E02 : Développement du projet RUBY : Développement d'algorithme d'intelligence artificielle pour la structuration des données des comptes rendus médicaux de patientes atteintes d'un cancer du sein

Titre

Français : Développement du projet RUBY : Développement d'algorithme d'intelligence artificielle pour la structuration des données des comptes rendus médicaux de patientes atteintes d'un cancer du sein

Anglais : RUBY : Artificial intelligence algorithm for automatic structuration of health record for patients with breast cancer

Auteurs

R SCHIAPPA (1), G UZBELGER (2), B THAMPHYA (1), A FABRE (2), S TOLEDANO (2), J GAL (1), O SCHNEIDER (3), C GARCIA (2), S BENAVIDES (2), J GOFDROY (3), J HAUDEBOURG (4), C BAILLEUX (5), JM FERRERO (5), O HUMBERT (6), E BARRANGER (5), E CHAMOREY (1)
(1) DEBDS, Centre Antoine Lacassagne, 33 avenue de Valombrose, 06189, Nice, France
(2) Artificial Intelligence – Advanced Analytics Solutions, IBM, 17 Avenue de L'Europe, 92275, BOIS-COLOMBES, France
(3) DSI, Centre Antoine Lacassagne, 33 avenue de Valombrose, 06189, Nice, France
(4) ACP, Centre Antoine Lacassagne, 33 avenue de Valombrose, 06189, Nice, France
(5) Oncologie-Sénologie, Centre Antoine Lacassagne, 33 avenue de Valombrose, 06189, Nice, France
(6) Médecine Nucléaire, Centre Antoine Lacassagne, 33 avenue de Valombrose, 06189, Nice, France

Responsable de la présentation

Nom : SCHIAPPA
Prénom : Renaud
Adresse professionnelle : 33 avenue de Valombrose
Code postal : 06189
Ville : Nice
Pays : France
Newsletter :

Mots clés

Français : Intelligence artificielle, machine Learning, deep Learning, cancer du sein
Anglais : artificial intelligence, machine Learning, deep Learning, breast cancer

Spécialité

Principale : Autres

Texte

Contexte
La prise en charge des patients génère de grande quantité de données dont 80% sont enregistrées dans des comptes rendus (CR) textuels non structurés. Au Centre Antoine Lacassagne (CAL), c'est plus de 3 millions de CR qui constituent un réservoir d'informations très peu exploité. A l'heure du Big Data et de l'Intelligence Artificielle (IA), la création d'une plateforme de données de santé structurées et exploitables est un important challenge pour les établissements de santé.

Objectifs
L'objectif de cette première étape de RUBY était de développer, en collaboration avec IBM, des algorithmes d'IA capables de structurer les données des CR des patientes atteintes d'un cancer du sein et de les intégrer automatiquement dans un fichier de données structurées. Cette étude de type « preuve de concept » a été réalisée sur la première consultation (PCONS), la première biopsie (PBIO), la première chirurgie (PCHIR) et le premier CR d'anatomopathologie (PANA) des patientes du CAL.

Méthodes
Deux bases de données de cancer du sein ont été fusionnées pour créer une base de données structurées (BDDS). La base SEIN-CAL (patientes prises en charge avant 2008) et la base ESME-CSM (patients prises en charge après 2008). La population a été scindée en deux, permettant de créer une cohorte d'entrainement (CE=70% des patientes et leurs CR associés) et une cohorte de test (CT=30%). Les CR ont nécessité un prétraitement, et une segmentation a été effectuée afin de faciliter l'identification des données à extraire. Les CR ont été annotés manuellement avec BRAT, puis des algorithmes d'apprentissage utilisant le réseau neuronal convolutif ont été exécutés avec SpaCy®. Un fichier de données structurées (.csv) est produit et les indicateurs de performances de RUBY ont été évalués et comparés aux performances d'une structuration manuelle des données par un attaché de recherche clinique (ARC).

Résultats
Plus de 2300 patientes ont été incluses dans les deux bases fusionnées. Pour les variables PCONS, sur 8 variables testées, la précision de RUBY

était comprise entre 64% et 98,18% respectivement pour les variables « N clinique » et « indication de la première venue ». Pour PBIO, sur 10 variables, la précision
de RUBY variait de 93 à 100%. Pour PCHIR, la précision de RUBY était >93% pour 6 des 7 variables d'intérêt et de 79% pour l' « indication de la chirurgie ». Pour PANA, la précision de RUBY était >90% pour 15/19 variables dont 11 étaient >95%, la précision était >75% pour les 4 autres variables. A ce stade de niveau d'apprentissage, les performances de RUBY sont supérieures à celles d'un ARC dans 43% des cas. Le temps nécessaire pour structurer automatiquement les 2036 CR de la CT a été de 9,7 minutes avec RUBY versus 30 jours par un ARC en structuration manuelle.

Discussions
Les premiers résultats de RUBY sont très encourageants et une nouvelle phase d'annotation est en cours afin d'améliorer les résultats de l'algorithme. Les performances de RUBY sont meilleures sur les CR semi structurés comme PBIO, PCHIR et PANA. Les PCONS sont les plus difficiles à structurer automatiquement car restent médecins-dépendants.

Conclusion
Le démonstrateur RUBY a permis de progresser dans la structuration automatique des données du cancer du sein au CAL. Il est maintenant nécessaire d'optimiser les algorithmes afin d'améliorer les performances de RUBY et de déployer cette application à d'autres types de CR et d'autres pathologies.