

P089 - Création d'un système d'aide informatisée au codage CIM: choix de métrique et utilisation de la hiérarchie CIM

Q. Marcou^{1,2,*}, L. Berti-Equille³, N. Novelli²

¹ Aix-Marseille Université, Faculté des sciences médicales et paramédicales, Marseille, France

² Aix-Marseille Université, UMR7020 CNRS, LIS, Marseille, France

³ IRD, UMR 228 ESPACE-DEV, Montpellier, France

*Auteur correspondant

E-mail address: quentin.marcou@ap-hm.fr

Introduction : Différentes versions de la CIM (Classification Internationale des Maladies) sont routinemement utilisées dans plusieurs dizaines de pays pour l'annotation d'évènements de soins. En France notamment, l'annotation des séjours hospitaliers à l'aide de la CIM10-FR, par les cliniciens et/ou les services d'information médicale, est le pilier central du financement hospitalier à travers la tarification à l'activité (T2A). Pourtant le codage CIM est une tâche à la fois chronophage et difficile. Cette difficulté se traduit par une faible qualité des données, avec par exemple, une concordance inter-codeur estimée entre moyenne et pauvre pour les diagnostics principaux au niveau de détail requis pour la tarification. Il existe donc un triple enjeu pour la création de systèmes d'aide au codage: un gain de temps clinique, ainsi que l'amélioration de la qualité des données épidémiologiques pour la recherche et pour le financement du système de soin. Avec la publication de jeux de données hospitalières ouverts, de nombreux articles s'intéressant au codage CIM à partir de données du dossier clinique ont été publiés dans la dernière décennie. Néanmoins, ces publications étudient une grande variété d'algorithmes au moyen de métriques de performance et sur des sous-ensembles de données hétérogènes ciblant l'automatisation du codage plutôt que l'assistance au codage. A contre-pied de cette tendance, nous cherchons, par notre étude, à dégager plutôt les principes généraux pour la construction d'un système d'assistance informatisée au codage.

Méthodes : Nous utilisons le jeu de données ouvert de réanimation MIMIC-III pour construire un système d'aide au codage CIM-9 basé sur les prescriptions médicamenteuses à l'aide d'un réseau de neurones séquentiel. Nous investigons dans un premier temps le choix de la métrique de performance à optimiser pour produire le système de recommandation de codes le plus performant. Dans un second temps, nous interrogeons l'intérêt de la pratique de filtrage (basé sur la fréquence et/ou catégories) des codes pour la construction de système d'aide au codage. Enfin, nous investigons également l'utilisation des propriétés hiérarchiques de la CIM à l'aide de réseaux de neurones pour la création d'un système d'aide au codage, en tirant parti de la corrélation des labels, en considérant le problème d'équilibre des classes et en utilisant explicitement les liens hiérarchiques.

Résultats : Nos expériences montrent que la métrique de performance la plus couramment utilisée, la micro F1, est aussi la moins adaptée à la création d'un système d'aide au codage. Nos résultats suggèrent également qu'utiliser l'ensemble des codes, plutôt qu'un sous-ensemble choisi, permet d'en retrouver un nombre plus important pour un même nombre de codes proposés. En effet, notre réseau de neurone permet d'extraire de l'information pour la quasi totalité des codes même lorsque leur fréquence est très faible ou leur lien avec la prise médicamenteuse insoupçonné. Enfin, nos résultats suggèrent également que l'utilisation de la hiérarchie des codes est bénéfique, notamment pour les niveaux les plus profonds de la hiérarchie.

Conclusion : Notre approche permet de dégager des principes généraux pour la construction d'un système d'aide au codage ainsi

que la quantification de ses performances. L'exploitation de ces principes nous permet de surpasser les performances de l'état de l'art utilisant les prescriptions médicamenteuses pour l'aide au codage.

Mots clés : Computer assisted coding (CAC) , Recommender systems , Hierarchical multilabel classification (HMC) , International classification of diseases (ICD) , MIMIC-III

Déclaration de liens d'intérêts : Les auteurs n'ont pas précisé leurs éventuels liens d'intérêts.

<https://doi.org/10.1016/j.respe.2023.101733>

P090 - L'approche fédérée permet-elle de préserver la valeur scientifique et statistique des résultats ?

D. Pau^{1,*}, R. Jegou², M. Boucher², I. Diabate¹, C. Monteil¹, C. Bachot¹

¹ Roche, Centre de données médicales, Boulogne Billancourt, France

² Keyrus Life Science, Biométrie, Nantes, France

*Auteur correspondant

E-mail address: david.pau@roche.com

Introduction : Les hôpitaux ont de plus en plus la volonté de valoriser leurs données et souhaitent limiter au maximum leur sortie en dehors de leur environnement. L'approche fédérée permet de générer des résultats agrégés à partir de plusieurs établissements de soin sans transfert des données. L'objectif de ce projet est d'évaluer si cette approche permet de conserver la valeur scientifique et statistique des résultats obtenus.

Méthodes : Trois bases de données étanches (en silos) ont été constituées à partir des données de l'étude de vie réelle KADOR dans le cancer du sein précoce. Aucune donnée individuelle n'a quitté la plateforme de données intégrant les bases de données étanches et seules des données agrégées ont été générées à partir des fonctions d'analyse sous R et des fonctions d'agrégation disponibles dans DATASHIELD (Fig. 1). Statistiques descriptives, matrices de corrélation, analyses de survie et modèles de régression ont été réalisées sur les données brutes (base complète), reproduites sur chaque échantillon indépendant puis les résultats ont été agrégés afin de comparer les résultats obtenus.

Résultats : Au total, 315 patients ont été inclus dans l'étude KADOR, 3 échantillons aléatoires de taille différente ont été générés (N1=157, N2=94, N3=64). Onze variables dérivées et 24 tables (103 items individuels/variables) ont été générées dans les rapports statistiques ; 100 % des variables dérivées ont pu être reproduites à partir de DATASHIELD et 93,9 % des items individuels ont été reproduits à partir des fonctions implémentées dans DATASHIELD: 100 % des 73 variables de statistiques descriptives, 89 % des 9 paramètres de la matrice de corrélation, 67 % des 12 paramètres des analyses de survie et 78 % des 9 paramètres du modèle de régression. Selon les méthodes d'agrégation disponibles dans DATASHIELD, les résultats obtenus peuvent différer par paramètre (par exemple les pourcentages et moyennes sont similaires, néanmoins les valeurs des quartiles peuvent différer).

Conclusion : La comparaison des résultats a permis de montrer que la plupart des analyses ont pu être reproduites par requête fédérée implémentée dans DATASHIELD. Les résultats obtenus avec l'approche fédérée sont dépendants de la méthode d'agrégation utilisée. La création de fonction spécifique dans DATASHIELD peut permettre l'ajout de méthode d'agrégation supplémentaire (courbes de Kaplan-Meier notamment).

Mots clés : Approche fédérée , Données de vie réelle , Datashield

Déclaration de liens d'intérêts : D.Pau, C. Monteil, I. Diabate, C. Bachot sont employés par Roche SAS.