Infarctus du myocarde : quelles sont les trajectoires de soins pronostiques du décès à l'hôpital?

Jessica Pinaire[1,2,3], Jerôme Azé[1], Sandra Bringay[1,4], Paul Landais[2,3]

[1] LIRMM, UMR 5506, Université Montpellier, France
prenom.nom@lirmm.fr

[2] BESPIM, CHU de Nîmes, France
Jessica.Pinaire@chu-nimes.fr

[3] ÉQUIPE D'ACCUEIL 2415, Institut Universitaire de Recherche Clinique, Université Montpellier, Montpellier, France
Paul.Landais@umontpellier.fr

[4] AMIS, Université Paul Valéry, Montpellier, France
Sandra.Bringay@univ-montp3.fr

Résumé : Les maladies cardiovasculaires représentent la première cause de mortalité dans le monde. En France, environ 120 000 personnes sont atteintes d'infarctus du myocarde par an ; 12 000 en décèdent lors de la crise, et 18 000 personnes en seront décédées un an après. Prévenir le risque de décès lié à l'infarctus du myocarde est un des objectifs que nous nous sommes fixés. Nous proposons une méthode pour identifier les parcours de soins les plus pronostiques du décès hospitalier. À partir des données médico-administratives issues du PMSI (Programme Médicalisé des Systèmes d'Information), nous extrayons des motifs séquentiels fréquents et nous les intégrons dans un processus de prédiction du décès par un score mesurant la similarité entre la trajectoire du patient et chacun des motifs extraits. Les résultats obtenus nous ont permis de mettre en évidence l'importance de la surveillance et du suivi de ces patients longtemps après leur infarctus.
Mots-clés : Infarctus du myocarde, PMSI, Trajectoires de patients, Fouille de données, Prédiction, Décès.

1 Introduction

Avec 17,5 millions de morts par an, les maladies cardiovasculaires représentent la première cause de mortalité dans le monde [1]. L'Organisation Mondiale de la Santé (OMS) estime que d'ici 2030 près de 24 millions de personnes mourront de maladies cardiovasculaires et ces affections demeureront la première cause de mortalité. Le risque majeur associé à l'infarctus du myocarde (IM) est le décès. En France, environ 120 000 personnes sont atteintes d'infarctus du myocarde par an ; 12 000 en décèdent lors de la crise, et 18 000 personnes en seront décédées un an après. Par ailleurs, les maladies cardiovasculaires constituent une part importante de la consommation des soins ; elles représentent le poste de dépenses le plus important de la consommation de soins et de biens médicaux. En France, les dépenses pour l'année 2002 concernant les maladies cardiovasculaires ont représenté 13,6% des dépenses publiques de santé soit 15,3 milliards d'euros (Heijink *et al.*, 2008). À mesure que la population vieillit, ces dépenses devraient augmenter considérablement (Heidenreich *et al.*, 2011).

Compte tenu de ces enjeux, de nombreux chercheurs universitaires s'intéressent à la consolidation et à l'enrichissement des connaissances médicales, mais aussi à la prévision des risques

1. http://www.who.int/cardiovascular_diseases/global-hearts/Global_hearts_initiative/en/

de mortalité associés aux maladies cardiovasculaires (Freemantle *et al.*, 2013; Fox *et al.*, 2006). Étant donné le nombre de patients impliqués et la quantité de données à exploiter, les chercheurs ont également utilisé des méthodes de fouille de données (Rajalakshmi & Dhenakaran, 2015; Austin *et al.*, 2012), parfois combinées à des méthodes statistiques plus classiques pour étudier des motifs ou évaluer le risque de mortalité. D'autres auteurs se sont intéressés aux données séquentielles pour construire des modèles prédictifs, notamment en se basant sur des règles d'association, pour prédire les divers cheminements du patient entre les différentes unités médicales (Dart *et al.*, 2003), ou la prochaine étape du traitement médicamenteux (Wright *et al.*, 2015). Enfin, des chercheurs ont développé des modèles pronostiques du décès à partir des données médico-administratives (Freemantle *et al.*, 2013; Aylin *et al.*, 2007) et ont obtenus des résultats similaires à ceux des données cliniques voire, dans certains cas, de meilleurs résultats.

Dans cet article, nous proposons d'identifier les parcours de soins les plus pronostiques du décès hospitalier. À partir des données médico-administratives issues du PMSI (Programme Médicalisé des Systèmes d'Information), nous avons extraits des motifs fréquents dans des sous-populations (ou contextes). Ces motifs sont particulièrement intéressants car ils sont facilement interprétables par les experts. Nous démontrons également leur puissance prédictive pour prédire le risque de décès.

Nous avons intégré ces motifs dans les modèles prédictifs à l'aide d'un score. Nous avons comparé entre-elles les méthodes prédictives les plus utilisées dans la littérature et notre approche donne, non seulement des résultats compétitifs avec ceux de la littérature, mais également un modèle interprétable par l'expert.

Dans la section 2, nous introduisons le vocabulaire spécifique au domaine de la recherche de motifs séquentiels et nous l'illustrons avec un exemple d'applications à partir des données du PMSI. Ensuite, nous présentons le processus d'extraction de motifs, puis le processus de prédiction dans la section 3. Dans la section 4, nous identifions les parcours les plus à risque pour deux contextes particuliers. Enfin, nous discutons des résultats obtenus et de notre méthode dans la section 5.

2 Définitions préliminaires

2.1 Motif séquentiel

Définition 1 (Itemset, séquence et sous-séquences d'évènements)
Soit $I = \{i_1, i_2..., i_k\}$ l'ensemble de tous les items. Un sous-ensemble de I est appelé un ***itemset****. Une* ***séquence*** *$s = \langle e_1 e_2 ... e_m \rangle$ est une liste ordonnée d'itemsets, où $e_i \subset I$ pour $1 \leq i \leq m$. Une séquence $s' = \langle r_1 r_2 ... r_p \rangle$ est une* ***sous-séquence*** *de s, s'il existe des entiers $1 \leq n_1 \leq n_2 \leq ... \leq n_p \leq m$, tels que $r_1 \subseteq e_{n_1}, r_2 \subseteq e_{n_2}, ..., r_m \subseteq e_{n_p}$.*

Définition 2 (Support)
Soit une base de séquences $B = \{s_1, ..., s_n\}$, le support de la séquence s, $Freq_B(s)$, est le nombre de séquences dans B ayant s comme sous-séquence.

Définition 3 (Motif séquentiel fréquent)
Une séquence s est fréquente et appelée motif séquentiel si son support est supérieur ou égal à un support minimum $k > 0$ fixé : $Freq_B(s) \geq k$.

2.2 Motif séquentiel contextuel

Définition 4 (Contexte)
*Un **contexte** c est une catégorie ou une modalité d'une variable (e.g. Homme ou Femme pour le sexe). L'ensemble de tous les contextes muni d'une relation d'ordre partiel, $\leq$, constitue la **hiérarchie des contextes** H. Les **contextes feuilles** de H sont appelés les **contextes minimaux**. A contrario, plus on remonte dans l'arborescence de H plus le contexte est dit **général**.*

Définition 5 (Motif séquentiel contextuel fréquent)
Soit c un contexte, H la hiérarchie des contextes et s une séquence. Une séquence s est fréquente dans un contexte c si son support dans c est supérieur ou égal à un support minimum $k > 0$ fixé : $Freq_c(s) \geq k$.

2.3 Exemple

Dans cet exemple, nous allons considérer les évènements (les séjours hospitaliers), représentés par les GHM (Groupes Homogènes de Malades), de 14 patients sur une période de 4 mois. Le temps est divisé en estampilles temporelles représentées par les mois. Supposons qu'à chaque mois, il ne peut se produire qu'un seul évènement (*i.e* un patient a un seul séjour par mois). Ces informations sont contenues dans la base de données présentée dans le tableau 1. Elle décrit différents GHM (05M13 : Douleurs thoraciques ; 05M06 : Angine de poitrine ; 05M16 : Athérosclérose coronarienne ; 05M04 : IM aigu ; 05M09 : Insuffisance cardiaque) associés au cours du temps par des professionnels de santé à des séjours de patients.

TABLE 1 – Mise en valeur du motif $\langle$(05M13)(05M06)$\rangle$ (en gras) soit le GHM 05M13 suivi du GHM 05M06. Ce motif est fréquent dans la base pour un support minimum de 50%.

Patients	Janvier	Février	Mars	Avril
P_1		**05M13**	05M04	**05M06**
P_2	**05M13**	05M09	**05M06**	
P_3	**05M13**	05M13		**05M06**
P_4	05M16	**05M13**	**05M06**	05M16
P_5	05M04	**05M13**	**05M06**	05M04
P_6		05M06		05M13
P_7		**05M13**	**05M06**	05M13
P_8	05M04	**05M13**	05M16	**05M06**
P_9		**05M13**	05M13	**05M06**
P_{10}		05M06	05M16	05M04
P_{11}		05M06	05M04	05M13
P_{12}	05M09	05M06	05M04	05M13
P_{13}	05M06	05M04	05M09	
P_{14}	05M06		05M13	05M09

Ces données sont **séquentielles** car elles présentent des évènements (les GHM) disposés suivant un ordre (le temps). Par exemple, pour le patient P_{12}, le GHM 05M09 associé en janvier, le GHM 05M06 a été associé au séjour de février, puis le GHM 05M04 associé au séjour de mars, enfin le GHM 05M13 associé au séjour d'avril. Une **sous-séquence** de la séquence du patient P_{12} est par exemple, la séquence $\langle$(05M09)(05M06)$\rangle$. Elle est également présente dans

la séquence du patient P_2 : son **support** est donc de 14% (2 sur 14) [2]. En examinant le tableau 1, nous constatons que le motif 05M13 suivi plus tard par 05M06 est vérifié par plus de 50% des patients (8 sur 14). En supposant que le professionnel de santé précise qu'il est intéressé par des GHM qui apparaissent dans au moins 50% des cas (support minimum) présents dans la base alors il s'avère que la sous-séquence $\langle$(05M13)(05M06)$\rangle$ est un **motif séquentiel fréquent**.

Jusqu'à présent, nous avons considéré la base comme un ensemble indivisible pour la recherche des motifs. Maintenant, nous allons prendre en compte les circonstances liées aux données : **les contextes**. Nous intégrons des informations supplémentaires, dans le tableau 2, qui associent à chaque patient son âge (*jeune* ou *âgé*) et son sexe (*homme* ou *femme*).

Ces informations contextuelles peuvent avoir une influence non négligeable sur la séquence d'évènements. Ainsi, l'extraction de motifs doit rendre cette influence perceptible pour l'utilisateur afin de lui offrir une vue contextualisée des données. Considérons maintenant la séquence $\langle$(05M13,05M06)$\rangle$ dans le tableau 2, nous constatons que :

- cette séquence de GHM est fréquente dans la population âgée (7 personnes âgées sur 8) mais pas dans la population jeune (seulement 1 personne sur 6) ;
- cette séquence de GHM demeure fréquente chez les âgés quel que soit leur sexe (5 hommes âgés sur 5 et 2 femmes âgées sur 3).

TABLE 2 – Mise en valeur du motif $\langle$(05M13)(05M06)$\rangle$ (en gras) avec les informations contextuelles sur l'âge et le sexe. Ce motif est spécifique aux personnes âgées. Une seule personne jeune est concernée.

Patients	**Age**	**Sexe**	**Janvier**	**Février**	**Mars**	**Avril**
P_1	âgé	homme		**05M13**	05M04	**05M06**
P_2	âgé	homme	**05M13**	05M09	**05M06**	
P_3	âgé	homme	**05M13**	05M13		**05M06**
P_4	âgé	homme	05M16	**05M13**	**05M06**	05M16
P_5	âgé	homme	05M04	**05M13**	**05M06**	05M04
P_6	âgé	femme		05M06		05M13
P_7	âgé	femme		**05M13**	**05M06**	05M13
P_8	âgé	femme	**05M13**	05M16	**05M06**	
P_9	jeune	homme		**05M13**	05M13	**05M06**
P_{10}	jeune	homme		05M06	05M16	05M04
P_{11}	jeune	homme		05M06	05M04	05M13
P_{12}	jeune	femme	05M09	05M06	05M04	05M13
P_{13}	jeune	femme	05M06	05M04	05M09	
P_{14}	jeune	femme	05M06		05M13	05M09

3 Protocole de prédiction

La première étape de notre protocole consiste à extraire puis à trier des motifs contextuels fréquents à partir des données issues du PMSI. L'étape suivante de notre protocole de prédiction a été construite à l'aide des recommandations établies pour élaborer des modèles prédictifs

2. Notons que dans notre cas les itemsets sont réduits à un item.

à des fins de pronostic ou de diagnostic : la méthode TRIPOD (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) (Collins *et al.*, 2015).

Ces étapes sont plus précisément expliquées dans la suite de cette section et sont schématisées dans la figure 2. Les expérimentations ont été réalisées à l'aide du logiciel R version 3.3.1.

3.1 Étape 1. Extraction de motifs séquentiels contextuels

L'objectif de cette première étape est d'extraire des profils de parcours de soins fréquents pour l'IM qui prennent en compte des informations contextuelles fréquemment associées aux données séquentielles. Ainsi, nous identifions des profils de parcours de soins spécifiques d'une sous-population donnée.

Le processus de fouille, s'effectue de la façon suivante :

1. Nous sélectionnons les patients ayant eu un IM à l'aide d'une requête SQL sur la base PMSI. Pour chaque patient nous récupérons l'ensemble de ses séjours sur la période 2009 à 2014, excepté les séjours pour séances (*e.g.* radiothérapie, dialyses, chimiothérapie, *etc.*) et les prestations inter-établissements [3]. Selon les règles de codage du PMSI, ces séjours ont le même motif d'admission ;
2. Nous créons des sous-populations appelées contextes, à l'aide de covariables associées aux patients. Pour ce faire, nous avons pris en compte le genre (Homme/Femme), la classe d'âge du patient au moment de sa première apparition dans la période d'observation : -45 ans, 45-65 ans et +65 ans [4]. Enfin, nous avons retenu le nombre de séjours selon trois catégories : les 3-5 séjours, les 5-60 séjours et enfin les plus de 60 séjours [5]. Nous obtenons 18 contextes minimaux. La hiérarchie de nos contextes est représentée dans la figure 1 ;

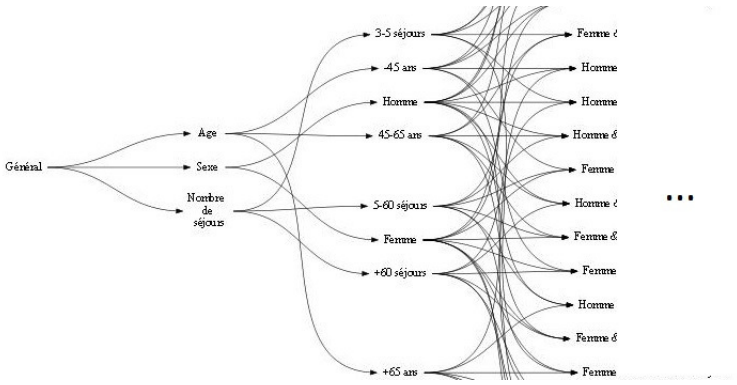

FIGURE 1 – Hiérarchie des contextes.

3. Nous construisons la base séquentielle en nous intéressant aux GHM, c'est-à-dire aux codes utilisés par les professionnels pour catégoriser un séjour et donnant lieu à tarifica-

3. Transfert d'un patient dans un autre établissement pour lui faire faire un acte (*e.g.* une coronarographie) car le premier n'a pas l'équipement pour le réaliser.

4. Ces classes d'âge ont été construites après concertation de l'expert médical.

5. Classes déterminées par l'expert médical : plus un patient est hospitalisé, plus il a des complications médicales associées à sa maladie.

tion. À chaque patient est associé une séquence de GHM, de longueur égale au nombre de séjours effectués pendant ces six années. Pour chaque patient, nous avons nettoyé automatiquement les données, en conservant la première hospitalisation du patient puis tous ses séjours d'hospitalisation liés à la cardiopathologie. Par exemple, pour un patient dont la séquence d'hospitalisation est : *IM, diabète, obésité, fracture du poignet et angine de poitrine*. Dans cette séquence d'évènements, nous identifions un séjour non lié à la cardiologie : *fracture du poignet*. Cet évènement est retiré de la séquence du patient. Les séquences de GHM ainsi triées constituent **la trajectoire du patient** ;

4. Nous effectuons la recherche de motifs séquentiels contextuels à l'aide de l'algorithme CFPM (Contextual Frequent Pattern Mining) (Rabatel, 2011) avec un support de 1%. Cela signifie que les motifs sont extraits s'ils concernent au moins 1% des patients d'un contexte ;
5. Nous procédons à la suppression de l'information que nous souhaitons prédire : le décès. Il s'agit des codes GHM contenant cette information de façon intrinsèque [6]. D'autre part, nous ne conservons que les motifs maximaux (Gouda & Zaki, 2005) non inclus dans un autre motif.

Nous passons ensuite à l'étape de prédiction.

3.2 Étape 2. Prédiction

Modéliser consiste à représenter un phénomène ou une situation observée afin de mieux l'étudier. Il s'agit généralement de trouver une représentation qui, à partir de paramètres, permet d'obtenir une retranscription qui soit la plus en adéquation possible avec les observations.

Ici, nous souhaitons prédire la mortalité hospitalière suivant le parcours du patient, ainsi la variable binaire à expliquer est l'état final du patient : présumé vivant ou décédé dans un établissement de soins.

Le processus de prédiction, s'effectue en 4 étapes :

1. Nous constituons des échantillons équilibrés [7] par contexte à partir des données issues du point 3 de l'étape 1, en ne conservant que les patients ayant eu au moins 4 évènements durant la période d'observation. Ceci nous permet de constituer une base de données avec des patients ayant un historique suffisant pour améliorer la capacité prédictive d'un modèle. De plus, nous écartons les patients originaires des régions Sud Méditerranée et Nord-Ouest [8] que nous conservons pour la validation externe [9] ;
2. Nous mesurons un score de similarité, entre les différents motifs et la trajectoire du patient. Nous intégrons la notion de distance dans le choix du modèle. Nous avons retenu les distances suivantes : la plus longue sous-chaîne commune (LCS), la distance de Levenshtein, la distance d'alignement optimal, la distance de Damerau-Levenshtein, les distances q-gramme, Jaccard, cosinus, Jaro et Jaro-Winckler ;

6. *e.g.* 05M21 signifiant Infarctus aigu du myocarde avec décès : séjours de moins de 2 jours.

7. L'équilibrage se fait sur la variable à prédire : autant de patients décédés que de patients vivants.

8. La région d'origine du patient est déterminée selon les territoires médicaux de la carte des inter-régions du CeNGEPS (`http://www.cengeps.fr/fr`).

9. Selon le principe d'une validation dite "géographique".

3. Nous prédisons avec plusieurs modèles : régression logistique (RL), Naïf Bayes (NB), k plus proches voisins (KNN), arbre de régression et Séparateur à Vastes Marges (SVM). Nous comparons les modèles et sélectionnons le meilleur selon les critères : accuracy, taux d'erreur, précision, F-mesure et aire sous la courbe ROC (Receiver operating characteristic (AUC)) ;
4. Nous validons le modèle sélectionné à partir de l'échantillon des patients ayant des trajectoires de longueur 4 et originaires des régions Sud Méditerranée et Nord-Ouest. Les critères d'évaluation sont : l'AUC et le score de Brier (Brier, 1950).

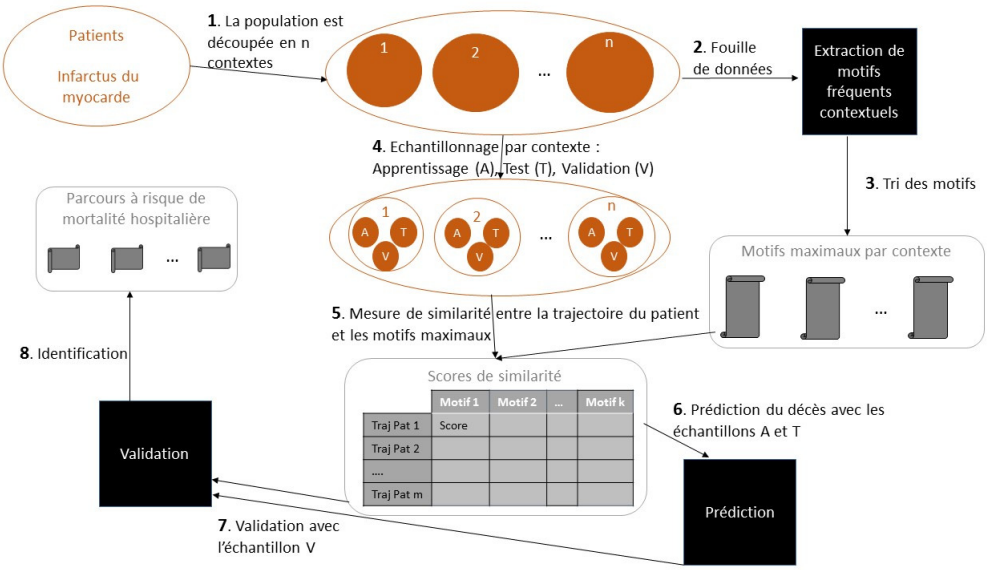

FIGURE 2 – Schéma du protocole de prédiction.

4 Expérimentations

Suite à l'application du protocole décrit dans la section 3, les modèles retenus sont ceux de la RL couplés à une distance d'édition. Dans cette section, nous identifions les parcours les plus à risque de la mortalité hospitalière pour les contextes suivants : les patients ayant 3-5 séjours et les patients ayant 5-60 séjours.

Pour identifier les parcours les plus à risque de la mortalité hospitalière, nous avons analysé l'influence des variables impliquées dans les modèles. Les résultats sont présentés dans le tableau 3.

Les résultats d'une RL sont interprétés en termes de facteurs de risque si l'OR est supérieur à 1, de facteurs protecteurs si l'OR est inférieur à 1 ou encore absence d'association entre l'évènement d'intérêt et la variable si l'OR est égal à 1.

Après examen de la première partie du tableau 3 nous identifions les motifs 05M04-05K06 (IM aigu suivi de pose de stent) et 05K10 (Actes diagnostiques par voie vasculaire) comme étant des parcours augmentant le risque du décès hospitalier pour le contexte 5-60 séjours. Tandis que les motifs 05K06-05K06-05K06 (3 séjours pour pose de stent) et 05K13 (Actes

TABLE 3 – Modèles logistiques pour les trajectoires de GHM.

		Modèle 5-60 séjours	
Variables	**Coefficient**	**OR**	**IC 95%**
Constante	-3,13***		
Scores			
05M04-05K06	9,63 **	1,52e+04	38,83 à 9,37e+06
05K10	43,08***	5,12e+18	5,13e+09 à 2,56e+28
05K06-05K06-05K06	-8,34 ***	2,39e-04	5,33e-06 à 0,009
05K13	-20,87*	8,62e-10	4,82e-17 à 0,007
		Modèle 3-5 séjours	
Variables	**Coefficient**	**OR**	**IC 95%**
Constante	-18,26***		
Classe d'âge			
45-65 ans	-2,26**	0,1	0,009 à 0,57
-45 ans	-2,96	0,05	3,89e-04 à 1,54
Scores			
05K10	137,38***	4,62e+59	7,37e+32 à 2,33e+101
05K13	-66,74**	1,03e-29	8,37e-56 à 6,41e-10
05M04	15,23	4,13e+06	4,24e-01 à 7,31e+14

Variables : variables retenues dans le modèle ;
Coefficient : valeur des β_i du modèle
et test de nullité des β_i avec *p<0,05 ; **p<0,01 ; ***p<0,001 ;
OR : Odds-ratio ; **IC 95%** : intervalle de confiance à 95% des OR.

thérapeutiques par voie vasculaire) sont des facteurs protecteurs. Pour le contexte 3-5 séjours, nous retrouvons des résultats assez similaires. Dans la deuxième partie du tableau 3, les motifs 05M04 et 05K10 sont des facteurs de risque du décès alors que le motif 05K13 est identifié comme un facteur protecteur. Par ailleurs, l'examen des résultats concernant les classes d'âge indique une augmentation du risque pour la classe des +65 ans.

5 Discussion

Dans cette section, nous discutons de la méthode et de ses performances, dans la partie 5.1, nous nous comparons à d'autres et nous envisageons d'autres outils. Ensuite, dans la partie 5.2, nous faisons une synthèse des résultats obtenus et nous argumentons sur les limites de l'interprétation de ces résultats.

5.1 Choix du modèle : performances et extensions possibles

La comparaison des différents modèles montre que la RL couplée à une distance d'édition obtient les meilleures performances dans la plupart des contextes. D'autres auteurs (Austin, 2007; Steyerberg *et al.*, 2001) obtiennent des résultats similaires concernant la compétitivité de la RL comparée à d'autres modèles dans le cas de la prédiction de la mortalité à 30 jours après un IM aigu. En outre, le seul cas où nous obtenons un modèle avec une distance q-gramme est le contexte des femmes avec 5-60 séjours. Or, de manière générale, les femmes ont des trajectoires plus courtes (dans notre sélection des données). Le choix de la distance est donc lié à la longueur des séquences.

Par ailleurs, une étude comparative (Siontis *et al.*, 2012) des travaux de modélisation du risque de la mortalité, dans le cas de maladies cardiovasculaires, réalisés à partir de données cliniques, montre que les performances selon l'AUC varient de 0,71 à 0,88. Or nous obtenons dans le cas de la sélection des meilleurs modèles des performances variant de 0,6 à 0,98. Nos modèles ont donc des performances comparables à ceux évoqués plus haut. En outre, nous constatons que les résultats sont meilleurs pour des contextes à faibles effectifs. En effet, dans ces contextes, l'échantillonnage arrive à recouvrir plus de situations diverses. Ainsi, les échantillons sont plus représentatifs de la population et de fait les modèles n'en sont que meilleurs.
Toutefois, nous pourrions affiner nos résultats en intervenant sur différentes étapes de notre protocole, notamment à l'aide de techniques pour la sélection de prédicteurs (Guyon & Elisseeff, 2003; Claeskens *et al.*, 2008). En outre, d'autres approches pourraient être envisagées pour étudier les trajectoires de patients, comme l'analyse formelle de concept utilisée dans (Jay *et al.*, 2013) pour obtenir une représentation hiérarchique de l'information. Cette dernière pourrait être combinée à une méthode de calcul d'évènements (Event Calculus) (Mueller, 2008) afin de déterminer les évènements liés au décès. Une autre approche possible est celle employée dans (Fabregue *et al.*, 2011). Les motifs sont classés selon l'état final du patient et utilisés comme descripteurs d'un classifieur classique pour déterminer de quel groupe (Vivant/Décédé) un patient est le plus proche en fonction de sa trajectoire.

5.2 Les parcours à risque : résultats et limites de l'interprétation

La modélisation du décès à l'aide des motifs fréquents permet de distinguer les évènements hospitaliers favorisant une augmentation du risque de décès de ceux qui, au contraire, ont un effet protecteur. D'après les résultats décrit dans la section 4, les motifs préservant du décès sont les actes thérapeutiques et le parcours IM aigu suivi d'une pose de stent. Ceci atteste de l'efficacité de la prise en charge (Falconnet *et al.*, 2009) avec un suivi des soins de la dilatation artérielle par divers moyens : endoprothèse, angioplastie... associés à un traitement médicamenteux. En revanche, suivant l'état de gravité de la pathologie, un acte exploratoire, comme une artériographie ou une coronarographie, représentant déjà un risque pour le patient (comme tout acte invasif), favorisera d'autant plus une augmentation du risque (Mottier & Baba-Ahmed, 2006). Ceci explique la présence de 05K10 dans les motifs à risque. Les motifs fréquents identifiés et intégrés dans un modèle prédictif du décès hospitalier viennent souligner l'importance du suivi des patients atteints de cette pathologie sur une période d'un an voire au-delà, comme en atteste d'ailleurs la littérature sur ce sujet (Neff, 2004).

D'autres auteurs ont également utilisé les bases médico-administratives pour de la prédiction. Par exemple, (Aylin *et al.*, 2007) ont comparé les modèles pronostiques du décès hospitalier à partir des bases administratives et des bases cliniques (registres nationaux vasculaires et cancers). Ils ont obtenu des résultats similaires avec les deux types de bases et ont ainsi démontré l'intérêt d'utiliser les bases administratives pour construire des modèles pronostiques. Un autre exemple d'étude est celui de (Jensen *et al.*, 2014) dans le cas des maladies chroniques afin de prévoir les flux de patients et d'envisager les infrastructures.

Néanmoins, nous pouvons formuler quelques critiques au regard de cette méthode, notamment dans le choix de la base de données. En effet, le PMSI est un outil d'allocation budgétaire, mais il a des limites dans le domaine épidémiologique car il expose à des imprécisions et à des erreurs tenant, entre autres raisons, à l'insuffisance de l'information, à des difficultés de

codage et à des nomenclatures inappropriées (adéquation de la codification de la maladie avec la réalité). Par voie de conséquence, la fiabilité du codage des séjours via les bases médico-administratives est controversée (Lombrail *et al.*, 1994). Pourtant, elles représentent indéniablement une source importante d'informations. Elles couvrent la majorité des établissements de santé privés et publics sur le plan national et recèlent de données médicales sur tous les séjours hospitaliers. Une alternative pour réduire ce biais intrinsèque au codage pourrait être l'appariement des informations avec les bases de données du Sniiram (Système nationale d'information inter-régimes de l'assurance maladie).

6 Conclusion

En utilisant les motifs séquentiels nous avons élaboré des modèles pour prédire le décès au sein d'un établissement de santé. Ces motifs ont été intégrés par des scores en mesurant la similarité entre la trajectoire du patient et les motifs. Le choix du score n'étant pas évident, nous avons choisi de mettre en concurrence différentes mesures entre chaînes de caractères de la littérature. Nous avons construit un protocole de prédiction qui s'articule en plusieurs étapes en nous appuyant sur la méthode TRIPOD. Nous avons introduit les modèles prédictifs les plus couramment employés afin de les comparer. *In fine*, notre objectif était double : 1) déterminer le couple *(modèle, score)* ayant les meilleures performances pour chacun des contextes ; 2) identifier les motifs favorisant une augmentation du risque de décès.

Notre avons atteint notre premier objectif. Il résulte de la comparaison entre les différents modèles que le modèle logistique couplé à une distance d'édition est le modèle offrant les meilleures performances avec la conservation des scores de similarités en variables continues. D'autres perspectives de comparaisons et de modélisation sont envisageables à l'aide des modèles de survie tels que Cox (Timsit *et al.*, 2005) ou encore des modèles prédictifs basés sur les séquences.

Pour affiner ces investigations, nous prévoyons d'employer d'autres types de motifs comme par exemple, les motifs obtenus avec la r-confiance (Mercadier *et al.*, 2016) qui permettent d'identifier les parcours les plus probables à partir d'un ou plusieurs premiers évènements. Nous souhaitons construire des modèles prédictifs en sélectionnant les parcours les plus représentatifs à la fois en fréquence, en taille et en confiance. Une autre approche pourrait également être envisagée, en tenant compte de l'état final du patient à la fin de la période d'observation dans la répartition des contextes. Cette approche consisterait alors à mettre en évidence des motifs qui soient spécifiques du décès. Ainsi, nous pourrions, soit intégrer ces motifs dans notre protocole de prédiction, soit mettre en œuvre une méthode de classification des patients à partir de leur trajectoire comme dans (Fabregue *et al.*, 2011) afin de prédire le décès.

Nous avons également atteint notre deuxième objectif en distinguant les motifs présentant un risque accru de décès. Ces motifs difficiles à interpréter par des experts médicaux, s'avèrent, en revanche utiles pour prédire le risque de mortalité hospitalière d'un patient. En effet, à l'issue de la fouille, les résultats obtenus mettent en évidence les motifs fréquents dans des sous-populations mais ne permettent pas à l'expert d'interpréter les résultats car ils n'ont pas de particularité populationnelle. L'implémentation de ces résultats dans un modèle favorise leur

interprétation dans la mesure où ils sont imputables à une cause : le décès. Nous retenons de nos expérimentations, présentées dans la partie 4, que le risque de décès est fortement influencé par le profil d'évolution de la maladie et le suivi du patient après IM. Ceci témoigne de l'importance des recommandations de la société française de cardiologie (Delahaye *et al.*, 2001) sur la surveillance régulière des patients après un IM au moins durant une année. En effet, le risque de rechute et de décès est encore très élevé durant cette période et même encore au-delà.

Références

AUSTIN P. C. (2007). A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statistics in Medicine*, **26**(15), 2937–2957.

AUSTIN P. C., LEE D. S., STEYERBERG E. W. & TU J. V. (2012). Regression trees for predicting mortality in patients with cardiovascular disease : what improvement is achieved by using ensemble-based methods ? *Biometrical Journal*, **54**(5), 657–673.

AYLIN P., BOTTLE A. & MAJEED A. (2007). Use of administrative data or clinical databases as predictors of risk of death in hospital : comparison of models. *British Medical Journal*, **334**(7602), 1044–1052.

BRIER G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**(1), 1–3.

CLAESKENS G., CROUX C. & KERCKHOVEN J. V. (2008). An Information Criterion for Variable Selection in Support Vector Machines. *Journal of Machine Learning Research*, **9**(Mar), 541–558.

COLLINS G. S., REITSMA J. B., ALTMAN D. G. & MOONS K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) : the TRIPOD statement. *BioMed Central Medicine*, **13**(1), g7594.

DART T., CUI Y., CHATELLIER G. & DEGOULET P. (2003). Analysis of hospitalised patient flows using data-mining. *Studies in Health Technology and Informatics*, **95**, 263–268.

DELAHAYE F., BORY M., COHEN A., DANCHIN N., DE GEVIGNEY G., DELLINGER A., FRABOULET J.-Y., GAYET J.-L., GUIZE L., IUNG P., MABO C., MONPÈRE P.-G., STEG D. & THOMAS (2001). Recommandations de la société française de cardiologie concernant la prise en charge de l'infarctus du myocarde après la phase aiguë. *Archives des maladies du cœur et des vaisseaux*, **94**(7), 697–738.

FABREGUE M., BRINGAY S., PONCELET P., TEISSEIRE M. & ORSETTI B. (2011). Mining microarray data to predict the histological grade of a breast cancer. *Journal of Biomedical Informatics*, **44**(Suppl. 1), S12–S16.

FALCONNET C., PERRENOUD J.-J., CARBALLO S., ROFFI M. & KELLER P.-F. (2009). Syndrome coronarien aigu : guidelines et spécificité gériatrique. *Revue médicale suisse*, **5**(204), 1137–1147.

FOX K. A. A., DABBOUS O. H., GOLDBERG R. J., PIEPER K. S., EAGLE K. A., WERF F. V. D., AVEZUM A., GOODMAN S. G., FLATHER M. D., ANDERSON F. A. & GRANGER C. B. (2006). Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome : prospective multinational observational study (GRACE). *British Medical Journal*, **333**(7578), 1091–1094.

FREEMANTLE N., RICHARDSON M., WOOD J., RAY D., KHOSLA S., SUN P. & PAGANO D. (2013). Can we update the Summary Hospital Mortality Index (SHMI) to make a useful measure of the quality of hospital care ? An observational study. *British Medical Journal Open*, **3**(1), e002018.

GOUDA K. & ZAKI M. J. (2005). GenMax : An Efficient Algorithm for Mining Maximal Frequent Itemsets. *Data Mining and Knowledge Discovery*, **11**(3), 223–242.

GUYON I. & ELISSEEFF A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**(Mar), 1157–1182.

HEIDENREICH P. A., TROGDON J. G., KHAVJOU O. A., BUTLER J., DRACUP K., EZEKOWITZ M. D., FINKELSTEIN E. A., HONG Y., JOHNSTON S. C., KHERA A., LLOYD-JONES D. M., NELSON S. A., NICHOL G., ORENSTEIN D., WILSON P. W. F. & WOO Y. J. (2011). Forecasting the Future of Cardiovascular Disease in the United States. *Circulation*, **123**(8), 933–944.

HEIJINK R., NOETHEN M., RENAUD T., KOOPMANSCHAP M. & POLDER J. (2008). Cost of illness : An international comparison. *Health Policy*, **88**(1), 49–61.

JAY N., NUEMI G., GADREAU M. & QUANTIN C. (2013). A data mining approach for grouping and analyzing trajectories of care using claim data : the example of breast cancer. *BioMed Central Medical Informatics and Decision Making*, **13**(130).

JENSEN A. B., MOSELEY P. L., OPREA T. I., ELLESØE S. G., ERIKSSON R., SCHMOCK H., JENSEN P. B., JENSEN L. J. & BRUNAK S. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, **5**, 4022.

LOMBRAIL P., MINVIELLE E., COMAR L. & GOTTOT S. (1994). Programme de médicalisation des systèmes d'information et épidémiologie : une liaison qui ne va pas de soi. *Revue d'épidémiologie et de santé publique*, **42**(4), 334–344.

MERCADIER Y., PINAIRE J., AZÉ J., BRINGAY S. & TEISSEIRE M. (2016). La r-confiance pour l'identification de trajectoires de patients. In *Proceedings des 16ème Journées Francophones Extraction et Gestion des Connaissances*, p. 535–536.

MOTTIER D. & BABA-AHMED M. (2006). Anticoagulants et gestes invasifs. *Médecine thérapeutique*, **12**(1), 48–52.

MUELLER E. T. (2008). Event calculus. *Foundations of Artificial Intelligence*, **3**, 671–708.

NEFF M. J. (2004). Practice Guidelines : ACC/AHA Release Guidelines on Management of Patients with STEMI : Hospital and Long-Term Management. *American Family Physician*, **70**(10), 2011–2021.

RABATEL J. (2011). *Extraction de motifs contextuels : Enjeux et applications dans les données séquentielles*. PhD thesis, Université Montpellier II.

RAJALAKSHMI K. & DHENAKARAN S. S. (2015). Analysis of Datamining Prediction Techniques in Healthcare Management System. *International Journal of Advanced Research in Computer Science and Software Engineering*, **5**(4), 1343–1347.

SIONTIS G. C. M., TZOULAKI I., SIONTIS K. C. & IOANNIDIS J. P. A. (2012). Comparisons of established risk prediction models for cardiovascular disease : systematic review. *British Medical Journal (Clinical research edition)*, **344**, e3318.

STEYERBERG E. W., EIJKEMANS M. J., HARRELL F. E. & HABBEMA J. D. (2001). Prognostic modeling with logistic regression analysis : in search of a sensible strategy in small data sets. *Medical Decision Making : An International Journal of the Society for Medical Decision Making*, **21**(1), 45–56.

TIMSIT J.-F., ALBERTI C. & CHEVRET S. (2005). Le modèle de Cox. *Revue des maladies respiratoires*, **22**(6), 1058–1064.

WRIGHT A. P., WRIGHT A. T., MCCOY A. B. & SITTIG D. F. (2015). The use of sequential pattern mining to predict next prescribed medications. *Journal of Biomedical Informatics*, **53**, 73–80.