Revue de la littérature : entrepôts de données biomédicales et traitement automatique de la langue

Adrien Bazoge

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Revue de la littérature : entrepôts de données biomédicales et traitement automatique de la langue

Adrien Bazoge[1, 2]
(1) LS2N, UMR CNRS 6004, Université de Nantes, France
(2) CHU de Nantes, INSERM CIC 1413, Pôle Hospitalo-Universitaire 11 : Santé Publique, Clinique des données, Nantes, France
adrien.bazoge@univ-nantes.fr

RÉSUMÉ

La quantité de données de santé informatisées ne cesse de croître et ouvre de nouvelles possibilités pour la recherche scientifique. L'accès à ces données passe très souvent par l'utilisation d'entrepôts de données biomédicales, déployés pour cet usage. Parmi les données stockées dans ces entrepôts, on peut trouver des données textuelles, en plus ou moins grande quantité. Le traitement automatique de la langue (TAL) est le domaine de prédilection pour l'exploitation des données textuelles. Cet article propose une revue de la littérature qui s'intéresse, à travers les publications sur PubMed, ACL Anthology et Google Scholar, à l'interaction entre deux thématiques : les entrepôts de données biomédicales et le traitement automatique des langues. Cette revue montre que l'intérêt pour les données de santé et les entrepôts de données biomédicales est en constante croissance dans la littérature. Elle montre également que le TAL devient peu à peu un outil indispensable afin d'exploiter au mieux les entrepôts de données biomédicales.

ABSTRACT

Literature review : biomedical data warehouse and natural language processing

The amount of electronic health data continues to grow and its availability for research purposes opens up new era of secondary uses of biomedical data. In care organisations, access to these data is mediated by biomedical data warehouses. Biomedical Data warehouses have been recently deployed and recognized as an value creation instrument in care organisations through data. Among the large diversity of data, textual information can be found, in great quantities. To facilitate handling of textual information, natural language processing (NLP) algorithm are increasingly used in biomedical data warehouse. This review present a systematic literature analysis of publications from sources (PubMed, ACL Anthology and Google Scholar) in the interaction between three themes : computerization of health data, biomedical data warehouses and natural language processing. This review shows that the interest in health data and biomedical data warehouses is exponentially growing in the literature. It also shows that NLP is a pivotal tool of data access, extraction and transformation in biomedical data warehouses in all fields of modern medicine.

MOTS-CLÉS : revue de la littérature, entrepôt de données biomédicales, traitement automatique de la langue.

KEYWORDS: literature review, biomedical data warehouse, natural language processing.

Volume 2 : 23e REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL), pages 96–109.

1 Introduction

Depuis 20 ans, les données de santé issues du soin des patients sont systématiquement archivées. Les bases de données ainsi constituées, souvent électroniquement, rassemblent à la fois des données structurées (biologie, démographies, etc.) et des données non structurées (comptes rendus textuels d'hospitalisation ou de consultation). L'intentionnalité première de ces données est l'acte de soin au sens large, leur usage est celui du soin et non de la recherche biomédicale. La recherche biomédicale est un secteur où la source traditionnelle de données chez l'homme est un essai clinique ou un registre de pathologie. Cette masse de données est au carrefour de multiples contributions : celle du patient, pour lequel les données sont collectées lors de l'hospitalisation ou de la consultation ; celles des soignants, qui s'occupent des patients et permettent la collection de ces données ; et celles de l'établissement de santé, qui organise toute la logistique opérationnelle et financière autour du soin et de ses données. Dans un premier temps utilisées pour le soin, ces données peuvent désormais être exploitées à des fins secondaires, pour la recherche et l'évaluation des soins, postérieurement à leur enregistrement, grâce aux avancées technologiques en matière d'intelligence artificielle (IA) appliquées sur des grandes masses de données (*big data*). Cette grande quantité de données qui devient accessible, et notamment les données textuelles, renforce l'intérêt de l'application du traitement automatique des langues (TAL) qui met en oeuvre des algorithmes permettant d'opérer à une échelle aussi massive que les données elles-mêmes (Daille & Nazarenko, 2017).

Dans cette revue de la littérature, nous étudions l'évolution de l'application du TAL dans les entrepôts de données biomédicales depuis l'informatisation des données de santé, à travers les publications sur des moteurs de recherche de la littérature scientifique : PubMed [1], ACL Anthology [2] et Google Scholar [3]. Bien que l'application du TAL soit répandue sur les dossiers patient informatisés de manière générale, nous nous intéressons ici uniquement à l'application du TAL sur des données issues des entrepôts de données de santé. En effet, la recherche sur données de santé est de plus en plus réglementée afin de mieux préserver la confidentialité des patients à l'origine de ces données. Pour permettre cette réglementation, l'exploitation de ces données pour la recherche passe désormais davantage par les entrepôts de données de santé, conçus pour cet usage. L'objectif de cette revue de la littérature est donc d'analyser l'évolution de l'application du TAL sur les données de santé issues des entrepôts de données biomédicales. Cet article est structuré en trois sections. La première section définit les thématiques. Notre méthodologie est présentée dans la deuxième section et aborde la construction des requêtes sur les moteurs de recherche et la définition des axes de classification des publications. Ensuite, la dernière section rassemble les résultats des requêtes et leur analyse.

2 Entrepôts de données biomédicales

Cette revue de littérature est restreinte au champ des entrepôts de données biomédicales pour laquelle nous appliquons la définition suivante :

Un entrepôt de données de santé (*Health Data Warehouse*), aussi appelé entrepôt de données biomédicales ou entrepôt de données cliniques (*Biomedical Data Warehouse*, *Clinical Date Warehouse*), est une base de données relationnelle regroupant une partie ou l'ensemble des données d'une base

1. https://pubmed.ncbi.nlm.nih.gov/
2. https://www.aclweb.org/anthology/search/
3. https://scholar.google.com/

de données opérationnelle dans un établissement de soin. Les entrepôts de données peuvent être construits à partir de plusieurs sources de données *via* un processus dit ETL (*extract, transform, load*). Les entrepôts de données sont ensuite utilisés pour le pilotage de l'activité ou son évaluation à travers les statistiques et l'analyse de données. L'explosion de la production de données numériques a été le facteur permettant de démocratiser la construction et l'utilisation des entrepôts de données.

Le domaine de la santé a également tardé à intégrer en profondeur cette transition numérique. Bien que les entrepôts de données soient installés dans le paysage clinique anglo-saxon depuis plus de dix ans, ce n'est qu'après l'obtention de l'autorisation de la CNIL [4] que les premiers entrepôts de données biomédicales voient le jour en France pour une utilisation à des fins de recherche. L'AP-HP [5] est le premier établissement à obtenir cette autorisation en janvier 2017, suivie par le CHU de Nantes en juillet 2018 et le CHU de Lille en septembre 2019. Les entrepôts de données biomédicales rassemblent les données de millions de patients traités dans les établissements hospitaliers. Les données contenues dans ces entrepôts sont de natures diverses : des données démographiques, des données du PMSI [6], des résultats de biologie et d'imagerie, des prescriptions de médicament ou encore des comptes rendus médicaux de consultation ou d'hospitalisation. À titre d'exemple, l'Entrepôt de Données Biomédicales Nantais (EDBN) rassemble les données de 2,7 millions de patients pour 30 millions de documents.

Dans le cadre de l'exploitation des entrepôts de données biomédicales, ces données sont utilisées à des fins de recherche et peuvent permettre d'améliorer l'efficacité des systèmes de santé, la vigilance et la veille sanitaire.

3 Thèmes de l'analyse

À l'aide des moteurs de recherche de la littérature scientifique PubMed, ACL Anthology et Google Scholar, nous nous intéressons aux travaux publiés entre 1995 et 2020, à travers trois thématiques : (i) l'informatisation des données de santé, (ii) les entrepôts de données de santé et (iii) le traitement automatique de la langue. La thématique d'« informatisation des données de santé » fait référence à la transition numérique, la constitution de bases de données pour stocker les données de santé des patients (Moore *et al.*, 2021). Pour chacune de ces thématiques, une liste de mots clés a été établie :

1. Informatisation des données de santé : *electronic medical record, EMR, electronic health record, EHR, real world evidence, real world data*

   Les mots clés « *electronic medical record* » et « *electronic health record* » font références aux dossiers patient informatisés qui peuvent être exploités dans les études, tandis que les mots clés « *real world data* » et « *real world evidence* » correspondent plutôt aux données de soin des patients qui sont générées au cours de la pratique clinique de routine.

2. Entrepôts de données de santé : *clinical data warehouse, biomedical data warehouse, health data warehouse*

   Les mots clés sélectionnés pour représenter la thématique « Entrepôts de données de santé » correspondent aux appellations les plus couramment utilisées pour désigner les entrepôts de données de santé.

4. Commission Nationale Informatique et Libertés
5. Assistance Publique - Hôpitaux de Paris
6. Programme de Médicalisation des Systèmes d'Information

3. Traitement automatique de la langue : *natural language processing, NLP, text mining*

   Le mot clé « *text mining* » vient ici compléter le mot clé « *natural language processing* ». En effet, la fouille de textes apparait comme étant l'application du TAL la plus utilisé dans le domaine médical. C'est pourquoi le terme « *natural language processing* » peut parfois être eclipsé par le terme « *text mining* ».

À partir de ces listes de mots clés, plusieurs requêtes, présentées dans la section suivante, ont été faites sur les différents moteurs de recherche.

4 Collecte des publications

Trois moteurs de recherche bibliographiques ont été utilisés pour cette étude : PubMed, ACL Anthology et Google Scholar. PubMed est spécialisé dans la médecine et la biologie, son mode de requêtage permet de construire des requêtes qui s'appuient à la fois sur des descripteurs MeSH (*Medical Subject Headings*) et sur du langage naturel. ACL Anthology couvre la bibliographie liée à la linguistique informatique et au traitement automatique des langues. Le moteur de recherche ACL Anthology fonctionne avec Google Custom Search [7]. Google Scholar, quant à lui, n'a pas de domaine de spécialité particulier pour les publications qu'il référence. Toutes les requêtes présentées dans cette section ont été exécutées le 18 mai 2021. Les publications PudMed et ACL Anthology ont été récupérées après exécution manuelle des requêtes sur les sites web respectifs de ces bases de données bibliographiques. Quant aux publications Google Scholar, elles ont été collectées à l'aide du logiciel libre « Publish or Perish » [8].

PubMed est le moteur de recherche le plus sophistiqué parmi les trois utilisés. Nous avons donc pu croiser facilement les trois thématiques présentées précédemment et construire les requêtes suivantes :

- Requête 1 - Informatisation des données de santé

"electronic health records"[MeSH Terms] OR ("electronic health record") OR ("electronic medical record") OR ("EMR") OR ("EHR") OR ("real world data") OR ("real world evidence")

- Requête 2 - Entrepôts de données de santé

("data warehousing"[MeSH Terms] OR ("data warehouse")) AND (("clinical") OR ("biomedical") OR ("health")))

- Requête 3 - Informatisation des données de santé et Traitement automatique de la langue

(("electronic health records"[MeSH Terms] OR ("electronic health record") OR ("electronic medical record") OR ("EMR") OR ("EHR")
AND
(("natural language processing") OR ("NLP") OR ("text mining"))

- Requête 4 - Entrepôts de données de santé et Traitement automatique de la langue

("data warehousing"[MeSH Terms] OR ("data warehouse")) AND (("clinical") OR ("biomedical") OR ("health")))

7. https://developers.google.com/custom-search/
8. https://harzing.com/resources/publish-or-perish

AND
(("natural language processing") OR ("NLP") OR ("text mining"))

Pour les moteurs de recherche Google Scholar et ACL Anthology, qui ne proposent pas de requêtage avancé comme PubMed, il était plus difficile de combiner les mots clés et croiser les thématiques. De ce fait, nous nous sommes concentrés sur l'intersection entre les thématiques d'entrepôts de données biomédicales et de traitement automatique de la langue, ce qui équivaut à la requête 4 faite sur PubMed.

Pour le moteur de recherche ACL Anthology, trois requêtes ont été construites. ACL Anthology ayant un domaine de base bibliographique couvrant le TAL, les mots clés liés à cette thématique n'ont pas été pris en compte pour les requêtes. Les résultats de ces requêtes ont été concaténés et les doublons de publications ont été filtrés :

- Requête 1 - *"clinical data warehouse"*

- Requête 2 - *"health data warehouse"*

- Requête 3 - *"biomedical data warehouse"*

Pour le moteur de recherche Google Scholar, trois requêtes ont été construites. Pour chacune de ces requêtes, nous lançons une requête similaire en remplaçant « *natural language processing* » par son acronyme « *nlp* ». Les résultats de ces requêtes ont été concaténés et les doublons de publications ont été filtrés :

- Requête 1 - *"clinical data warehouse" "natural language processing"*

- Requête 2 - *"biomedical data warehouse" "natural language processing"*

- Requête 3 - *"health data warehouse" "natural language processing"*

5 Classification des publications

Les requêtes croisant les thématiques « entrepôts de données biomédicales » et « traitement automatique de la langue » ont fait l'objet d'une analyse plus approfondie grâce à revue manuelle de publications. Les publications retenues pour cette revue ont été classifiées en cherchant à répondre à 6 questions :

1. *Quel est le sujet principal de la publication ?* :
   — extraction d'informations : un des objectif de la publication est d'extraire des informations dans des données textuelles, souvent avec l'utilisation du TAL.
   — exploitation des données d'un entrepôt : utilisation des données d'un entrepôt pour une étude précise (souvent médicale).

— revue de la littérature : articles de revue de la littérature sur des thématiques précise
— présentation d'outils : articles de présentation d'outils commerciaux ou libre de droit, ces outils proposent généralement des solutions d'entreposage de données ou des applications de techniques de TAL.
— autres : publications dont le nombre de publications par catégorie est trop faible pour constituer une catégorie à part entière. Les sujets abordés dans ces publications : techniques d'entreposage de données (construction entrepôt de données, intégration de données, etc.), requêtage des entrepôts.

2. *Quel est le cas d'usage de l'entrepôt de données dans la publication ?* :
   — exploitation de données structurées
   — exploitation de données non structurées
   — construction d'entrepôts de données
   — structuration de données : structuration des données non structurées en des formats de données existants, basés sur des lexiques et/ou ontologies.
   — autres (nombre d'occurrences trop faible pour en faire une catégorie à part entière)

3. *Est-ce qu'au moins une méthode TAL est mentionné dans la publication ? Si oui, quel(s) type(s) de méthode(s) ?* :
   — linguistique
   — apprentissage automatique
   — apprentissage profond
   — inconnu : l'utilisation du TAL est mentionné mais la méthode n'est pas précisée.

4. *Quelle est la langue des données exploitées ?* (si une méthode TAL est mentionnée dans la publication)

5. *Quel est l'objectif médical dans la publication ?* (si une méthode TAL est mentionnée dans la publication) :
   — Médecine interventionnelle : étude d'un acte fort de médecine (opérations, traitements, etc.)
   — Médecine de spécialité : étude d'une maladie dans son ensemble

6. *À quelle spécialité médicale se rattachent les données exploitées ?* (si une méthode TAL est mentionnée dans la publication) : neurologie, oncologie, pneumologie, etc.

Pour la question 1, les sujets ont été obtenus de manière itérative lors de la revue manuelle des publications. Lorsqu'une publication ne pouvait être associée à un sujet existant, un nouveau sujet était créé. Sur les questions 2 et 3, les publications peuvent recevoir plusieurs réponses.

6 Analyse des publications

Depuis l'informatisation des données de santé, les données issues du soin des patients sont davantage utilisées pour la recherche clinique. La distribution des résultats obtenus avec la première requête PubMed (cf. figure 1) montre que les articles mentionnant des dossiers patient informatisés dans la littérature ont augmenté de manière exponentielle ces dix dernières années, passant 1 850 mentions en 2010 à 7 915 en 2020.

La croissance d'utilisation des entrepôts de données biomédicales se reflète également dans la littérature scientifique (cf. figure 2).

La croissance des données de santé informatisées et l'utilisation des entrepôts de données figurent parmi les facteurs qui favorisent l'usage du TAL (cf. figures 3, 4 et 5), que ce soit pour extraire de l'information, ou pour pré-traiter des données textuelles. Le TAL permet de rendre plus accessible des informations uniquement présentes dans le texte. Ces informations peuvent ensuite être utilisées dans des études de recherche clinique, ou plus généralement, ajoutées dans les entrepôts de données biomédicales afin de les enrichir. L'extraction d'informations à l'aide du TAL peut aussi permettre de récupérer des données déjà présentes de manière structurée dans les entrepôts, afin de consolider ces données, mais également de compléter les données manquantes pour certains patients.

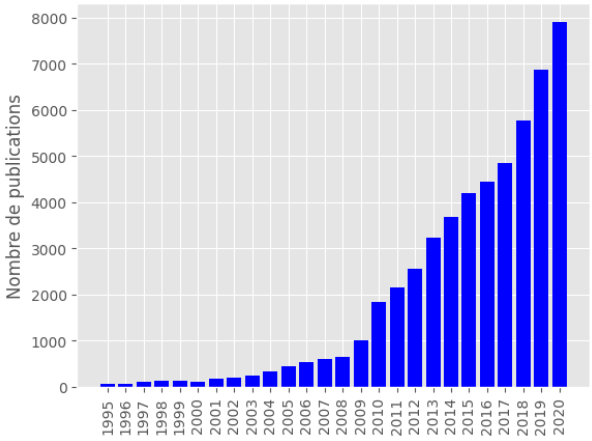

FIGURE 1 – Requête 1 - PubMed - Informatisation des données de santé

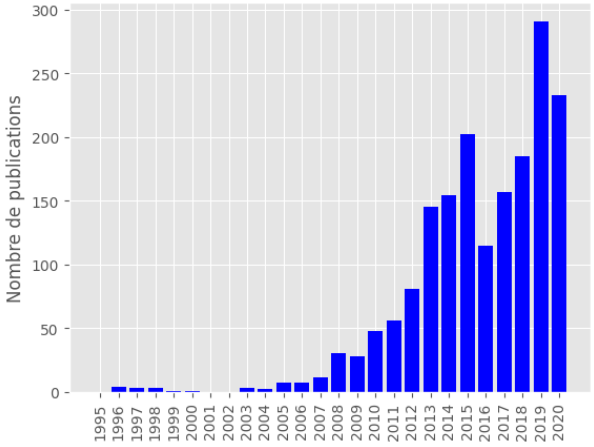

FIGURE 3 – Requête 3 - PubMed - Informatisation des données de santé et TAL

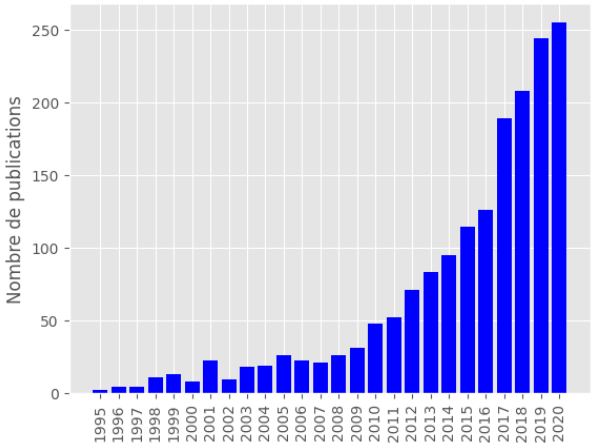

FIGURE 2 – Requête 2 - PubMed - Entrepôts de données de santé

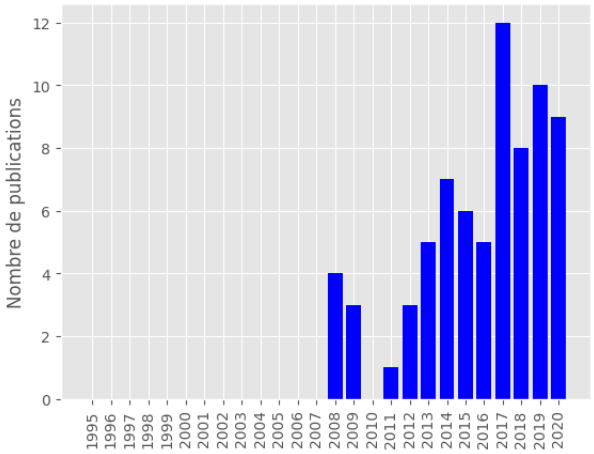

FIGURE 4 – Requête 4 - PubMed - Entrepôts de données de santé et TAL

Les résultats des requêtes croisant les thématiques « entrepôts de données biomédicales » et « traitement automatique de la langue » comptent 69 publications sur PubMed, 918 publications sur Google Scholar et seulement 3 publications pour ACL Anthology. Les publications issues de PubMed se trouvant également dans les résultats des requêtes Google Scholar ont été supprimées des résultats Google Scholar (11 publications Pubmed). Un échantillon de 80 publications de la requête Google Scholar ainsi que les 69 publications de la requête 4 PubMed ont été manuellement analysées, tandis que les publications de la requête ACL Anthology ont été abandonnées. Elles étaient trop peu nombreuses (seulement 3 publications) pour pouvoir être comparées avec les publications des autres moteurs de

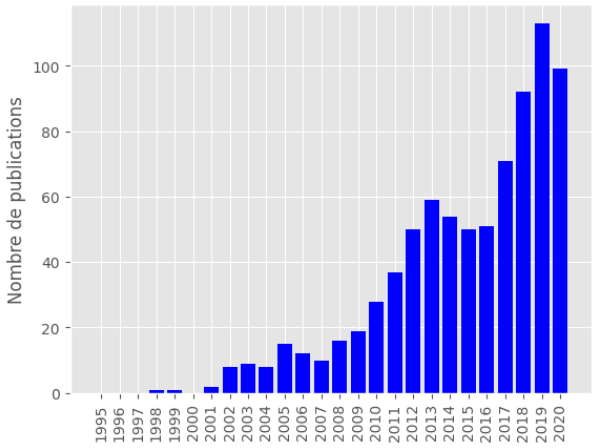

FIGURE 5 – Requête Google Scholar - Entrepôts de données de santé et TAL

recherches. Ce manque de publications sur ACL Anthology peut s'expliquer par le fait que les mots clés choisis soient trop stricts. Les publications présentes sur la base bibliographique ACL Anthology mettent généralement en avant des méthodes TAL. L'origine des données a donc moins d'importance dans ces publications, et la notion d'entrepôts de données peut paraître éloignée pour les auteurs. Les sujets traités dans ces publications sont variés (cf. figure 6). L'extraction d'informations est la thématique dominante parmi toutes ces publications puisqu'elle figure dans 74 publications (soit environ 50 % des publications revues manuellement). La thématique d'exploitation d'entrepôt de données est uniquement présente dans les publications PubMed (6 publications). Le manque d'articles sur cette thématique dans Google Scholar peut s'expliquer par le fait que cette thématique est proche du domaine médical, puisque cela correspond aux études sur données de santé. Par conséquent, on retrouve ces publications plus facilement sur Pubmed que sur Google Scholar. Parmi les publications résultants de cette requête se trouve également 33 revues de la littérature, dont la majorité provient de Google Scholar. Ces revues de la littérature portent sur différents sujets : le *big data* (Singh, 2019; Schoenthaler *et al.*, 2019), le traitement automatique de la langue (Sheikhalishahi *et al.*, 2019; Névéol *et al.*, 2018) ou plus généralement les données de santé informatisées (Safran, 2017). D'autres publications présentent différents outils ou logiciels prêts à l'emploi, tels que des outils d'entreposage de données ou d'extraction d'informations.

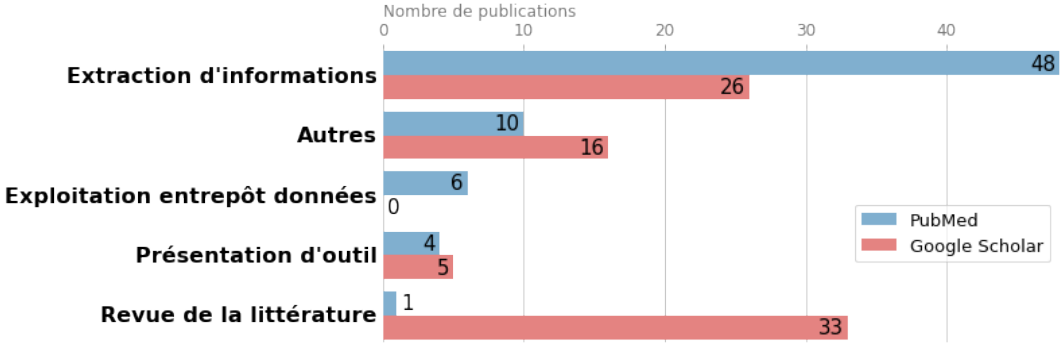

FIGURE 6 – Sujets traités dans les publications « Entrepôts de données de santé et TAL »

Les entrepôts de données peuvent avoir différents rôles (cf. figure 7). L'exploitation d'entrepôts de données est le cas d'usage le plus fréquent, avec d'un côté l'exploitation des données non structurées, présent dans 57 publications et, de l'autre, l'exploitation des données structurées, présent dans

20 publications. En amont de l'exploitation des entrepôts, la conception des entrepôts de données est également importante, avec la définition des données et des architectures qui composeront ces entrepôts. Entre ces deux tâches de conception et d'exploitation se place l'amélioration des entrepôts, avec notamment la structuration des données (Thoroddsen *et al.*, 2017; Chiudinelli *et al.*, 2019; Afshar *et al.*, 2019) déjà présentes dans l'entrepôt mais encore l'intégration de nouveaux flux de données (Delamarre *et al.*, 2015; Hernandez *et al.*, 2009).

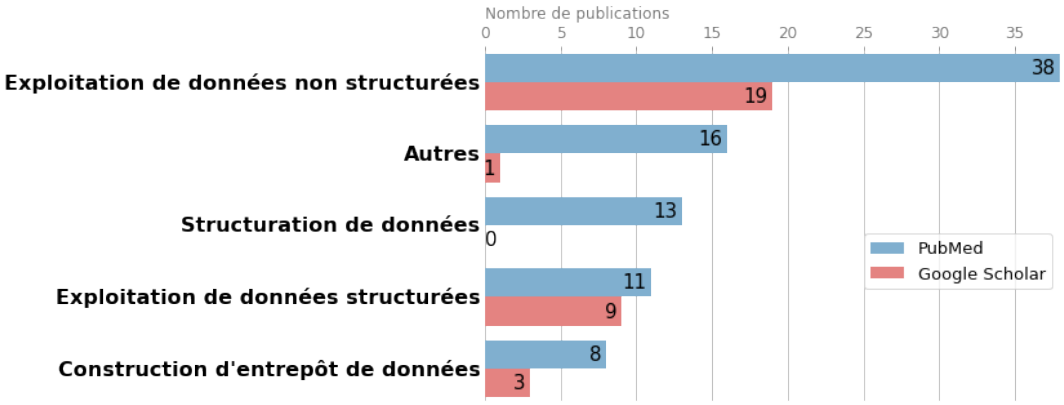

FIGURE 7 – Cas d'usage des entrepôts de données biomédicales dans les publications « Entrepôts de données de santé et TAL »

L'engouement de ces dernières années autour des méthodes à base d'apprentissage se reflète dans la littérature, la majorité des articles exploitant ces méthodes ont été publiés entre 2016 et 2020 (cf. figures 9, 10, 11 et 12). La régression (Quéroué *et al.*, 2019) et la classification (Osborne *et al.*, 2016; Chase *et al.*, 2017) comptent parmi les méthodes d'apprentissage automatique utilisées, tandis que les méthodes d'apprentissage profond s'appuient sur des réseaux de neurones (Zhao *et al.*, 2019; He *et al.*, 2019; Neuraz *et al.*, 2020). Malgré l'intérêt porté à ces méthodes, les méthodes linguistiques restent les approches les plus courantes dans la littérature médicale (cf. figures 8, 9 et 10). Parmi les méthodes linguistiques utilisées, on peut citer les approches à base de règles (Upadhyaya *et al.*, 2017; Lee *et al.*, 2020; Ryu & Zimolzak, 2020; Luther *et al.*, 2017), les expressions régulières (Wang *et al.*, 2019; Glaser *et al.*, 2018; Atti *et al.*, 2020; Kim *et al.*, 2017), ou encore les approches s'appuyant sur des lexiques (Campillo-Gimenez *et al.*, 2013; Lowe *et al.*, 2009; Evans *et al.*, 2016). Un pic de publications en 2017 portant sur les méthodes linguistiques. L'analyse des publications en question n'a pas permis d'expliquer ce pic.

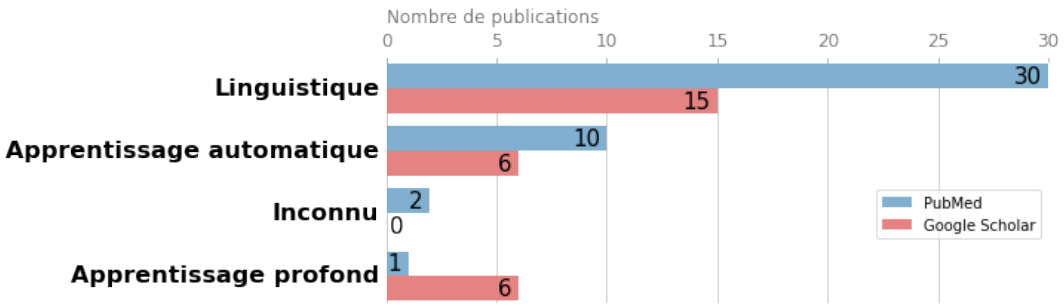

FIGURE 8 – Méthodes TAL présentes dans les publications « Entrepôts de données de santé et TAL »

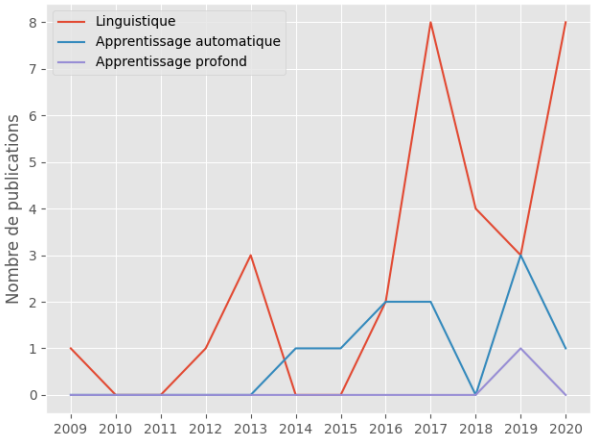

FIGURE 9 – Méthodes TAL par année des publications « Entrepôts de données de santé et TAL » mentionnant une méthode TAL sur PubMed

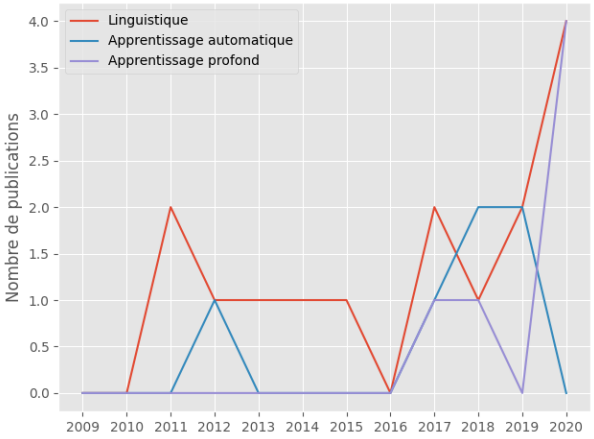

FIGURE 10 – Méthodes TAL par année des publications « Entrepôts de données de santé et TAL » mentionnant une méthode TAL sur Google Scholar

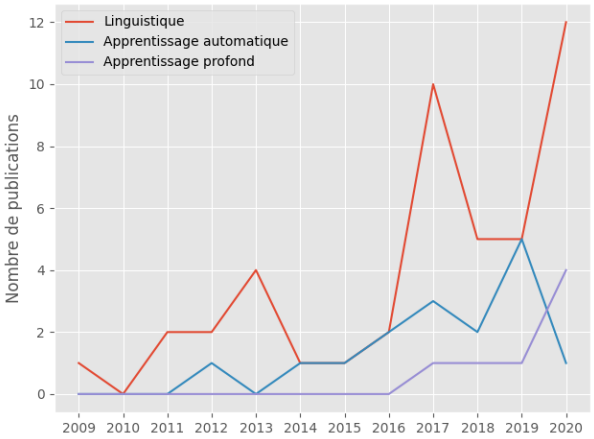

FIGURE 11 – Méthodes TAL par année des publications « Entrepôts de données de santé et TAL » mentionnant une méthode TAL (Google Scholar et PubMed cumulés)

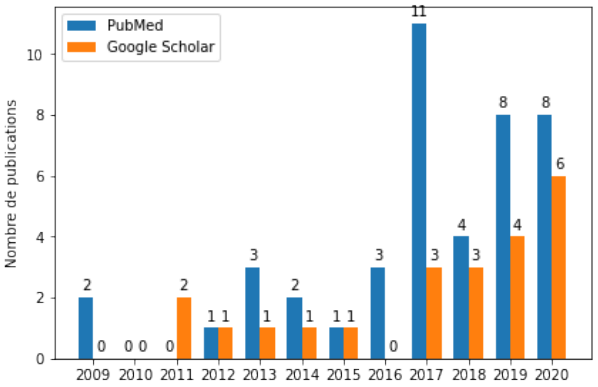

FIGURE 12 – Années de publication des publications « Entrepôts de données de santé et TAL » mentionnant une méthode TAL sur Google Scholar et PubMed

Les précédentes méthodes sont appliquées à des données médicales de différentes langues (cf. figure 13), avec une sur-représentation de la langue anglaise, mais aussi à diverses spécialités médicales (cf. figure 14). L'oncologie est la spécialité la plus traitée, suivie par la cardiologie et la neurologie. La modalité « Autres » rassemblent les spécialités médicales qui correspondent qu'à une seule publication. Parmi ces spécialités médicales, on peut retrouver la génomique, la psychiatrie, la radiologie ou encore l'endocrinologie.

L'objectif médical des études qui appliquent les méthodes de TAL peut être varié. Certaines publications portent sur la médecine interventionnelle, elles cherchent à améliorer les pratiques médicales liées à des actes forts lors de la prise en charge de patients (opérations, traitements, prélèvements biologiques, etc.). D'autres publications s'intéressent à l'étude de maladies ou de pathologies dans leur ensemble et sont classées comme médecine de spécialité. Les publications notées comme non classées traitent globalement de tâches de TAL sur des problématiques autres que le médical. L'aspect

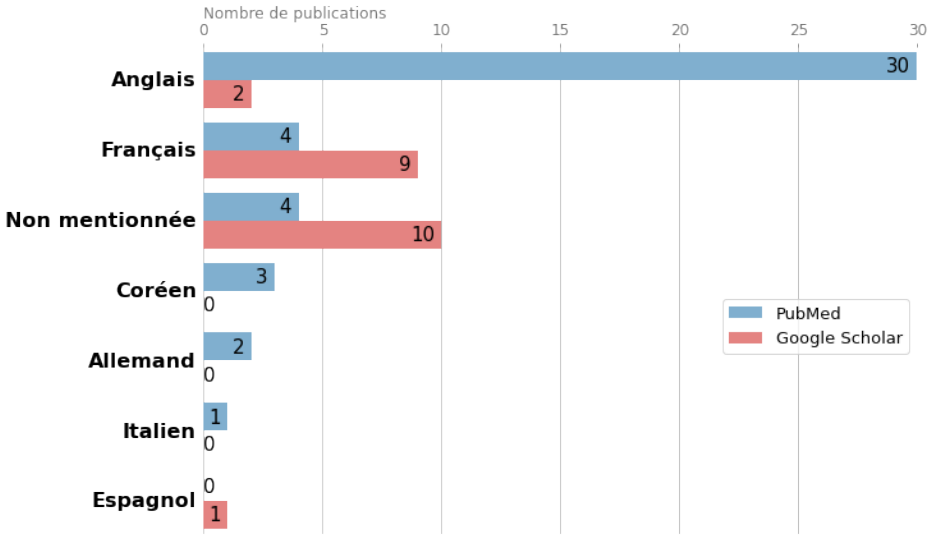

FIGURE 13 – Langue des données exploitées dans les publications « Entrepôts de données de santé et TAL » mentionnant une méthode TAL

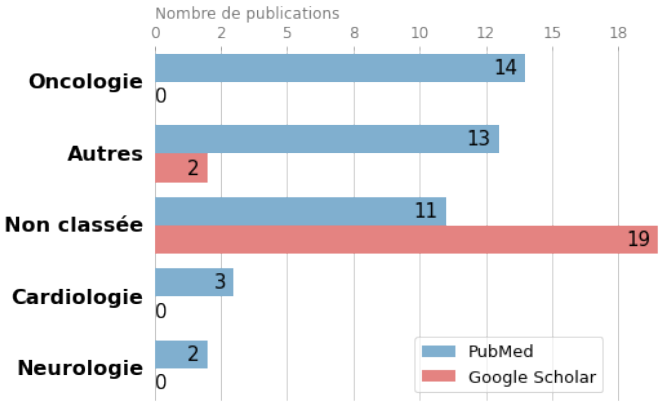

FIGURE 14 – Spécialité médicale des données exploitées dans les publications « Entrepôts de données de santé et TAL » mentionnant une méthode TAL

médical est présent dans ces publications, mais au second plan. C'est le cas pour la majorité des publications analysées qui ont été extraites de Google Scholar.

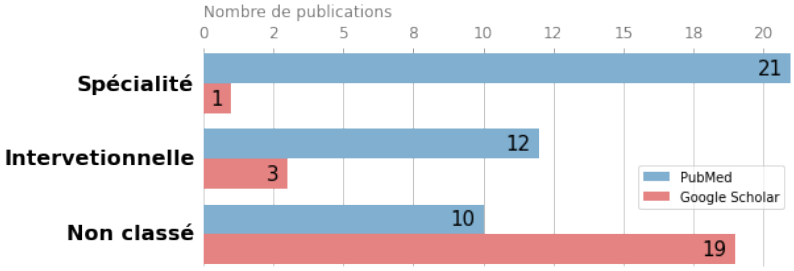

FIGURE 15 – Objectifs médicaux des publications « Entrepôts de données de santé et TAL » mentionnant une méthode TAL

7 Discussion et Conclusion

Cette revue a montré l'intérêt croissant porté aux données de santé informatisées dans la littérature biomédicale et la grande hétérogénéité des abords du TAL dans les publications. Les entrepôts de données sont au cœur de l'exploitation de ces données à des fins de recherche. Le panel de méthodes appliquées aux données textuelles médicales dans la littérature exploite bien le potentiel du traitement automatique de la langue. De plus en plus d'articles sur ces thématiques sont publiés, et ce, dans tous les champs de la santé. Sans surprise, en ayant recours à plusieurs moteur de recherche, nous avons pu également remarquer que PubMed répertorie principalement les publications où l'aspect médical est au premier plan. Les publications où les problématiques sont liés aux méthodes de TAL figurent peu sur PubMed, malgré le contexte médical présent dans ces publications. L'engouement autour du TAL et de la Santé, que l'on retrouve notamment dans le dernier numéro de la revue TAL [9], montre qu'il y a de l'intérêt pour accéder aux connaissances des données médicales, bien que l'accès à ces données soit parfois la première difficulté. Le développement du TAL dans le domaine médicale passera assurément par bien une coopération entre les experts du domaine de la santé et experts du TAL.

Remerciements

Ce travail a reçu le soutien du projet AIBy4 [10]. Nous tenons aussi à remercier les relecteurs anonymes pour leurs conseils avisés sur ce travail.

Références

AFSHAR M., DLIGACH D., SHARMA B., CAI X., BOYDA J., BIRCH S., VALDEZ D., ZELISKO S., JOYCE C., MODAVE F. & PRICE R. (2019). Development and application of a high throughput natural language processing architecture to convert all clinical documents in a clinical data warehouse into standardized medical vocabularies. *Journal of the American Medical Informatics Association*, **26**(11), 1364–1369. DOI : 10.1093/jamia/ocz068.

ATTI M., PECORARO F., PIGA S., LUZI D. & RAPONI M. (2020). Developing a surgical site infection surveillance system based on hospital unstructured clinical notes and text mining. *Surgical Infections*, **21**. DOI : 10.1089/sur.2019.238.

CAMPILLO-GIMENEZ B., GARCELON N., JARNO P., CHAPPLAIN J. & CUGGIA M. (2013). Full-text automated detection of surgical site infections secondary to neurosurgery in Rennes, france. *Studies in health technology and informatics*, **192**, 572–5. DOI : 10.3233/978-1-61499-289-9-572.

CHASE H. S., MITRANI L. R., LU G. G. & FULGIERI D. J. (2017). Early recognition of multiple sclerosis using natural language processing of the electronic health record. *BMC medical informatics and decision making*, **17**(1), 24–24. 28241760[pmid], DOI : 10.1186/s12911-017-0418-4.

CHIUDINELLI L., GABETTA M., CENTORRINO G., VIANI N., TASCA C., ZAMBELLI A., BUCALO M., GHIRARDI A., BARBARINI N., SFREDDO E., TONDINI C., BELLAZZI R. & SACCHI L. (2019).

9. https://www.atala.org/content/traitement-automatique-des-langues-et-santé
10. https://aiby4.ls2n.fr/

Ontology-driven real world evidence extraction from clinical narratives. *Studies in health technology and informatics*, **264**, 1441–1442. DOI : 10.3233/SHTI190474.

DAILLE B. & NAZARENKO A. (2017). Le tournant des données en traitement automatique des langues. . In M. BOUZEGHOUD & R. MOSSERI., Éds., *Les Big Data à découvert*, p. 118–119. CNRS editions. HAL : hal-01693019.

DELAMARRE D., BOUZILLE G., DALLEAU K., COURTEL D. & CUGGIA M. (2015). Semantic integration of medication data into the EHOP clinical data warehouse. *Studies in health technology and informatics*, **210**, 702–6. DOI : 10.3233/978-1-61499-512-8-702.

EVANS R. S., BENUZILLO J., HORNE B., LLOYD J., BRADSHAW A., BUDGE D., RASMUSSON K., ROBERTS C., BUCKWAY J., GEER N., GARRETT T. & LAPPÉ D. (2016). Automated identification and predictive tools to help identify high-risk heart failure patients : Pilot evaluation. *Journal of the American Medical Informatics Association*, **23**, ocv197. DOI : 10.1093/jamia/ocv197.

GLASER A., JORDAN B., COHEN J., DESAI A., SILBERMAN P. & MEEKS J. (2018). Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. *JCO Clinical Cancer Informatics*, **2**, 1–8. DOI : 10.1200/CCI.17.00128.

HE T., PUPPALA M., EZEANA C., HUANG Y.-S., CHOU P.-H., YU X., CHEN S., WWANG L., YIN Z., DANFORTH R., ENSOR J., CHANG J., PATEL T. & WONG S. (2019). A deep learning–based decision support tool for precision risk assessment of breast cancer. *JCO Clinical Cancer Informatics*, **3**, 1–12. DOI : 10.1200/CCI.18.00121.

HERNANDEZ P., PODCHIYSKA T., WEBER S., FERRIS T. & LOWE H. (2009). Automated mapping of pharmacy orders from two electronic health record systems to rxnorm within the stride clinical data warehouse. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, **2009**, 244–8.

KIM Y., YOON D., BYUN J., PARK H., LEE A., KIM I., LEE S., LIM H.-S. & PARK R. W. (2017). Extracting information from free-text electronic patient records to identify practice-based evidence of the performance of coronary stents. *PLOS ONE*, **12**, e0182889. DOI : 10.1371/journal.pone.0182889.

LEE K. H., KIM H. J., KIM Y.-J., KIM J. & SONG E. (2020). Extracting structured genotype information from free-text hla reports using a rule-based approach. *Journal of Korean Medical Science*, **35**. DOI : 10.3346/jkms.2020.35.e78.

LOWE H., HUANG Y. & REGULA D. (2009). Using a statistical natural language parser augmented with the umls specialist lexicon to assign snomed ct codes to anatomic sites and pathologic diagnoses in full text pathology reports. *AMIA Annual Symposium proceedings*, **2009**, 386–90.

LUTHER S. L., THOMASON S. S., SABHARWAL S., FINCH D. K., MCCART J., TOYINBO P., BOUAYAD L., MATHENY M. E., GOBBEL G. T. & POWELL-COPE G. (2017). Leveraging electronic health care record information to measure pressure ulcer risk in veterans with spinal cord injury : A longitudinal study protocol. *JMIR research protocols*, **6**(1), e3–e3. 28104580[pmid], DOI : 10.2196/resprot.5948.

MOORE N., BLIN P., LASSALLE R., THURIN N., BOSCO-LEVY P. & DROZ C. (2021). *National Health Insurance Claims Database in France (SNIRAM), Système Nationale des Données de Santé (SNDS) and Health Data Hub (HDH)*, In M. STURKENBOOM & T. SCHINK, Éds., *Databases for Pharmacoepidemiological Research*, p. 131–140. Springer International Publishing : Cham. DOI : 10.1007/978-3-030-51455-6_10.

NEURAZ A., LERNER I., DIGAN W., PARIS N., TSOPRA R., ROGIER A., BAUDOIN D., COHEN K., BURGUN A., GARCELON N. & RANCE B. (2020). Natural language processing for rapid response to emergent diseases : Case study of calcium channel blockers and hypertension in the covid-19 pandemic. *Journal of Medical Internet Research*, **22**, e20773. DOI : 10.2196/20773.

NÉVÉOL A., ZWEIGENBAUM P. & ON CLINICAL NATURAL LANGUAGE PROCESSING S. E. F. T. I. Y. S. (2018). Expanding the diversity of texts and applications : Findings from the section on clinical natural language processing of the international medical informatics association yearbook. *Yearbook of medical informatics*, **27**(1), 193–198. 30157523[pmid], DOI : 10.1055/s-0038-1667080.

OSBORNE J., WYATT M., WESTFALL A., WILLIG J., BETHARD S. & GORDON G. (2016). Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. *Journal of the American Medical Informatics Association*, **23**, ocw006. DOI : 10.1093/jamia/ocw006.

QUÉROUÉ M., LASHÉRAS-BAUDUIN A., JOUHET V., THIESSARD F., VITAL J.-M., ROGUES A.-M. & COSSIN S. (2019). Automatic detection of surgical site infections from a clinical data warehouse.

RYU J. H. & ZIMOLZAK A. J. (2020). Natural language processing of serum protein electrophoresis reports in the veterans affairs health care system. *JCO Clinical Cancer Informatics*, (4), 749–756. PMID : 32813561, DOI : 10.1200/CCI.19.00167.

SAFRAN C. (2017). Update on data reuse in health care. *Yearbook of medical informatics*, **26**(1), 24–27. 29063535[pmid].

SCHOENTHALER M., BOEKER M. & HORKI P. (2019). How to compete with google and co. : big data and artificial intelligence in stones. *Current Opinion in Urology*, **29**(2).

SHEIKHALISHAHI S., MIOTTO R., DUDLEY J. T., LAVELLI A., RINALDI F. & OSMANI V. (2019). Natural language processing of clinical notes on chronic diseases : Systematic review. *JMIR medical informatics*, **7**(2), e12239–e12239. 31066697[pmid], DOI : 10.2196/12239.

SINGH S. (2019). *Big Data Meets Real World! The Use of Clinical Informatics in Biomarker Research*, p. 345–352. DOI : 10.1007/978-3-030-11446-6_29.

THORODDSEN A., GUÐJÓNSDÓTTIR H. & GUDMUNDSDOTTIR E. (2017). From capturing nursing knowledge to retrieval of data from a data warehouse.

UPADHYAYA S. G., MURPHREE JR. D. H., NGUFOR C. G., KNIGHT A. M., CRONK D. J., CIMA R. R., CURRY T. B., PATHAK J., CARTER R. E. & KOR D. J. (2017). Automated diabetes case identification using electronic health record data at a tertiary care facility. *Mayo Clinic proceedings. Innovations, quality & outcomes*, **1**(1), 100–110. 30225406[pmid], DOI : 10.1016/j.mayocpiqo.2017.04.005.

WANG Y., MEHRABI S., SOHN S., ATKINSON E., AMIN S. & LIU H. (2019). Natural language processing of radiology reports for identification of skeletal site-specific fractures. *BMC Medical Informatics and Decision Making*, **19**, 73. DOI : 10.1186/s12911-019-0780-5.

ZHAO S.-C., WEI R., XIE Y.-M., WANG L.-X., WANG Q. & YI D.-H. (2019). Analysis of qingkailing injection in treatment of combined medication features of 2 147 cases of upper respiratory tract infection. *Zhongguo Zhong yao za zhi = Zhongguo zhongyao zazhi = China journal of Chinese materia medica*, **44**, 5207–5216. DOI : 10.19540/j.cnki.cjcmm.20191115.501.