SIMI v3
Une liste de cas patients similaires pour la télé expertise médicale

Pierre Jourlin[1] Marc-Antoine Sulmon[2] David Bensoussan[2,3] Émilie Mercadal[2]
(1) LIA, Avignon Université, 339, chemin des Meinajariès, 84 911,Avignon, France
(2) ROFIM, 22 cours Puget, 13 006, Marseille
(3) Chirurgie vasculaire, Centre Hospitalier du Pays d'Aix
Pierre.Jourlin@univ-avignon.fr,
{Marc-Antoine, David.Bensoussan, Emilie.Mercadal}@rofim.doctor

RÉSUMÉ

Cet article présente SIMI v3, une brique logicielle hybridant deux approches d'IA, l'une symbolique et l'autre connexionniste intégrée dans la plateforme web ROFIM, une solution de télé-expertise, e−RCP et téléconsultation médicale. Lors d'une télé-expertise, SIMI v3 permet de rechercher automatiquement des cas patients issus de la littérature scientifique, similaires à celui décrit par le requérant. Une fois cette recherche documentaire accomplie, il propose au médecin requis de les consulter avant de produire son expertise. Ce logiciel, dont les aspect fondamentaux ont été développés au Laboratoire d'Informatique d'Avignon et qui a fait l'objet d'un programme de transfert technologique soutenu par la SATT Sud-Est est aujourd'hui en phase de déploiement sur la plateforme. Nous espérons qu'il permette en définitive de réduire l'errance diagnostique, de raccourcir les échanges entre médecin requérant et médecin requis et d'alerter ce dernier sur la possible existence de maladies rares dont les symptômes pourraient être confondus avec ceux de pathologies plus courantes.

ABSTRACT

SIMI v3 : A list of similar patient cases for medical tele-expertise

This article presents SIMI v3, a software brick that combines two AI approaches, one symbolic and the other connectionist, integrated into the ROFIM web platform, a tele-expertise, e-RCP and tele-consultation solution. During a tele-expertise, SIMI v3 automatically searches for patient cases from the scientific literature that are similar to the one described by the requester. Once this documentary research has been completed, it suggests that the requested doctor consult them before producing his expertise. This software, whose fundamental aspects were developed at the Avignon Computer Science Laboratory and which was the subject of a technology transfer program supported by SATT Sud-Est, is now in the process of being deployed on the platform. We hope that it will ultimately reduce diagnostic errors, shorten exchanges between the requesting doctor and the requested doctor and alert the latter to the possible existence of rare diseases whose symptoms could be confused with those of more common pathologies.

MOTS-CLÉS : Extraction d'information ; Fouille de texte ; Recherche d'information ; Classification de documents médicaux ; Approches hybrides en TALN ; Textes biomédicaux.

KEYWORDS: Information extraction ; Text mining ; Information retrieval ; Medical document clustering ; NLP hybrid approach ; Biomedical texts.

1 Introduction

ROFIM est une entreprise fondée en 2018 qui développe et commercialise une plateforme web destinée aux professionnels de santé pour la pratique de la téléconsultation, de la téléexpertise, et la tenue de réunions de concertation pluridisciplinaires en ligne. À la date de soumission de cet article, la plateforme était utilisée par 65000 professionnels de santé répartis sur 1500 établissements, donnant lieu à la réalisation de plus d'$1,5$ millions d'actes médicaux.

Lors d'éditions antérieures de la conférence TALN, nous avons présenté sous forme de démonstration la première version de SIMI (Jourlin, 2022), conçue à partir d'un outil d'annotation sémantique développé l'année précédente (Murata *et al.*, 2021). Lors d'une télé-expertise, SIMI permet de rechercher automatiquement des cas patients issus de la littérature scientifique, similaires à celui décrit par le requérant. Une fois cette recherche documentaire accomplie, il propose au médecin requis de les consulter avant de produire son expertise. L'objectif recherché est une amélioration du diagnostic et une réduction de l'errance médicale, particulièrement importante et impactante dans le cas des maladies rares.

Ces travaux ont donné lieu à la signature d'une convention de concours scientifique [1] entre Avignon Université et ROFIM qui a permis de poursuivre ces recherches ainsi que d'encadrer l'effort d'ingénierie logicielle indispensable à une mise en production sur la plateforme ROFIM. Dans cet article, nous présenterons le contexte de ce développement sous l'angle de la médecine, du TALN et de l'informatique industrielle au sein de cette jeune entreprise. Nous mettrons ensuite en lumière les différentes difficultés rencontrées lors des phases d'intégration de SIMI, ainsi que les stratégies que nous avons déployées pour les dépasser.

2 Contexte médical et économique

L'errance diagnostique (Valarmathi, 2021) et le défi spécifique induit par les maladies rares constituent des problématiques majeures au sein des systèmes de santé, particulièrement en France. Elle est définie par une période prolongée durant laquelle un patient, malgré la présence de symptômes, ne parvient pas à obtenir un diagnostic pertinent. Ce parcours, souvent semé d'embûches, est marqué par des consultations multiples, des examens répétés et un sentiment d'incertitude croissant pour le patient. Ces situations engendrent non seulement une souffrance considérable pour les personnes affectées et leurs familles, mais de plus, elles imposent un fardeau économique substantiel aux structures de soins.

Le cas des patients atteints de maladies rares est particulièrement critique. Ces maladies qui, par définition, touchent 1 personne sur 2000, affectent 350 millions de personnes à travers le monde, dont 30 millions en Europe et 3 millions en France [2]. Or, le délai moyen pour obtenir un diagnostic pour les patients européens atteints de maladies rares est de 5 ans et 70% des personnes atteintes de maladies rares attendent plus d'un an pour obtenir un diagnostic confirmé après avoir consulté un médecin (Faye *et al.*, 2024)

Selon un rapport de la H.A.S. (Haute Autorité de Santé, 2024b) s'appuyant sur plusieurs publications

1. Article 25-2 de la loi du 12 juillet 1999
2. https://www.maladiesraresinfo.org/informer/information-generale-sur-les-maladies-les-chiffres-clefs-des-maladies-rares.html consulté le 13 mai 2025

scientifiques, aux États-Unis, les erreurs diagnostiques seraient à l'origine de 25% des décès dus à des erreurs médicales, celles-ci étant la 3e cause de mortalité dans le pays. Le coût annuel des erreurs diagnostiques y est estimé à plus de 100 milliards de dollars. En comparaison, le coût des erreurs thérapeutiques n'y est estimé qu'à 20 milliards de dollars. Le rapport cite également une étude américaine récente établissant que le diagnostic retardé d'appendicite représente 2,7% des cas annuels et que sa réduction pourrait permettre d'économiser plus de 21 millions de dollars par an.

3 Contexte technologique

3.1 IA et aide au diagnostic

Les grands modèles de langage (LLM) sont de plus en plus explorés pour leur capacité à assister les cliniciens dans la prise de décision et l'établissement de diagnostics. Plusieurs études ont évalué les performances des LLMs pour cette tâche et pour diverses affections, telles que les maladies gastro-intestinales et les troubles neurologiques. Bien que certaines expérimentations aient montré des résultats prometteurs, elles soulignent également la nécessité de poursuivre la recherche et de développer des critères d'évaluation standardisés (Yang *et al.*, 2025).

Cependant, les avancées majeures de la recherche en intelligence artificielle (IA) pour la santé, le déploiement et l'adoption des technologies d'IA restent limitées dans la pratique clinique. (Lekadir *et al.*, 2025) établissent ainsi des recommandations basées sur 6 principes directeurs : l'équité, l'universalité, la traçabilité, la facilité d'utilisation, la robustesse et l'explicabilité. Ils définissent également un ensemble de 30 bonnes pratiques, abordant les dimensions techniques, cliniques, socio éthiques et juridiques. Ces systèmes prédictifs sont donc considérés aujourd'hui comme des outils complémentaires et non comme des substituts aux cliniciens humains, dont l'expertise et le jugement clinique restent essentiels.

C'est la raison principale pour laquelle nous avons choisi de conserver pour l'instant un cadre strict de suggestion documentaire, auxquelles s'ajoutent évidemment des considérations d'ordre réglementaire (Solaiman & Malik, 2024) et économique, voire écologiques, induites par l'utilisation des IA dites « génératives » (Dhar, 2020).

3.2 IA et recherche de cas patients similaires

L'étude de la similarité entre patients et la conception de systèmes informatiques permettant d'automatiser cette recherche peuvent prendre en compte des données hétérogènes (métadonnées, images, textes, bilans biomédicaux, etc.) et dépassent le seul cadre du traitement automatique des langues. L'atelier de l'AMIA (Seligson *et al.*, 2020), a permis d'établir un cadre commun pour la définition et la communication autour de la similarité des patients dans le contexte de la médecine de précision. Les participants, issus de divers horizons (académique, industrie, régulateurs, pratique clinique), ont convergé vers une classification en quatre classes principales de similarité des patients :

- Similarité par caractéristiques : basée sur des propriétés statiques ou dynamiques du patient, comme le type de maladie, le stade, ou des antécédents médicaux.
- Similarité par résultats : basée sur des mesures de l'évolution ou de la réponse au traitement.
- Similarité par exposition : basée sur les interventions ou expositions environnementales,

souvent temporelles.
— Similarité mixte : interaction de plusieurs classes, intégrant par exemple des caractéristiques de base, des expositions et des résultats.

Ces problématiques ont donné récemment lieu à de riches échanges au sein de notre communauté TAL (Coulet *et al.*, 2023). Les actes de cette journée d'étude font toutefois apparaître la rareté des corpus et des bancs d'évaluation librement accessibles. Les plus cités sont MIMIC-III (Johnson *et al.*, 2016) et ses versions ultérieures (Johnson *et al.*, 2024) qui sont en anglais et spécifiques au domaine du soin intensif. Pour ce qui est du français, les jeux de données pertinents et ouverts sont encore plus limités, parfois même restreints aux seuls usages académiques : Par exemple, (Grabar *et al.*, 2018), propose un corpus de cas patients de 39700 occurrences de mots et (Hiebel *et al.*, 2022) offre un jeu 1000 paires de phrases annotées manuellement en scores de similarité.

Si ces jeux de données de référence peuvent se révéler très utiles pour les phases de test dans les toutes premières étapes du développement, les données traitées dans l'environnement de production sont souvent de bien moins bonne qualité du point de vue typographique, lexical et sont la plupart du temps simplement inaccessibles en raison de la réglementation sur les données personnelles de santé [3]. L'effort financier nécessaire à l'anonymisation de ces données reste, en l'absence de système automatique suffisamment fiable, un frein très important pour le développement de tels jeux d'évaluation. On peut noter des tentatives de produire des jeux de données annotés automatiquement à partir des graphes de citation de cas patients publiés dans des revues en accès libre et disponibles sur Pubmed (Zhao *et al.*, 2023). Cependant, les meilleurs résultats obtenus par une recherche de cas patients similaires dans ce cadre restent très faibles (précision de 6.51% et rappel de 46.23%), ce qui suggère un niveau de corrélation entre proximité relative aux citations et proximité du point de vue médical relativement faible.

3.3 Transposition en recherche documentaire

C'est pourquoi nous avons projeté la problématique de la recherche de similarité de cas patients sur l'axe de la recherche documentaire classique dans laquelle une première phase consiste à transformer un cas patient en *requête*. Cela ne change en rien la méthodologie et les problématiques d'évaluation, mais cela permet de séparer le cas patient à traiter, protégé par certification HDS [4], des cas patients publiés dans la littérature scientifique et qui sont eux accessibles aux équipes de recherche et développement.

Cependant, la recherche documentaire médicale traditionnelle (Haute Autorité de Santé, 2024a) repose souvent sur l'utilisation de mots-clés et de vocabulaires contrôlés tels que le MeSH (Medical Subject Headings), voir (Kim *et al.*, 2016). Bien que ces méthodes se montrent très efficaces dans certains cas d'usages, elles présentent des limitations importantes dans le cadre d'une recherche de similarité. Il est par exemple évident que la recherche par mots-clés peut passer à côté d'informations pertinentes en raison de variations dans la terminologie utilisée par les auteurs ou de l'incapacité à saisir les relations sémantiques nuancées entre les concepts (Arsenault, 2006). À titre d'exemple simple, les termes « présence de », « absence de », « antécédent de », « pas d'antécédent de » peuvent modifier profondément le sens du concept qui suit dans la description clinique et donc altérer radicalement

3. Journal officiel de l'Union européenne, L 119, 4 mai 2016
4. La certification HDS (Hébergeur de Données de Santé) est une norme instaurée par le décret n°2018-137, visant à garantir la protection et la sécurité des données de santé à caractère personnel hébergées par divers professionnels et établissements de santé.

l'opérationnalité de la distance de similarité entre deux cas patients.

C'est pourquoi nous avions dans un premier temps développé le système SIMI qui, en prenant en compte le contexte linguistique des mentions de concepts médicaux, permettait d'extraire dans une description en langue naturelle, les entités sémantiques répertoriées dans le metathesaurus UMLS (NIH, 2009). Ce système présentait des atouts majeurs en ce qui concerne la frugalité, l'explicabilité, la précision, l'interopérabilité. Il reste très performant pour la classification et l'indexation sémantique des cas patients. Nous avons en revanche observé qu'il était faiblement adapté aux échanges textuels entre des médecins requis et requérants qui sont en permanence soumis à d'importantes contraintes d'efficacité. Ces contraintes peuvent en effet induire une forte proportion de fautes de frappe, d'écriture abrégée, de notions évoquées implicitement, etc. Or, la nécessaire prise en compte de ces variations dans les lexiques et règles grammaticales de SIMI engendrait des coûts de développement importants, induisant un report significatif de la date de mise en production.

4 SIMI v3

Le schéma en figure 1 résume l'architecture logicielle de SIMI v3. Nous allons détailler chaque composante dans les sections suivantes.

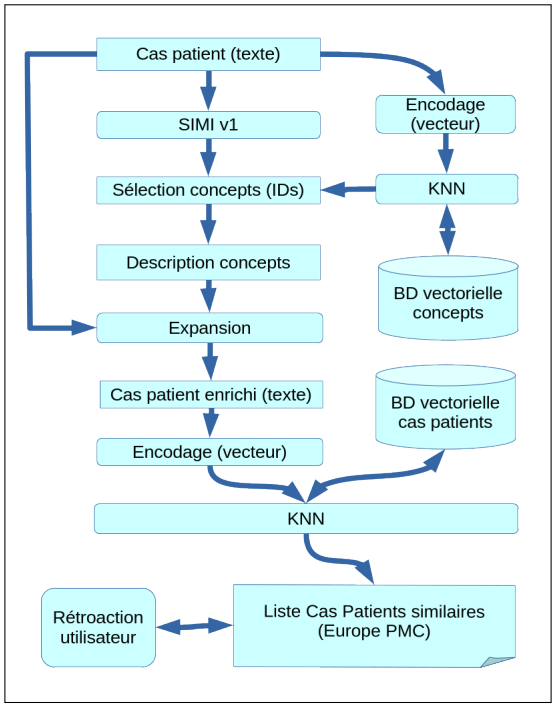

FIGURE 1 – Architecture logicielle de SIMI v3

4.1 Approche symbolique

SIMI v1 (Jourlin, 2022) permettait de représenter chaque cas patient comme un ensemble d'entités conceptuelles de l'ontologie UMLS en automatisant l'extraction, la désambiguïsation et la traduction en anglais des termes médicaux présents dans les cas patients. SIMI est fondé sur une technologie propriétaire qui s'appuie sur des règles syntaxico-lexicales construites semi-automatiquement. La recherche documentaire pouvait donc facilement être mise en œuvre dans des systèmes classiques et frugaux en ressources matérielles : par exemple, de type vectoriel, au sens de (Salton, 1975), ou probabiliste, au sens de (Spärck Jones *et al.*, 1998).

Comme c'est souvent le cas avec les approches formelles et symboliques, ce système permettait d'obtenir une excellente précision, mais un rappel trop insatisfaisant. Nous l'avons néanmoins conservé, mais cette fois en tant que sous-système permettant de suggérer un ensemble de concepts médicaux à associer au cas patient visé par la télé expertise.

4.2 Approche connexionniste

SIMI v3 est quant à lui, fondé sur une représentation du cas patient sous forme d'un plongement lexical, produit à partir de la description textuelle en langue naturelle, par un modèle pré entraîné et affiné sur des textes médicaux multilingues. Nous avons, très empiriquement, testé plusieurs modèles à poids ouverts, disponibles sur le site HuggingFace et avons sélectionné « PubMedBERT Embeddings Matryoshka » [5] avec une dimension de 768 composantes. Ce modèle s'appuie sur les techniques décrites dans (Kusupati *et al.*, 2022) et obtient un score moyen de 95.73% sur 3 bancs d'évalution dans le domaine médical : *PubMed QA*, *PubMed Subset* et *PubMed Summary*.

Il existe de nombreux algorithmes permettant de classifier un ensemble de vecteurs en classes de similarité (arbres de décision, forêts aléatoires, machines à vecteurs de support (SVM), réseaux neuronaux, etc). Cependant, comme nous l'avons vu en section 3.2, il est assez difficile de trouver un jeu de données annoté suffisamment large pour pouvoir entraîner et évaluer des approches fondées sur un apprentissage supervisé. D'autres part, d'un point de vue industriel, nous sommes contraints par les infrastructures matérielles et logicielles de notre hébergeur ainsi que par les questions de coût.

Nous avons ainsi opté pour une recherche documentaire finale fondée sur un algorithme non-supervisé de recherche des K plus proches voisins (KNN) et plus spécifiquement l'algorithme HNSWG (Malkov & Yashunin, 2018). L'efficience étant une condition déterminante en termes d'ergonomie, de coûts variables en production ainsi que du point de vue des émissions de gaz à effet de serre, nous continuons d'étudier d'autres algorithmes KNN présentés comme encore plus efficients dans la littérature récente, par ex. : (Palpanas, 2025).

La base documentaire est constituée par tous les articles complets du corpus Europe PMC (Rosonovski *et al.*, 2023) qui décrivent des cas patients (nommés ici *Case Reports*). Le texte de chaque article est transformé en plongement lexical en utilisant les mêmes paramètres que ceux utilisés pour le cas patient qui fait l'objet de la télé expertise.

5. https://huggingface.co/NeuML/pubmedbert-base-embeddings-matryoshka

4.3 Hybridation et interaction

Une phase intermédiaire donne la possibilité au médecin requis de sélectionner des concepts UMLS pertinents au regard du cas patient, et proposés soit par SIMI v1, soit par SIMI v3. Dans le cas de SIMI v3, la technique d'extraction consiste à représenter chaque concept de l'ontologie par sa liste de synonymes, de traductions ainsi que ses définitions dans toutes les langues disponibles. Le texte ainsi collecté pour chaque concept est transformé en plongement lexical et injecté dans une base vectorielle. Les concepts suggérés pour l'extension du cas patient sont simplement les K plus proches voisins du vecteur obtenu pour le cas patient visé. Une fois les concepts sélectionnés par le médecin requis, leurs différentes variantes lexicales (synonymes, traductions, définitions) sont ajoutées au cas patient avant transformation en plongement lexical.

Cette *extension de requête* semi-automatique, optionnelle, a permis à la fois d'augmenter le rappel et la précision dans des tests menés dans l'environnement de développement. Au moment de la rédaction de cet article, nous n'avons malheureusement pas encore pu effectuer de test utilisateur dans aucun des autres environnements de (pré)production.

Ce retour d'utilisateurs sera en effet déterminant dans le choix de proposer ou non cette interaction. Dans la négative, nous pourrons initier une nouvelle réflexion pour tenter rendre cette phase entièrement automatique en explorant différentes techniques d'expansion de requêtes *en aveugle* comme celles décrites dans (Jourlin *et al.*, 2000). Cette expansion pourra bien évidemment être combinée à d'autres types d'interactions, par exemple avec une prise en compte des jugements de pertinence fournis par le médecin requis, afin filtrer et affiner la liste des cas patients similaires.

L'intégration de la version actuelle de SIMI v3 dans la plateforme ROFIM est conçue pour être d'un accès simple et intuitif, comme le montre la figure 2. Dans ce cas d'utilisation, les résultats de la recherche documentaire doivent apparaître en *temps réel*[6], le médecin requis voit une fenêtre modale s'ouvrir et qui présente sous une forme synthétique (titre, auteur, revue, année), une liste d'entrées bibliographiques pertinentes et un lien pour consulter l'article complet.

5 Conclusion et Perspectives

Dans cet article, nous avons cherché à présenter dans un cas très concret, comment des recherches académiques peuvent aujourd'hui se transformer en innovations dans des logiciels « métier ». Le monde de la recherche en TAL et le monde de l'informatique industrielle possèdent leurs propres lots de contraintes, de difficultés et peuvent même assez souvent poursuivre des objectifs à court terme très différents. Le croisement de ces deux mondes est loin d'aller de soi et il est même assez évident qu'il peut engendrer de nouvelles difficultés à surmonter, tant du point de vue du personnel de recherche que de celui des acteurs du développement logiciel.

Toutefois, nous jugeons qu'au regard de l'impact positif que ces innovations pourraient avoir sur la vie de centaines de millions de patients dans le monde, affectés par l'errance médicale, c'est un jeu certes complexe et difficile, mais qui en vaut très largement la chandelle. À la date de rédaction de cet article et au stade actuel du développement de SIMI v3, nous ne pouvons malheureusement pas encore présenter d'indicateurs objectifs permettant d'estimer précisément cet impact sur les patients. Notre approche du croisement entre recherche académique et informatique industrielle reste

6. typiquement moins de 2 secondes

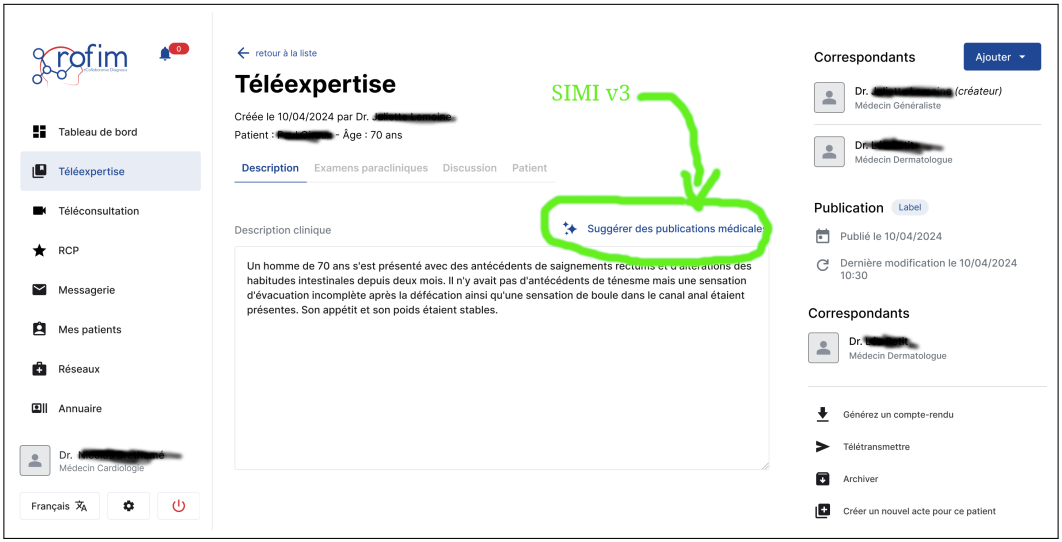

FIGURE 2 – Intégration SIMI v3 dans ROFIM

encore très empirique et pragmatique, mais nous espérons que cet article contribuera à la production d'échanges fertiles au sein de la communauté de recherche et développement en TAL et en définitive à la conception et à l'amélioration de nouveaux dispositifs pour aider le personnel médical à accomplir au mieux ses missions.

Références

ARSENAULT C. (2006). L'utilisation des langages documentaires pour la recherche d'information. *Documentation et bibliothèques*, **52**(2), 139–148. DOI : https ://doi.org/10.7202/1030017ar.

COULET A., GÉRARDIN C., NÉVÉOL A. & TANNIER X., Éds. (2023). *Actes de la journée d'étude sur la Similarité entre Patients, SimPa 2023*. HAL : hal-04080808.

DHAR P. (2020). The carbon impact of artificial intelligence. *Nature Machine Intelligence*, **2**(8), 423–425. DOI : 10.1038/s42256-020-0219-9.

FAYE F., CROCIONE C., ANIDO DE PEÑA R., BELLAGAMBI S., ESCATI PEÑALOZA L., HUNTER A., JENSEN L., OOSTERWIJK C., SCHOETERS E., DE VICENTE D., FAIVRE L., WILBUR M., LE CAM Y. & DUBIEF J. (2024). Time to diagnosis and determinants of diagnostic delays of people living with a rare disease : results of a rare barometer retrospective patient survey. *European Journal of Human Genetics*, **32**(9), 1116–1126. DOI : 10.1038/s41431-024-01604-z.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). Cas : French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, p. 122–128 : Association for Computational Linguistics. DOI : 10.18653/v1/w18-5614.

HAUTE AUTORITÉ DE SANTÉ (2024a). *Guide méthodologique de recherche documentaire*. Saint-Denis La Plaine : HAS.

HAUTE AUTORITÉ DE SANTÉ (2024b). *Les erreurs diagnostiques en médecine*. Haute Autorité de Santé.

HIEBEL N., FORT K., NÉVÉOL A. & FERRET O. (2022). CLISTER : Un corpus pour la similarité sémantique textuelle dans des cas cliniques en français (CLISTER : A corpus for semantic textual similarity in French clinical narratives). In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. ZANON BOITO, Éds., *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 287–296, Avignon, France : ATALA.

JOHNSON A., BULGARELLI L., POLLARD T., GOW B., MOODY B., HORNG S., CELI L. A. & MARK R. (2024). Mimic-iv. DOI : 10.13026/KPB9-MT58.

JOHNSON A. E., POLLARD T. J., SHEN L., LEHMAN L.-W., FENG M., GHASSEMI M. M., MOODY B. E., SZOLOVITS P., CELI L. A. G. & MARK R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, **3**, 160035. DOI : 10.1038/sdata.2016.35.

JOURLIN P. (2022). SIMI : un système de suggestion de littérature médicale. In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. ZANON BOITO, Éds., *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 3 : Démonstrations*, p. 9–11, Avignon, France : ATALA. HAL : hal-03704019.

JOURLIN P., JOHNSON S. E., SPÄRCK JONES K. & WOODLAND P. C. (2000). Spoken document representations for probabilistic retrieval. *Speech Communication*, **32**(1), 21–36. Accessing Information in Spoken Audio, DOI : https ://doi.org/10.1016/S0167-6393(00)00021-2.

KIM S., YEGANOVA L. & WILBUR W. J. (2016). Meshable : searching pubmed abstracts by utilizing mesh and mesh-derived topical terms. *Bioinformatics*, **32**(19), 3044–3046. DOI : 10.1093/bioinformatics/btw331.

KUSUPATI A., BHATT G., REGE A., WALLINGFORD M., SINHA A., RAMANUJAN V., HOWARD-SNYDER W., CHEN K., KAKADE S., JAIN P. & FARHADI A. (2022). Matryoshka representation learning. In S. KOYEJO, S. MOHAMED, A. AGARWAL, D. BELGRAVE, K. CHO & A. OH, Éds., *Advances in Neural Information Processing Systems*, volume 35, p. 30233–30249 : Curran Associates, Inc.

LEKADIR K., FRANGI A. F., PORRAS A. R., GLOCKER B., CINTAS C., LANGLOTZ C. P., WEICKEN E., ASSELBERGS F. W., PRIOR F., COLLINS G. S., KAISSIS G., TSAKOU G., BUVAT I., KALPATHY-CRAMER J., MONGAN J., SCHNABEL J. A., KUSHIBAR K., RIKLUND K., MARIAS K., AMUGONGO L. M., FROMONT L. A., MAIER-HEIN L., CERDÁ-ALBERICH L., MARTÍ-BONMATÍ L., CARDOSO M. J., BOBOWICZ M., SHABANI M., TSIKNAKIS M., ZULUAGA M. A., FRITZSCHE M.-C., CAMACHO M., LINGURARU M. G., WENZEL M., DE BRUIJNE M., TOLSGAARD M. G., GOISAUF M., CANO ABADÍA M., PAPANIKOLAOU N., LAZRAK N., PUJOL O., OSUALA R., NAPEL S., COLANTONIO S., JOSHI S., KLEIN S., AUSSÓ S., ROGERS W. A., SALAHUDDIN Z. & STARMANS M. P. A. (2025). Future-ai : international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, **388**. DOI : 10.1136/bmj-2024-081554.

MALKOV Y. A. & YASHUNIN D. A. (2018). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, **42**(4), 824–836.

MURATA J., CARRETTE R. & JOURLIN P. (2021). SIDRES : A Novel Annotation Tool For The Automatic Detection of Semantic Entities. In P. DENIS, N. GRABAR, A. FRAISSE, R. CARDON, B. JACQUEMIN, E. KERGOSIEN & A. BALVET, Éds., *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 3 : Démonstrations*, p. 15–17, Lille, France : ATALA. HAL : hal-03265913.

NIH (2009). *UMLS® Reference Manual*. National Library of Medicine (US), public domain.

PALPANAS T. (2025). Scalable vector analytics : A story of twists and turns. *Revue des Nouvelles Technologies de l'Information*, **Extraction et Gestion des Connaissances, RNTI-E-41**, 5–6.

ROSONOVSKI S., LEVCHENKO M., BHATNAGAR R., CHANDRASEKARAN U., FAULK L., HASSAN I., JEFFRYES M., MUBASHAR S. I., NASSAR M., JAYAPRABHA PALANISAMY M., PARKIN M., POLURU J., ROGERS F., SAHA S., SELIM M., SHAFIQUE Z., IDE-SMITH M., STEPHENSON D., TIRUNAGARI S., VENKATESAN A., XING L. & HARRISON M. (2023). Europe pmc in 2023. *Nucleic Acids Research*, **52**(D1), D1668–D1676. DOI : 10.1093/nar/gkad1085.

SALTON G. (1975). A vector space model for information retrieval. *Journal of the ASIS*, p. 613–620.

SELIGSON N. D., WARNER J. L., DALTON W. S., MARTIN D., MILLER R. S., PATT D., KEHL K. L., PALCHUK M. B., ALTEROVITZ G., WILEY L. K., HUANG M., SHEN F., WANG Y., NGUYEN K. A., WONG A. F., MERIC-BERNSTAM F., BERNSTAM E. V. & CHEN J. L. (2020). Recommendations for patient similarity classes : results of the amia 2019 workshop on defining patient similarity. *Journal of the American Medical Informatics Association*, **27**(11), 1808–1812. DOI : 10.1093/jamia/ocaa159.

SOLAIMAN B. & MALIK A. (2024). Regulating algorithmic care in the european union : evolving doctor–patient models through the artificial intelligence act (ai-act) and the liability directives. *Medical Law Review*, **33**(1), fwae033. DOI : 10.1093/medlaw/fwae033.

SPÄRCK JONES K., WALKER S. & ROBERTSON S. (1998). *A probabilistic model of information and retrieval : development and status*. Rapport interne, University of Cambridge, Computer Laboratory.

VALARMATHI M. T. (2021). *Rare Diseases*. Rijeka : IntechOpen. DOI : 10.5772/intechopen.92919.

YANG X., LI T., SU Q., LIU Y., KANG C., LYU Y., ZHAO L., NIE Y. & PAN Y. (2025). Application of large language models in disease diagnosis and treatment. *Chinese Medical Journal*, **138**(2). DOI : 10.1097/CM9.0000000000003456.

ZHAO Z., JIN Q., CHEN F., PENG T. & YU S. (2023). A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific Data*, **10**(1). DOI : 10.1038/s41597-023-02814-8.