Alexandre Batisse[1], Pauline Lemeille[1], Ludovic Lamarsalle[1], Martin Prodel[1]

[1]HEVA, 186 avenue Thiers 69465 Lyon Cedex 06, France

Colloque Données de santé en vie réelle 2021 | AFCROS

28 juin 2021

Analyse et exploitation des Résumés des Caractéristiques du Produit : création d'un moteur de recherche de médicaments

Contexte

Tout médicament autorisé ou l'ayant été, en France, dispose d'un Résumé des Caractéristiques du Produit (RCP), mis à disposition par l'ANSM ou l'EMA. Il comporte les informations destinées aux professionnels de santé et est annexé à la décision d'Autorisation de Mise sur le Marché.

Objectif

Créer automatiquement une unique base de données :
- Qui contienne de multiples caractéristiques de médicaments
- Qui soit utilisable par des pharmaciens, consultants, etc..

Trouver des médicaments sur la base du croisement d'informations décrites dans les RCP (indications cliniques, contre-indications, fabricants, dates, substance active, conditionnement, excipient).

Méthode

1. L'ANSM fournit sur son site des RCP au format HTML, format balisé et permettant donc d'extraire en routine les sections d'intérêt. Les médicaments autorisés par le circuit européen ont leurs RCP hébergés au format PDF sur le site de l'EMA. Ce format a été traité par des expressions régulières pour délimiter les sections d'intérêt.

2. Les contenus textuels extraits des sections ont été organisés dans une base de données orientée documents avec l'outil Elasticsearch. Cette base crée un système d'index qui classe les documents par mots clés.

3. La robustesse aux erreurs typographiques est gérée par la distance de Levenshtein. Les médicaments renvoyés lors d'une requête sont classés par ordre de pertinence selon la méthode de pondération TF-IDF (term frequency-inverse document frequency).

4. La validité des résultats du moteur de recherche a été établie par une validation manuelle sur trois aires thérapeutiques.

Résultats

Ont été extraits :
- L'indication thérapeutique
- La composition
- Le laboratoire pharmaceutique titulaire de l'AMM

Les 5 faux positifs retrouvés pour le cancer du sein sont des RCP organisés différemment, avec un paragraphe relatif aux recommandations pour les populations particulières (dont les patients atteints du cancer du sein) dès la section relative à la posologie. L'inclusion de la section posologie dans les indications cliniques nous permet d'obtenir une plus grande exhaustivité des résultats. L'exclusion de ces RCP des résultats fera l'objet de travaux futurs.

Les faux négatifs sont :
- **Myélome multiple :** 1 référence non indexée dans la base ;
- **Cancer du sein :** 15 références non retrouvées à cause de l'utilisation de synonymes (e.g. carcinome mammaire pour cancer du sein), et sont donc tous rattrapés dans un 2ème temps ;
- **Psoriaris :** 1 RCP qui ne mentionne pas explicitement l'indication cherchée.

Conclusion

Les RCP forment une source fiable d'information sur les médicaments. Ils sont accessibles via l'ANSM et l'EMA, mais ils restent difficilement exploitables en routine à cause de leur nombre et de leur format.
L'utilisation de techniques d'indexation de corpus documentaire a permis d'unifier l'extraction d'informations clés de ces RCP.

Son intérêt dépasse sa possibilité actuelle de croiser les critères de recherche sur les caractéristiques des médicaments, à partir d'une source consolidée et à jour.

Cette base indexée pourra également servir à mener des analyses médico-économiques et épidémiologiques à grande échelle et déclinées sur de multiples spécialités thérapeutiques.

HEVA