# PREDICTIVE MODELING OF HOSPITAL READMISSION RATES USING ELECTRONIC MEDICAL RECORD-WIDE MACHINE LEARNING :**A CASE-STUDY USING MOUNT SINAI HEART FAILURE COHORT**

**AUTEURS DE L'ARTICLE :** KHADER SHAMEER, KIPP W JOHNSON, ALEXANDRE YAHI, RICCARDO MIOTTO, LI LI, DORAN RICKS, JEBAKUMAR JEBAKARAN, PATRICIA KOVATCH, PARTHO P. SENGUPTA, ANNETINE GELIJNS, ALAN MOSKOVITZ, BRUCE DARROW, DAVID L REICH, ANDREW KASARSKIS, NICHOLAS P. TATONETTI, SEAN PINNEY AND JOEL T DUDLEY.

**AUTEUR DU RAPPORT :** MEDOU Daniel Magloire, étudiant en Master recherche en informatique. IFI, Décembre 2017

**SUPERVISEUR ACADEMIQUE :** Dr. HO Tuong Vinh

**Institut Francophone International (IFI)**

## I. INTRODUCTION

La réadmission hospitalière est définie comme étant (confère wiipédia). Ce phénomène est un défi que bon nombre d'organisations de règlementation comme les Centers for Medicaid and Medicare Services (CMS) aux Etats-Unis pour ne citer que ceux-là, accordent une grande importance voir capitale car il en est de la survie des êtres humains. De ce faite, l'étude sur la réadmission hospitalière aux Etats-Unis résulte comme le spécifie les auteurs de cet article des affections chroniques ou aigues et ont énuméré les maladies entrant dans ce package à savoir les accidents vasculaires cérébraux, l'insuffisance cardiaque, l'infarctus du myocarde et la pneumonie. En effet, l'observation avait été faite sur un cas de maladie bien précis parmi tous ceux qui cités ci-dessus à savoir les crises cardiaques (insuffisance cardiaque) donc le taux des patients réadmis est relativement élevés. Malgré la mise en œuvre de lignes directrices d'opérations de prestation de soins de santé de haute qualité créées par les autorités de réglementation (CMS), les patients sont réadmis en hospitalisation avant les trente (30) suivant leur hospitalisation initiale. Le problème reste donc à connaitre comment diminuer au maximum le taux de réadmission hospitalière de cette pathologie qu'est l'insuffisance cardiaque ? Cependant, dans l'article soumis à notre étude dans le cadre de l'unité d'enseignement intitulé ≪ Bibliographie et Etude des Cas ≫, KHADER SHAMEER et al [1], par le biais de l'apprentissage automatique prédire les probabilités de réadmission en se basant des variables du dossier médicale électronique des patients atteints d'insuffisance cardiaque. Le présent rapport nous donnera un résumé détaillé de l'article suivant le plan que nous avons arrêté.

***Keywords—Réadmission hospitalière, AUC (Area Under the Curve), Cohorte, Accuracy.***

## II. CONTEXT, OBJECTIFS ET PROBLEMATIQUE DE L'ARTICLE

KHADER SHAMEER et al, dans leur article présentent les points essentiels abordés dans la thématique de l'apprentissage automatique et de la fouille des données donc l'objectif principal est de développer une approche de sélection des caractéristiques basées sur des données électroniques, à l'échelle du dossier médical et de l'apprentissage automatique afin de prédire les probabilités de réadmissions hospitalières des affections chroniques ou aigues. Les deux domaines cités à savoir l'apprentissage automatique et de la fouille des données ont fait l'objet d'une étude sous forme de module respectivement du Master 2 et du Master 1 durant notre parcours académique à L'IFI (Institut Francophone International). Etude faite sur deux niveaux. Le premier étant théorique et l'autre pratique avec la mise en œuvre des travaux pratiques au moins trois par domaine. Il est question (objectif) pour les auteurs de cet article de mettre en place un modèle pour prédire les taux de réadmission chez les patients souffrant d'insuffisance cardiaque et ce modèle est basé sur les données prélevées dans le dossier médical électronique. **La sortie de ce modèle est de classifier les patients Réadmis (RA) et Non-Réadmis (NonRA) qui est classée comme une tâche de classification binaire**.

## III. PROPOSITION DE LA SOLUTION

Pour résoudre ce problème de réadmission qui est lié à l'insuffisance cardiaque, les auteurs de cet article ont proposé une solution qui est l'une des l'une des premières tentatives d'utiliser les données à l'échelle du phénomène pour identifier de nouveaux facteurs de réadmission liés à l'insuffisance cardiaque et développer des modèles de prédiction à l'échelle du Dossier Médical Electronique (DME). Pour ce faire, deux étapes sont mises en place pour la stratégie de modélisation. La première qui est une approche basée sur la sélection des caractéristiques basées sur les données (variables) fournies par le DME. Dans ce cas, la sélection des variables qui contribueront à la construction du modèle sont retenus par la méthode de sélection des caractéristiques basées sur la corrélation. La deuxième partie est la construction des modèles en utilisant une méthode de validation croisée 5 fois (k=5).

## IV. PRESENTATION DES RESULTATS

Les résultats de cette études sont liés à plusieurs éléments à savoir ; les caractéristiques de la cohorte, la sélection des caractéristiques à l'échelle de DME et modélisation prédictive à l'aide de cinq modalités de données différentes, la réduction de

fonctionnalités et affinement du modèle et enfin la comparaison avec les modèles existants de réadmission pour l'insuffisance cardiaque. Pour une bonne compréhension, nous avons trouvé nécessaire de présenter selon chaque étape mise en place par les auteurs.

### *A. La cohorte*

La cohorte, elle est constituée de 4205 variables extraites du DME. Celle-ci se trouvant être classées en cinq (05) catégories à savoir :

- Les codes de diagnostic (codes de la CIM-9 et IMO), avec un total de n = 1763 variables pour CIM-9. Il est à noter que tous les codes ont été unifiés selon CIM-9 ;
- Les procédures (codes CIM-9, SNOMEDCT et CPT), avec n = 564 variables et toutes les procédures ont été utilisées. Donc celles codées SNOMED-CT ou CIM-9-CM ;
- Les médicaments, avec n = 1028 variables et toutes les données sur les médicaments avaient été normalisées à l'aide de RxNorm ;
- Les mesures du laboratoire, avec n = 846 ;
- Les signes vitaux, avec n = 4 constitués du ≪ pouls, le taux de respiration, la pression artérielle systolique, les battements cardiaques et la température.

Lorsque nous additionnons ces quantités toutes les variables (1763 + 564 + 1028 + 846 + 4 = 4205), nous avons le total exacte de la cohorte qu'avait été utilisé par les auteurs de cet article.

### *B. La sélection des caractéristiques*

La sélection des caractéristiques à l'échelle du DME et la modélisation prédictive à l'aide de cinq modalités différentes est inspirée de la figure 1 de l'article dans sa page 4 comme le montre la figure 1 dans notre rapport.

Cette figure nous montre les différentes étapes à suivre pour mettre en place le modèle qui permettra de faire la prédiction des patients souffrant de l'insuffisance cardiaque ayant déjà été admis en hospitalisation pour au moins une fois. Les auteurs de l'article cherchent à évaluer les taux potentiels de réadimission de ces patients dans les 30 jours qui suivent leur première hospitalisation. **La sortie de ce modèle est LA CLASSIFICATION DES PATIENTS READMIS ET NON-READMIS**. Dans la première étape, K. SHAMEER et al. passent tout d'abord au prétraitement du jeu de données pour le cas des données manquantes et des doublons si elles existent. Une fois ce prétraitement terminé, s'en suit la sélection des variables qui permettrons de mettre en place le modèle. Et cette sélection est faite en fonction des cinq variables retenues du DME, chacune faisant objet d'un modèle généré qui permet de sélectionner les caractéristiques pertinentes. Ces dernières ont été comparées à l'aide de trois méthodes à savoir la régression logistique, PCA (Analyse en Composants Principal) et la mesure orthogonales ceci dans le but de comprendre l'espèce variable et leur relation inhérentes ou leur corrélation. La deuxième étape consiste à générer un modèle qui permettra d'effectuer des prédictions. La troisième partie permet de voir la précision du modèle ainsi que la qualité de ce modèle avec AUC (Area Under the Curve) comprise entre 0.5 et 1. La quatrième concerne la validation du modèle avec les données de tests. La figure 2 est le tableau des résultats du modèle mis en place par KHADER SHAMEER et al en utilisant l'algorithme Naïve Bayes.

(FIGURE 1).

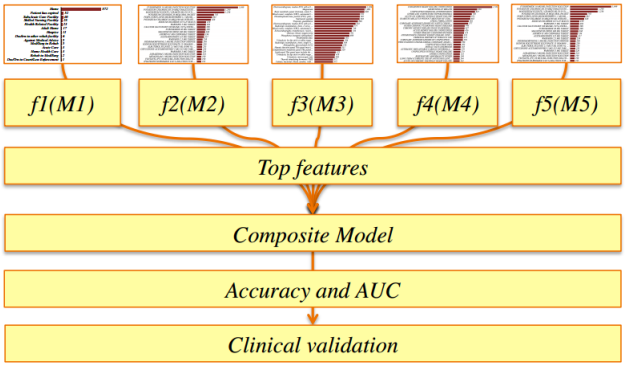


FIGURE 1. Architecture d'apprentissage automatique à l'échelle du DME et stratégie de modélisation prédictive pour trouver les facteurs déterminants des taux de réadmission

Au vue de ce tableau, Le modèle final est développé en utilisant 105 sur les 4205 caractéristiques avec une AUC = 0,78, proche de 1 ce qui montre la bonne qualité du modèle et une précision de test de validation croisée de 83,19 % confirmant le résultat de l'AUC.

(FIGURE 2).

| Data-element | Type | Encoding | Accuracy | AUC | Features |
|---|---|---|---|---|---|
| Diagnosis | ICD-9 Diagnosis | Binary | 70.3297% | 0.605 | 34/1763 |
| Procedures | ICD-9-Procedure | Binary | 77.907% | <0.50 | 4/273 |
| Procedure | CPT-codes | Binary | 72.9858% | 0.553 | 8/564 |
| Medications | Medication name and dosage | Binary | 81.9048% | 0.615 | 26/1028 |
| Labs | Non-descriptive lab measurements | Continuous | 73.9336% | 0.535 | 29/846 |
| Composite model | Combined features | Hybrid | 83.9000% | 0.780 | 105 |

FIGURE 2. Récapitulatif des différents prédicteurs et caractéristiques bayésiens compilés à l'aide de la méthode CFS

### *C. Comparaison avec les modèles actuels de réadmission pour l'insuffisance cardiaque*

Pour une étude comparative, les auteurs ont présenté les résultats obtenus à partir de certains critères de sélection des caractéristiques et avec un algorithme d'apprentissage automatique bien précis (Naïve Bayes) et ont obtenus comme résultat AUC = 0.78 et Accuracy = 83.90 %. Ces derniers avaient été comparés à une modélisation prédictive e réadmissions hospitalières utilisant les données de santé du Québec, canada par Hosseinzadeh et al la méthode de sélection des caractéristiques pour la construction du modèle reste inconnue et les algorithmes utilisés étaient Naïve Bayes (0.65) et Random Forest (0.64). Pour la cohorte de diabétiques d'un hôpital en Inde, Duggal et al ont montré que Naïve Bayes (0,67) présentait des économies associées à la réadmission plus élevées que la régression logistique (0,67), Forêts aléatoires (0,68), Adaboost (0,67) et réseaux neuronaux (0,62) . Futoma et.al ont montré

que Random Forests (0,68) et l'apprentissage en profondeur en utilisant des réseaux de neurones (0,67) ont un taux de précision similaire inférieur à 1 million de patients et inférieur à 3 millions d'admission.

## V. CONCLUSION ET PERSPECTIVES

Parvenu au terme de notre analyse des travaux de recherche de **KHADER SHAMEER et al.** qui, nous présentent dans cet article une nouvelle procédure de développement d'une approche de sélection de caractéristiques basées sur les données électroniques à l'échelle du DME et l'apprentissage automatique pour prédire le taux de réadmission hospitalière des patients atteints de l'insuffisance cardiaque dans les 30 jours. Les données provenant de la base de données MySQL de l'hopital du Mont Sinaï avaient tout d'abord été soumis à un prétraitement ensuite de chaque modalité avait été généré un modèle dans le but de sélectionner les caractéristiques inhérentes. Pour cette sélection des caractéristiques, avait été appliqué des méthodes comme la régression logistique, ACP (Analyse des Composants Principaux). Seules celles des caractéristiques corrélées avaient été retenues pour former le modèle prédictif à mettre en place. **105 sur 4205 variables** (confère FIGURE 2) que forment la cohorte sont retenues après l'extraction des meilleures caractéristiques pour former le modèle et **La sortie de ce modèle est la classification des patients Réadmis (RA) et Non-Réadmis (NonRA)**. Cependant, **KHADER SHAMEER et al.** pour leurs expérimentations avaient divisé les données en deux jeux « train et test » avec 70 % pour le train et 30 % pour le test. Ce modèle soumis à un algorithme d'apprentissage automatique « **Naïve Bayes** » a donné un taux de précision de test de validation croisée de **83.90 %** et **AUC de 0.78**. Ce qui présente un meilleur résutat comparativement à ceux des autres chercheurs présentés dans cet article. En ce qui est de notre ressort, nous avons comme perspective la confirmation de ces résultats sera faite après une analyse comparative des jeux de données de **KHADER SHAMEER et al.** et ceux des autres chercheurs ainsi que les méthodes et algorithmes utilisés pour en tirer une meilleure conclusion et si possible faire de nous même une impléméntation du même modèle tout en respectant la même architecture, les mêmes librairies python, le même algorithme utilisés par **KHADER SHAMEER et al.** et en fin expérimenter avec d'autres d'apprentissage automatique dans le but de comparer les différents résultats à celui des auteurs de l'article.

## RÉFÉRENCES

[1] K. SHAMEER, K. W. JOHNSON, A. YAHI, R. MIOTTO et. al. *Predictive Modeling of Hospital Readmission Rates Using Electronic Medical Record-wide Machine Learning : a case-study using Mount Sinai heart failure cohort* . Pacific Symposium on Biocomputing 2017. Page 1-10