleur d'août 2003 en France. Paris: Institut de Veille Sanitaire, 2003.

13. Lalande F, Legrain S, Valleron AJ, Meyniel D. Mission d'expertise et d'évaluation du système de santé pendant la canucule de 2003. Paris: Ministère de la santé, de la famille et des personnes handicapées, 2003.

14. Proctor ME, Blair KA, Davis JP. Surveillance data for waterborne illness detection: an assessment following a massive waterborne outbreak of Cryptosporidium infection. Epidemiol Infect. 1998;120(1):43-54.

15. James W. Buehler RSH, J. Marc Overhage, Daniel M. Sosin,Van Tong. Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks MMWR Morb Mortal Wkly Rep. 2004;53:1-11.

16. Triple S Project. Assessment of syndromic surveillance in Europe. Lancet. 2011;378(9806):1833-4.

17. Espino JU, Wagner MM, Tsui F-C, Su H-D, Olszewski RT, Liu Z, et al., editors. The Rods open source project: removing a barrier to syndromic surveillance. Medinfo; 2004.

18. Moore KM, Edgar BL, McGuinness D. Implementation of an automated, real-time public health surveillance system linking emergency departments and health units: rationale and methodology. Canadian Journal of Emergency Medicine. 2008;10(2):114-9.

19. Wu T-SJ, Shih F-YF, Yen M-Y, Wu J-SJ, Lu S-W, Chang KC-M, et al. Establishing a nationwide emergency department-based syndromic surveillance system for better public health responses in Taiwan. BMC Public Health. 2008;8(1):18.

20. Lombardo J, Burkom H, Elbert E, Magruder S, Lewis SH, Loschen W, et al. A systems overview of the electronic surveillance system for the early notification of community-based epidemics (ESSENCE II). J Urban Health. 2003;80(1):i32-i42.

21. CDC. Syndromic surveillance for bioterrorism following the attacks on the World Trade Center--New York City, 2001. MMWR Morb Mortal Wkly Rep. 2002;51 Spec No:13-5.

22. Leonardi G, Hajat S, Kovats R, Smith G, Cooper D, Gerard E. Syndromic surveillance use to detect the early effects of heat-waves: an analysis of NHS direct data in England. Sozial-und Präventivmedizin. 2006;51(4):194-201.

23. Elliot AJ, Hughes HE, Hughes TC, Locker TE, Shannon T, Heyworth J, et al. Establishing an emergency department syndromic surveillance system to support the London 2012 Olympic and Paralympic Games. Emerg Med J. 2012;29(12):954-60.

24. Meynard JB, Chaudet H, Texier G, Queyriaux B, Deparis X, Boutin JP. [Real time epidemiological surveillance within the armed forces: concepts, realities and prospects in France]. Rev Epidemiol Sante Publique. 2008;56(1):11-20.

25. Daudens E, Langevin S, Pellegrin L, Texier G, Dupuy B, Chaudet H, et al. Assessment of a military real-time epidemiological surveillance system by its users in French Guiana. Public Health. 2008;122(7):729.

26. Triple S Project. Inventory of Syndromic Surveillance Systems in Europe. Europe: 2012.

27. Euromomo hub. Euromomo [updated 25/05/201925/05/2019]. Available from: http://www.euromomo.eu/methods/rationale.html.

28. Caserio-Schönemann C, Bousquet V, Fouillet A, Henry V. Le système de surveillance syndromique SurSaUD (R) (The syndromic surveillance system SurSaUD). Bull Epidémiol Hebd. 2014;3:38-44.

29. Fouillet A, Franke F, Bousquet V, Durand C, Henry V, Golliot F, et al. Principe du traitement des données du système de surveillance syndromique SurSaUD® : indicateurs et méthodes d'analyse statistique. Numéro thématique. La surveillance syndromique en France en 2014. Bull Epidemiol Hebd. 2014(3-4):45-52.

30. Xu H, Fu Z, Shah A, Chen Y, Peterson NB, Chen Q, et al. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. AMIA Annual Symposium proceedings AMIA Symposium. 2011;2011:1564-72.

31. Sada Y, Hou J, Richardson P, El-Serag H, Davila J. Validation of Case Finding Algorithms for Hepatocellular Cancer From Administrative Data and Electronic Health Records Using Natural Language Processing. Med Care. 2016;54(2):e9-e14.

32. Savova GK, Fan J, Ye Z, Murphy SP, Zheng J, Chute CG, et al. Discovering peripheral arterial disease cases from radiology notes using natural language processing. AMIA Annual Symposium proceedings AMIA Symposium. 2010;2010:722-6.

33. Kim Y, Garvin JH, Heavirland J, Meystre SM, editors. Improving heart failure information extraction by domain adaptation. Medinfo; 2013.

34. Ludvigsson JF, Pathak J, Murphy S, Durski M, Kirsch PS, Chute CG, et al. Use of computerized algorithm to identify individuals in need of testing for celiac disease. Journal of the American Medical Informatics Association : JAMIA. 2013;20(e2):e306-e10.

35. Roch AM, Mehrabi S, Krishnan A, Schmidt HE, Kesterson J, Beesley C, et al. Automated pancreatic cyst screening using natural language processing: a new tool in the early detection of pancreatic cancer. HPB : the official journal of the International Hepato Pancreato Biliary Association. 2015;17(5):447-53.

36. Flynn RW, Macdonald TM, Schembri N, Murray GD, Doney AS. Automated data capture from free-text radiology reports to enhance accuracy of hospital inpatient stroke codes. Pharmacoepidemiol Drug Saf. 2010;19(8):843-7.

37. Zuccon G, Wagholikar AS, Nguyen AN, Butt L, Chu K, Martin S, et al. Automatic Classification of Free-Text Radiology Reports to Identify Limb Fractures using Machine Learning and the SNOMED CT Ontology. AMIA Joint Summits on Translational Science proceedings AMIA Joint Summits on Translational Science. 2013;2013:300-4.

38. Yang H, Spasic I, Keane JA, Nenadic G. A text mining approach to the prediction of disease status from clinical discharge summaries. J Am Med Inform Assoc. 2009;16(4):596-600.

39. Chen L, Vallmuur K, Nayak R. Injury narrative text classification using factorization model. BMC Med Inform Decis Mak. 2015;15 Suppl 1:S5.

40. McKenzie K, Campbell MA, Scott DA, Discoll TR, Harrison JE, McClure RJ. Identifying work related injuries: comparison of methods for interrogating text fields. BMC Med Inf Decis Making. 2010;10:19-.

41. Tvardik N, Kergourlay I, Bittar A, Segond F, Darmoni S, Metzger M-H. Accuracy of using natural language processing methods for identifying healthcare-associated infections. Int J Med Inf. 2018;117:96-102.

42. Branch-Elliman W, Strymish J, Kudesia V, Rosen AK, Gupta K. Natural Language Processing for Real-Time Catheter-Associated Urinary Tract Infection Surveillance: Results of a Pilot Implementation Trial. Infect Control Hosp Epidemiol. 2015;36(9):1004-10.

43. Gundlapalli AV, Divita G, Redd A, Carter ME, Ko D, Rubin M, et al. Detecting the presence of an indwelling urinary catheter and urinary symptoms in hospitalized patients using natural language processing. J Biomed Inform. 2017;71s:S39-s45.

44. Gerbier S, Yarovaya O, Gicquel Q, Millet A-L, Smaldore V, Pagliaroli V, et al. Evaluation of natural language processing from emergency department computerized medical records for intra-hospital syndromic surveillance. BMC Med Inf Decis Making. 2011;11(1):50.

45. Carrell DS, Cronkite D, Palmer RE, Saunders K, Gross DE, Masters ET, et al. Using natural language processing to identify problem usage of prescription opioids. Int J Med Inform. 2015;84(12):1057-64.

46. Iyer SV, Harpaz R, LePendu P, Bauer-Mehren A, Shah NH. Mining clinical text for signals of adverse drug-drug interactions. J Am Med Inform Assoc. 2014;21(2):353-62.

47. Davis K, Staes C, Duncan J, Igo S, Facelli JC. Identification of pneumonia and influenza deaths using the Death Certificate Pipeline. BMC Med Inform Decis Mak. 2012;12:37.

48. Koopman B, Karimi S, Nguyen A, McGuire R, Muscatello D, Kemp M, et al. Automatic classification of diseases from free-text death certificates for real-time surveillance. BMC Med Inf Decis Making. 2015;15:53.

49. Koopman B, Nguyen A, Cossio D, Courage M, Francois G. Extracting cancer mortality statistics from free-text death certificates. Proceedings of the 23rd Australasian Document Computing Symposium, ADCS'18. 2018:6:1–6:4.

50. Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N. Extracting cancer mortality statistics from death certificates: A hybrid machine learning and rule-based approach for common and rare cancers. Artif Intell Med. 2018.

51. Trinidad JP, Warner M, Bastian BA, Minino AM, Hedegaard H. Using Literal Text From the Death Certificate to Enhance Mortality Statistics: Characterizing Drug Involvement in Deaths. Natl Vital Stat Rep. 2016;65(9):1-15.

52. CLEF Initiative. CLEF 2018 Conference and Labs of the Evaluation Forum

Information Access Evaluation meets Multilinguality, Multimodality, and Visualization 2018 [29/06/2019]. Available from: http://clef2018.clef-initiative.eu/index.php?page=Pages/labs_info.html.

53. Névéol A, Anderson RN, Cohen KB, Grouin C, Lavergne T, Rey G, et al., editors. CLEF eHealth 2017 Multilingual Information Extraction task overview: ICD10 coding of death certificates in English and French. CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS; 2017.

54. Neveol A, Cohen KB, Grouin C, Hamon T, Lavergne T, Kelly L, et al. Clinical Information Extraction at the CLEF eHealth Evaluation lab 2016. CEUR workshop proceedings. 2016;1609:28-42.

55. Zweigenbaum P, Lavergne T, editors. Hybrid methods for ICD-10 coding of death certificates. Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis; 2016.

56. Zweigenbaum P, Lavergne T, editors. Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates2017: CLEF.

57. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128-44.

58. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. J Biomed Inf. 2018;77:34-49.

59. Alpaydin E. Introduction to Machine Learning: The MIT Press; 2010. 584 p.

60. Cover T, Hart P. Nearest neighbor pattern classification. IEEE transactions on information theory. 1967;13(1):21-7.

61. Mitchell TM. Machine learning. WCB. McGraw-Hill Boston, MA:; 1997.

62. Schölkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond: MIT press; 2002.

63. Personnaz L, Rivals I. Réseaux de neurones formels pour la modélisation, la commande et la classification: CNRS; 2003.

64. Berkhin P. A survey of clustering data mining techniques. Grouping multidimensional data: Springer; 2006. p. 25-71.

65. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA. 2010;17(5):507-13.

66. Aronson AR, editor Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp; 2001: American Medical Informatics Association.

67. German RR, Lee L, Horan J, Milstein R, Pertowski C, Waller M. Updated guidelines for evaluating public health surveillance systems. MMWR Recomm Rep. 2001;50(1-35).

68. Gergonne B, Mazick A, ODonnell J. A European algorithm for a common monitoring of mortality across Europe 2011 [cited 7]. Available from: http://www.euromomo.eu/methods/pdf/wp7_report.pdf#page=1&zoom=auto,-274,842.

69. Jackson ML. Confounding by season in ecologic studies of seasonal exposures and outcomes: examples from estimates of mortality due to influenza. Ann Epidemiol. 2009;19(10):681-91.

70. Eilstein D, Zeghnoon A, Le Tertre A, Cassadou S, Declercq C, Filleul L, et al. [Short-term modeling of the effect of air pollution on health: analytical methods of time series data]. Rev Epidemiol Sante Publique. 2004;52(6):583-9.

71. Cox NJ, Subbarao K. Global epidemiology of influenza: past and present. Annu Rev Med. 2000;51:407-21.

72. Paget J, Marquet R, Meijer A, van der Velden K. Influenza activity in Europe during eight seasons (1999-2007): an evaluation of the indicators used to measure activity and an assessment of the timing, length and course of peak activity (spread) across Europe. BMC Infect Dis. 2007;7:141.

73. Hajat S, Kovats RS, Lachowycz K. Heat-related and cold-related deaths in England and Wales: who is at risk? Occup Environ Med. 2007;64.

74. Équipes de surveillance de la grippe. Surveillance épidémiologique, clinique et virologique de la grippe en France métropolitaine : saison 2011-2012 (Epidemiologic, virologic and clinic Influenza surveillance in metropolitan region in France: 2011-2012 season). Bull Épidemiol Hebd. 2012;38:424-7.

75. Équipes de surveillance de la grippe. Surveillance épidémiologique et virologique de la grippe en France, saison 2012-2013 (Epidemiologic and virologic Influenza surveillance in France. 2012-2013 season). Bull Épidemiol Hebd. 2013;32:394-401.

76. Équipes de surveillance de la grippe. Surveillance épidémiologique et virologique de la grippe en France métropolitaine. Saison 2013-2014. (Epidemiologic and virologic Influenza surveillance in metropolitan regions in France: 2013-2014 season). Bull Epidémiol Hebd. 2014;28:460-5.

77. Équipes de surveillance de la grippe. Surveillance de la grippe en France métropolitaine. Saison 2014-2015 (Influenza surveillance in metropolitan regions in France. 2014-2015 season). Bull Epidémiol Hebd. 2015;32-33:593-8

78. Équipes de surveillance de la grippe. Surveillance de la grippe en France métropolitaine, saison 2015-2016 (Influenza surveillance in metropolitan regions in France: 2015-2016 season). Bull Epidémiol Hebdomadaire. 2016;32-33 558-63.

79. Carrat F, Flahault A, Boussard E, Farran N, Dangoumau L, Valleron A-J. Surveillance of Influenza-Like Illness in France. The Example of the 1995/1996 Epidemic. Journal of Epidemiology and Community Health (1979-). 1998;52:32S-8S.

80. Santé Publique France. Chaleur et Santé/Actualité/Archive (Heat and Health /News/ Archive) Paris CoreTechs; [22/08/2017].

81. Santé publique France. Bulletin de veille sanitaire Antilles-Guyane. n°3-4-5 - Septembre-Novembre 2014. (Health monitoring bulletin for the French West Indies and Guiana n°3-4-5 - September-November 2014) Paris: CoreTechs; 2014 [cited 2017 27/07/2017].

82. Baghdadi Y, Gallay A, Caserio-Schönemann C, Fouillet A. Evaluation of the French reactive mortality surveillance system supporting decision making. 2018.

83. Molbak K, Espenhain L, Nielsen J, Tersago K, Bossuyt N, Denissov G, et al. Excess mortality among the elderly in European countries, December 2014 to February 2015. Euro Surveill. 2015;20(11).

84. Wagner V, Ung A, Calmet C. Évolution des vagues de chaleur et de la mortalité associée en France, 2004-2014. Bull Epidémiol Hebd. 2018:16-7.

85. Matias G, Taylor R, Haguinet F, Schuck-Paim C, Lustig R, Shinde V. Estimates of mortality attributable to influenza and RSV in the United States during 1997-2009 by influenza type or subtype, age, cause of death, and risk status. Influenza Other Respir Viruses. 2014;8(5):507-15.

86. Thompson WW, Shay DK, Weintraub E, Brammer L, Cox N, Anderson LJ, et al. Mortality associated with influenza and respiratory syncytial virus in the United States. JAMA. 2003;289(2):179-86.

87. Donaldson GC, Keatinge WR. Excess winter mortality: influenza or cold stress? Observational study. BMJ : British Medical Journal. 2002;324(7329):89-90.

88. Mayor S. Cold weather kills far more people than hot weather, study shows. BMJ. 2015;350:h2740.

89. Équipes de surveillance de la grippe. Surveillance de la grippe en France, saison 2017-2018. . Bull Épidémiol Hebd 2018;34:664-74.

90. Basu R, Samet JM. Relation between Elevated Ambient Temperature and Mortality: A Review of the Epidemiologic Evidence. Epidemiol Rev. 2002;24(2):190-202.

91. Bunker A, Wildenhain J, Vandenbergh A, Henschke N, Rocklöv J, Hajat S, et al. Effects of Air Temperature on Climate-Sensitive Mortality and Morbidity Outcomes in the Elderly; a Systematic Review and Meta-analysis of Epidemiological Evidence. EBioMedicine. 2016;6:258-68.

92. Gasparrini A, Guo Y, Hashizume M, Lavigne E, Zanobetti A, Schwartz J, et al. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. The Lancet. 2015;386(9991):369-75.

93. Basu R. High ambient temperature and mortality: a review of epidemiologic studies from 2001 to 2008. Environ Health. 2009;8(1):40.

94. Ye X, Wolff R, Yu W, Vaneckova P, Pan X, Tong S. Ambient Temperature and Morbidity: A Review of Epidemiological Evidence. Environ Health Perspect. 2012;120(1):19-28.

95. Pialoux G, Gauzere BA, Jaureguiberry S, Strobel M. Chikungunya, an epidemic arbovirosis. Lancet Infect Dis. 2007;7(5):319-27.

96. Economopoulou A, Dominguez M, Helynck B, Sissoko D, Wichmann O, Quenel P, et al. Atypical Chikungunya virus infections: clinical manifestations, mortality and risk factors for severe disease during the 2005–2006 outbreak on Reunion. Epidemiol Infect. 2009;137(04):534-41.

97. Josseran L. Chikungunya Disease Outbreak, Reunion Island-Volume 12, Number 12—December 2006-Emerging Infectious Disease journal-CDC. 2006.

98. Lefeuvre D, Pavillon G, Aouba A, Lamarche-Vadel A, Fouillet A, Jougla E, et al. Quality comparison of electronic versus paper death certificates in France, 2010. Population health metrics. 2014;12(1):3.

99. Lassalle M, Caserio-Schönemann C, Gallay A, Rey G, Fouillet A. Pertinence of electronic death certificates for real-time surveillance and alert, France, 2012–2014. Public Health. 2017;143:85-93.

100. Fouillet A, Pavillon G, Vicente P, Caillere N, Aouba A, Jougla E, et al. La certification électronique des décès, France, 2007-2011. Bull Epidemiol Hebd. 2012(1):7-10.

101. Ministère des affaires sociales et de la santé. Instruction DGS/DAD/BSIIP no 2013-291 du 12 juillet 2013 relative au déploiement dans les établissements de santé de la certification électronique en matière de certificats de décès. 2013.

102. Ministère des affaires sociales et de la santé, Ministère des Familles dlEedDdF. Instruction no DGS/DAD/BSIIP/DGOS/2016/302 du 7 octobre 2016 relative au déploiement dans les établissements de santé de la certification électronique en matière de certificats de décès. In: Santé, editor. 2016.

103. Gabet A, Danchin N, Juilliere Y, Olie V. Acute coronary syndrome in women: rising hospitalizations in middle-aged French women, 2004-14. Eur Heart J. 2017;38(14):1060-5.

104. Gabet A, Juilliere Y, Lamarche-Vadel A, Vernay M, Olie V. National trends in rate of patients hospitalized for heart failure and heart failure mortality in France, 2000-2012. Eur J Heart Fail. 2015;17(6):583-90.

105. Robert M, Juilliere Y, Gabet A, Kownator S, Olie V. Time trends in hospital admissions and mortality due to abdominal aortic aneurysms in France, 2002-2013. Int J Cardiol. 2017;234:28-32.

106. Dalkey N, Helmer O. An experimental application of the Delphi method to the use of experts. Management science. 1963;9(3):458-67.

107. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K, Al-Garadi MA. Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. PLoS One. 2017;12(2):e0170242-e.

108. Butt L, Zuccon G, Nguyen A, Bergheim A, Grayson N. Classification of cancer-related death certificates using machine learning. The Australasian medical journal. 2013;6(5):292-9.

109. Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N. Automatic ICD-10 classification of cancers from free-text death certificates. Int J Med Inf. 2015;84(11):956-65.

110. Miftahutdinov Z, Tutubalina E, editors. KFU at CLEF eHealth 2017 Task 1: ICD-10 Coding of English Death Certificates with Recurrent Neural Networks. CLEF (Working Notes); 2017.

111. Cabot C, Soualmia LF, Darmoni SJ, editors. SIBM at CLEF eHealth Evaluation Lab 2017: Multilingual Information Extraction with CIM-IND. CLEF (Working Notes); 2017.

112. Atutxa A, Casillas A, Ezeiza N, Goenaga I, Fresno V, Gojenola K, et al., editors. IxaMed at CLEF eHealth 2018 Task 1: ICD10 Coding with a Sequence-to-Sequence approach2018: CLEF.

113. Duarte F, Martins B, Pinto CS, Silva MJ. Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text. J Biomed Inform. 2018;80:64-77.

114. Thompson WW, Shay DK, Weintraub E, et al. Mortality associated with influenza and respiratory syncytial virus in the united states. JAMA. 2003;289(2):179-86.

115. Muscatello DJ, Morton PM, Evans I, Gilmour R. Prospective surveillance of excess mortality due to influenza in New South Wales: feasibility and statistical approach. Communicable diseases intelligence quarterly report. 2008;32(4):435-42.

116. Shah AD, Martinez C, Hemingway H. The freetext matching algorithm: a computer program to extract diagnoses and causes of death from unstructured text in electronic health records. BMC Med Inform Decis Mak. 2012;12:88.

117. Baghdadi Y, Gallay A, Caserio-Schönemann C, Thiam M-M, Fouillet A. Towards real-time mortality surveillance by medical causes of death: A strategy of analysis for alert. Rev Epidemiol Sante Publique. 2018;66:S402.

118. Désesquelles A, Gamboni A, Demuru E. On ne meurt qu'une fois... mais de combien de causes? Population & Societies. 2016(534):1.

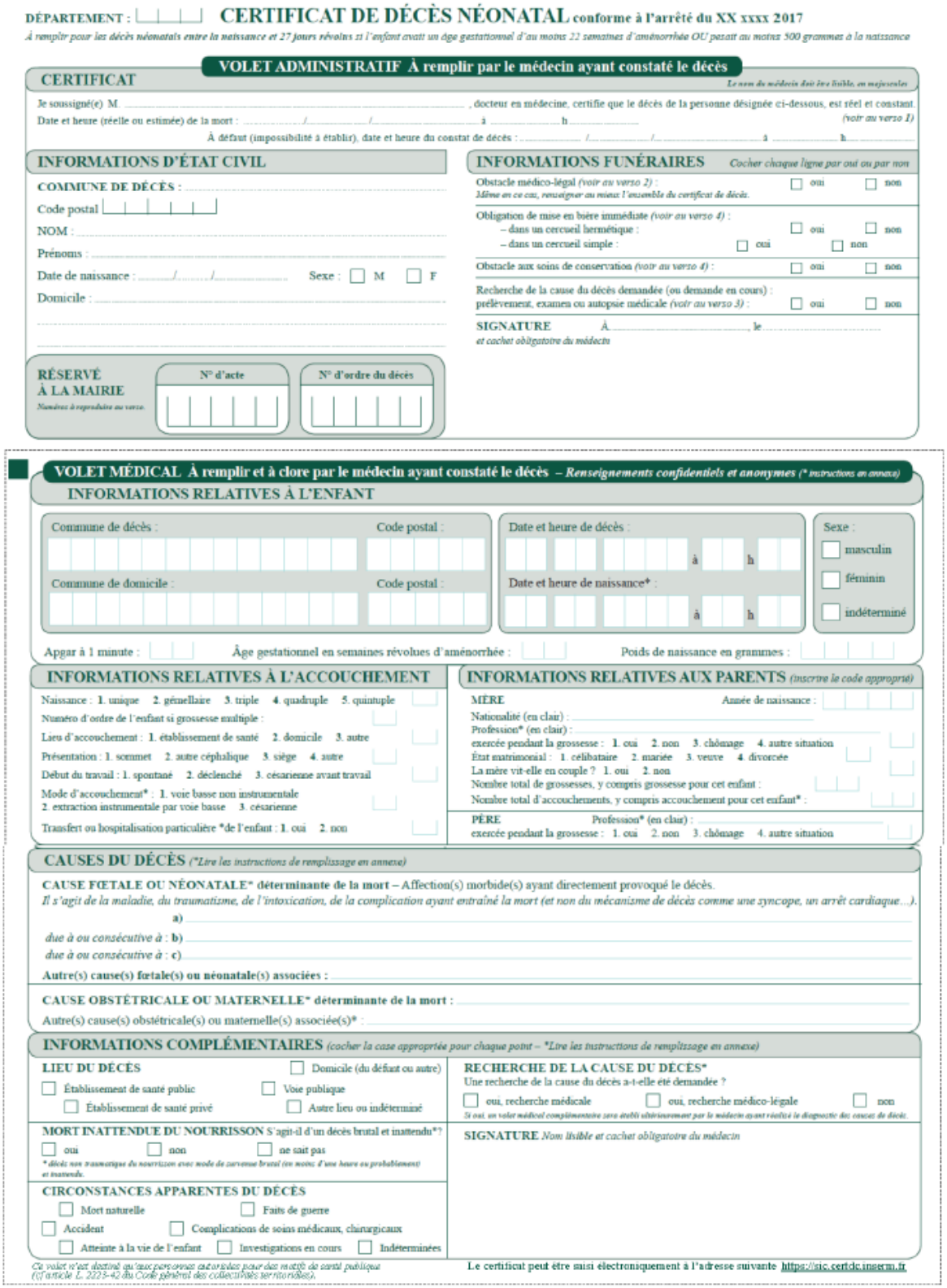
119. Désesquelles A, Meslé F. Intérêt de l'analyse des causes multiples dans l'étude de la mortalité aux grands âges : l'exemple français. Cahiers québécois de démographie. 2004;33(1):83-116.

120. Piffaretti C, Moreno-Betancur M, Lamarche-Vadel A, Rey G. Quantifying cause-related mortality by weighting multiple causes of death. Bull WHO. 2016;94(12):870-9.

121. Fouillet A, Pigeon D, Carton I, Robert A, Pontais I, Caserio-Schönemann I, et al. Evolution de la certification électronique des décès en France depuis 2011. Bulletin Epidémiol Hebdomadaire. 2019;Soumis.

Annexes

Annexe 1 : Certificat de décès néonatal avant 2018



DÉPARTEMENT : | | | CERTIFICAT DE DÉCÈS NÉONATAL conforme à l'arrêté du XX xxxx 2017

À remplir pour les décès néonatals entre la naissance et 27 jours révolus si l'enfant avait un âge gestationnel d'au moins 22 semaines d'aménorrhée OU pesait au moins 500 grammes à la naissance

VOLET ADMINISTRATIF À remplir par le médecin ayant constaté le décès

CERTIFICAT

Le nom du médecin doit être lisible, en majuscules

Je soussigné(e) M. ..... , docteur en médecine, certifie que le décès de la personne désignée ci-dessous, est réel et constant.

Date et heure (réelle ou estimée) du décès : ... / ... / ... à ... h ... *(voir au verso 1)*

À défaut (impossibilité à établir), date et heure du constat de décès : ... / ... / ... à ... h ...

INFORMATIONS D'ÉTAT CIVIL

COMMUNE DE DÉCÈS : .....

Code postal | | | | | |

NOM : .....

Prénoms : .....

Date de naissance : ... / ... / ... Sexe : ☐ M ☐ F

Domicile : .....

INFORMATIONS FUNÉRAIRES *Cocher chaque ligne par oui ou par non*

Obstacle médico-légal *(voir au verso 2)* : ☐ oui ☐ non
Même en ce cas, renseigner au mieux l'ensemble du certificat de décès.

Obligation de mise en bière immédiate *(voir au verso 4)* :
– dans un cercueil hermétique : ☐ oui ☐ non
– dans un cercueil simple : ☐ oui ☐ non

Obstacle aux soins de conservation *(voir au verso 4)* : ☐ oui ☐ non

Recherche de la cause du décès demandée (ou demande en cours) : prélèvement, examen ou autopsie médicale *(voir au verso 3)* : ☐ oui ☐ non

SIGNATURE À ..... le .....
et cachet obligatoire du médecin

RÉSERVÉ À LA MAIRIE
Numéros à reproduire au verso.

N° d'acte | | | | | | |

N° d'ordre du décès | | | | | | | | |

VOLET MÉDICAL À remplir et à clore par le médecin ayant constaté le décès – *Renseignements confidentiels et anonymes (* instructions en annexe)*

INFORMATIONS RELATIVES À L'ENFANT

Commune de décès : ..... Code postal : .....

Commune de domicile : ..... Code postal : .....

Date et heure de décès : ... à ... h ...

Date et heure de naissance* : ... à ... h ...

Sexe : ☐ masculin ☐ féminin ☐ indéterminé

Apgar à 1 minute : .... Âge gestationnel en semaines révolues d'aménorrhée : .... Poids de naissance en grammes : .....

INFORMATIONS RELATIVES À L'ACCOUCHEMENT

Naissance : 1. unique 2. gémellaire 3. triple 4. quadruple 5. quintuple ....
Numéro d'ordre de l'enfant si grossesse multiple : ....
Lieu d'accouchement : 1. établissement de santé 2. domicile 3. autre ....
Présentation : 1. sommet 2. autre céphalique 3. siège 4. autre ....
Début du travail : 1. spontané 2. déclenché 3. césarienne avant travail ....
Mode d'accouchement* : 1. voie basse non instrumentale 2. extraction instrumentale par voie basse 3. césarienne ....
Transfert ou hospitalisation particulière *de l'enfant : 1. oui 2. non ....

INFORMATIONS RELATIVES AUX PARENTS *(inscrire le code approprié)*

MÈRE Année de naissance : .....
Nationalité (en clair) : .....
Profession* (en clair) : .....
exercée pendant la grossesse : 1. oui 2. non 3. chômage 4. autre situation ....
État matrimonial : 1. célibataire 2. mariée 3. veuve 4. divorcée ....
La mère vit-elle en couple ? 1. oui 2. non ....
Nombre total de grossesses, y compris grossesse pour cet enfant : ....
Nombre total d'accouchements, y compris accouchement pour cet enfant* : ....

PÈRE Profession* (en clair) : .....
exercée pendant la grossesse : 1. oui 2. non 3. chômage 4. autre situation ....

CAUSES DU DÉCÈS *(*Lire les instructions de remplissage en annexe)*

CAUSE FŒTALE OU NÉONATALE* déterminante de la mort – Affection(s) morbide(s) ayant directement provoqué le décès.
Il s'agit de la maladie, du traumatisme, de l'intoxication, de la complication ayant entraîné la mort (et non du mécanisme de décès comme une syncope, un arrêt cardiaque...).

a) .....
due à ou consécutive à : b) .....
due à ou consécutive à : c) .....
Autre(s) cause(s) fœtale(s) ou néonatale(s) associées : .....

CAUSE OBSTÉTRICALE OU MATERNELLE* déterminante de la mort : .....
Autre(s) cause(s) obstétricale(s) ou maternelle(s) associée(s)* : .....

INFORMATIONS COMPLÉMENTAIRES *(cocher la case appropriée pour chaque point – *Lire les instructions de remplissage en annexe)*

LIEU DU DÉCÈS
☐ Domicile (du défunt ou autre) ☐ Établissement de santé public ☐ Voie publique ☐ Établissement de santé privé ☐ Autre lieu ou indéterminé

MORT INATTENDUE DU NOURRISSON S'agit-il d'un décès brutal et inattendu*? ☐ oui ☐ non ☐ ne sait pas
** décès non traumatique du nourrisson avec mode de survenue brutal (en moins d'une heure ou probablement) et inattendu.*

CIRCONSTANCES APPARENTES DU DÉCÈS
☐ Mort naturelle ☐ Faits de guerre ☐ Accident ☐ Complications de soins médicaux, chirurgicaux ☐ Atteinte à la vie de l'enfant ☐ Investigations en cours ☐ Indéterminées

RECHERCHE DE LA CAUSE DU DÉCÈS*
Une recherche de la cause du décès a-t-elle été demandée ?
☐ oui, recherche médicale ☐ oui, recherche médico-légale ☐ non
Si oui, un volet médical complémentaire sera établi ultérieurement par le médecin ayant réalisé le diagnostic des causes de décès.

SIGNATURE *Nom lisible et cachet obligatoire du médecin*

Ce volet n'est destiné qu'aux personnes autorisées pour des motifs de santé publique (cf. article L. 2223-42 du Code général des collectivités territoriales).

Le certificat peut être saisi électroniquement à l'adresse suivante https://sic.certdc.inserm.fr

Source : Arrêté du 17 juillet 2017 relatif aux deux modèles du certificat de décès (https://www.legifrance.gouv.fr/eli/arrete/2017/7/17/PRMX1720890A/jo/texte)

Annexe 2 : Couverture de la mortalité par département et par région de France métropolitaine entre 2011 et 2013

Régions/départements	2011 (%)	2012 (%)	2013 (%)
ILE-DE-FRANCE	89,1	88,9	89,6
75 Paris	97,5	97,7	98,7
77 Seine et Marne	80,6	81,1	72,3
78 Yvelines	70,9	69,5	71,1
91 Essonne	80,4	79,3	81,1
92 Hauts de Seine	94,0	95,0	98,1
93 Seine-Saint-Denis	91,8	91,2	94,6
94 Val de Marne	94,8	94,3	94,3
95 Val-d'Oise	90,5	91,3	92,4
PROVENCE-ALPES-COTE-D'AZUR	87,1	87,2	87,7
4 Alpes-de-Hautes-Provence	58,2	58,5	56,9
5 Hautes-Alpes	69,0	68,9	71,7
6 Alpes-Maritimes	89,6	89,1	90,0
13Bouches-du-Rhône	91,0	91,4	91,3
83 Var	85,5	86,0	86,6
84 Vaucluse	84,5	84,4	85,4
GRAND-EST	82,0	81,9	80,0
8 Ardennes	80,5	80,5	81,3
10 Aube	81,3	81,5	82,5
51 Marne	88,8	88,6	89,2
52 Haute Marne	78,7	80,8	80,4
54 Meurthe et Moselle	86,5	87,4	87,2
55 Meuse	74,8	76,1	76,6
57 Moselle	84,1	82,2	70,3
67 Bas-Rhin	79,7	79,7	79,9
68 Haut-Rhin	79,8	79,4	79,6
88 Vosges	74,7	75,8	76,9
BRETAGNE	82,6	81,6	81,4
22 Côtes-d'Armor	74,9	74,1	74,4
29 Finistère	86,6	86,8	86,4
35 Ille et Vilaine	84,9	85,1	85,7
56 Morbihan	81,5	77,1	76,4
PAYS-DE-LA-LOIRE	78,3	77,5	79,2
44 Loire-Atlantique	86,7	84,4	87,5
49 Maine-et-Loire	75,5	76,9	75,9
53 Mayenne	68,4	68,3	69,6
72 Sarthe	77,5	75,5	77,8
85 Vendée	71,0	71,3	72,9

Régions/départements	2011 (%)	2012 (%)	2013 (%)
NORMANDIE	77,5	77,2	78,5
14 Calvados	78,4	77,2	78,7
27 Eure	64,0	64,2	65,6
50 Manche	72,1	72,3	74,0
61 Orne	69,1	69,1	70,4
76 Seine-Maritime	86,0	86,1	87,1
HAUT-DE-FRANCE	76,8	77,2	77,1
2 Aisne	70,9	71,0	71,1
59 Nord	83,8	84,4	83,6
60 Oise	69,6	66,6	69,8
62 Pas-de-Calais	72,5	73,4	73,3
80 Somme	71,2	73,2	72,5
BOURGOGNE-FRANCHE-COMTE	73,4	71,6	73,1
21 Côted'Or	78,8	75,8	78,4
25 Doubs	79,9	79,7	81,2
39 Jura	68,5	67,1	68,6
58 Nièvre	75,3	72,4	72,9
70 Haute-Saône	65,7	61,3	64,4
71 Saône et Loire	70,9	70,3	70,6
89 Yonne	65,7	63,5	65,3
90 Territoire de Belfort	78,3	76,8	78,2
CENTRE-VAL-DE-LOIRE	72,8	72,0	72,8
18 Cher	75,1	74,6	74,6
28 Eure et Loir	76,1	74,7	77,5
36 Indre	69,4	70,0	71,1
37 Indre et Loire	77,2	76,0	77,1
41 Loir et Cher	62,9	61,2	64,0
45 Loiret	72,2	71,9	70,4
CORSE	68,3	65,8	68,0
2A Corse du Sud	78,1	77,3	79,4
2B Haute Corse	59,1	55,9	58,0

Régions/départements	2011 (%)	2012 (%)	2013 (%)
NOUVELLE-AQUITAINE	70,7	70,4	71,3
16 Charente	66,5	66,2	68,4
17 Charente-Maritime	69,6	70,4	70,5
19 Corrèze	63,9	62,7	63,8
23 Creuse	47,0	49,3	46,8
24 Dordogne	53,7	53,2	54,1
33 Gironde	78,1	78,3	78,9
40 Landes	61,9	64,3	65,9
47 Lot et Garonne	72,8	71,2	71,8
64 Pyrénées-Atlantiques	77,3	75,6	78,1
79 Deux-sèvres	66,1	63,3	64,1
86 Vienne	78,6	78,1	78,0
87 Haute-Vienne	76,5	75,3	76,6
AUVERGNE-RHONE-ALPES	69,1	68,1	68,6
1 Ain	50,9	50,0	50,8
3 Ardèche	69,9	70,5	70,2
7 Allier	58,3	58,5	58,2
15 Cantal	66,3	66,5	65,9
26 Drôme	65,8	65,9	66,9
38 Isère	66,5	67,5	68,6
42 Loire	76,0	76,7	76,3
43 Haute-Loire	40,2	40,4	40,6
63 Puy de Dôme	64,4	59,0	61,5
69 Rhône	80,9	81,1	81,0
73 Savoie	68,4	69,2	69,7
74 Haute-Savoie	73,4	65,2	65,4

Régions/départements	2011 (%)	2012 (%)	2013 (%)
OCCITANIE	67,5	67,1	67,2
9 Ariège	59,6	59,9	61,6
11 Aude	64,3	62,3	63,0
12 Aveyron	57,0	56,3	55,5
30 Gard	66,0	66,0	68,1
31 Haute-Garonne	77,7	76,5	77,2
32 Gers	42,1	44,6	45,7
34 Hérault	77,4	77,5	76,6
46 Lot	58,7	58,5	58,1
48 Lozère	43,8	39,5	42,9
65 Hautes-Pyrénées	67,8	68,2	67,7
66 Pyrénées-Orientales	58,8	57,2	55,1
81 Tarn	69,4	71,5	71,1
82 Tarn et Garonne	63,5	62,1	62,9

Annexe 3 : Différence entre la couverture max et min par région sur la période 2011 à 2013

	Différences des couvertures max-min						
	≤ 1 an	**2-14 ans**	**15-44 ans**	**45-64 ans**	**65-74 ans**	**75-84 ans**	**≥ 85 ans**
ILE-DE-FRANCE	1,9	3,9	2,8	0,1	1,3	0,3	1,9
PROVENCE-ALPES-COTE D'AZUR	1,9	8,4	0,4	1,8	0,2	1,2	0,3
GRAND-EST	5,3	3,2	2,0	2,1	3,8	2,4	1,3
BRETAGNE	2,4	11,3	2,7	0,6	0,7	2,3	1,1
PAYS-DE-LA-LOIRE	2,6	6,3	2,4	1,7	1,0	1,5	2,0
NORMANDIE	1,8	9,4	0,9	0,4	1,1	1,0	2,0
HAUT-DE-FRANCE	0,5	1,8	2,3	0,3	1,0	0,9	1,0
BOURGOGNE-FRANCHE-COMTE	3,3	13,8	2,9	3,2	2,1	0,2	1,6
CENTRE-VAL-DE-LOIRE	1,8	15,1	2,3	1,7	0,4	1,7	0,4
NOUVELLE-AQUITAINE	5,9	8,1	1,1	1,6	0,7	0,7	0,7
AUVERGNE RHONE-ALPES	4,4	5,5	0,9	0,8	0,5	0,7	1,1
CORSE	0,0	75,0	10,0	3,8	5,6	1,8	4,4
OCCITANIE	2,3	4,8	0,1	1,2	1,1	0,6	0,1
MARTINIQUE	5,3	9,1	2,8	2,0	2,3	1,4	1,5
LA REUNION	2,4	10,5	0,6	2,0	2,7	3,0	1,3
GUADELOUPE	13,6	11,4	2,8	3,1	5,0	5,3	2,1
GUYANE	3,7	13,4	12,0	4,4	9,9	9,9	12,0

Annexe 4 : Définition des 7 regroupements syndromiques et exemples d'expressions de causes de décès

Regroupement syndromique	Liste de codes CIM-10	Extrait des expressions des causes de décès inclues dans le regroupement syndromique
Grippe	J09, J10, J11	**109 expressions, dont :** affection grippale, affection respiratoire grippale, bronchite aigüe grippale, bronchite aigüe post-grippale, bronchite grippale, bronchite grippale aigüe, bronchite post-grippale, bronchopathie grippale, bronchopneumopathie grippale, complication grippale, complications grippe, contexte épidémie grippe, contexte grippal, décompensation grippale, encéphalite grippale, épisode grippal, épisode viral grippal, état grippal, état grippal aigu, état grippal épidémique, état infectieux grippal, grippe, grippe A, grippe A contexte épidémie, grippe aigüe, grippe aviaire, grippe B, grippe décompensée, grippe endémique, grippe épidémique, grippe espagnole, infection broncho-pulmonaire type grippal, infection grippale, infection virale type grippal, invasion grippale, syndrome grippal, tableau grippal, terrain grippal, virose grippale….
Insuffisance respiratoire aigue	J96	**60 expressions, dont :** anoxie pulmonaire, atteinte respiratoire aigüe, atteinte ventilatoire, complication respiratoire aigüe, complications ventilatoires, décompensation respiratoire aigüe spastique, défaillance ventilatoire, dépression ventilatoire, détresse pulmonaire, détresse respiratoire, détresse respiratoire aigüe, détresse ventilatoire, détresses respiratoires, détresses respiratoires itératives, difficulté respiratoire aigüe, DRA, état ventilatoire précaire, exacerbation insuffisance respiratoire sévère, inefficacité respiratoire aigüe brutale, insuffisance pulmonaire aigüe, insuffisance pulmonaire subaigüe, insuffisance respiratoire aigüe, insuffisance respiratoire aigüe anoxique, insuffisance respiratoire obstructive aigüe, insuffisance respiratoire subaigüe, insuffisance ventilatoire aigüe, IR aigüe, IRA, IRA obstructive, poumon choc, poussée aigüe insuffisance respiratoire, poussée insuffisance respiratoire aigüe, souffrance respiratoire aigüe, syndrome défaillance respiratoire aigüe, syndrome insuffisance respiratoire aigüe, troubles respiratoires aigus…

Infections respiratoires aigües basses	J12, J13, J14, J15, J16, J17, J18, J20, J22, J40, J440, J851,J852	**2402 expressions, dont :** abcédation lobe supérieur, abcédation poumon, abcédation pulmonaire, abcès bronchopulmonaire, abcès lobe inférieur poumon, abcès lobe supérieur poumon, abcès multiples poumon, abcès poumon, accès bronchitique, accident infectieux pulmonaire aigu, acute chest syndrome, affection aigüe saisonnière, affection bronchique aigüe, affection virale respiratoire, BPCO chronique surinfectée, BPCO infectée, BPCO infectieuse, bronchio-alvéolite, bronchiolo-alvéolite bilatérale, bronchite, bronchite Acinetobacter, bronchite adénovirus, bronchite aigüe, bronchite aigüe surinfectée, bronchite aigüe virale, bronchite infectée, bronchite infectieuse, bronchite infectieuse aigüe, broncho-alvéolite aigüe, bronchopathie aigüe, bronchopathie aigüe hypoxémique, bronchopathie aigüe pulmonaire, broncho-pleuro-pneumopathie, bronchopneumonie, bronchopneumopathie, bronchopneumopathie aigüe, broncho-pneumopathie fébrile aigüe, bronchopneumopathie infectée, bronchopneumopathie unilatérale, choc infectieux pulmonaire, choc septique pulmonaire, colonisation pulmonaire, complication bronchique, complication broncho-pulmonaire décubitus, décompensation bronchique, décompensation bronchite aigüe, encombrement bronchique aigu fébrile, encombrement pulmonaire infectieux, encombrement pulmonaire surinfecté, encombrement purulent voies respiratoires, épisode aigu pulmonaire, épisode viral broncho-pulmonaire, épisode viral respiratoire, épisodes bronchitiques, état pulmonaire septique, état septique broncho-pulmonaire, état septique pulmonaire, foyer infectieux bronchique, foyer infectieux broncho-pulmonaire, infection aigüe respiratoire, infection bronchique aigüe, infection broncho-pneumopathique, infection bronchopulmonaire, infection broncho-respiratoire, infection poumon, infection pulmonaire sévère, infection purulente poumon, infection respiratoire aigüe, infection voies respiratoires basses, infection voies respiratoires inférieures, infections pulmonaires multiples, maladie respiratoire infectieuse, pathologie broncho-pulmonaire, pathologie infectieuse respiratoire, pleuro-broncho-pneumonie, pleuro-broncho-pneumopathie, pleuropneumonie, pleuropneumopathie, pneumonie aigüe, pneumopathie, sepsis pulmonaire, surinfection bronchique aigüe, surinfection broncho-pulmonaire, syndrome bronchique aigu, syndrome bronchique infectieux, trachéo-bronchite, trachéo-bronchite adénovirus, virose bronchique, virus respiratoire….
Asphyxie/anomalie de la respiration	R06, R061, R09, R090, R091, R093, R098, T71	**356 expressions, dont :** accès hypoxémique, ACR anoxique, altération pulmonaire, anoxie, anoxie aigüe, anoxie brutale, anoxie, anoxie prolongée, antécédents spasmes sanglot, apnée, apnée centrale, arrêt anoxique, arrêt hypoxémique, asphyxie, asphyxie cardio-

		pulmonaire, bradypnée, choc anoxique, choc hypoventilatoire, crise asphyxique, décompensation respiratoire aigüe hypercapnique, désaturation oxygène, désaturation pulmonaire, épisodes désaturation, état asphyxique, état asphyxique aigu, état hypoxémique, état hypoxique, gasp, gêne respiratoire, gêne ventilatoire, hoquet chronique, hypercapnie, hyperventilation, hypoventilation aigüe, hypoventilation chronique, hypoxémiant, hypoxie; hypoxygénie, obstacle asphyxique, polypnée, râles agoniques, sepsis cavité pleurale, spasme respiratoire, syndrome anoxique aigu, syndrome asphyxique, tableau anoxique, ventilation faible, ventilation impossible, vomique…
Sepsis	A40,A41, R572	**928 expressions, dont :** Bactériémie, bactériémie anaérobie, bactériémie anaérobies, bactériémie BLSE, bactériémie Campylobacter, bactériémie E coli, bactériémie Enterobacter, bactériémie entérocoque, bactériémie Escherichia coli, choc bactérien, choc bactérien anaérobique, choc infectieux, choc septicémique, choc septique, choc septique aigu, choc septique brutal, choc septique E coli, choc septique gram+, choc septique multi-bactérien, choc toxico-septique, choc toxi-infectieux, embol septique, embols infectieux, état choc toxi-infectieux, état septicémique, état septique, états septiques multiples, foyers infectieux multiples, généralisation infectieuse, infection générale, infection généralisée, infection septicémique, infections multiples, sepsis, septicémie, septicémique, syndrome infectieux général, syndrome septique, tableau septicémique…
Maladies chroniques endocriniennes	E00-E030, E031, E034-E039, E04, E050-E054, E056-E059, E062, E063, E065-E069, E07, E10-E149, E20-E35, E40-E46, E50-E90	**2308 expressions, dont :** Acanthocytose, accident métabolique, acido-diabète, acromégalie, Addison, adrénaline, alcaptonurie, amylose, aphasie Gayet-Wernicke, artériopathie diabétique, athérosclérose diabétique, avitaminose, Bartter, Basedow, calcinose, cardiopathie carentielle, cérébrolipofuscinose, cholestérol, coma acidocétosique, coronarite diabétique, coronaropathie diabétique, Cushing, cystinurie, décompensation surrénalienne, décompensation thyroïdienne, déminéralisation diffuse rachis, dénutrition calorico-azotée, diabète, diabète 2 insulino-requérant, diabète 2 multicompliqué, DID, dyscalcémie, dyslipémie, dysthyroïdie, enzymopathie, goitre obstructif, goitre thyroïdien, hémochromatose, hypercalcémie, insuffisance hypophysaire globale, Launois-Bensaude, leucodystrophie, lipidose, lysinurie, maladie Basedow, maladie Batten, maladie Fabry, maladie Hurler, maladie peroxysomiale, maladie Sanfilippo, maladie Schindler, marasme, mucoviscidose, myocardiopathie amyloïde, nanisme, nécrose hémorragique bilatérale surrénale, neurolipidose, obésité, parathyroïdie, pathologie mitochondriale, polyneuropathie amyloïde, polysurcharge, pré-

		coma diabétique, rachitisme, syndrome Cushing, syndrome Gayet-Wernicke, syndrome lyse, syndrome ravine, trouble dysmétabolique, troubles thyroïdiens, xanthomatose…
Maladies chroniques de l'appareil digestif	K00-K22 (except K222, K223, K226), K23-K31, K36, K38-K55 (except K550), K56-K63 (except K631), K64-K71, K73-K81(except K810), K82-K84, K87-K92 (except K922), K93, R12-R19	**6666 expressions, dont :** abcès cholédocien, abcès colique, bulbite, bulbo-duodénite, calcul biliaire, calcul Wirsung, cholangite aigüe, cholangite chronique, cholangite ischémique, cholécystite, cholécystite calculeuse, cholestase, cirrhose, infection rectale, infiltration anses grêle, inondation péritonéale, intestin radique, ischémie caecum, pathologie gastrique ulcéreuse, pathologie pancréas, péliose hépatique, perforation cholédoque, perforation duodénale, perforation hernie étranglée, perforation pylore, perforation ulcère, péritonite asthénique, perte possibilités alimentation naturelle, perturbation état digestif, plaie nécrotique œsophage, pneumatose colique, presbyphagie, trouble moteur œsophage, troubles épigastriques, troubles pancréatiques, troubles vidange gastrique, ulcération anastomique gastro-jéjunale, ulcération colique, vessie pseudo-tumorale, volvulus duodénal, volvulus œsophage, watermelon gastrique, wirsungo-pancréatectomie

Annexe 5 : Exemple de deux vecteurs d'apprentissage (champs de certificat de décès) en utilisant les unigrammes et bigrammes de mots en caractéristique

Vecteur 1 (Champ 1) : « rupture anevrysme cerebral »

→ Regroupement syndromique associé (la classe): AVC/Hémorragie cérébrale

Vecteur 2 (Champ 2) : « demence vasculaire, malaidie parkinson, insuffisance renale chronique »

→ Regroupements syndromiques associés (la classe) : *Autres troubles mentaux ; Maladies chroniques du système nerveux, Maladies chroniques de l'app. Génito-urinaire*

$\boldsymbol{Vecteur_{X1}}$	$\boldsymbol{Vecteur_{X2}}$	$\boldsymbol{Caractéristiques}$
1	0	$rupture$
1	0	$anevrysme$
1	0	$cerebrale$
0	1	$demence$
0	1	$vasculaire$
0	1	$malaidie$
0	1	$parkinson$
0	1	$insuffisance$
0	1	$renale$
0	1	$chronique$
1	0	$\#rupture$
1	0	$rupture\ anevrysme$
1	0	$anevrysme\ cerebral$
1	0	$cerebral\$$
0	1	$\#demence$
0	1	$demence\ vasculaire$
0	1	$vasculaire\ malaidie$
0	1	$malaidie\ parkinson$
0	1	$parkinson\ insuffisance$
0	1	$insuffisance\ renale$
0	1	$renale\ chronique$
0	1	$chronique\$$

$\boldsymbol{Vecteur_{X1}}$	$\boldsymbol{Vecteur_{X2}}$	$\boldsymbol{Classes}$
1	0	$AVC, Hémorragie\ cérébrale$
0	1	$Autres\ troubles\ mentaux$
0	1	$Maladies\ chroniques\ du\ système\ nerveux$
0	1	$Maladies\ chroniques\ de\ l'appareil\ genito-urinaire$

Annexe 6 : Exemple de deux vecteurs d'apprentissage (champs de certificat de décès) en utilisant les trigrammes de caractères en caractéristique

Vecteur 1 (Champ 1) : « rupture anevrysme cerebral »

→ Regroupement syndromique associé (la classe): AVC/Hémorragie cérébrale

Vecteur 2 (Champ 2) : « demence vasculaire, malaidie parkinson, insuffisance renale chronique »

→ Regroupements syndromiques associés (la classe) : *Autres troubles mentaux ; Maladies chroniques du système nerveux, Maladies chroniques de l'app. Génito-urinaire*

$Vecteur_{X1}$	$Vecteur_{X2}$	$Caractéristiques$
1	0	#. #. r
1	0	#. #. r
1	0	r. u. p
1	0	u. p. t
1	0	p. t. u
1	0	t. u. r
1	0	r. e. _
1	0	e. _. a
1	0	_. a. n
1	0	a. n. e
1	0	e_c
0	1	#. #. m
0	1	m. a. l
0	1	a. l. a
0	1	l. a. i
0	1	a. i. d
0	1	i. d. i
0	1	i. e_
0	1	e. _. p
0	1	p. a. r
0	1	r. k. i
0	1	k. i. n

$Vecteur_{X1}$	$Vecteur_{X2}$	$Classes$
1	0	$AVC, Hémorragie\ cérébrale$
0	1	$Autres\ troubles\ mentaux$
0	1	$Maladies\ chroniques\ du\ système\ nerveux$
0	1	$Maladies\ chroniques\ de\ l'appareil\ genito-urinaire$

Annexe 7 : Exemple de deux vecteurs de caractéristique contenant les regroupements syndromiques attribués par la méthode par règles aux deux champs de certificat de décès

Vecteur 1 (Champ 1) : « rupture anevrysme cerebral »

→ Regroupement syndromique attribué (la classe) : AVC/Hémorragie cérébrale

Vecteur 2 (Champ 2) : « demence vasculaire, malaidie parkinson, insuffisance renale chronique »

→ Regroupements syndromiques attribués (la classe) : *Autres troubles mentaux ; Maladies chroniques du système nerveux, Maladies chroniques de l'app. Génito-urinaire*

$$\begin{pmatrix} \boldsymbol{Vecteur_{X1}} & \boldsymbol{Vecteur_{X2}} \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \quad \begin{matrix} \boldsymbol{Caractéristiques} \\ AVC/Hémorragie\ cérébrale \\ Autres\ troubles\ mentaux \\ Maladies\ chroniques\ du\ système\ nerveux \\ Maladies\ chroniques\ de\ l'app.Génito-urinaire \end{matrix}$$

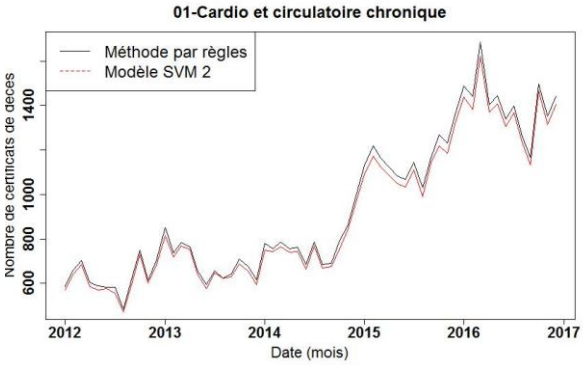
$$\begin{pmatrix} \boldsymbol{Vecteur_{X1}} & \boldsymbol{Vecteur_{X2}} \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \quad \begin{matrix} \boldsymbol{Classes} \\ AVC, Hémorragie\ cérébrale \\ Autres\ troubles\ mentaux \\ Maladies\ chroniques\ du\ système\ nerveux \\ Maladies\ chroniques\ de\ l'appareil\ genito-urinaire \end{matrix}$$

Annexe 8 : Liste des 60 regroupements syndromiques sélectionnés pour l'évaluation des performances de classification des causes médicales de décès par la méthode par règles et par le modèle SVM 2

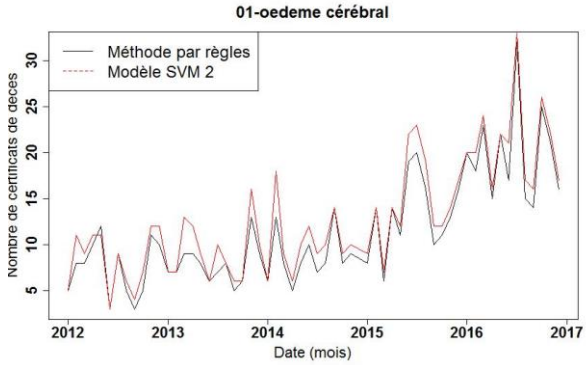
1-Cardio et cérébrovasculaire
1-Non informatif
1-Maladies cérebrovasculaires chronique
1-Rupture d'anévrisme non cérébral
1-Cardio et circulatoire chronique
1-Syndrome coronarien aigue
1-Décompensation/insuffisance cardiaque aiguë
1-Oedeme cérébral
1-Troubles du rythme de la conduction / fibrillation
1-Cardio Infectieux
1-Embolie pulmonaire
1-AVC/Hémorragie cérébrale
2-Cancers/Tumeurs
2- Cancers/Tumeurs
3-Maladies respiratoires
3-Maladies respiratoires chroniques
3-Maladies respiratoires aigües
3-Asthme
3-Dyspnée
4-Symptomes
4-Grabataire/ Sénile
4-Chocs cardio et hypovolémique
4-Troubles cognitifs et de l'humeur
4-Coma
4-Symptômes généraux
4-Malaise et fatigue
4-Autres chocs
4-Mort subite
4-Symptomes du trouble de l'alimentation
5-Maladies infectieuses
5-Infections parasitaires
5-Résistance
5-Infections bactériennes
5-Gastro-entérite - Diarrhée
5-Maladies à Déclaration Obligatoire infectieuses
5-Hépatites virales
6-Maladies endocriniennes
6-Anomalies métaboliques
6-Déshydratation
7-Maladies du système nerveux central
7-Maladies chronique du système nerveux
7-Tableaux neuropsychologiques aigus
7-Epilepsie
7-Maladies inflammatoire du système nerveux central

8-Autres facteurs et motifs de recours aux soins
8-Antécédants
8-Autres facteurs et recours aux soins
9-Trauma et empoisonnement
9-Complication
9-Effet toxique lié à des substances médicinales
9-Effet toxique lié à l'alcool
9-Lesions traumatiques et présence de corps étrangers
10-Troubles mentaux et du comportement
10-Troubles mentaux chroniques
10-Autres troubles mentaux
10-Troubles mentaux liés à des substances
11-Maladies de la peau et du tissu cellulaire sous cutané
11-Maladies chroniques de la peau et du tissu
12-Maladies du sang et des organes hématopoïétiques
12-Maladies chroniques du sang
12-Maladies aiguës du sang
12-Purpura et affections hémorragiques
13-Maladies de la grossesse, néonatales et congénitales
13-Maladies congénitales
15-Maladies de l'appareil digestif
15-Maladies aigües de l'appareil digestif
16-Autres causes de mortalité mal définies et non précisées
16-Autres causes de mortalité mal définies et non précisées
17-Causes externes
17-Accidents et autres causes externes
17-Accident de transport
17-Suicides
18-Maladies de l'appareil génito-urinaire
18-Maladies aigües de l'appareil génito-urinaire
18-Insuffisance rénale aiguë
18-Maladies chroniques de l'appareil génito-urinaire
20-Maladies du système ostéo-articulaire
20-Maladies du système ostéo-articulaire

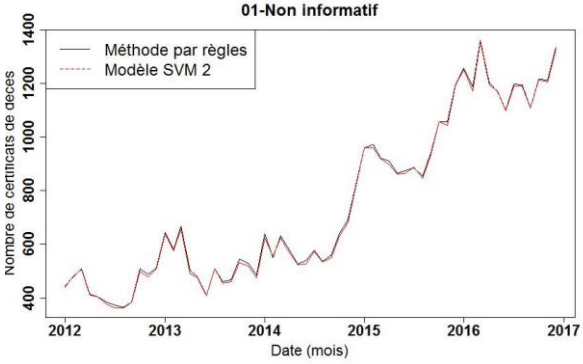
Annexe 9 : Dynamique mensuelle des RS du groupe 1 pour lesquels les performances de la méthode par règles et du modèle SVM 2 étaient supérieures à 0,95



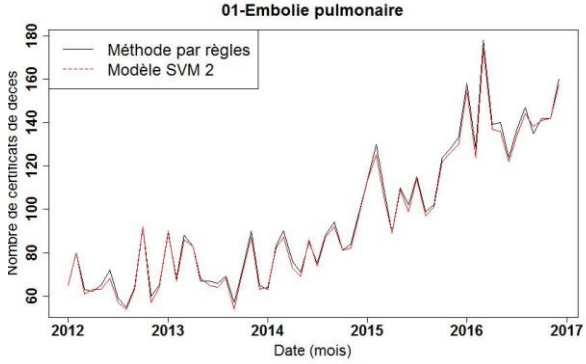
Δi= 2,93% ; Coef correlation= 1,00



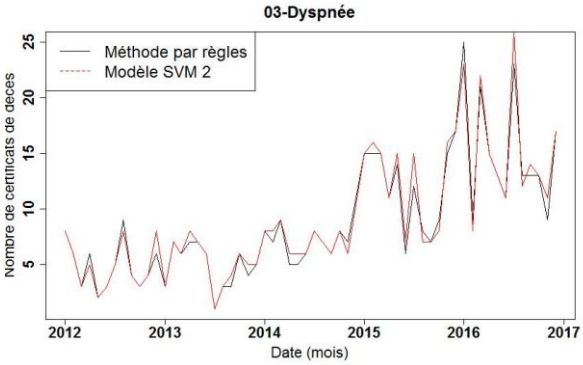
Δ_i= -10,79% ; Coef correlation= 0,98



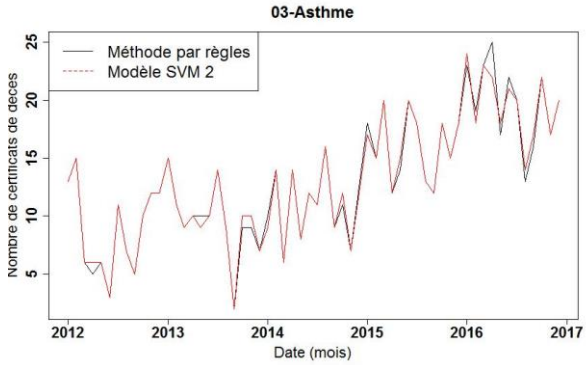
Δi= 0,79% ; Coef correlation= 1,00



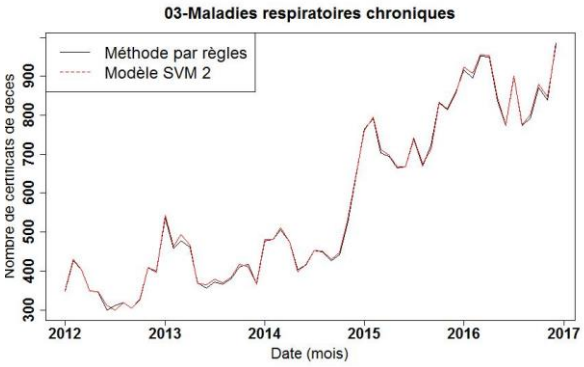
Δi= 1,66%, Coef correlation= 1,00



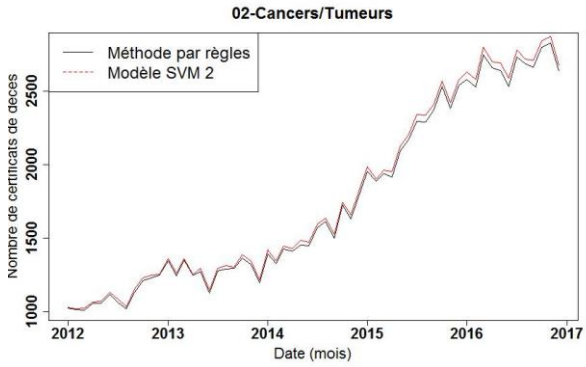
Δi= -2,47%, Coef correlation=0,99



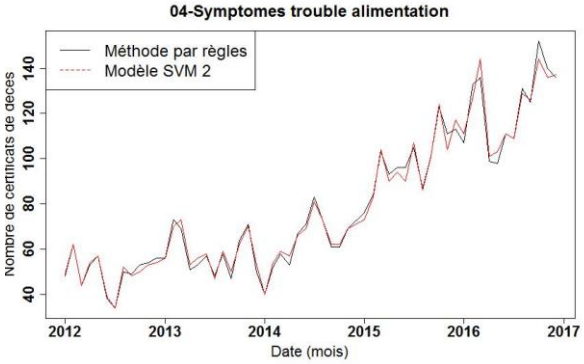
Δi= 0,13%, Coef correlation= 0,99



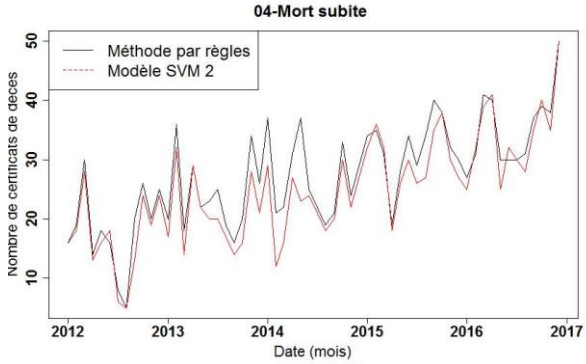
Δ_i= -0,42%, Coef correlation= 1,00



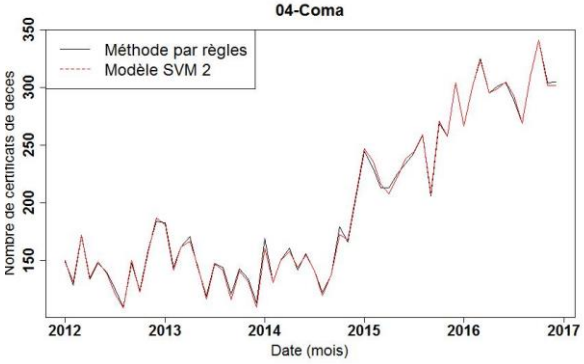
Δ_i= -1,55%, Coef correlation= 1,00



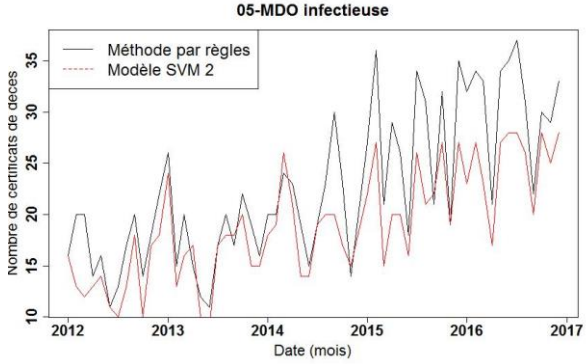
Δ_i= 0,04%, Coef correlation=1,00



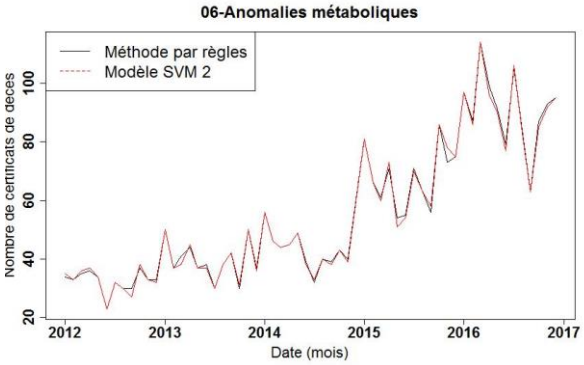
Δ_i= 9,39%, Coef correlation=0,95



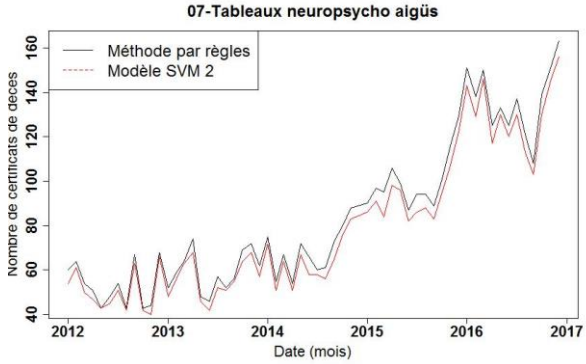
Δ_i= 0,27%, Coef correlation= 1,00



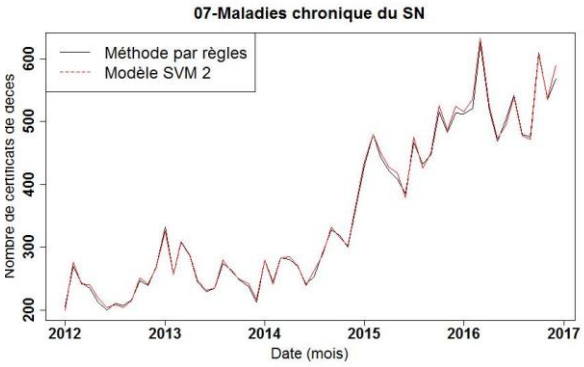
Δ_i= 17,78%, Coef correlation= 0,91



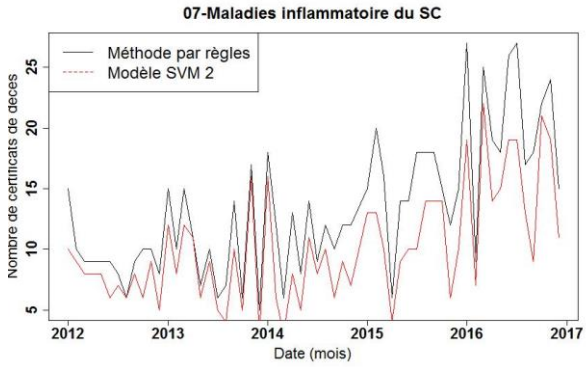
Δ_i= 0,37%, Coef correlation= 1,00



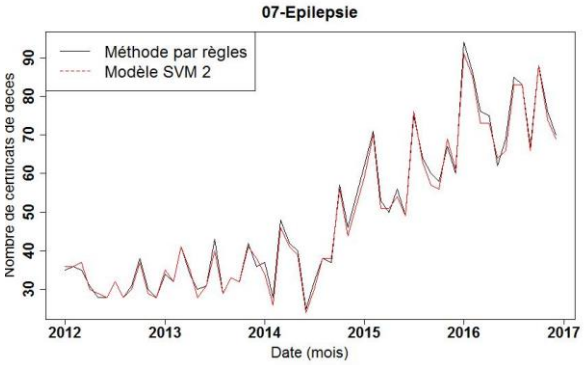
Δ_i= 5,92%, Coef correlation= 1,00



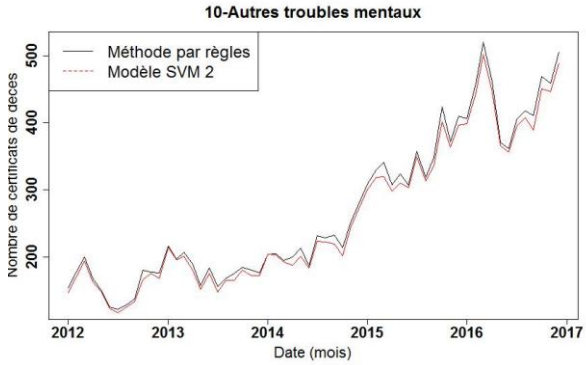
Δ_i= - 0,61%, Coef correlation=1,00



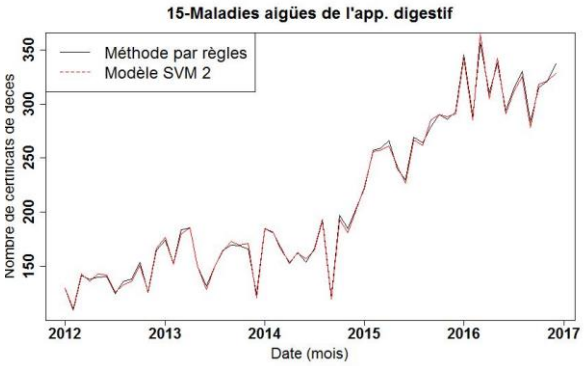
Δ_i= 5,92%, Coef correlation= 1,00



Δ_i= 1,43, Coef correlation= 1,00

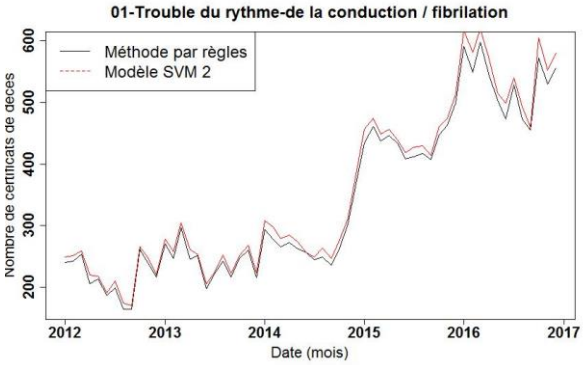


Δ_i= 3,28%, Coef correlation= 1,00

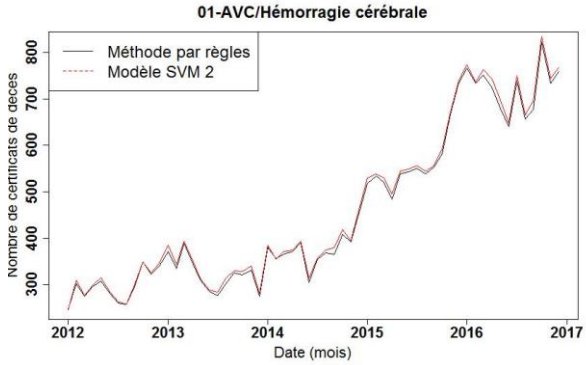


Δ_i= 0,27%, Coef correlation= 1,00

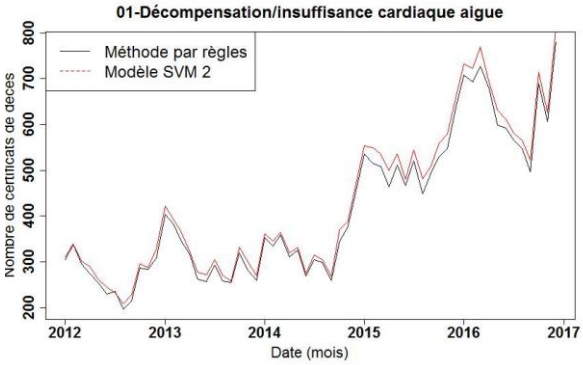
Annexe 10 : Dynamique mensuelle des RS du groupe 2 pour lesquels les performances de la méthode par règles et du modèle SVM 2 étaient supérieures à 0,95



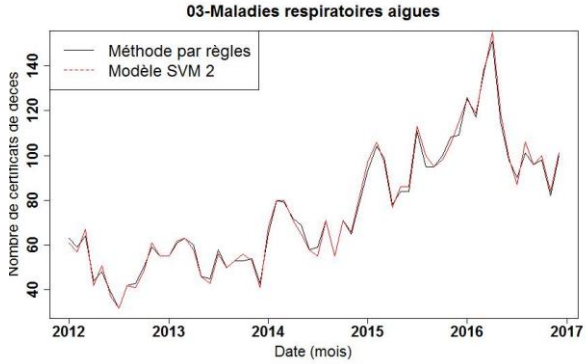
Δ_i= -3,43%, Coef correlation= 1,00%



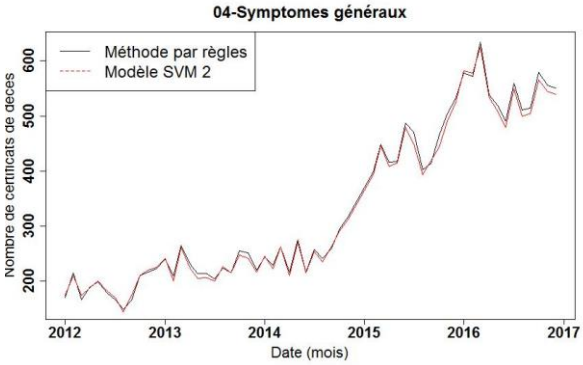
Δ_i= -1,52%, Coef correlation= 1,00



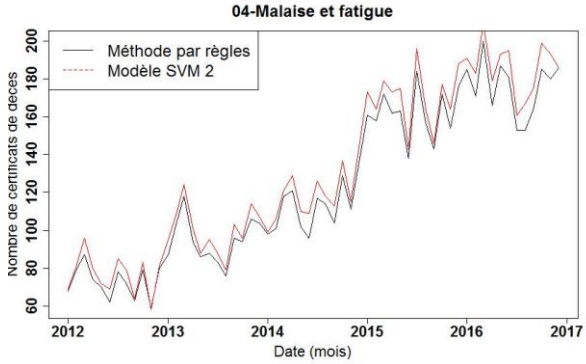
Δ_i= -3,89%, Coef correlation= 1,00



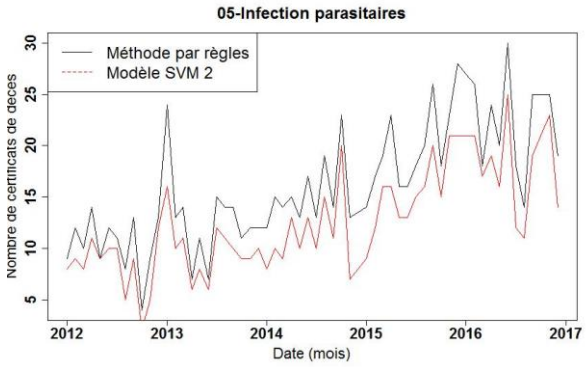
Δ_i= -0,38%, Coef correlation= 1,00



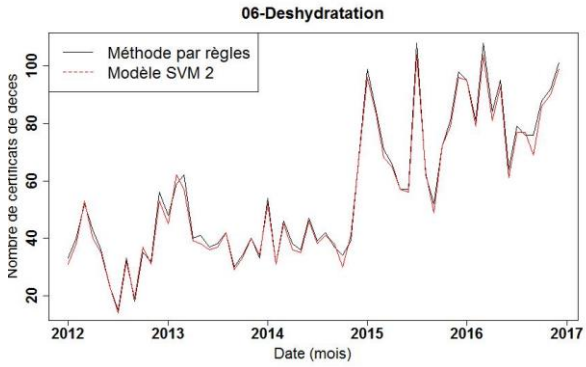
Δ_i= 1,27%, Coef correlation= 1,00



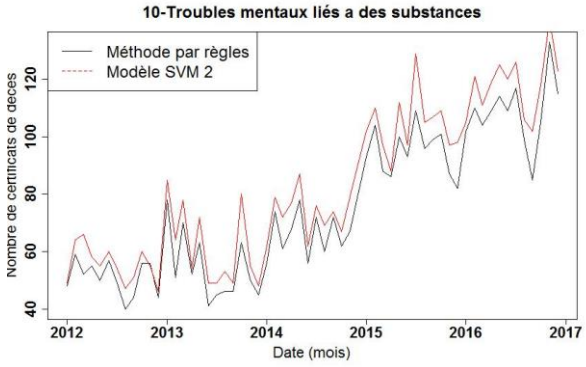
Δ_i= -5,39%, Coef correlation= 1,00



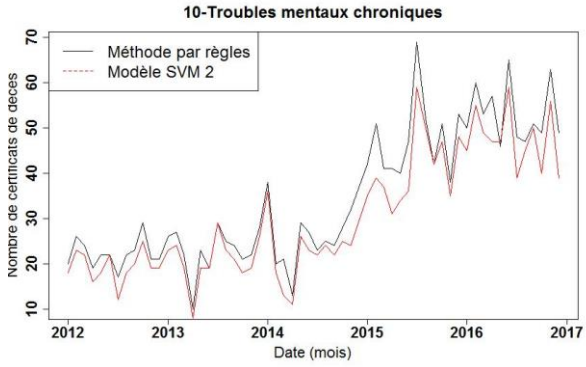
Δ_i= 24,31%, Coef correlation= 0,96



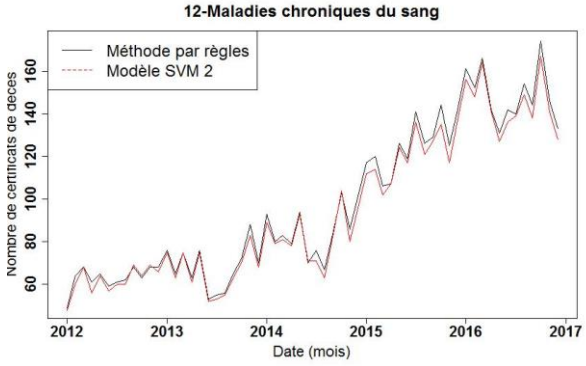
Δ_i= 2,42%, Coef correlation= 1,00



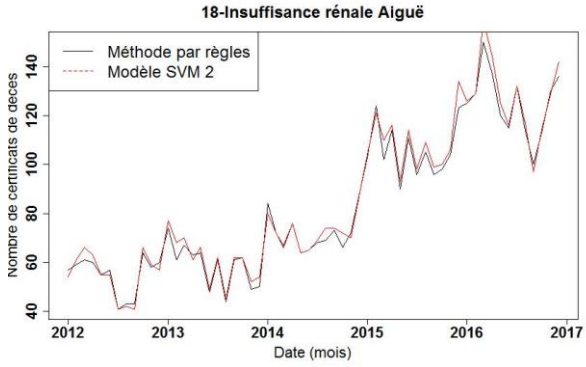
Δ_i= -9,61%, Coef correlation= 0,99



Δ_i= 12,53%, Coef correlation= 0,98



Δ_i= 2,86%, Coef correlation=1,00



Δ_i= -2,04%, Coef correlation= 1,00

Articles

Evaluation of the French reactive mortality surveillance system supporting decision making

European Journal of Public Health, 1–7

doi:10.1093/eurpub/cky251

Evaluation of the French reactive mortality surveillance system supporting decision making

Yasmine Baghdadi[1], Anne Gallay[2], Céline Caserio-Schönemann[1], Anne Fouillet[1]

1 Santé Publique France, Division for Data Science, Saint-Maurice, France
2 Santé Publique France, Division of Non communicable diseases and Injuries, Saint-Maurice, France

Correspondence: Yasmine Baghdadi, Santé Publique France, Division for Data Science, 12 rue du Val d'Osne, Saint-Maurice, France, Tel: +33 (0) 1 41 79 68 16, Fax: +33 (0) 1 41 79 68 11, e-mail: yasmine.baghdadi@santepubliquefrance.fr

Background: In France, a mortality syndromic surveillance system was set up with objectives of early detection and reactive evaluation of the impact of expected and unexpected events to support decision makers. This study aims to describe the characteristics of the system and its usefulness for decision makers. **Methods:** Anonymized data from the administrative part of death certificates were daily collected from 3062 computerized city halls and were transmitted to Santé publique France in routine. Coverage of the system was measured as the proportion of deaths registered by the system among the complete number of deaths and analyzed by age, month and region. Deaths were described by gender, age and geographical level using proportion. The excess periods of deaths were described based on the comparison of the weekly observed and expected numbers of deaths between 2012 and 2016. **Results:** The system recorded 77.5% of the national mortality covering the whole territory. About 81% of deaths were aged 65 years old and more. The surveillance system identified mortality variations mainly during winter and summer, for some concomitant with influenza epidemic or heatwave period, and thus provided information for decision makers. **Conclusion:** The ability of the system to detect and follow mortality outbreaks in routine in the whole territory has been demonstrated. It is a useful tool to provide early evaluation of the impact of threats on mortality and alert decision makers to adapt control measures. However, the absence of information on medical causes of death may limit the ability to target recommendations.

Introduction

Mortality is one of the main indicators traditionally used to measure population health and the severity of public health events, since death is the most severe outcomes of disease. Mortality is of fundamental significance to health surveillance.

The need to acquire an early and routinely surveillance system of mortality emerged after the sanitary crisis that occurred during the 2003 extreme heatwave in Europe.

Syndromic surveillance, defined as 'a real time (or near real time) collection, analysis, interpretation and dissemination of health related data to enable the early identification of the impact (or absence of impact) of potential human public health threats that require effective public health action'[1] constitutes an adapted tool to reach this requirement.

In 2004, Santé publique France, the French public health agency, set up the syndromic surveillance system SurSaUD® with the objectives to ensure a reactive detection of expected and unexpected threats and to provide an early impact assessment of public health events. The system combines both morbidity and mortality data sources.[2]

As part of the SurSaUD® system, the mortality surveillance system is based on the anonymized civil-status data collected from the administrative part of the death certificate.

This routine surveillance allows production of weekly (daily if necessary) analyses and reports of all-cause mortality variations at national and regional levels which are transmitted to the health authorities to support decision making. Timely communication through reports is a strong lever for health authorities to adapt counter measures and prevention messages.

The aim of this study is to describe the main characteristics and performance of the French mortality surveillance system and its usefulness for decision makers.

Methods

Data for surveillance

The administrative component of the death certificates is recorded by the registry office of city halls. The French National Institute for Statistic and Economic Studies (Insee) receives data from the computerized city halls (i) with an automatic and timely transmission ('Routine sample'); (ii) with a non-automatic transmission. This second transmission takes longer. Thus the total database is consolidated after several weeks ('Insee total').

The mortality surveillance system is based on the information from the 'Routine sample'. That information are provided in routine by Insee to Santé publique France. Initially including 1042 computerized city halls, the sample was extended to 3062 city halls in 2011 ('Routine sample'). This study focuses on the 3062 city halls sample.

This automatic and secure transmission enables to deliver individual data six days a week (no transmission on Saturday to Sunday night). Transmitted data contain year of birth, date and place of death, gender and date of data transmission but, medical causes of death are not provided. No missing values were found in these data.

On average about 90% of the deaths registered in the 3062 municipalities is received within seven days at the agency.[3,4] Due to the organization of registry offices, this delay may vary according to several factors: day of week, day-off, holidays… Thus, mortality is considered to be robustly analyzed with at least a lag of 10 days in terms of trend analysis and with at least a three-week delay in terms of excess quantification.

The use of mortality data has been authorized by the French National Commission for Data protection and Liberties.

Coverage of the system

Coverage of the system is defined as the number of deaths registered by the system ('Routine sample') divided by the total number of deaths provided by the Epidemiology center of the medical causes of death (Inserm-CépiDc).[5] Inserm-CépiDc is in charge of the national complete mortality database of the medical causes of death, which constitutes a reference for epidemiological studies. Due to validation delay, the complete database was only available from 2011 to 2013 at the date of the study.

Coverage of the system was calculated between 2011 and 2013 at national and regional levels, by age group [≤1 year old, 2–14, 15–44, 45–64, 65–84 and ≥85 years old (yo)] and by month to evaluate its stability over time.

In order to assess the stability of the coverage between 2014 and 2016, the monthly number of deaths recorded by the system ('Routine sample') was extrapolated using the monthly coverage calculated on 2011–13 to provide a monthly estimated total number of deaths on 2014–16 period at national level for all ages. These estimations were then compared to the monthly total number of deaths ('Insee total').[6] It must be noted that the Insee national mortality database was still provisional for 2015 and 2016, at the date of the study.

Population description

The distribution of deaths recorded by the system from 2011 to 2016 ('Routine sample') was described by gender and age group (0–14, 15–64, 65–84 and 85 yo and more), both at national and regional levels, using proportion. The average age at death was also calculated by year at national and regional levels.

Detection method of an excess of deaths and performance of the system

The French mortality surveillance system is based on the comparison of the weekly observed number of deaths recorded by the system to an expected number of deaths ('baseline'). The baseline is calculated using a time-series Poisson regression model with a five-year historic period of data. The model is adjusted for a trend and sinusoidal seasonal variation depending on the geographical level and age group:

- A linear trend and a seasonality:
  - National and metropolitan regional levels for all ages and 15–64, 65–84, ≥85 yo
- A linear trend and no seasonality:
  - National and metropolitan regional levels for age group 0–14 yo.
  - Overseas regions for all ages and 0–14, 15–64, 65–84, ≥85 yo.

This model has been set up by the Euromomo consortium and is performed weekly for mortality surveillance in about 20 European countries and at a pooled European level.[7]

We used the standardized indicator *Z*-score (defined by the difference between the observed number of deaths and the baseline, divided by the standard-deviation (SD) of the baseline) to compare mortality patterns among different regions, age groups and time periods.

A statistical alarm was considered when the observed number of deaths for one week was exceeding the threshold (baseline + 2 SD or *Z*-score ≥2) and is by was systematically investigated. *Z*-score levels were defined as followed:[7]

- *Z*-score <2: no excess of deaths.
- *Z*-score comprised between [2–4[: low excess of deaths.
- *Z*-score comprised between [4–6[: medium excess of deaths.
- *Z*-score ≥6: high excess of deaths.

Excess of deaths periods were analyzed between 2012 and 2016.

Performance to detect outbreaks were measured by calculating the sensitivity and specificity of the system in France and regions between 2012 and 2013. The number of weeks with alarm detected by the Euromomo model using the complete database provided by Inserm-CépiDc was compared to the number of weeks with alarm detected by the system using the 'routine sample'.

Usefulness of the system

The study focused on different outbreaks, which required a specific analysis and are of particular interest to the agency and public health decision makers. To do this, additional data have been collected: (i) the period of influenza epidemic between 2012 and 2016;[8–12] (ii) the weekly number of visits with clinical diagnostic of influenza of the Sentinelles* network from 2012 to 2016;[13] (iii) the occurrence of heatwave periods;[14] (iv) the weekly number of chikungunya cases reported in Guadeloupe during the 2014 epidemic.[15]

All the analyses were performed with R software Rx64 3.3.0.

Results

Coverage of the system

About 77.5% of the total French deaths was recorded by the system between 2011 and 2013 (figure 1a). In the metropolitan regions (Supplementary Material annex 1), this proportion varied from 67% in Occitanie to 89% in Ile-de-France. It was higher in the overseas (from 86% in French Guyana to 96% in Martinique). It remained very stable from 2011 to 2013, both at national (±0.5% in France) and regional levels (less than ±2% in regions), except for French Guyana where it varied from ±5.5% during this three-year period (figure 1a).

The coverage at national level varied according to the age groups (figure 1b). While the system recorded an average of 93% of the national mortality for children ≤1 year of age, the coverage was lower for young adults aged 15–44 and for the elderly aged 85 years and older (respectively 71% and 74% of the national mortality). The coverage for the other age groups was about 77% in average. This distribution of the coverage by age group was similar in the metropolitan regions. In three of four overseas, the mortality coverage was highly comparable among all age groups (figure 1b).

The monthly mortality coverage between 2011 and 2013 was almost constant within a year both at the national level (±1.5%) and in the metropolitan regions (less than ±2.5%) except for Corsica (results not shown). Noticeable variations were observed in the overseas regions (from ±2 to ±11%).

The extrapolated numbers of deaths estimated monthly from 2014 to 2016 (based on the 'Routine sample' and the mortality coverages calculated on 2011–13 period) were close to the national numbers of deaths ('Insee total') (less than ±1%—results not shown).

Population description

From 2011 to 2016, 49% of deaths recorded by the system were women and 81.6% of people were aged 65 yo and more (table 1). Among them, a half was aged more than 85 yo. Less than 1% was children under 15 years. The mean age of death increased of 1.4 years between 2011 and 2016. The mean age of death in the overseas (French Guyana and Ia Réunion) was lower than in the whole territory and particularly marked in French Guyana (58.8 vs. 77.7 yo for the 2011–16 period). Also 2–11% of mortality were children <15 yo.

Detection of the excess of deaths and performance of the system

At national level, three major excess of deaths episodes were observed between 2012 and 2016, as well as isolated periods with

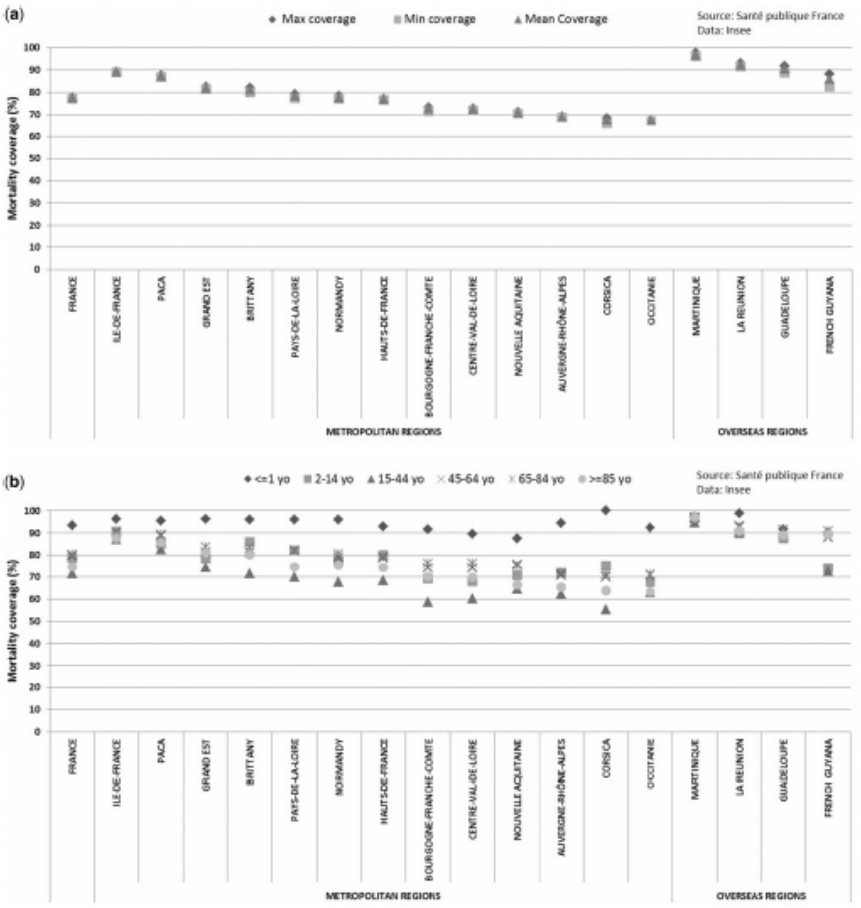


Figure 1 Mortality coverage from 2011 to 2013 in France at national and regional levels: mean, maximum and minimum all ages (a), mean by age group (b)

one or two consecutive weeks with a significant increase in mortality above the threshold (*Z*-score ≥2) (figure 2a).

The major excess deaths episodes occurred during winters 2011–12, 2012–13 and 2014–15, lasted respectively 6, 12 and 13 consecutive weeks and reaching a peak of more than a *Z*-score of 6. The analysis by age group showed that the elderly were the most frequently concerned, even if a low excess was also observed in winters for people aged 15–64 yo (figure 2a).

Five isolated excess periods were identified during the summer months at national level (figure 2a) with a maximum *Z*-score comprised between 2 and 4. These excesses mainly affected the elderly.

The sensitivity and specificity of the system to detect weeks with alarm were respectively 0.96 and 0.99 at national level and for all ages. The sensitivity varied in France from 0.65 for the 15–64 yo to 0.86 for the elderly. The specificity varied from 0.93 for the 15–64 yo to 0.99 for the 0–14 yo. At regional level the sensitivity varied between 0.42 and 1, except for Corsica (Supplementary Material annex 2). The specificity was over 0.89 for all regions.

Usefulness of the system

From 2012 to 2016, winter excesses of deaths were highly correlated with flu epidemic periods particularly for people aged 65 years and older (figure 3a and b). During 2015 winter, mortality surveillance also demonstrated that the number of deaths began increasing before influenza epidemic, suggesting the role of other potential factors.

In summer, only 3 out of 5 outbreaks were concomitant with a heatwave period (figure 3a and b).

Focusing on the 2014–15 period at regional level, figure 2b shows that mortality increased in all metropolitan regions during the 2014–15 winter with different magnitude and duration except for Corsica.

In overseas, only isolated limited increases of mortality were observed during the study period, except in Guadeloupe. From June to July 2014, an excess mortality occurred in weeks 22–25 and 27, with a *Z*-score comprised between 4 and 6 in week 25 (figure 2b). This excess occurred during a major chikungunya epidemic that lasted from week 13 to 35, 2014. It affected all age

Table 1 Proportion of deaths recorded by the system from 2011 to 2016 in France by gender and age group, mean age of death at national and regional levels

	N	Gender (%)	Age group (%)				
Sociodemographic variables/years		Female	0–14 yo	15–64 yo	65–84 yo	≥85 yo	Mean age of death (years)
France							
2011	424 802	48.6	0.9	19.2	40.8	39.2	76.9
2012	441 656	49.2	0.8	18.1	40.0	41.0	77.5
2013	442 540	49.1	0.9	17.8	39.6	41.8	77.6
2014	437 637	49.1	0.8	17.5	39.5	42.1	77.7
2015	462 939	49.4	0.8	16.6	38.8	43.8	78.3
2016	460 545	49.4	0.8	16.3	38.6	44.3	78.3
2011–16	2 670 119	49.2	0.8	17.5	39.5	42.1	77.7
Region (2011–16)							
Ile-de-France	396 383	50	1.5	20.7	38.3	39.6	76.0
Auvergne-Rhône-Alpes	268 399	48.9	0.9	16.0	39.9	43.3	78.3
Nouvelle-Aquitaine	262 490	48.3	0.5	15.7	38.9	44.9	79.0
Provence-Alpes-Côtes-d'Azur	256 629	49.9	0.6	15.1	38.8	45.5	79.1
Hauts-de-France	248 753	49.1	0.7	21.1	41.4	36.7	76.1
Grand-Est	243 274	50.1	0.7	17.3	41.7	40.3	77.6
Occitanie	222 614	48.4	0.7	15.9	38.9	44.5	78.6
Brittany	161 591	49.9	0.6	16.6	39.0	43.8	78.5
Pays-de-la-Loire	152 361	48.3	0.7	17.1	38.6	43.6	78.3
Normandy	147 892	49.2	0.6	18.1	39.8	41.6	77.8
Bourgogne-Franche-Comte	125 101	49.0	0.6	15.6	39.9	43.9	78.8
Centre-Val de Loire	108 413	48.3	0.6	15.3	38.8	45.3	79.1
Corsica	12 071	48.6	0.2	15.4	42.3	42.1	78.8
La Réunion	25 721	45.3	3.2	28.8	42.3	25.6	69.9
Guadeloupe	16 613	46.2	1.9	23.7	39.8	34.5	73.9
Martinique	17 775	48.9	1.5	18.5	41.1	38.9	76.5
French Guyana	4039	42.5	10.6	39.7	32.0	17.6	58.8

groups and was concomitant with the peak of the epidemic which occurred in week 24 (figure 3c).

Finally, mortality variations observed through the system were concomitant with clearly identified events, such as the dramatic terrorist attacks occurring in Paris region on 13 November 2015. A sudden high peak of the mortality was observed for people aged 15–64 on figure 3d.

Discussion

The system recorded 77.5% of the total number of deaths, covering the whole territory and all age groups. This coverage remained stable on the study period. The stability of the coverage within each year suggests that the coverage is not modified when a period of excess of deaths occurs.

This system was able to detect and timely evaluate expected and unusual mortality outbreaks across the whole territory. These different outbreaks affected all age groups and more particularly the elderly.

Almost all the mortality outbreaks occurring at national level were identified by the system. However at regional level, performances of detection were heterogeneous. The sensitivity depends on the weekly number of deaths (the more the number of deaths is low, the more the sensitivity is low) and to a lesser degree, on the coverage of the system.

The routine surveillance and analysis of the mortality conducted at regional and national levels, using a similar methodology and visualization tools, facilitated the evaluation of the health situation and supported decision making for the health authorities. Besides, the analysis using the same methodology is performed simultaneously in more than 20 European countries, enabling comparison between them.[16]

Such reactive system is a useful tool to provide an early evaluation of the impact of threats on mortality and alert decision makers at local, national and European levels. It is also useful to adapt, maintain or reinforce recommendation for control and prevention measures.

Limitations of the system

The system allowed identifying periods with significant increase of deaths, mainly occurring with winter flu epidemic or heatwaves. However the contribution of flu epidemics and heatwaves to the overall all-cause mortality could not be determined because of the absence of medical causes of death.

Influenza is known to contribute substantially to winter excess mortality, particularly during influenza A (H3N2) virus epidemics.[17,18] But other concomitant occurrence of several winter factors (other seasonal respiratory viral outbreaks or cold weather) over the same periods may also, even in part, contribute to the increase in mortality.[19,20] Likewise, heatwaves are known to affect vulnerable populations like the elderly or children, despite the prevention measures set up in the early part of the event.[21–25]

The absence of medical causes of death is an even more important issue (i) when an excess death is detected outside the occurrence of known public health events influencing mortality; (ii) to quantify the part to be attributable to different factors or specific events. That is the case for isolated excess periods. Investigation alongside the health professionals may be conducted when excess deaths concerned a specific age group or a limited geographical perimeter. A recent example was an excess death for the 0–14 yo in June 2017 observed in three specific regions.

When a larger excess of deaths occurs, only hypotheses might be formulated. That was the case with the excess deaths observed during the chikungunya epidemic in Guadeloupe in 2014. Chikungunya was known to be a viral disease transmitted by the bite of *Aedes* mosquitoes.[26] But it was not considered life threatening, even until a previous experience of an excess mortality during a chikungunya outbreak in La Réunion in 2005–06.[27,28] If the causal relation between the epidemic and the excess mortality was not

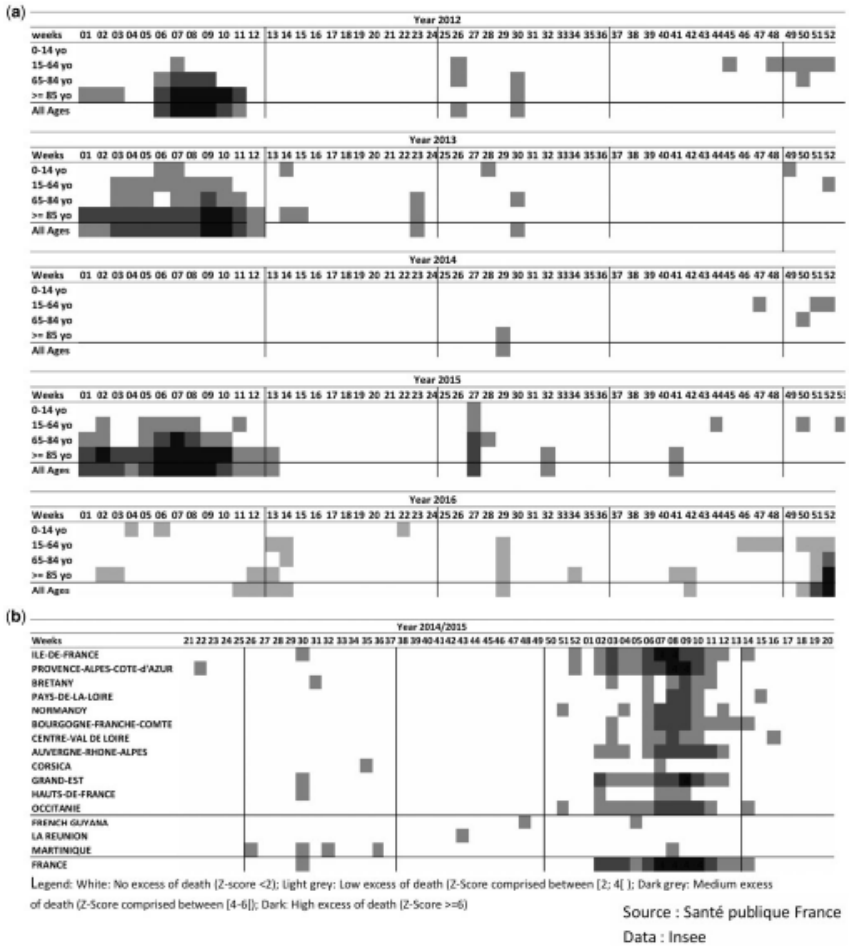


Figure 2 Level of weekly standardized deviation of the mortality from the baseline (Z-score) in France, from 2012 to 2016 by age group (a), focused on 2014–15 period by regions, all ages (b)

formally established, the surveillance system nonetheless enabled to underline an unexpected potential effect of this epidemic and contributed to improve the knowledge of such arboviruses.

Finally, in absence of medical causes of death, the system is also able to identify and measure the impact of very specific events like the terrorist attack in Paris in November 2015, even if other dedicated systems were pertinent to precisely quantify the impact of such event.

Towards a more reactive and complete surveillance system of mortality

In 2015, the sample of city halls participating to the system was extended and data from each new city hall are systematically transmitted to the agency. In 2017, the sample includes more than 6000 city halls, increasing the coverage. However, this extended sample will be used for surveillance when a sufficient historical period of data will be available, since the statistical model for estimating the baseline requires 3–5-year data.

This statistical model used to support the objectives of early detection and impact assessment was developed and commonly used at a European level. Its flexibility allowed taking account the different trends and seasonality, according to the age groups and geographical levels. An evolution to include the age structure of the population as an offset is considered to improve the estimation of the mortality baseline.

In 2007, the French syndromic surveillance system has been enriched with a new Electronic Death Registration System (EDRS). EDRS enables the physician to certify deaths electronically via a secured web application. Information from the medical component of the certificates (including the medical causes of death) are automatically transmitted within a few minutes to Santé publique France[29] with 95% of the e-certificates available on the day following the death.[3] Indeed, one current limit of the system

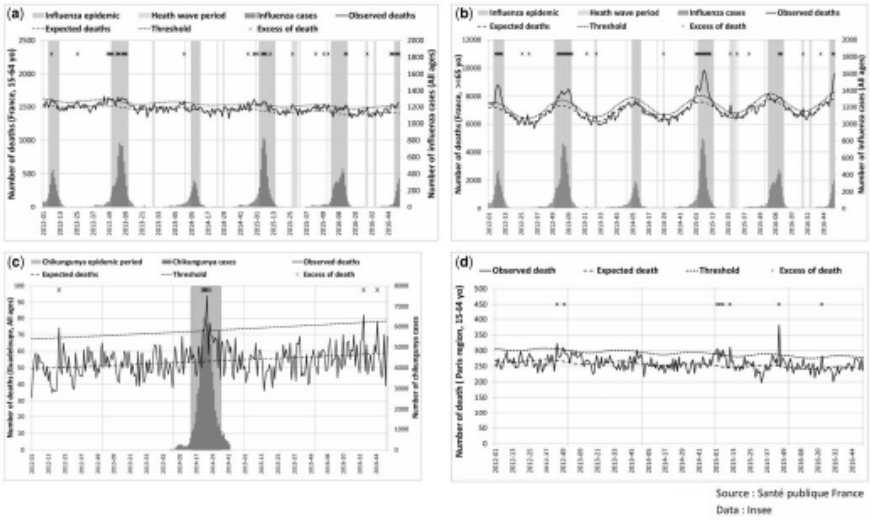


Figure 3 Weekly fluctuations of observed and expected mortality from 2012 to 2016, with influenza epidemic periods and heatwaves in France, for 15–64 yo (a), for ≥65 yo (b), with weekly number of chikungunya cases in Guadeloupe all ages (c), in Paris region for 15–64 yo (d)

is its reporting delay in registration: 90% of the mortality is registered within seven days. A first study demonstrated the pertinence of data from EDRS for the objectives of the syndromic surveillance system[3] and this new data source will enable to improve the reactivity of the surveillance. However the deployment of this system is still too limited (12% of the national mortality recorded by EDRS in 2017).

The future use of medical causes of death recorded by EDRS for the real-time surveillance in routine would substantially improve the current mortality surveillance but will face major challenges.

Supplementary data

Supplementary data are available at *EURPUB* online.

Acknowledgements

The authors thank the French Institute for Statistics and Economic Studies (Insee) for their contribution to the syndromic surveillance system SurSaUD, by transmitting mortality data daily. They also thank the Inserm-CépiDc for the transmission of the complete mortality database. Finally, the authors are also grateful to the members of Santé publique France who participated and made this study possible.

Funding

This research received no specific grant from any funding agency commercial or not-for-profit sectors. This surveillance is undertaken as part of the national surveillance functions of Santé publique France.

Conflicts of interest: None declared.

Key points

- A reactive all-cause mortality surveillance system for mortality outbreaks detection and impact assessment for decision making.
- About 77.5% of the total mortality caught by the system, covering the whole territory and all age groups.
- Excess mortality periods identified mainly during winters and summers and affected elderly.
- French reactive mortality surveillance system enables to provide information and recommendation through weekly report for decision makers.
- The system is being improved with information on medical causes of death from Electronic death certification.

References

1 Triple S Project. Assessment of syndromic surveillance in Europe. *Lancet* 2011;378:1833–4.

2 Caserio-Schönemann C, Bousquet V, Fouillet A, Henry V. Le système de surveillance syndromique SurSaUD (R) (The syndromic surveillance system SurSaUD). *Bull Epidémiol Hebd* 2014;3:38–44.

3 Lassalle M, Caserio-Schönemann C, Gallay A, et al. Pertinence of electronic death certificates for real-time surveillance and alert, France, 2012–2014. *Public Health* 2017;143:85–93.

4 Fouillet A, Pavillon G, Vicente P, et al. La certification électronique des décès, France, 2007–2011. *Bull Epidemiol Hebd* 2012;2012:7–10.

5 World Health Organization. Health statistics and information systems-Completeness and coverage of death registration data: WHO, 2018. Available from: http://www.who.int/healthinfo/statistics/mortcoverage/en/ (9 July 2018, date last accessed).

6 National Institute for Statistics and Economic Studies. Démographie—Nombre de décès—France métropolitaine Paris. Insee, 2017. Available at: https://www.insee.fr/fr/statistiques/serie/000436394? idbank=000436394 (16 August 2019, date last accessed).

7 Gergonne B, Mazick A, O'Donnell J. A European algorithm for a common monitoring of mortality across Europe 2011. Available at: http://www.euromomo.eu/methods/pdf/wp7_report.pdf#page=18&zoom=auto,-274, 842 (25 July 2017, date last accessed).

8 Équipes de surveillance de la grippe. Surveillance épidémiologique, clinique et virologique de la grippe en France métropolitaine: saison 2011–2012 (Epidemiologic, virologic and clinic Influenza surveillance in metropolitan region in France: 2011–2012 season). *Bull Épidemiol Hebd* 2012;38:424–7.

9 Équipes de surveillance de la grippe. Surveillance épidémiologique et virologique de la grippe en France, saison 2012–2013 (Epidemiologic and virologic Influenza surveillance in France. 2012–2013 season). *Bull Épidemiol Hebd* 2013;32:394–401.

10 Équipes de surveillance de la grippe. Surveillance épidémiologique et virologique de la grippe en France métropolitaine. Saison 2013-2014 (Epidemiologic and virologic Influenza surveillance in metropolitan regions in France: 2013–2014 season). *Bull Epidémiol Hebd* 2014;28:460–5.

11 Équipes de surveillance de la grippe. Surveillance de la grippe en France métropolitaine. Saison 2014–2015 (Influenza surveillance in metropolitan regions in France. 2014–2015 season). *Bull Epidémiol Hebd* 2015;32–33:593–8.

12 Équipes de surveillance de la grippe. Surveillance de la grippe en France métropolitaine, saison 2015–2016 (Influenza surveillance in metropolitan regions in France: 2015–2016 season). *Bull Epidémiol Hebd* 2016;32–33: 558–63.

13 Carrat F, Flahault A, Boussard E, et al. Surveillance of Influenza-Like Illness in France. The Example of the 1995/1996 Epidemic. *J Epidemiol Community Health* 1998;52:32S–8.

14 Santé Publique France. Chaleur et Santé/Actualité/Archive (Heat and Health/News/Archive). Paris: CoreTechs. Available at: http://invs.santepubliquefrance.fr/Dossiers-thematiques/Environnement-et-sante/Climat-et-sante/Chaleur-et-sante/Actualites (22 August 2017, date last accessed).

15 Santé publique France. Bulletin de veille sanitaire Antilles-Guyane. n°3-4-5 - Septembre-Novembre 2014 (Health monitoring bulletin for the French West Indies and Guiana n°3-4-5 - September-November 2014). Paris: CoreTechs; 2014. Available at: http://invs.santepubliquefrance.fr/Publications-et-outils/Bulletin-de-veille-sanitaire/Tous-les-numeros/Antilles-Guyane/Bulletin-de-veille-sanitaire-Antilles-Guyane.-n-3-4-5-Septembre-Novembre-2014 (27 July 2017, date last accessed).

16 Mølbak K, Espenhain L, Nielsen J, et al. Excess mortality among the elderly in European countries, December 2014 to February 2015. *Eurosurveillance* 2015;20:21065.

17 Matias G, Taylor R, Haguinet F, et al. Estimates of mortality attributable to influenza and RSV in the United States during 1997–2009 by influenza type or subtype, age, cause of death, and risk status. *Influenza Other Respir Viruses* 2014;8:507–15.

18 Thompson WW, Shay DK, Weintraub E, et al. Mortality associated with influenza and respiratory syncytial virus in the United States. *JAMA* 2003;289:179–86.

19 Donaldson GC, Keatinge WR. Excess winter mortality: influenza or cold stress? Observational study. *BMJ* 2002;324:89–90.

20 Mayor S. Cold weather kills far more people than hot weather, study shows. *BMJ* 2015;350:h2740.

21 Basu R, Samet JM. Relation between elevated ambient temperature and mortality: a review of the epidemiologic evidence. *Epidemiol Rev* 2002;24:190–202.

22 Bunker A, Wildenhain J, Vandenbergh A, et al. Effects of air temperature on climate-sensitive mortality and morbidity outcomes in the elderly; a systematic review and meta-analysis of epidemiological evidence. *EBioMedicine* 2016;6:258–68.

23 Gasparrini A, Guo Y, Hashizume M, et al. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. *Lancet* 2015;386:369–75.

24 Basu R. High ambient temperature and mortality: a review of epidemiologic studies from 2001 to 2008. *Environ Health* 2009;8:40.

25 Ye X, Wolff R, Yu W, et al. Ambient temperature and morbidity: a review of epidemiological evidence. *Environ Health Perspect* 2012;120:19–28.

26 Pialoux G, Gaüzère B-A, Jauréguiberry S, Strobel M. Chikungunya, an epidemic arbovirosis. *Lancet Infect Dis* 2007;7:319–27.

27 Economopoulou A, Dominguez M, Helynck B, et al. Atypical Chikungunya virus infections: clinical manifestations, mortality and risk factors for severe disease during the 2005–2006 outbreak on Reunion. *Epidemiol Infect* 2009;137:534–41.

28 Josseran L. Chikungunya Disease Outbreak, Reunion Island. *EID J* 2006;12.

29 Lefeuvre D, Pavillon G, Aouba A, et al. Quality comparison of electronic versus paper death certificates in France, 2010. *Popul Health Metr* 2014;12:3.

Automatic classification of free-text medical causes from death certificates for reactive mortality surveillance in France.

International Journal of Medical Informatics 131 (2019) 103915

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Automatic classification of free-text medical causes from death certificates for reactive mortality surveillance in France

Yasmine Baghdadi[a,*,1], Alix Bourrée[a,1], Aude Robert[b], Grégoire Rey[b], Anne Gallay[c], Pierre Zweigenbaum[d], Cyril Grouin[d], Anne Fouillet[a]

[a] Santé publique France, Division for Data Science, Saint-Maurice, France
[b] CépiDc–Inserm, Epidemiology Center on Medical Causes of Death, Kremlin-Bicêtre, France
[c] Santé publique France, Division of Non communicable Diseases and Injuries, Saint-Maurice, France
[d] LIMSI, CNRS, Université Paris-Saclay, Orsay, France

ARTICLE INFO

Keywords:
Automatic classification
Rule-based method
SVM
Evaluation performance
Medical causes of death
Syndromic surveillance

ABSTRACT

Background: Mortality surveillance is of fundamental importance to public health surveillance. The real-time recording of death certificates, thanks to Electronic Death Registration System (EDRS), provides valuable data for reactive mortality surveillance based on medical causes of death in free-text format. Reactive mortality surveillance is based on the monitoring of mortality syndromic groups (MSGs). An MSG is a cluster of medical causes of death (pathologies, syndromes or symptoms) that meets the objectives of early detection and impact assessment of public health events. The aim of this study is to implement and measure the performance of a rule-based method and two supervised models for automatic free-text cause of death classification from death certificates in order to implement them for routine surveillance.

Method: A rule-based method was implemented using four processing steps: standardization rules, splitting causes of death using delimiters, spelling corrections and dictionary projection. A supervised machine learning method using a linear Support Vector Machine (SVM) classifier was also implemented. Two models were produced using different features (*SVM1* based solely on surface features and *SVM2* combining surface features and MSGs classified by the rule-based method as feature vectors). The evaluation was conducted using an annotated subset of electronic death certificates received between 2012 and 2016. Classification performance was evaluated on seven MSGs (Influenza, Low respiratory diseases, Asphyxia/abnormal respiration, Acute respiratory disease, Sepsis, Chronic digestive diseases, and Chronic endocrine diseases).

Results: The rule-based method and the *SVM2* model displayed a high performance with F-measures over 0.94 for all MSGs. Precision and recall were slightly higher for the rule-based method and the *SVM2* model. An error-analysis shows that errors were not specific to an MSG.

Conclusion: The high performance of the rule-based method and *SVM2* model will allow us to set-up a reactive mortality surveillance system based on free-text death certificates. This surveillance will be an added-value for public health decision making.

1. Introduction

Mortality is the indicator generally used to measure the impact of an event on a given population; it is of fundamental importance to health surveillance. Syndromic surveillance is "a real-time (or near real-time) collection, analysis, interpretation, and dissemination of health-related data to enable the early identification of the impact (or absence of impact) of potential human or veterinary public health threats that require effective public health action" [1]. As part of the French syndromic surveillance system SurSaUD, the reactive mortality surveillance is carried out by Santé publique France, the French public health agency [2].

Abbreviations: CépiDc, epidemiology center on medical causes of death; EDRS, Electronic Death Registration System; ICD10, International Classification of Diseases 10th revision; MSG, Mortality Syndromic Group; NLP, Natural Language Processing; OOV, out-of-vocabulary; SVM, Support Vector Machine; WHO, World Health Organization

* Corresponding author.
E-mail address: yasmine.baghdadi@santepubliquefrance.fr (Y. Baghdadi).

[1] As first authors (Equal contribution).

https://doi.org/10.1016/j.ijmedinf.2019.06.022
Received 4 March 2019; Received in revised form 14 May 2019; Accepted 24 June 2019

Mortality surveillance is based on death certificates filled out by medical practitioners. Death certificates in France are modelled on a template recommended by the World Health Organization (WHO) [3], they are split into two sections: an administrative section and a medical section. The medical section is made up of two different parts: in Part I, the morbid sequence is reported in descending order from the immediate to the underlying causes of death (4 fields); in Part II, the unrelated but contributory causes of death are reported (2 fields).

Each medical part consists of free-text fields where physicians write down the one or more causes of death with one or more words which are frequently misspelled. The French Epidemiology Center on Medical Causes of Death (Inserm-CépiDc), the institution responsible for the French national mortality database, then codes the medical causes of death. The coding follows WHO rules, using the 10th revision of the International Classification of Diseases (ICD10) [3]. Based on these rules, the coding of one cause of death is determined according to the other causes stated in the certificate. Thus, the coding of the same cause can be different between two death certificates. For the mortality syndromic surveillance purpose we need to follow the causes of death independently of one another over time. Furthermore, it can take 6 to 24 months for the coded medical causes of the death to be made available for epidemiologists. Both the coding delay and the attribution rules of the ICD10 codes are not suitable to meet the objectives of reactive mortality surveillance in routine.

Since 2007, an Electronic Death Registration System (EDRS) allows the physician to certify deaths electronically. The added-value of EDRS for reactive mortality surveillance is to send free-text causes of death within a few minutes after the validation of the death certificate. A pilot study demonstrated that E-death certificates constitute reactive and valuable data for real-time mortality surveillance [4].

Reactive mortality surveillance is based on the routine monitoring of a hundred syndromic groups [5]. A Mortality Syndromic Group (MSG) is defined as a cluster of medical causes of death (pathologies, syndromes or symptoms) that meets the objectives of early detection and impact assessment of public health events. As E-death certificates are filledin a free-text format, the automatic classification of medical causes of death into MSGs requires the use of Natural Language Processing (NLP) methods. We have observed a constant increase of the use of such methods to classify medical information for health surveillance purposes, in the last two decades [6,7]. These methods are used to classify chief complaints and medical texts [8–12] or identify particular diseases in death certificates for surveillance purposes [13–17]. They are also used to automatically assign ICD10 codes to death certificates [15,18–20]. These studies provide evidence that NLP methods are useful tools for health surveillance. However, to the best of our knowledge, no study has so far proposed an automatic classification of each cause of death into Mortality Syndromic Groups for real-time mortality surveillance.

The objective of this study is to implement and evaluate the performance of a rule-based method and two supervised models chosen for their demonstrated high performances for classification. Those methods were tuned to classify free-text medical causes of death directly into selected MSGs.

2. Material and methods

The purpose of this study is to implement a method capable of automatically classifying each cause of death written in the fields of the death certificates into Mortality Syndromic Groups. We define the term "entity" as an expression of a cause of death in a field of the medical section of the death certificate.

Table 1 is a simple example of the medical section of a death certificate. In this certificate, the physician reported the morbid sequence that led to a patient's death (part I of the death certificate) in fields 1–3 and the unrelated but contributory causes of death (part II of the death certificate) in field 5. Thus, the physician filled up 4 out of 6 fields of the death certificate (3 out of 4 fields in part I and 1 out of 2 fields in part II). Each field of this certificate contains one or more words. Some are misspelled, like "parquison" (instead of "Parkinson") in field 5. Some are abbreviated (see fields 1 and 3). One field contains multiple entities (Field 5). These entities are not systematically separated by punctuation. The content of death certificates is a little more complex than the one presented here.

Based on the medical section of the death certificate described in Table 1, Table 2 illustrates what we want to achieve through automatic classification.

We will describe hereafter the characteristics of the free-text medical causes of death and the data collection performed in this study. This helps to understand the design of the classification methods. The classification methods and their evaluation measures are then described.

2.1. Data source

Data consisted of the medical section of e-death certificates received routinely at Santé publique France between 2012 and 2016. These data were available anonymized. Each certificate contained free-text medical causes of death. For this study we used 4500 death certificates.

2.2. Corpus characteristics: training and test split

The data was split by date of death (year) because this split simulates the realistic setting in which the mortality surveillance system will be used: a first period composed of the years 2012 to 2015, and a second period composed of 2016. In a real-world setting, the classifiers will only be trained on retrospective data to classify the data of the current period of analysis. The training set (2000 death certificates) and the development set (1500 death certificates) were extracted from the first period and the test set (1000 death certificates) was extracted from the second period. All sets were extracted according to a random sampling without replacement.

These 4500 death certificates were doubly-annotated manually by two annotators: both are native French-speaking epidemiologists who previously worked on the definition of MSGs. Both annotators used the detailed definition of MSGs in order to assign one or more MSGs to each field of the medical section of each death certificate. Each MSG consists of a list of ICD10 codes (conversely, each code belongs to a unique MSG). The annotation was also facilitated by the Inserm-CépiDc dictionary, which contains every single expression of medical causes of death, starting from the early 2000s and their associated ICD10 codes. This dictionary was revised in order to fit for the objectives of syndromic surveillance.

All double annotations were then compared. The inconsistencies were discussed between annotators; MSG definitions and the Inserm-CépiDc dictionary were both used to resolve conflicts. The agreement rate (proportion of identical annotations) was 0.90 in the test set. The final annotated sets are the ground truth against which the methods tested were compared and evaluated.

2.3. Data preprocessing

The data was preprocessed in 6 different steps: 1/ converting letters to lowercase, 2/ removing diacritics, 3/ standardizing compound words (replacing hyphens with a space to obtain two distinct words), e.g. *"cardio-vasculaire" becomes "cardio vasculaire" and "cardiovasculaire" becomes "cardio vasculaire"*, 4/ replacing punctuation (except comma, simple quote, semicolon)[2] with a space, 5/ removing prepositions (de, d, du, le, les, la…), 6/ removing the additional spaces at the beginning

[2] These punctuations were not removed because they are useful in the segmentation step

Table 1
Example of an e-death certificate.

Certificate ID	Date of death	Age	Gender	Fields/rank	Free-text causes of death
22203632	13.09.2012	41	M	1-0	IDM[1]
22203632	13.09.2012	41	M	2-0	cardiopathie ischémique[2]
22203632	13.09.2012	41	M	3-0	insuffisance resp aigüe[3]
22203632	13.09.2012	41	M	5-0	Surpoids /diabète évoluée malaidie parquinson[4]

[1] MI: Myocardial infarction.
[2] Ischemic heart disease.
[3] Acute respiratory failure.
[4] Overweight/ advanced diabetes Parkinson disease.

Table 2
Example of the output of automatic classification of medical causes of death into MSGs.

Certificate ID	Date of death	Age	Gender	Fields/rank	Free-text causes of death	attributed MSGs
22203632	13.09.2012	41	M	1-0	IDM[1]	Acute coronary syndrome
22203632	13.09.2012	41	M	2-0	cardiopathie ischémique[2]	Cardiac and circulatory chronic diseases
22203632	13.09.2012	41	M	3-0	insuffisance resp aigüe[3]	Acute respiratory failure
22203632	13.09.2012	41	M	5-0	Surpoids /diabète évoluée malaidie parquinson[4]	Chronic endocrine diseases, Chronic digestive diseases, Chronic diseases of the nervous system

[1] MI: Myocardial infarction.
[2] Ischemic heart disease.
[3] Acute respiratory failure.
[4] Overweight/ advanced diabetes Parkinson disease.

and at the end of the text, in order to obtain a narrower set of more relevant terms.

2.4. Rule-based method

The rule-based method relies on the dictionary provided by Inserm-CépiDc on which the preprocessing steps were applied. It contains over 150,000 expressions found in death certificates since the early 2000s and their assigned ICD10 codes. This dictionary was initially built to facilitate and homogenize the coding. It provides a wide variety of expressions of a same cause with spelling variations (Table 3). Since each MSG is defined by a list of ICD10 codes, it was possible to automatically associate an MSG to every expression presented in the dictionary, thereby obtaining an MSG dictionary.

The principle of the rule-based method was to assign an MSG to an entity found in a death certificate when it matched an expression found in the dictionary (Fig. 1). To detect which entities in a death certificate were closest to the expressions of the dictionary, we applied 3 processing steps that normalize and split the fields into entities and a dictionary projection (fourth step). After each step, if an entity entirely matches a dictionary expression, an MSG is assigned and the processing stops for that expression.

The first step used about nine hundred standardization rules to normalize entities. These rules were initially created by Inserm-CépiDc for the coding task [21], and were modified for the needs of the study. These rules were based on regular expressions that detect and replace each expression's synonymous wording or abbreviation with a single standardized expression. For the standardization to occur satisfactorily, the rules have to be ordered in a precise manner.

The second step used surface cues to segment text into distinct entities. These cues included punctuations (",", ";"), words ("avec", "après", "sur", "puis", "pour", "en rapport avec", "en raison de" and "suite à"[3]) and other signs which marked out entity boundaries within a same field. Two coordination conjunctions ("et", "ou") [4] were often used to factorize entities (e.g., "Hepatite A et B"): we expanded such

[3] "with", "after", "on", "then", "for", "in connection with", "because of" and "following"
[4] "and", "or"

Table 3
Sample of the Inserm-CépiDc dictionary, illustrating different expressions of causes of death with their related ICD10 code, grouped by mortality syndromic group.

Mortality syndromic groups	Number of entities in the dictionary	Example of entities	ICD10 codes
Stroke/cerebral hemorrhage	3200	accident cérébral	I64.09
		accident cérébral vasculaire	I64.09
		rupture anévrisme cérébrale	I60.79
		rupture anévrisme méningé	I60.80
		...	
Influenza	109	grippe	J11.10
		affection grippale	J11.10
		état grippal	J11.10
		syndrome grippal	J11.10
		...	
Viral hepatitis	331	hépatite A	B15.9
		hépatite B	B18.1
		hépatite C	B18.2
		hépatite VHB	B18.1
		HVB	B18.1
		...	
Acute respiratory failure	78	insuffisance respiratoire aigüe	J96.0
		IRA	J96.0
		détresse respiratoire aigüe	J96.0
		...	
Chronic respiratory diseases	3533	détresse respiratoire chronique	J96.1
		insuffisance bronchique	J98.0
		insuffisance respiratoire	J96.9
		...	
Chronic diseases of the nervous system	4192	maladie dégénérative type Parkinsonnienne	G2009
		maladie Parkinson	G20.09
		...	

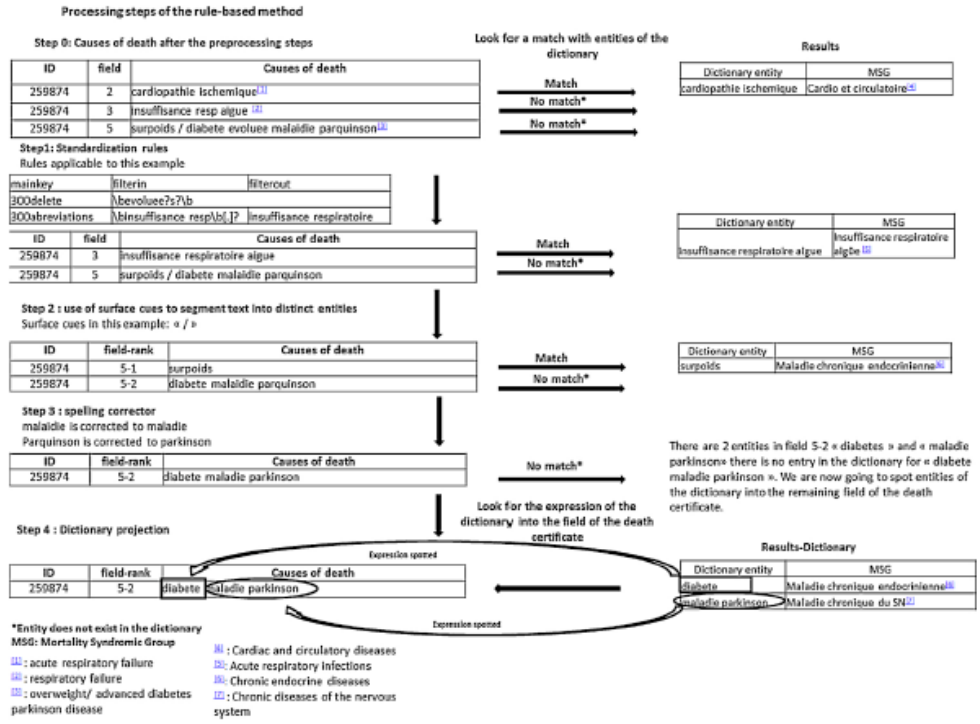


Fig. 1. Processing steps of the rule-based method.

factorizations into two separate entities (e.g., "Hepatite A", "Hepatite B"). The word "sous" was also used as a specific delimiter for expressions such as "ACFA sous anticoagulant"[5] . The entity was segmented before the word "sous", and "sous" was kept with the second part of the entity.

The third step applied a spelling corrector to out-of-vocabulary (OOV) words. The spelling corrector was implemented using an information retrieval method: both dictionary words and OOVs were represented as bags-of-character-unigrams and bigrams. Each OOV was used as a query, and the closest dictionary word was selected and returned.

The fourth step, text scanning, aims at spotting mentions of one or more expressions of the dictionary within each remaining entity of the death certificate obtained after the 3 previous steps. It performed a left-to-right scan of each field, aimed at finding exact matches of the longest group of words with the dictionary expressions, without having any overlaps between groups.

2.5. Machine learning method

Among the possible supervised machine learning methods, we chose a linear Support Vector Machine (SVM) classifier [20,22]. For classification tasks, this method demonstrated a high performance. For instance Butt et al. [14] studied a number of machine learning classifiers to detect cancer cases from free-text death certificates and obtained the best performance with a SVM. Likewise, Koopman and al. [16] developed key-word matching rules and a SVM model to classify free-text death certificates into four diseases. Both methods had a high performance score with F-measures over 0.95. This method also achieved a high performance when using death certificates for an ICD10 coding task during the CLEF e-Health challenges [18,23].

We used a SVM with linear kernel to classify each field of each death certificate, using the training data to set up the parameters. The default hyperparameter settings were selected (C parameter equal to 1). Since it may be possible to associate 0, 1 or several MSGs to each field of a death certificate, we performed a multi-label classification using the one-versus-all strategy [22], which internally trains one SVM for each class.

We implemented two models: *SVM1* for which only features are the surface features, whereas *SVM2* used a combination of surface features and the MSGs previously attributed using the rule-based method as its features. To extract the features we first tokenized the fields of the death certificates using spaces and non-alphanumeric characters as delimiters. Among the usual surface features, we selected bags-of-word-unigrams (set of unique words for each example in the training set), bags-of-word-bigrams (set of unique sequences of two consecutive words for each example in the training set) and character trigrams (providing a degree of robustness to spelling errors). Previous work showed that these surface features provided the best classification performance in a similar task with the same source of data [20]. Different combinations of these surface features were tested during the development stage. The performance of the two combinations of surface features that achieved the highest performance for most of the MSGs (a SVM model with bags-of-word-unigrams and bigrams and *SVM1*) are presented as well as the performance of *SVM2*. All features were modelled using word count, which indicates the number of features contained in each field of death

[5] CAAF (Cardiac arrhythmia by atrial fibrillation) on anticoagulant

certificates.

2.6. Evaluation measures

Three evaluation measures were considered: precision, recall, and F-measure. The study focused on the evaluation of the classification performances of seven MSGs chosen for their three main dimensions: their epidemiological relevance to mortality surveillance, their frequency of occurrence in death certificates, and the closeness of their definition (mainly acute vs chronic diseases). We then verified if the distributions of MSGs in the training, development and test sets were similar to the distribution of MSGs (Appendix A) in the whole dataset of death certificates. The seven MSGs were: Influenza, Low acute respiratory infections, Asphyxia and abnormal respiration, Acute respiratory failure, Sepsis, Chronic digestive diseases, and Chronic endocrine diseases (Detailed definitions in Appendix B).

The statistical differences of performance between pairwise methods were measured with a two-tailed Z-test. Classifiers were implemented using the Scikit-learn library [24] and its LinearSVC class.

3. Results

3.1. Final test set

Table 4 presents the detailed classification performance results for each selected MSG obtained at each step of the rule-based method and obtained with the SVM models combining different features.

Performance was measured on the test set of one thousand death certificates. In this set six hundred and twenty six fields (22%) contained at least one entity belonging to the selected MSGs. Models were trained with three thousand five hundred death certificates (training and development sets).

We observed that the successive steps of the rule-based method mainly improved the recall (Table 4). Two rules were particularly involved in the increase of the classification performance: the text segmentation (F-measure increased of +1 to +13 Pt according to the MSGs from step 1 to step 2) and the dictionary projection (F-measure increased of +1 to +20 Pt according to the MSGs from step 3 to step 4). The rule-based method finally obtained an F-measure superior or equal to 0.96 for all MSGs. Precision was higher than recall, respectively, superior or equal to 0.98 and superior or equal to 0.95 (Table 4).

The *SVM1* model achieved an F-measure slightly higher than the

Table 4
Evaluation of classifaction performance of the rule-based method combining one step at a time and the SVM models combining different features-Test set (1000 E-death certificates, year 2016)-France.

	All 7 MSGs	Influenza	Acute respiratory failure	Low acute respiratory infections	Asphyxia / Abnormal breathing	Sepsis	Chronic endocrine diseases	Chronic digestive diseases
Number of fields	**626**	**3**	**86**	**115**	**36**	**119**	**134**	**133**
				Rule-based method				
Without steps [a]								
Precision	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
Recall	0.61	0.67	0.81	0.61	0.64	0.60	0.37	0.57
F-measure	0.75	0.80	0.90	0.76	0.78	0.75	0.54	0.72
Step 1 [b]								
Precision	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
Recall	0.64	0.67	0.81	0.77	0.64	0.61	0.39	0.59
F-measure	0.77	0.80	0.90	0.87	0.78	0.75	0.56	0.74
Step 1 + 2 [c]								
Precision	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.99
Recall	0.78	0.67	0.91	0.86	0.72	0.79	0.75	0.78
F-measure	0.87	0.80	0.95	0.92	0.84	0.88	0.86	0.87
Step 1 + 2 +3[d]								
Precision	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.99
Recall	0.80	0.67	0.94	0.90	0.72	0.79	0.76	0.82
F-measure	0.87	0.80	0.97	0.94	0.84	0.79	0.86	0.90
Rule-based method [e]								
Precision	0.99	1.00	1.00	0.98	1.00	0.99	1.00	0.98
Recall	0.97	1.00	0.97	0.95	0.97	0.95	0.99	0.95
F-measure	0.98	1.00	0.98	0.96	0.98	0.97	0.99	0.96
				Machine learning method				
Bags-of-word- unigrams and bigrams								
Precision	0.97	1.00	1.00	0.94	0.97	0.98	1.00	0.91
Recall	0.89	0.60	0.95	0.95	0.83	0.99	0.94	0.87
F-measure	0.92	0.75	0.97	0.94	0.90	0.98	0.97	0.89
Model *SVM1* [f]								
Precision	0.96	1.00	0.98	0.96	0.96	0.98	1.00	0.86
Recall	0.94	1.00	0.96	0.95	0.86	0.96	0.94	0.90
F-measure	0.95	1.00	0.97	0.95	0.91	0.97	0.97	0.88
Model *SVM2* [g]								
Precision	0.99	1.00	1.00	0.99	1.00	0.98	1.00	0.94
Recall	0.97	1.00	0.96	0.99	0.91	0.98	0.97	0.94
F-measure	0.98	1.00	0.98	0.99	0.95	0.98	0.98	0.94

[a] Look for a match in a dictionary without any rules.
[b] Standardization rules.
[c] Standardization rules and Segmentation of the text.
[d] Standardization rules and Segmentation of the text and Spelling corrector.
[e] Standardization rules and Segmentation of the text and Spelling corrector and dictionary projection.
[f] Bags-of-word- unigrams and bigrams + characters trigrams.
[g] Bags-of-word-unigrams and bigrams + characters trigram + mortality syndromic groups classified by the rule-based method.

Table 5
Break down of the classification errors according to different categories for the 3 models-Test set (1000 death certificates, year 2016)-France.

	All Models		Rule-based method		*SVM1* model		*SVM2* model	
Categories	Total errors	% of records	Total errors	% of records	Total errors	% of records	Total errors	% of records
Classification errors	113	96.6	22	18.0	63	53.8	28	23.9
Word variation (spelling errors)	18	15.4	6	5.1	7	6.0	5	4.2
Word combinations	23	19.7	0	0.0	18	15.5	5	4.2
Class confusion errors (Closeness of the definition of the MSG)	38	32.5	4	3.4	23	19.7	11	9.4
Absence of entity in the dictionary	12	10.3	12	10.3	0	0.0	0	0.0
Absence of the entity of the training set	22	18.8	0	0.0	15	12.8	7	6.0
Ground truth issues								
Annotation errors	4	3.4	2	1.7	1	0.9	1	0.8

SVM model with the combination of bags-of-words-unigrams and bi-grams as surface features, for all the MSGs except for Chronic digestive diseases and Sepsis.

The *SVM1* model achieved an F-measure superior or equal to 0.95 for 5 out of 7 MSGs (Influenza, Acute respiratory failure, Low acute respiratory infections, Sepsis and Chronic endocrine diseases). Their precision and recall were either superior or equal to 0.95 except for Chronic endocrine disease (recall equal to 0.94). F-measures of the model *SVM1* for the two other MSGs (Asphyxia/abnormal breathing and Chronic digestive diseases) were respectively 0.88 and 0.91. Precision was higher than the recall except for Chronic endocrine diseases where the recall was higher than the precision (Table 4).

The model *SVM2* demonstrated a high performance, F-measure were over or equal to 0.98 except for Chronic digestive diseases (F-measure equal to 0.94) and Asphyxia/abnormal breathing (F-measure equal to 0.95). The precision and recall for these two MSGs were respectively 0.94 and 0.94 for Chronic digestive diseases and 1 and 0.91 for Asphyxia/abnormal breathing (Table 4).

3.2. Error analysis

To further understand the classification issues, we conducted a review of the classification errors. The lead author reviewed one hundred and seventeen fields of death certificates that contained at least one entity incorrectly classified: twenty-four for the rule-based method, sixty four for the model *SVM1* and twenty nine for the model *SVM2*. Among those fields, some were misclassified by more than one method.

Two main categories of errors were identified: actual classification errors (96.6%) and ground truth errors (3.4%). These errors were not specific to an MSG. Among the classification errors we observed (Table 5):

- **Word variation or spelling errors (15.4%):** Lexical variant of an entity or spelling variation that led to misclassification or absence of classification. For example, "detreses respiratoire aigue"[6] contained a spelling error and should have been written "detresse respiratoire aigue".
- **Word combination (19.7%):** Words which should have been taken separately but were combined, to constitute a single entity with an alternative meaning than the words intended. For instance "fausse route alimentaire asphyxiante"[7] contained two entities ("fausse route alimentaire", "asphyxiante") that should have been classified into two different MSGs "Other external causes of death" and "Asphyxia and abnormal breathing" but weren't.
- **Class confusion (32.5%):** The entity was detected by the model, but the incorrect class was assigned. This error was mainly due to closeness of the definition of two MSGs. For example "volvulus sigmoide"[8] was classified as an "Acute digestive diseases" while it should have been classified in "Chronic digestive diseases".
- **Absence of the entity in the dictionary (10.3%):** This type of error is specific to the rule-based method as it relies on the dictionary. If an entity was missing in the dictionary, the rule-based method could not have assigned an MSG to it. For example, "ischemie duodeno pancreatique"[9] or "sepsis point depart pulmonaire"[10] were entities that did not exist in the dictionary.
- **Absence of the entity in the training set (18.8%):** This error is specific to the SVM models. The training set was constituted of 3500 annotated death certificates. It was not a sufficient amount to cover all the different ways to write an entity. For example, "surpoids"[11] was missing in the training set. The classifier did not assign any MSG to this entity.
- **Annotation errors (3.4%):** There were entities for which the ground truth appeared incorrect. For example "Broncho pneumopathie sur inhalation"[12] should have been classified by the annotator into "Chronic respiratory diseases" but was classified into "Low acute respiratory infections". Even though the classifiers accurately assigned the MSG to the expression, it remains nonetheless a classification error.

4. Discussion and related work

The performance of the rule-based method and of the *SVM2* model was high, with F-measures over 0.94. The precision was higher than the recall for both methods, showing that false negatives were more common than false positives. The fact that precision is higher than recall may be reflective of fields of death certificates that contained many entities, some of which could not be detected by the system. The model *SVM1* had slightly lower scores than the other two methods, with F-measures varying from 0.91 to 1 except for Chronic digestive diseases (0.88). However no significant statistical differences were observed between the performance of the rule-based method and the model *SVM1* or between the model *SVM1* and the model *SVM2*. Precision was also higher than recall for this model *SVM1*. The size of the training set and the variability of expressions of the causes of death belonging to an MSG could be a plausible explanation for these results. The error analysis showed that errors were not specific to the targeted MSGs and might occur with any MSG. Remaining errors might be due to the complexity of human languages and their use by speakers, and the technical difficulty to capture this complexity.

These results are consistent with those found by Koopman [16], Muscatello [25], and Shah [26]. However, it is important to note that

[6] Acute respiratory distress.
[7] Asphyxiating ingestion of food through the wrong track.
[8] Sigmoid volvulus.
[9] Pancreatic duodeno ischemia.
[10] Sepsis with pulmonary starting point.
[11] Overweight.
[12] Bronchopneumopathy related to inhalation.

they identified specific pathologies (Pneumonia, Influenza, HIV, and Diabetes) whereas we identified MSGs that included a wider range of diseases (Chronic endocrine diseases, Chronic digestive diseases).

For real-time mortality surveillance both precision and recall are important. Indeed, to meet the objective of a reactive detection of events, high precision is needed to limit false alarms. To measure the impact of an event, the surveillance system must have a high recall, in order to avoid underestimating the impact of an event. This is especially true for rarer diseases.

When choosing a model, it would be too simplistic to only consider the overall effectiveness as our sole criterion. In the case of daily health surveillance, the ease of which a model can be updated in terms of new rules for the rule-based method, or in terms of new data for the training stage for supervised learning, is of considerable importance. As rules are defined manually, to update the rules is to change the computer program. Supervised models require a sufficient quantity of suitably labeled training data in order to function correctly. Rules are computationally quite simple and use few resources. Machine learning methods require the extraction of the features of the death certificates in order to train the classifier; this stage can be computationally expensive for larger datasets [16,17]. Finally, these two methods seem to be complementary, which supports the use of a combined model.

Furthermore, the use of these methods for daily reactive mortality surveillance also implies that the running time of said methods must be short. We measured the running times of both methods by simulating their use in a daily setting : we trained the models with the whole set of annotated death certificates (4500 death certificates) ; and asked the methods to predict the MSGs of a sample of death certificates approximating the daily number of deaths observed in France (2000 death certificates).

For the training phase, the *SVM1* model took 2 min to train itself, while the *SVM2* model took 18 min. For the classification of the sample the *SVM1* model took less than 30 s, while the *SVM2* model took 4 min 50 s to predict the MSGs. Finally, the rule-based method took 4 min 40 s to predict the sample. These running times were measured using the same conditions as in the rest of the study (128 Gb RAM, 32 threads) and are suitable for daily usage, whichever model is chosen. But we can expect that the running time would be longer for *SVM* models with a larger training sample of death certificates.

The use of electronic death certificates for real time mortality surveillance will complete the reactive surveillance currently based on administrative data enabling a quantitative surveillance [2]. The routine analysis of classified MSGs will enable to detect an unusual increase of mortalityand an alert may be trigger for decision makers to set up an adapted response. It will also enable to estimate the portion attributable to an identified event (epidemics, environmental events) that impact mortality.

5. Limitation and future works

In this study, we focused on the evaluation of classification performance for seven MSGs among a hundred identified for reactive mortality surveillance [5].

In the two manually annotated sets, every MSGs was assigned. Through error analyses, the models were designed to perform optimally for seven selected MSGs during the training stage. In order to set up the French reactive mortality surveillance system based on medical causes of death, we have to verify that this performance holds up for other MSGs.

Furthermore, we are reminded that the training set contained 3500 annotated death certificates. This might be too small a number of death certificates to accurately represent the various expressions of causes of death. An additional number of annotated death certificates could enable future studies to improve the results of the SVM models.

We focused our study on the development of a rule-based method and one machine learning method because they meet our goal and respect major requirements: methods easy to run for a daily use, methods easily understandable for epidemiologists and methods that can be regularly and timely updated. Besides, these methods were known to achieve high performance. More recent studies explored the performance of deep learning methods to automatically code causes of death in ICD10 [27,28]. The authors found that the SVM model outperformed some of the simpler neural architectures, but that a combined neural network outperformed the SVM [27]. These methods could constitute an interesting perspective to compare and improve our classification performance and could also simplify the manual features engineering necessary for the SVM.

6. Conclusion

The daily analysis of free-text medical causes of death is essential to upgrade the current mortality surveillance system, presently solely based on administrative information from civil registry offices [2]. Such a system would be able to provide specific information to public health authorities regarding causes of deaths during public health crises (e.g. Dengue fever outbreaks experienced in Reunion Island since September 2018) or during periods of excess mortality (e.g. winter epidemics), this can help decision makers, to adapt public service announcements and adopt new counter measures in the future. Approximately fifteen percent of French mortality has been reported with EDRS in 2018. This rate is likely to increase in the coming months, thanks to the actions taken to encourage hospitals as well as general practitioners, to use the EDRS.

This study measured the performance of two methods to classify free-text medical causes of death into MSGs specifically defined for reactive mortality surveillance and alert. The high performance of the rule-based method and of the *SVM2* model for classification supports the relevance of implementation of these methods for real-time French mortality surveillance.

Authors' contributions

AB, YB and AF prepared data and conducted the analyses. AB wrote the code and YB wrote the manuscript. AR and GR provided medical resources, particularly the dictionary and standardization rules. CG and PZ provided methodological support for free-text analysis. AG and AF provided epidemiological support on syndromic surveillance. All authors contributed to the interpretation of results. All have revised and validated the manuscript.

Funding

This research did not receive any specific grant from funding agencies of the public, commercial, or not-for-profit sector.

Summary table

What was already known on the topic

- Medical causes of death from E-death certificates are valuable data for reactive mortality surveillance systems.
- Support Vector Machine achieved high performance for classification tasks.

What this study added to our knowledge

- The study reinforces the evidence that a traditional method and SVM models are highly effective in achieving classification of medical data in free-text for different objectives.
- The use of these methods is suitable for the daily analysis of causes of death in free-text format.
- The high performance of these methods reinforces the relevance for the creation of a reactive mortality surveillance system based on free-text causes of death.

A new approach to compare performance of two classification methods of causes of death for timely surveillance in France

Yasmine Baghdadi[a], Alix Bourrée[a,b], Aude Robert[c], Grégoire Rey[c], Anne Gallay[a], Pierre Zweigenbaum[b], Cyril Grouin[b], Anne Fouillet[a]

[a] *Santé publique France, Saint-Maurice, France*
[b] *LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France*
[c] *Epidemiology Center on Medical Causes of Death (Inserm-CépiDc), Kremlin-Bicêtre, France*

Abstract

Timely mortality surveillance in France is based on the monitoring of electronic death certificates to provide information to health authorities. This study aims at analyzing the performance of a rule-based and a supervised machine learning methods to classify medical causes of death into 60 mortality syndromic groups (MSGs). Performance was first measured on a test set. Then we compared the evolutions of the monthly numbers of deaths classified into MSGs from 2012 to 2016 using both methods. Among the 60 MSGs, 31 achieved recall and precision over 0.95 for either one or the other method on the test set. On the whole dataset, the correlation coefficient of the monthly numbers of deaths obtained by the two methods were close to 1 for 21 of 31 MSGs. This approach is useful to analyze a large number of categories or when the annotated resources are quite limited.

Keywords: Classification, Machine learning, Syndromic surveillance, Mortality

Introduction

Syndromic surveillance is "a real-time (or near real-time) collection, analysis, interpretation, and dissemination of health-related data to enable the early identification of the impact (or absence of impact) of potential human or veterinary public health threats that require effective public health action" [1].

Mortality is the indicator traditionally used to measure the severity of the impact of an event on the population; it is of fundamental significance to health surveillance. As part of the French syndromic surveillance system SurSaUD, timely mortality surveillance is carried out by Santé publique France, the French public health agency [2].

Mortality surveillance is based on death certificates. They consist of an administrative part and a medical part based on a model developed by the World Health Organization (WHO) [3]. The medical part is divided into two sections: first, the morbid sequence is reported in Part I in descending order from immediate to underlying causes of death (4 fields); second, the unrelated but contributory causes of death are reported in Part II (2 fields).

The medical part consists of free-text fields where physicians write down the one or more causes of death with one or more words which are frequently misspelled. The French Epidemiology Center on Medical Causes of Death (Inserm-CépiDc), the institution responsible for the French national mortality database, then codes the medical causes of death. The coding follows WHO rules, using the 10th revision of the International Classification of Diseases (ICD10). It can take 6 to 24 months for the coded medical causes of the death to be made available for epidemiologists. Efforts at automating this coding task to reduce this delay are currently made through several shared tasks during the CLEF eHealth campaigns [4,5].

Since 2007, an Electronic Death Registration System (EDRS) allows physicians to certify deaths electronically. The added-value of EDRS for reactive mortality surveillance is to send free-text causes of death within a few minutes after the validation of the death certificate. A pilot study demonstrated that E-death certificates constitute valuable data for real-time mortality surveillance [6]. Timely mortality surveillance is based on the routine monitoring of syndromic groups [7]. A Mortality Syndromic Group (MSG) is defined as a cluster of medical causes of death (pathologies, syndromes or symptoms) that meets the objectives of early detection and impact assessment of public health events. Definitions of MSGs are based on a list of ICD-10 codes, each code belonging to a unique MSG. We defined a hundred MSGs: sixty MSGs gathered acute or severe causes of deaths and are monitored for early alert purposes and forty gathered. chronic diseases. All are used for impact evaluation of public health events. They span twenty different topics including Cardiac and circulatory conditions, Respiratory conditions, Cancer, Infectious diseases, Digestive conditions, Genito-urinary conditions, Symptoms, Poisoning and injuries, etc.

As E-death certificates are entered in free-text format, the automatic classification of medical causes of death into MSGs requires the use of Natural Language Processing (NLP) methods.

In a previous study, we have implemented and evaluated two methods to classify causes of death into MSGs: a rule-based method and a Support Vector Machine (SVM) method. Two models were developed using the SVM method: one was based on surface features and the other was a hybrid model combining surface features and features

obtained by the rule-based method. We evaluated the performance of the methods on seven out of one hundred MSGs (Influenza, Low respiratory diseases, Asphyxia and abnormal respiration, Acute respiratory disease, Sepsis, Chronic digestive diseases, and Chronic endocrine diseases). These MSGs were chosen to illustrate the variation within three main dimensions: their epidemiological relevance to mortality surveillance, their frequency of occurrence in death certificates, and the relatedness of their definitions (e.g., acute vs chronic diseases). The results showed that the rule-based method and the hybrid model displayed high performance with F-measures over 0.94 for all seven MSGs. The SVM model achieved lower performance compared to the other two models with F-measures varying from 0.91 to 1 except for Chronic digestive diseases (0.88).

The objectives of the present study were 1/ to analyze the performance of the rule-based method and the hybrid model to classify causes of death into sixty additional MSGs and 2/ to compare the evolution of the monthly numbers of certificates classified in MSGs from 2012 to 2016 using both methods. This second analysis was also performed on the seven MSGs from the previous study.

Materials and methods

1 Corpus characteristics

Data consisted of the medical section of e-death certificates received routinely between 2012 and 2016 (203,797 certificates). E-death certification has been gradually rolled out, recording 5% of the national mortality in 2012 and 12% in 2016. These data were available anonymized. They contained both administrative and demographic information, and free-text medical causes of death. For the evaluation of performance of the classification methods, we used 4,500 annotated death certificates. We define the term “entity” as an expression of a cause of death on a field of the medical part of the death certificate.

2 Training and test splits

Using a random sampling without replacement the dataset was split into two parts: a training set using death certificates from 2012 to 2015 and a test set using death certificates from 2016. Three thousand five hundred certificates were annotated for the training set and one thousand for the test set. For this annotation task, each annotator had the detailed definition of MSGs, based on lists of ICD-10 codes. Two annotators assigned MSGs to each field of the medical part of death certificates in all subsets. The two annotated subsets were compared to each other. Errors were discussed between the two annotators and corrected if necessary. The agreement rate was 0.90 on the test set.

Final annotated subsets represent the ground truth against which the methods tested were evaluated.

3 Data preprocessing

Data and dictionary were preprocessed by applying 6 basic steps: conversion of all characters to lowercase, removal of diacritics, standardization of compound words (replacement of hyphen by a space to obtain two distinct words), replacement of punctuation (except comma, simple quote, semicolon) by a space, stop words removal, and removal of multiple spaces in the text and at its beginning and its end. The preprocessing steps aim to obtain a narrower set of more relevant terms.

4 Rule-based method

The rule-based method consists in assigning MSGs to all entities from death certificates found in an ad-hoc dictionary, provided by the Inserm-CépiDc. This dictionary is composed of entities found in previous death certificates with their corresponding ICD-10 codes. The principle of the rule-based method is to assign a MSG based on ICD-10 codes to an entity found in a death certificate when it matches an entity found in the dictionary. Our method relies on four steps : i) entity normalization (removal of unnecessary words, replacement of synonyms, acronym expansion, etc.) based on more than 900 rules; (ii) text segmentation (based on coma and semi-colon punctuation marks, prepositions, and a few causal expressions) and factorization of coordinated elements (e.g., "hepatitis A and B" is replaced by entities "hepatitis A" and "hepatitis B"); (iii) spelling correction using a Levenshtein distance computed with words from the dictionary; and (iv) a dictionary matching. This method aims to find entities from the dictionary on death certificates, from the longest span to the shortest one, without overlap. Assignment of an MSG is attempted after each step.

5 Machine learning method

Among the supervised machine learning methods, we trained a linear Support Vector Machine (SVM) classifier [8, 9] using the default hyperparameter (C parameter equal to 1). Since it may be possible to associate 0, 1 or more MSGs to each field of a death certificate, we performed a multi-label classification using the one-versus-all strategy [8]. The SVM classifier was trained for all MSGs.

We implemented a hybrid model using a combination of surface features and the MSGs previously predicted using the rule-based method as its features. We selected, bags-of-word-unigrams (set of unique words for each example in the training set), bags-of-word-bigrams (set of unique sequences of two consecutive words for each example in the training set) and character trigrams (providing a degree of robustness to spelling errors) as surface features. Previous studies showed that

those surface features provided the best classification performance on French mortality data. [9]. This classifier was implemented using the scikit-learn library [10].

6 Evaluation metrics on the test set (1,000 certificates)

Both methods were previously trained and developed on the training set.

The study focused on the analysis of MSGs that were found in similar proportion and with at least three mentions in the training set, test set and data from 2012 to 2016. We analyzed the performance of the two methods on sixty MSGs which belonged to eighteen topics.

Three evaluation metrics were considered: precision, recall, and F-measure. In order to simplify the reporting of the results we defined three groups of MSGs depending on the performance of the methods on the test set:

- Group 1: MSGs with similar performance for both methods. We defined similar performance as the three evaluation metrics belonging to the same performance level. The performance levels of each evaluation metrics were defined using four categories: ≥0.95, [0.90-0.95[, [0.85-0.90[, and<0.85.
- Group 2: remaining MSGs with similar performance for either the rule-based method or the hybrid model.
- Group 3: remaining MSGs with heterogeneous performance for each method. We defined heterogeneous performance as the precision and recall belonging to different performance levels.

The cut off of the performance levels were determined depending on the objectives of the mortality surveillance system. A surveillance system implemented for timely detection of outbreaks must be based on methods with both high recall and high precision [11].

7 Comparison of the evolution of the monthly numbers of death certificates from 2012 to 2016 using the two methods

We applied the rule-based method and the hybrid model on the whole 2012-2016 dataset in order to verify how the performance measured on the test set would result on the whole set. All the causes of death contained in the certificates were classified using both methods. We then counted the number of death certificates by MSGs and by month for both methods ($\text{nb\_E-death}_{\text{RBM,month}}$ and $\text{nb\_E-death}_{\text{Hybrid,month}}$ for each MSG).

The comparison of the monthly numbers of MSGs was based on:

1/ a visual analysis of the monthly evolution of each MSG. We defined 3 patterns: a/ the monthly evolutions of a MSG using the rule-based method and the hybrid model overlap; b/ there is a small difference between the two curves; c/ there is a large difference between the 2 curves.

2/ the calculated difference (Δ_i) of the monthly numbers of death certificates between the two methods expressed as a proportion for the MSG *i* (formula below):

$$\Delta_i = \frac{\frac{\sum_{m1}^{m60}(\text{nb E-death}_{RBM,i,m}\text{-nb E-death}_{Hybrid,i,m})}{60}}{\frac{nbEdeath_{RBM,i} + nbEdeath_{Hybrid,i}}{2}}$$

where *m* corresponds to each month from 2012 to 2016 (*m*=60 months in total) and *i* refers to each analyzed MSG.

3/ the calculated correlation coefficient of $\text{nb\_E-death}_{\text{RBM,i}}$ and $\text{nb\_E-death}_{\text{Hybrid,i}}$. This comparison was performed first for the seven MSGs of the previous study and then for only MSGs with performance (3 metrics) of one or the other method over 0.95.

8. Comparison of the monthly evolution of MSGs to an external data source

For MSGs (such as "Influenza" or "Low acute respiratory infections" [12]) which are expected to have a similar temporal pattern than those of winter infectious epidemics, we visually compared the patterns of the monthly numbers of death certificates and the monthly numbers of attendances of the pathology. For this, we collected from the SurSaUD system the monthly numbers of attendances in emergency departments (EDs) with clinical diagnostic of "influenza" and "low acute respiratory infections" from 2012 to 2016. This analysis focused on two of the seven MSGs of the previous study for which an external data source was available.

Results

1 Analysis of performance of the rule-based method and hybrid model on the test set (1,000 death certificates)

1.1 MSGs with similar performance for both methods (Group 1)

Among the twenty-two MSGs of Group 1, performance (3 metrics) of the methods was over 0.95 for 19 MSGs (Table 1). Those excellent performances are due to distinct features (consistency of entities, useful cues and context for entity identification) depending on the MSG.

Table 1: Number of MSGs for which performance metrics belong to a performance level. France-2016

	Group 1	Group 2	
Performance	Both	Rule-	Hybrid

level	methods	based method	model
≥0.95	19	6	6
[0.90-0.95[	0	1	2
[0.85-0.90[	1	0	1
<0.85	2	3	1

They belonged to nine different topics (Figure 1): 1-Cardiac and circulatory conditions (4 MSGs/11 MSGs analyzed in this topic), 2-Cancer (1/1), 3-Respiratory conditions (3/4), 4-General symptoms (3/9), 5-Infectious diseases (1/6), 6-Nutritional and Endocrine conditions (1/2), 7-Nervous system conditions (4/4), 10-Mental and behavioral disorders (1/3), and 15-Digestive conditions (1/1). Two MSGs had performance below 0.85 ("5-Bacterial infections", "16-Other undefined and unspecified causes of death

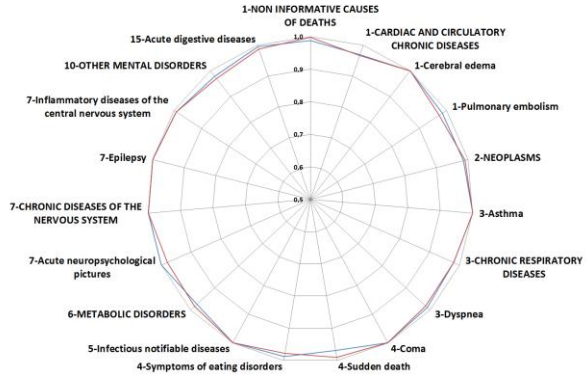


Figure 1: F-measure of the rule-based method (blue) and the hybrid model (red) for nineteen MSGs of Group 1 with performance over 0.95.France-2016

1.2 MSGs with similar performance for either one or the other method (Group 2)

Eighteen MSGs belonged to Group 2. Similar performance (3 metrics were observed solely with the rule-based method for 8 MSGs. Similar performance were also observed solely with the hybrid model for eight other MSGs. Similar performance belonging to different levels were provided by the two methods for 2 MSGs.

Performance of either the rule-based method or the hybrid model was over 0.95 for a majority of MSGs of this group (12/18) (Figure 2). Performance of the rule-based method or the hybrid model was also comprised between [0.90-0.95[for three MSGs, while lower performance (below 0.85) was measured for four MSGs (Table 1).

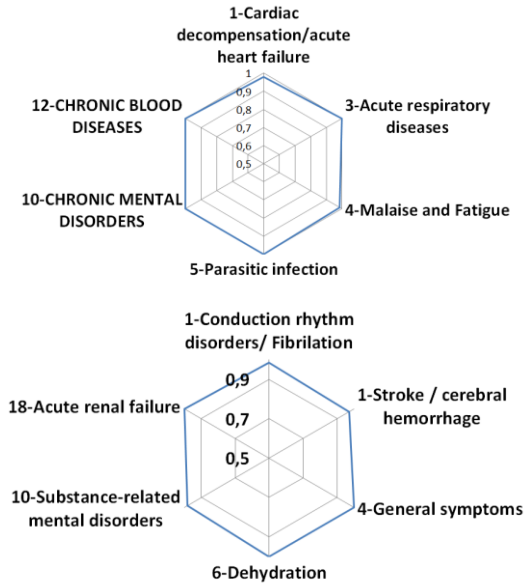


Figure 2: F-measure of the rule-based method (upper part) or hybrid model (lower part) for MSGs with performance (3 metrics) over 0.95 for one of both methods. France-2016

1.3.MSGs with heterogeneous performance (Group 3)

Group 3 contained twenty MSGs with heterogeneous levels of precision and recall. While the rule-based method achieved a higher recall than the hybrid model for more than half of the MSGs (11/20), the rule-based method achieved either higher or lower precision than the hybrid model for respectively six MSGs and seven MSGs (Table 2). Precision of both methods was similar for seven MSGs and recall was similar for three MSGs.

Table 2: Number of MSGs of Group 3 with performance level (PL) obtained using the rule-based method over, below or equal to the PL obtained using the hybrid model. France-2016

	Precision	Recall
$PL^{a}_{RBM} > PL_{Hybrid}$	6	11
$PL_{Hybrid} > PL_{RBM}$	7	6
$PL_{RBM} = PL_{Hybrid}$	7	3

[a] PL: Performance level

2 Comparison of the monthly numbers of death certificates classified into MSGs by the two methods

2.1 Seven MSGs of the previous study

For five out of seven MSGs evaluated in the previous study, the evolution of the monthly numbers of death certificates using the rule-based method overlapped the evolution of the monthly numbers of death certificates using the hybrid model. The monthly evolutions were closely correlated for the two other MSGs. They all had smallΔ_i. (Table 3).

Table 3: Number of MSGs according to the levels of Δ_i and the visual pattern of the monthly numbers of death certificates. France-2012-2016

		Visual patterns		
	Δ_i	Overlap	Good correlation	Low correlation
7 MSGs	Small	5	2	0
19 MSGs* Group 1	Small	7	6	1
	Intermediate	0	1	1
	High	0	0	3
12 MSGs* Group 2	Small	5	3	0
	Intermediate	0	2	0
	High	0	0	2

*MSGs with performance over 0.95

Small: Δ_i varying from -5% to 5%; Intermediate: Δ_i varying from [-10% to -5%] or to [5% to 10%]; High: Δ lower than -10% or over +10%

2.2 MSGs of Group 1 with performance over 0.95

As for the MSGs of the previous study, correlated evolutions were also observed for 7/19 MSGs of the Group 1 with performance over 0.95 in the test set. The Δ_i of those MSGs ranged from -5.0% to 5.0% (Table 3) and their correlation coefficient was very close to 1.

A good visual correlation was also observed between the monthly variations of deaths certificates for 7/19 other MSGs with performance over 0.95, even if a gap was noticed between the monthly numbers of deaths certificates provided by the two methods (Table 3). These seven MSGs had also Δ_i varying from -5.0% to 5.0%, except for "7-Acute neuropsychological pictures" for which Δ_i was 5.9% (Table 3). Correlation coefficients of these MSGs were over 0.99.

A poor visual correlation between the evolutions of the monthly numbers of deaths using the two methods was observed for five other MSGs which achieved performance over 0.95 for both methods in the test set. Δ_i of these MSGs were mostly lower than -10.0% or over 10.0% (Table 3).

2.3 MSGs of Group 2 with performance over 0.95

Overlapped evolution of the monthly numbers of deaths provided by the two classification methods was observed for 5/12 MSGs of Group 2 with performance of one or the other method over 0.95 (Table 3). Δ_i for these MSGs ranged from -3.9% to 1.7% and correlation coefficient was 1.

Five MSGs had also a good visual correlation between the evolution of the monthly numbers of deaths of the two methods (Table 3). Δ_i was small for 3 MSGs with a correlation coefficient over 0.99.

Poor visual correlation of the evolutions was observed for two MSGs ("10-Chronic mental disorders" and "5-Parasitic infections"). Δ were 12.5% and 24.3% and coefficient correlation were 0.98 and 0.96 respectively.

3. Comparison of the monthly evolution of MSGs to an external data source

The monthly evolution of the number of death certificates including a mention of "Influenza" was very close for the two methods. Compared to the evolution of the monthly numbers of EDs attendances for influenza, seasonality and peaks were concomitant even if magnitudes were different. We noted that the number of deaths increased only during winters, the epidemic period of influenza in France (Figure 3).

Likewise, the same observations were valid for the MSG "Low acute respiratory infections"(Figure 3).

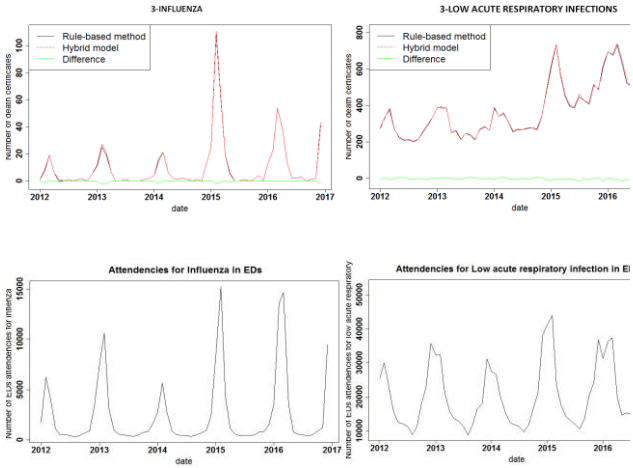


Figure 3: Evolution of the monthly numbers of death certificates classified by the rule-based methods and the hybrid model for "Influenza" and "Low acute respiratory infections"(upper part) and evolution of the monthly numbers of EDs attendances for the same pathologies (lower part). France-2012-2016

Discussion

The current study has shown that more than half of the sixty MSGs achieved both precision and recall over 0.95 at least for one of the two classification methods on the test set (nineteen MSGs from Group 1, six and six MSGs from Group 2 according to the methods).

When the classification methods were applied on the whole dataset we observed that 13/19 MSGs of Group 1 and 8/12 MSGs of Group 2 as well as the seven MSGs previously evaluated in the prior study had small differences between the monthly numbers of deaths certificates obtained using the two methods, with a correlation coefficient close to 1.

Those results suggest that both rule-based method and hybrid model initially implemented, tuned and evaluated to classify medical entities into seven MSGs, are suitable for the classification into a larger number of MSGs. That is of particular interest since 10/13 MSGs of Group 1 and 7/8 MSGs of Group 2 have been defined to be routinely monitored for timely detection of outbreaks.

Among the other 6/19 MSGs of Group 1 and 4/12 MSGs of Group 2 with performance over 0.95 measured on the test set, six had a low visual correlation associated with intermediate or high differences between the monthly numbers of deaths obtained using the two methods. The definition of those MSGs contained rare and/or a large variety of diseases, like "5-Infectious notifiable diseases", "5-Parasitic infections" or "10-Chronic mental disorders". The training and test sets included an insufficient amount of death certificates containing these medical entities to capture the variety of those rare pathologies.

For the two MSGs of the previous study, "Influenza" and "Low acute respiratory infections" with high performance and overlapped evolutions using both methods, the evolutions of MSGs were concomitant with those provided by the external data source. These MSGs built using NLP methods correctly reflect the pathologies we want to monitor in routine. This last step should be applied on the remaining MSGs in order to complete this study. The challenge is to find the most appropriate external data sources among specific surveillance system, notifiable diseases system or the Health National Data system .

Group 3 which contained MSGs with heterogeneous precision and recall for the same method, is more complex to analyze. To meet the objective of a reactive detection of events, a high precision is needed to limit false positive cases. To measure the impact of an event, the surveillance system should need a high recall, to avoid an underestimation of the impact. Results on the test set showed that recall was mostly higher with the rule-based method than with the hybrid model. However none of the two methods provided a mostly higher precision. As both high recall and precision are required to set up a performant surveillance system, we cannot recommend to routinely analyze MSGs of Group 3 whatever the classification method. A complementary study including an error analysis and a larger training set is required to improve the two methods to classify entities into MSGs from Group 3 and also into MSGs from Groups 1 and 2 that have lower performance.

The current study proposes an approach that goes beyond the traditional evaluation based on test set. This is interesting, especially with a large number of categories to evaluate or when the annotated resources are quite limited and does not capture the large variety of entities to classify. The analysis of the temporal pattern of the number of death certificates for each MSG (including the comparison with an external data source) enable to confirm that these indicators (MSGs) accurately capture the targeted health situation we would like to monitor. This approach is also a way to verify whether the misclassifications of one of the two methods are homogeneously distributed over time or are observed in specific periods. Indeed, such a timely mortality surveillance system based on MSGs will be used to provide specific information to health authorities during emergent public health events or during an excess period of deaths (e.g. winter epidemics) to help decision makers adapt counter measures and prevention messages.

Acknowledgements

The authors thank the IT department of LIMSI for computer support and help in computer setup.

References

[1] Triple S Project, Assessment of syndromic surveillance in Europe, *The Lancet* **378** (2011), 1833-1834

[2] Baghdadi Y, Gallay A, Caserio-Schönemann C, Fouillet A: Evaluation of the French reactive mortality surveillance system supporting decision making. *Eur J Public Health* 2018:cky251-cky251

[3] World Health Organization, *International statistical classification of diseases and related health problems 10th revision*, 2016.

[4] A. Névéol, R.N. Anderson, K.B. Cohen, C. Grouin, T. Lavergne, G. Rey, A. Robert, C. Rondet, and P. Zweigenbaum, CLEF eHealth 2017 Multilingual Information Extraction task overview: ICD10 coding of death certificates in English and French, in: *CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*, 2017, p. 17.

[5] P. Zweigenbaum and T. Lavergne, Multiple methods for multi-class, multi-label ICD-

10 coding of multi-granularity, multilingual death certificates, in, CLEF, 2017.
[6] M. Lassalle, C. Caserio-Schönemann, A. Gallay, G. Rey, and A. Fouillet, Pertinence of electronic death certificates for real-time surveillance and alert, France, 2012–2014, *Public Health* **143** (2017), 85-93.
[7] Y. Baghdadi, A. Gallay, C. Caserio-Schönemann, M.-M. Thiam, and A. Fouillet, Towards real-time mortality surveillance by medical causes of death: A strategy of analysis for alert, *Revue d'Épidémiologie et de Santé Publique* **66** (2018), S402.
[8] B. Schölkopf and A.J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press, 2002.
[9] P. Zweigenbaum and T. Lavergne, Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates, in, CLEF, 2017.
[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, Scikit-learn: Machine learning in Python, *Journal of machine learning research* 12 (2011), 2825-2830.
[11] J.W. Buehler, R.L. Berkelman, D.M. Hartley, and C.J. Peters, Syndromic surveillance and bioterrorism-related epidemics, *Emerging Infectious Diseases* **9** (2003), 1197-1204
[12] W.W. Thompson, D.K. Shay, E. Weintraub, L. Brammer, N. Cox, L.J. Anderson, and K. Fukuda, Mortality associated with influenza and respiratory syncytial virus in the United States, *JAMA* **289** (2003), 179-186

Address for correspondence

Yasmine Baghdadi:
yasmine.baghdadi@santepubliquefrance.fr

Résumé :

A partir des décès certifiés électroniquement de 2012 à 2016 en France, la thèse vise à mettre en œuvre et évaluer les performances de méthodes de traitement automatique des langues, pour classer les causes médicales de décès disponibles en texte libre dans des regroupements syndromiques (RS) pertinents pour la surveillance réactive de la mortalité à visée d'alerte et d'évaluation d'impact sanitaire.

Près de 100 RS répondant aux objectifs ont été définis. Deux méthodes de classification ont été développées : une méthode à base de règles linguistiques et une méthode par apprentissage supervisé (SVM). Deux modèles SVM ont été développés utilisant différentes combinaisons de caractéristiques. Le développement et l'évaluation des performances des méthodes se sont appuyés sur 4 500 certificats de décès annotés. L'évaluation a porté dans un premier temps sur 7 RS, puis a été étendue à 60 autres RS. Les évolutions mensuelles des RS attribués par les méthodes ont été comparées sur l'ensemble des décès de 2012 à 2016 (204 000 décès).

La méthode par règles et le modèle SVM incluant l'ensemble des caractéristiques ont obtenu des performances élevées (F-mesure≥0,95) pour la classification des causes dans 31 RS. L'évolution temporelle de ces RS obtenus par les deux méthodes était comparable. En moyenne, les causes de décès au sein d'un certificat sont classées dans 3,7 RS. Une méthode de pondération équilibrée des RS a été proposée pour prendre en compte ces causes multiples lors de l'analyse en routine de la mortalité pour la surveillance à visée d'alerte et d'évaluation d'impact.

Ces résultats permettent de compléter et enrichir la surveillance réactive actuellement fondée sur des données administratives.

Abstract:

Based on electronic death certificates from 2012 to 2016 in France, this thesis aims to implement and evaluate the performance of natural language processing methods, to classify medical causes of death in free text format into mortality syndromic groups (MSG) relevant for reactive mortality surveillance and health impact assessment.

Close to 100 MSGs meeting the objectives of the surveillance were defined. Two classification methods were developed: a rule-based method and a supervised machine learning method (SVM). Two SVM models were developed using different combinations of features. The development and evaluation of the performances of the methods were based on 4,500 annotated death certificates. The evaluation was initially based on 7 MSGs and was then extended to 60 other MSGs. The variations of the monthly number of MSGs assigned by the two methods were compared using the whole death certificates from 2012 to 2016 (204,000 deaths).

The rule-based method and the SVM model including all the features obtained high performances (F-mesure≥0.95) for the classification of causes into 31 MSGs. The monthly variations of those MSGs were comparable. In average, the causes of death within a certificate are classified into 3.7 MSG. We proposed a balanced weighting method of the MSG to take these multiple causes into account in the routine analysis of mortality for alert and impact assessment.

These results complete and enrich the reactive surveillance currently based on administrative data.