HAUTE AUTORITÉ DE SANTÉ

MESURER
& AMÉLIORER LA QUALITÉ

RAPPORT

Quel potentiel de l'IA pour le screening d'articles ? Evaluation à partir des bases de revues rétrospectives de la HAS.

Validé par le Collège le 29 janvier 2025

Descriptif de la publication

Titre	Quel potentiel de l'IA pour le screening d'articles ? Evaluation à partir des bases de revues rétrospectives de la HAS.
Méthode de travail	Evaluation de la performance d'un outil d'IA de revue priorisée pour aider la sélection de publications par l'humain lors du processus de revue de littérature
Objectif(s)	Evaluer l'intérêt de façon rétrospective en simulant un screening priorisé sur cinq revues de littérature déjà effectuées par la HAS Estimer le potentiel de quantité de travail évitable par le screening priorisé Identifier les éventuelles limites des bases de données disponibles pour évaluer les performances de l'outil
Cibles concernées	Professionnels, usagers, institutions
Demandeur	Auto-saisine Haute Autorité de Santé
Promoteur(s)	Haute Autorité de santé (HAS)
Pilotage du projet	Pierre-Alain Jachiet, Frédérique Pagès
Recherche documentaire	
Auteurs	Matthieu Doutreligne, Pavel Soriano, Noémie Jacquet, Sophie Despeyroux, Philippe Canet, Frédérique Pagès, Pierre-Alain Jachiet
Conflits d'intérêts	Les membres du groupe de travail ont communiqué leurs déclarations publiques d'intérêts à la HAS. Elles sont consultables sur le site https://dpi.sante.gouv.fr. Elles ont été analysées selon la grille d'analyse du guide des déclarations d'intérêts et de gestion des conflits d'intérêts de la HAS. Les intérêts déclarés par les membres du groupe de travail ont été considérés comme étant compatibles avec leur participation à ce travail.
Validation	Version du 29 janvier 2025
Actualisation	
Autres formats	

Ce document ainsi que sa référence bibliographique sont téléchargeables sur www.has-sante.fr

Haute Autorité de santé – Service communication et information
5 avenue du Stade de France – 93218 SAINT-DENIS LA PLAINE CEDEX. Tél. : +33 (0)1 55 93 70 00

Sommaire

Synthèse

➔ Contexte

La revue systématique de littérature occupe une place importante dans le processus de gestion de la connaissance pour l'élaboration des recommandations et évaluations à la Haute Autorité de Santé (HAS).

L'arrivée des systèmes d'intelligence artificielle générative ouvre des perspectives pour outiller ces procédures chronophages.

Ce rapport constitue une première étape vers la prise en main de ce type d'outils à la HAS.

Nous étudions l'intérêt de l'outil de revue priorisé open source ASReview (Van De Schoot et al. 2021) afin de rendre plus aisée et plus rapide l'étape d'inclusion sur titre/résumé d'une revue. Cet outil s'appuie sur les décisions d'inclusion et d'exclusion d'un évaluateur humain afin de prioriser la lecture des articles les plus pertinents. Le modèle de priorisation est dit actif car il s'améliore au fur et à mesure que l'évaluateur lui fournit des labels.

Nous évaluons cet outil de façon rétrospective, c'est à dire en simulant le screening assisté par attribution de labels d'inclusion grâce aux bases de données de la HAS.

➔ Méthodes

Nous adaptons à cette tâche les bases de données utilisées par le Service Documentation et Veille (SDV) pour cinq revues de littérature déjà menées par la HAS. Pour chaque revue, nous construisons automatiquement un jeu de données labellisées, où nous associons à chaque référence son statut d'inclusion selon son appartenance à la base de recherche (mère), la base d'inclusion sur titre/résumé (fille) ou la base finale (sélective). Nous nous appuyons sur le titre, le résumé et le premier auteur pour faire le lien entre ces bases.

Nous évaluons deux stratégies de screening : est-ce que l'outil est capable de prioriser les références incluses sur titre/résumé ou les références incluses en définitive sur texte complet ? Nous évaluons trois modèles de priorisation de complexité différentes. L'évaluation reporte pour un niveau de rappel demandé (proportion d'articles pertinents inclus par l'outil parmi l'ensemble des articles pertinents), le gain en nombre d'articles priorisés qu'il a fallu lire pour obtenir ce rappel par rapport à un choix aléatoire des articles (Worked Saved over Sampling).

➔ Résultats

Les performances varient beaucoup selon les revues.

Les résultats sont systématiquement meilleurs pour les revues avec un grand nombre de références, et un faible ratio d'inclusion des références. Cela s'explique notamment par le fait que le potentiel d'économie de travail est d'autant plus élevé que le ratio d'inclusion est faible.

Le screening priorisé a un meilleur potentiel de détection des références pertinentes pour la stratégie les identifiant sur texte complet versus la stratégie les identifiant sur titre/résumé. Dans les meilleurs cas, le screening priorisé permet d'identifier 95% des références pertinentes en ayant épargné la revue de 64% (revue TAVI, régression logistique) et 71% (revue TestsKSein, naive Bayes) des références pour la sélection sur texte complet, versus une économie de 50% (TAVI, régression logistique) et 45% (TestsKSein, régression logistique) des références pour la sélection sur titre/résumé.

Une première explication est le plus faible ratio d'inclusion à l'étape de sélection sur texte complet par rapport à la sélection sur titre résumé (moins de références pertinentes à identifier sur texte complet). Ce résultat suggère également que l'outil serait d'autant plus efficace qu'on lui fournit des revues véritablement pertinentes sur texte complet, l'utilisateur donnant un signal très spécifique au modèle. La sélection directe sur texte complet ne correspond pas à la pratique actuelle qui consiste en deux étapes de sélection. Mais le fait d'être exigeant dès la première étape (en examinant le texte complet dès qu'il est accessible et en cas de doute) peut être un moyen d'améliorer la performance de l'outil.

➔ Limites et ouverture

Deux limites apparaissent dans l'évaluation de la performance de l'outil ASReview. Premièrement la création des jeux de données labellisés à partir des bases du SDV introduit des erreurs de labels en raison de changements dans les clés de jointure entre les bases mère et fille. Deuxièmement, les bases bibliographiques utilisées pour le screening rétrospectif regroupent plusieurs questions de recherche, qui, en conditions réelles, conduiraient à des screenings distincts effectués par les chefs de projet.

Le gain de temps potentiel offert par l'outil ASReview devrait être étudié en se dotant d'un critère d'arrêt, déterminant quand l'humain s'arrête de revoir les articles proposés par l'outil, puisqu'il s'agit du premier bénéfice attendu pour un outil d'aide à la sélection d'articles.

Pour un humain qui revoit l'ensemble des articles, un autre bénéfice peut être le confort ressenti dans le processus de sélection, dans la mesure où les articles les plus pertinents sont priorisés par l'outil. Ce bénéfice reste à valider lors d'expérimentations prospectives.

Pour permettre le développement, l'évaluation et l'adoption des outils d'IA pour la revue de littérature, il serait nécessaire d'améliorer le système d'enregistrement de données lors des revues de la HAS. Cela pourrait passer par un outil spécifique à la revue de littérature dont l'utilisation directe par les chefs de projet permettrait de suivre les étapes de la revue, de gérer les doublons en attribuant un identifiant unique aux références, de conserver les références écartées, ainsi que d'enregistrer les critères d'inclusion et d'exclusion pour chaque article. Sous certaines conditions de transparence (notamment ouverture du code source), un tel outil permettrait d'expérimenter en situation prospective le screening priorisé, voire des systèmes d'IA générative.

1. Introduction

1.1. Enjeux et contexte

La revue systématique de littérature occupe une place importante dans le processus de gestion de la connaissance pour l'élaboration des recommandations et évaluations à la Haute Autorité de Santé (HAS).

Les progrès récents en traitement du langage naturel, notamment en intelligence artificielle générative, ouvrent des perspectives intéressantes pour l'automatisation de certaines étapes du processus de revue de littérature. Dans ce contexte, la HAS doit expérimenter ces nouvelles technologies afin d'établir si celles-ci peuvent permettre de faciliter ses processus de gestion de la connaissance. Peut-on assister la réalisation des revues de littérature avec des systèmes informatiques --avec ou sans IA ? Ces outils permettent-ils de gagner en efficacité ? Permettent-ils de gagner en qualité, par exemple pour la réalisation de *living reviews* (Elliott et al. 2014; 2017) ? Quels changements des processus à la HAS leur adoption induiraient-ils ?

1.2. Travaux existants

Les étapes classiques de revue systématique (Higgins et al. 2023), illustrées en Figure 1 sont :

- Définition de la question de recherche, des critères d'inclusion et d'exclusion ;

- Collecte des données. Définition de la stratégie de recherche (quelles bases de données choisir, quels termes de recherche, etc.), collecte et compilation des références ;

- Sélection des articles pertinents. Aussi appelé screening sur titre/résumé (pré-screening) puis sur texte complet ;

- Extraction des données pertinentes ;

- Synthèse des résultats ;

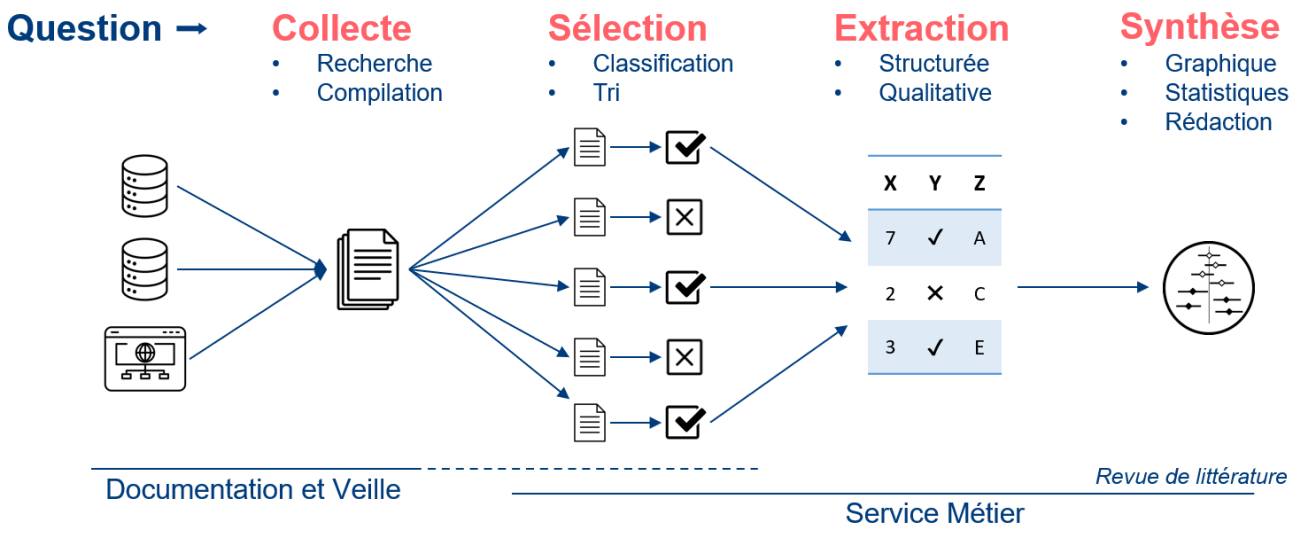

Figure 1 : Les étapes classiques de la revue systématique.

Les étapes de screening et d'extraction des données semblent particulièrement longues, pénibles et propices à une automatisation.

Des études récentes ont mis en évidence le potentiel des grands modèles de langage comme aide à la sélection de publications sur titre/résumé, en leur fournissant des instructions qui contiennent la description de la tâche de classification attendue en « pertinent » / « non pertinent » et les critères de sélection. Les résultats sont prometteurs, mais révèlent certaines limites : par exemple, la sensibilité des résultats aux variations des prompts et la faible spécificité pour un rappel élevé. Il est donc nécessaire de poursuivre les expérimentations avant de pouvoir intégrer ces modèles aux pratiques des services métiers. Nous ne testons pas les grands modèles de langage dans le présent projet mais nous nous concentrons sur un outil d'aide à la sélection de publications basé sur l'apprentissage actif.

Ce projet se focalise sur l'étape du screening des articles via une approche de screening priorisé, assisté par apprentissage actif, approche largement documentée par la littérature. Le screening priorisé consiste à proposer à l'utilisateur les articles les plus pertinents en premier (O'Mara-Eves et al. 2015). En général, ces propositions sont effectuées par un modèle d'apprentissage statistique attribuant à chaque référence une probabilité d'intérêt. L'apprentissage actif consiste à utiliser un certain type de processus d'apprentissage statistique consistant à demander à l'utilisateur de labelliser les références proposées pour améliorer le modèle de prédiction au cours du pré-screening (Settles 2009).

Différents outils permettant d'assister cette étape de screening existent : Covidence (Covidence 2024), EPPI-Reviewer (EPPI 2024), Rayyan (Ouzzani et al. 2016; Rayyan 2022), DistillereSR (DistillerSR 2024), et ASReview (Van De Schoot et al. 2021; ASReview LAB 2024). Ce projet est focalisé sur ASReview, choisi pour son caractère open source. Ce choix est guidé par le besoin de transparence du processus de revue de littérature à la HAS. De plus, ASReview dispose d'une documentation permettant d'extraire les résultats des screenings sous forme facilement exploitable. Cette fonctionnalité est nécessaire pour évaluer les performances de ces outils.

Le process d'apprentissage actif avec l'outil ASReview a déjà été évalué pour détecter en priorité les articles pertinents dans quatre revues de littérature (Van De Schoot et al. 2021). Dans ce travail, les revues considérées portent sur du séquençage génomique de virus pour le bétail, la détection de faute logicielle, le stress post traumatique, l'efficacité d'inhibiteurs enzymatiques. Elles contiennent 2 481, 8 911, 5 782 et 2 544 références avec des ratio respectifs d'inclusions de 4.84 %, 1.2%, 0.66 % et 1.60 %. Enfin, les performances laissent entendre une réduction du temps de screening de 67 % à 92 % pour identifier 95% des références pertinentes. La variabilité de ces résultats motive une évaluation sur les revues menées à la HAS.

L'utilisation d'ASReview par les agents à la HAS pourrait permettre d'apporter un bénéfice en termes de confort lors du processus de sélection, si l'utilisateur revoit l'ensemble des publications priorisées en termes de pertinence. L'outil ASReview pourrait en outre apporter un gain de temps si de futures expérimentations permettent de définir un critère d'arrêt dans le processus de revue des références. Des expérimentations prospectives sont en cours ; elles fourniront des bases de données permettant de tester les performances d'autres outils tels que les grands modèles de langage.

1.3. Processus de revue de littérature à la HAS

Le Service Documentation et Veille (SDV) de la HAS est constitué de 19 personnes, dont 10 documentalistes et 6 assistants documentalistes. En 2022, il a fourni 19 509 articles scientifiques aux services, principalement pour le Service des Bonnes Pratiques (SBP) (45 %), le Service Evaluation de

Santé Publique et Evaluation des Vaccins (SESPEV) (16 %) et le Service Evaluation des Actes Professionnels (SEAP) (15 %).

Pour chaque nouveau projet nécessitant une revue, un binôme documentaliste / assistante documentaliste est formé au sein du SDV. Une réunion de cadrage avec le chef de projet du service métier commanditaire permet de cibler le sujet, la question et l'équation de recherche. Puis, le SDV créé les équations de recherche, effectue la recherche initiale et en communique les résultats au service métier (par mail au format rtf ou excel). Le SDV limite le nombre de résultats à une quantité raisonnable de références, effectuant un compromis entre un objectif d'exhaustivité de la recherche et la charge de travail du service métier (en général, quelques centaines de références). Le chef de projet effectue une sélection des articles pertinents sur titre/résumé, et l'envoie par mail à l'assistante pour obtenir les textes complets. A cette recherche initiale, sont ajoutées en continu tout au long du projet, des références spécifiques identifiées par les chefs de projets et de la littérature grise. Enfin, après lecture des textes complets, la base bibliographique finale est établie. Presque toujours, pour une même publication, plusieurs recherches sont effectuées pour répondre à des questions différentes.

A chacune des trois étapes de la revue, le SDV construit une base EndNote[1] : une base mère, une base fille et un sous-groupe de la base fille appelée base sélective. La base mère contient les résultats d'une recherche bibliographique correspondant à une ou plusieurs équations de recherche. La base fille contient les références incluses en revoyant le titre et le résumé ainsi que de la littérature grise. La base sélective contient les références finalement incluses dans la revue (contenue originellement dans la base mère ou fille).

1.4. Objectifs

Nous évaluons l'outil de revue priorisé open source ASReview (Van De Schoot et al. 2021) de façon rétrospective, c'est à dire en simulant un screening priorisé sur cinq revues de littérature déjà effectuées par la HAS. Nous souhaitons estimer le potentiel de la quantité de travail évitable par screening priorisé. Nous détaillons également les limites et les enjeux d'une automatisation de l'étape de pré-screening (ex : non-exhaustivité de l'identification des références pertinentes, perspectives de living guidelines).

Les axes développés dans ce rapport sont :

- Comment construire des jeux de données labellisés adaptés à la simulation de screening priorisés à partir des bases de données du SDV ? Comment gérer les doublons dans les revues et attribuer les bons labels ? (Partie 2.2)

- Quelle est la variabilité des résultats pour différents choix des premiers articles labellisés et pour différentes revues rétrospectives ? (Partie 3)

- Le screening priorisé fonctionne-t-il mieux en ciblant les articles inclus sur titre/résumé (Partie 3.1) ou bien sur texte complet (Partie 3.2) ?

[1] EndNote est un logiciel de gestion bibliographique, destiné à la gestion des références de livres et travaux de recherche https://fr.wikipedia.org/wiki/EndNote

2. Méthodes

2.1. Données utilisées

Les données utilisées concernent cinq revues de littérature déjà menées par la HAS dans le cadre de cinq publications d'évaluation de technologie de santé. Trois revues proviennent du SEAP, une revue provient du SESPEV, et une du Service Evaluation des Dispositifs (SED). Les identifiants des revues sont de la forme `YY_TYPE_NNN_Titre` où `YY` est l'année de la revue, `TYPE` est le type de revue (TECHNO pour SEAP, ECO_SP pour le SESPEV et DM_Eval pour le SED) et `NNN` est un identifiant unique. Nous listons ces revues avec l'identifiant complet, puis y faisons référence uniquement par leur titre dans le reste du rapport :

- 17_ECO_SP_169_Test_HPV : Cette publication (Haute Autorité de Santé 2019b) évalue la place de la recherche des papillomavirus humains (HPV) et du double immuno-marquage p16/Ki67 dans la stratégie de dépistage du cancer du col de l'utérus. La revue est une revue systématique concentrée sur « l'analyse économique de l'utilisation des tests de dépistage, l'acceptabilité et les préférences pour les différentes modalités de dépistage du CCU (recherche du génome des HPV à haut risque (test HPV), auto-prélèvement pour dépistage par test HPV), les stratégies de dépistage envisagées dans les autres pays en fonction du statut vaccinal des femmes. ». La base mère de la revue contient 2882 références dont 441 retenues sur titre/résumé et 104 retenues finalement. C'est la plus grande des cinq revues évaluées dans cette expérimentation. La publication mentionne 2025 références initiales, 474 lues in extenso (après filtrage sur titre/résumé) et 268 références inclues in fine (ratio 0.04).

- 18_TECHNO_682_Fluoroscopie : Cette publication (Haute Autorité de Santé 2019a) précise les indications actuelles de l'acte fluoroscopie de l'oeil (examen du fond de l'oeil après injection de fluorescéine) chez les patients susceptibles d'être atteints d'une pathologie de la rétine. La recherche documentaire a d'abord ciblé l'acte de fluoroscopie mais s'est heurtée à une absence totale de littérature. La revue a donc porté sur les techniques exploratoires du fond de l'oeil réalisés en pratique courante et les techniques complémentaires de diagnostic et de suivi des pathologies. La base mère de la revue contient 284 références dédoublonnées dont 61 ont été retenues sur titre/résumé et 32 incluses (ratio 0.11). Le texte de la publication mentionne 345 références initiales, donc le même chiffre que les bases SDV si l'on ne dédoublonne pas. Il mentionne 57 références inclues sur titre/résumé et 24 références retenues in-fine selon la publication (cf. Annexe 2 pour une analyse plus fine de ces différences de chiffre). Dans la publication, 33 références écartées sur lecture complète sont mentionnées en annexe de la publication et font donc partie de la base sélective.

- 20_TECHNO_753_Endomicroscopie : Cette publication (Haute Autorité de Santé 2022) évalue le bénéfice/risque de l'endomicroscopie confocale par aiguille de ponction pour la caractérisation des tumeurs kystiques pancréatiques, définit ses conditions de réalisation et rend un avis sur le bien-fondé de l'inscription de cet acte à la Classification Commune des Actes Médicaux. La revue s'est concentrée sur l'apport diagnostique avant la caractérisation de la liaison et après la caractérisation de la lésion obtenu par le matériel de ponction. Des grilles PICOTS ont été utilisée pour chacune de ces deux questions. La base mère contient 339 références dédoublonnées dont 116 retenues sur titre/résumé et 87 incluses (ratio 0.26). Dans la publication, la revue de systématique de littérature mentionne 298 références initiales, 124 lues in extenso (après filtrage sur titre/résumé) et 16 références inclues in fine. L'écart sur les références finalement incluses est expliqué par la présence en annexe de la publication

(et donc dans la base sélective) des références non retenues, fournies pour information (cf. Annexe 2).

- 22_TECHNO_836_TestsKSein : Cette publication (Haute Autorité de Santé 2023) actualise l'évaluation de quatre signatures génomiques dans le cancer du sein au stade précoce, sensible à l'hormonothérapie et de statut HER2 négatif. La revue de littérature s'est focalisée sur les publications évaluant l'utilité clinique des signatures génomiques et sur les études évaluant le niveau de concordance entre les signatures chez une même patiente. La revue s'est appuyée sur un critère PICOTS et la publication mentionne des critères d'éligibilité. La base mère contient 778 références dédoublonnées dont 40 retenues sur titre/résumé et 12 incluses (ratio 0.02). Selon la publication, 520 références ont été incluses initialement, 30 conservées sur titre/résumé et 6 retenues finalement.

- 22_DM_Eval_314_TAVI : Cette publication (Haute Autorité de Santé 2024) revoit les critères d'éligibilité des centres implantant des TAVI. La revue de littérature comprend quatre revues portant chacune sur une question d'intérêt, totalisant 543 références dont 19 retenues selon le texte de la publication. La base mère contient 731 références dédoublonnées dont 81 retenues sur titre/résumé et 21 incluses (ratio 0.03). Dans la base labellisée, il manque trois des 19 références mentionnées dans la publication.

La restitution des méthodologies de revue décrites dans les publications est hétérogène selon les services et les objectifs des différentes revues. Lorsque le nombre de références retenues est mentionné dans la publication (quatre sur cinq), celui-ci est parfois différent du nombre obtenu par intersection de la base mère et de la base fille. En revanche, le nombre de références dans la section bibliographie de la publication correspond systématiquement au nombre de références dans la base sélective.

Il est difficile de relier les publications entre les différentes bases du SDV, notamment lorsqu'il s'agit de publications de société savantes pour lesquelles les auteurs ou le titre changent d'une base SDV à l'autre. De plus, l'ajout en annexe des publications HAS des références écartées sur lecture complète des articles fausse la base sélective qui contient alors toutes les références de la revue de littérature (par exemple pour Endomicroscopie ou Fluoroscopie). Ces deux problématiques mènent à la construction de jeux de données de screening non représentatifs des pratiques et objectifs réels des chefs de projet.

2.2. Construction des jeux de données labellisées

Afin d'évaluer l'outil ASReview, nous construisons pour chaque revue un jeu de données labellisées où nous associons à chaque référence son statut d'inclusion. Nous utilisons la même nomenclature de labels que Norman (Norman 2020) :

- Y (pour « yes ») pour les articles retenus après lecture des textes complets, soit les articles dans la base sélective finale.

- M (pour « maybe ») pour les articles inclus sur titre/résumé mais pas sur texte complets, soit les articles dans la base fille mais absents de la base sélective finale,

- N (pour « no ») pour les articles dans la base mère mais absents de la base sélective finale et de la base fille.

Plusieurs enjeux ont émergé lors de la construction de ces jeux de données.

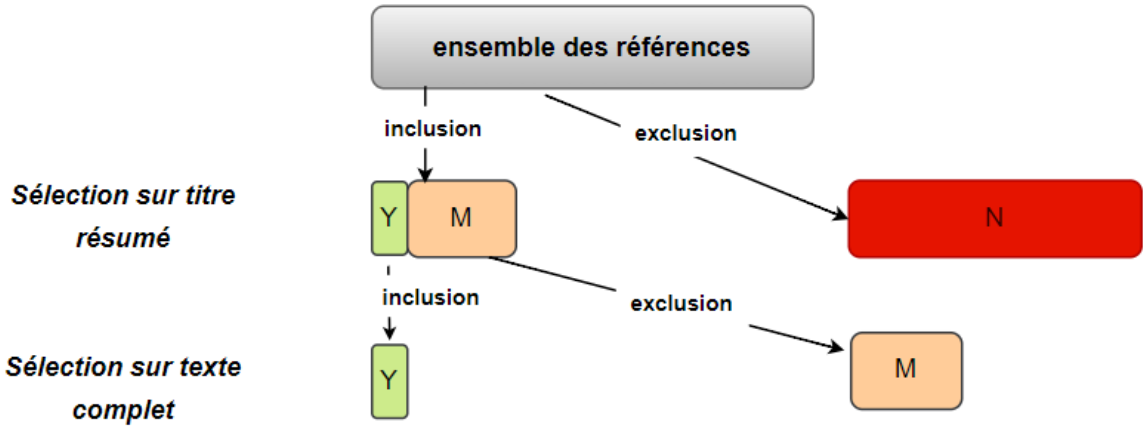

Figure 2 : Description du processus de sélection des références (Y pour « yes », M pour « maybe », N pour « no »)

2.2.1. Enjeu de déduplication

La base mère peut comporter des doublons. En revanche, sauf erreur manuelle, les bases filles et sélectives ne sont pas censées en contenir. Le choix de la clé de dédoublonnage conseillée par le SDV est : titre, premier auteur, année.

Nous proposons de faire l'appariement des différentes bases par cette triple clé et dans un second temps de chercher des appariements avec une clé plus permissive basée sur titre et année. Le processus de construction du jeu de données est détaillé dans l'Annexe 1.

2.2.2. Choix de la stratégie de screening

Nous pouvons répliquer l'étape de pré-screening en identifiant les articles retenus sur titre/résumé ou alors répliquer l'étape de screening en identifiant les articles retenus in-fine (sur texte complet). Formellement, une stratégie de labelling consiste à un regroupement de labels pour former une tâche de classification binaire : Par exemple (Y||MN) considère comme classe positive le label Y (articles retenus in fine), et comme classe négative le label M ou le label N (articles exclus in-fine).

Les travaux de Norman (Norman 2020) se concentrent sur le screening (Y||MN) au lieu de mener deux étapes de screening automatique (YM||N) puis (Y||M). Ce choix est justifié par des résultats empiriques avec des modèles (non actifs) TF-IDF[2] et régression logistique. En effet, distinguer (Y||M) est deux fois

[2] TF-IDF pour Term-Frequency-Inverse Document Frequency est une transformation du texte de la référence donnée en entrée donnant plus de poids au mot apparaissant souvent dans le corpus et moins d'importance à des mots apparaissant dans tous les documents. Le résultat de cette transformation est fourni au classifieur pour prédire le score de la référence donnée en entrée.

plus difficile que de distinguer (Y||N) ou (Y||MN), mesuré en WSS@95, 0.163 contre 0.423 ou 0.449 ((Norman et al. 2018), table 3.a).

Nous évaluons la stratégie condensant les deux étapes en une seule (Y||MN) ainsi que la première étape du processus usuel en deux étapes (YM||N) (nous n'avons pas évalué (Y||M)).

Dans tous les cas, nous utilisons comme données d'entrée pour les modèles uniquement les titre/résumé par simplicité mais nous pourrions théoriquement demander les bases avec les textes complets au SDV. Norman (Norman 2020) se base uniquement sur les résumés car il considère complexe voire impossible de récupérer les textes complets.

2.3. Description des jeux de données obtenus

2.3.1. Déduplication

Une étude des clés de déduplication [titre, auteur, année] montre que des clés sont manquantes pour quelques publications. Les titres sont présents pour toutes les publications. Sur l'ensemble des trois clés, on trouve 61 références dupliquées pour Fluoroscopie, 6 pour Endomicroscopie, 5 pour Test_HPV et aucune pour TestsKSein ou TAVI. On retient les références avec les résumés les plus longs.

Pour la revue Fluoroscopie, les auteurs sont manquants pour 3 arrêtés du Journal Officiel. Pour la revue Endomicroscopie, 4 références manquent d'auteurs sans que la raison soit identifiable. Une référence supplémentaire manque d'un auteur dans la base fille. Pour la revue Test_HPV, 29 références (sur 2882) manquent d'auteur dans la base mère sans raison apparente. Dans la base fille, seules 5 références manquent d'auteur dont 3 textes juridiques. Pour la revue TestsKSein, 3 références manquent d'un auteur dans la base mère, et dans la base fille 3 autres références manquent d'un auteur et une d'une année. Pour la revue TAVI, 4 références manquent d'un auteur dans la base mère dont 3 erratums. On retrouve un de ces erratums dans la base fille avec toujours l'auteur manquant.

Revue	Base	Nb références	Auteur manquant	Année manquante	Résumé manquant	Nb références dupliquées
Test_HPV	mère	2882	29	5	144	5
Test_HPV	fille	783	5	0	209	0
Test_HPV	sélective	267	3	0	109	0
Fluoroscopie	mère	345	3	1	83	61
Fluoroscopie	fille	102	3	0	49	0
Fluoroscopie	sélective	68	3	0	34	0
Endomicroscopie	mère	345	4	0	30	6
Endomicroscopie	fille	163	1	1	29	0
Endomicroscopie	sélective	104	1	0	8	0
TestsKSein	mère	788	3	0	19	0

TestsKSein	fille	327	3	1	123	0
TestsKSein	sélective	76	0	0	32	0
TAVI	mère	731	4	0	61	0
TAVI	fille	103	1	0	28	0
TAVI	sélective	36	0	0	12	0

Tableau 1 : Description des bases de données sources mère, fille et sélective.

La déduplication sur la clé [titre, auteur, année] permet de faire correspondre systématiquement toutes les références dans la base sélective à une référence dans la base mère ou fille. En revanche, plusieurs références de la base fille ne sont pas retrouvées dans la base mère :

- Test_HPV : 342 références (44%)
- Fluoroscopie : 35 références (34%)
- Endomicroscopie : 37 références (23%)
- TestsKSein : 246 références (75%). Pas de correspondance sur titre seul non plus.
- TAVI : 22 références (21%)

On supprime les références sans correspondance de la base fille dans la base mère pour l'expérience de screening rétrospective. En effet, ces références ne sont pas disponibles dans l'extraction initiale de données. Pour refléter au mieux le process actuel, nous avons donc décidé de les enlever. Ce choix est discutable car il peut introduire des erreurs dans les labels, si les clés de jointure ne sont pas fiables (ex. titre qui change entre la base mère et la base fille) alors une référence non appareillée de la base mère mais pertinente sera labellisé comme non pertinente.

On corrige par ailleurs les bases de trois revues du SEAP pour lesquelles les références écartées sur texte complet étaient incluses dans la base sélective (en raison de la nécessité de les citer en annexe du rapport).

2.3.2. Labels

review_id	nb refs	nb Y	nb M	nb N	ratio Y	ratio Y ou M	nb refs filles absentes de mere
Test_HPV	**2877**	**104**	**337**	**2436**	**0.04**	**0.15**	**0**
Fluoroscopie	**284**	**17**	**50**	**217**	**0.06**	**0.24**	**0**
Endomicroscopie	**339**	**16**	**100**	**223**	**0.05**	**0.34**	**0**
TestsKSein	**788**	**12**	**29**	**747**	**0.02**	**0.05**	**0**
TAVI	**731**	**21**	**60**	**650**	**0.03**	**0.11**	**0**

Tableau 2 : Nombre de références et de labels pour chaque revue.

Le Tableau 2 répertorie le nombre de références par type de label pour chaque revue. Le label Y correspond aux références incluses sur texte complet (dans la base sélective). Le label M correspond à celles incluses sur titre/résumé mais exclues sur texte complet (dans la base fille mais absentes de la base sélective finale). Le label N correspond aux références exclues dès le titre/résumé (dans la base mère mais absentes de la base fille ou sélective finale).

2.4. Protocole expérimental

Nous répliquons l'étape de screening sur titre/résumé de chaque revue en utilisant l'API python du logiciel d'ASReview[3] (Van De Schoot et al. 2021).

2.4.1. Métriques

La principale métrique pour le screening priorisé est le *work saved over sampling at a given recall* (WSS@recall). Cette métrique correspond au pourcentage de références que l'on peut éviter de screener afin d'obtenir un rappel donné comparé à lire les références dans un ordre aléatoire. C'est la quantité de travail évitée par le screening priorisé pour un niveau donné de rappel. Nous restituons WSS@95 et WSS@100. Enfin, nous restituons la courbe complète de rappel en fonction du nombre de références consultées.

2.4.2. Modèles évalués

Nous évaluons les caractéristiques extraites par TF-IDF suivies d'une régression logistique, d'un naive Bayes ou d'un SVC. Ce choix de modèles simples est justifié par les raisons suivantes :

- La régression logistique est utilisée dans le travail de (Norman et al. 2018),
- Le naive Bayes est l'algorithme par défaut de la bibliothèque ASReview,
- La SVC est un algorithme fonctionnant bien avec la représentation TF-IDF.

2.4.3. Variation des résultats selon l'aléa

Nous évaluons la variance des performances des modèles en fonction de la seed aléatoire. Celle-ci a un effet sur les choix des deux références initiales labellisées (« prior knowledge »), le modèle, si celui a une composante aléatoire (pas le cas pour la régression logistique et le naive bayes), la stratégie de repondération des classes. Chaque modèle est lancé sur 10 seeds.

2.4.4. Résumé du protocole expérimental

Nous évaluons les stratégies de labelling (Y||MN) et (YM||N) pour les cinq revues dédoublonnées sur DOI et titre. Nous évaluons avec 10 seeds différentes des modèles TF-IDF suivis de régression logistique, naive Bayes ou SVC en calculant les métriques WSS@100, WSS@95.

[3] Version du logicielle utilisée : https://github.com/asreview/asreview/tree/v1.6.1

3. Résultats

Nous reportons les résultats séparément pour les deux stratégies de labelling. Les résultats complets reportés en Worked Saved over Sampling sont disponibles dans l'Annexe 3. Les résultats en nombre de références évités (Worked Saved, WS) sont reportés dans les Figure 3 à 7.

3.1. Résultats pour la stratégie de labelling (YM||N)

Les performances des modèles de régression logistique et du naive Bayes sont semblables sauf pour la revue TestsKSein où le modèle naive Bayes est meilleur pour le WSS@100 (cf. Tableau 3). En raison de son coût computationnel plus élevé, le SVC n'a été évalué que pour la revue TestsKSein. Comme il n'apporte pas de gain significatif, nous ne l'avons pas évalué pour les autres revues.

Les résultats, mesurés en nombre de références évitées à un WSS donné (100 ou 95) diffèrent selon les revues.

Le WSS@95 est classiquement utilisé comme métrique d'évaluation de la quantité de travail qui peut être économisée, au prix de manquer 5% des références pertinentes. Nous proposons de considérer un seuil arbitraire de 30% pour le WSS@95 pour qualifier l'outil de pertinent pour une revue donnée.

En considérant ce seuil, deux revues ont un WSS@95 supérieur, égal dans les meilleurs cas à 50% pour la revue TAVI et 45% pour la revue TestsKSein (modèle de régression logistique).

Ces deux revues ont pour caractéristiques un grand nombre de références candidates (788 et 731) ainsi qu'un faible ratio d'inclusion sur titre résumé (0.11 et 0.05). De plus, la revue TestsKSein se base sur des critères PICOT, ce qui suppose un champ restreint de recherche.

Pour les revues Test_HPV, Endomicroscopie et Fluoroscopie, le WSS@95 est inférieur au seuil de 30% : il est compris entre 7% et 18% selon la revue et le modèle testé. Une caractéristique commune à ces trois revues est leur fort taux d'inclusion sur titre résumé (0.15 à 0.34). En outre, les moindres performances observées sur la revue Fluoroscopie pourraient être liées au fait que la revue ne porte pas sur un acte précis mais un ensemble d'actes.

Les résultats montrent une faible variation des performances en fonction de la graine aléatoire utilisée pour choisir les articles initiaux. Quel que soit le choix des deux articles initiaux labellisés, les performances restent donc stables.

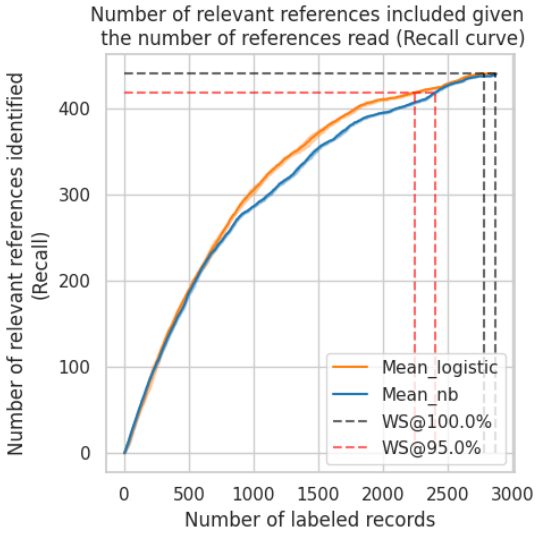

Figure 3 : Test_HPV, stratégie (YM||N)

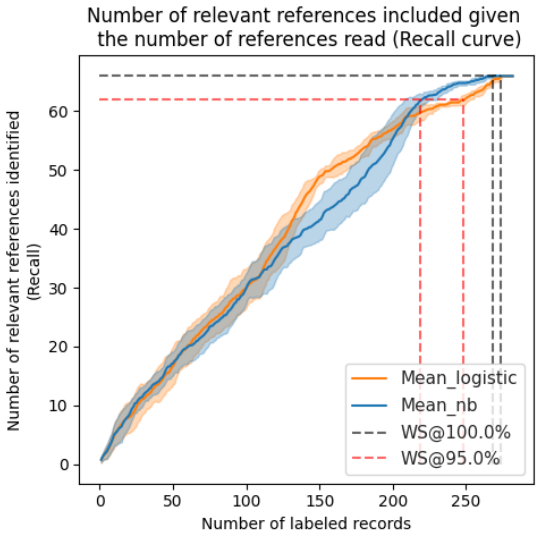

Figure 4 : Fluoroscopie, stratégie (YM||N)

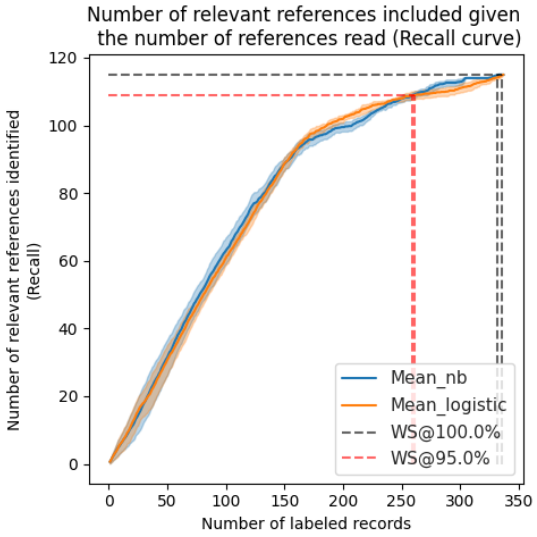

Figure 5 : Endomicroscopie, stratégie (YM||N)

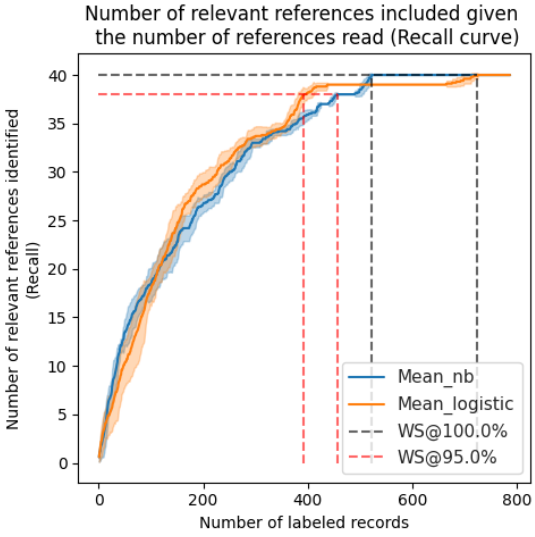

Figure 6 : TestsKSein, stratégie (YM||N)

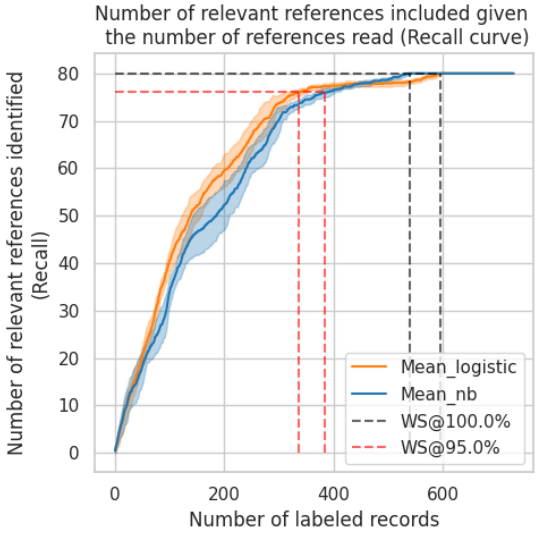

Figure 7 : TAVI, stratégie (YM||N)

3.2. Résultats pour la stratégie de labelling (Y||MN)

Les performances de la régression logistique et du naive Bayes sont semblables sauf pour deux revues où l'un des modèles est meilleur pour découvrir les derniers articles (WSS@100) (naive Bayes pour Fluoroscopie, régression logistique pour TestsKSein).

Le nombre de références évitées à un WSS donné est plus marqué pour toutes les revues pour la sélection sur texte complet (Y||MN) versus la sélection sur titre/résumé (YM||N). Ce résultat peut notamment s'expliquer par le plus faible taux d'inclusion des références sur texte complet pour les cinq revues (entre 0.02 et 0.06).

En considérant le seuil arbitraire de 30% pour le WSS@95 pour représenter une économie significative de travail, le WSS@95 dépasse ce seuil pour quatre des cinq revues testées. Dans les meilleurs cas le WSS@95 atteint 44% pour la revue Test_HPV (modèle naive Bayes), 58% pour Endomicroscopie (régression logistique), 64% pour TAVI (régression logistique) et 71% pour TestsKSein (modèle naive Bayes).

La revue Fluoroscopie a un WSS@95 inférieur, atteignant 27% pour le modèle naive Bayes. Cette moindre performance est peut-être à relier au fait que la revue ne porte pas sur un acte précis mais un ensemble d'actes.

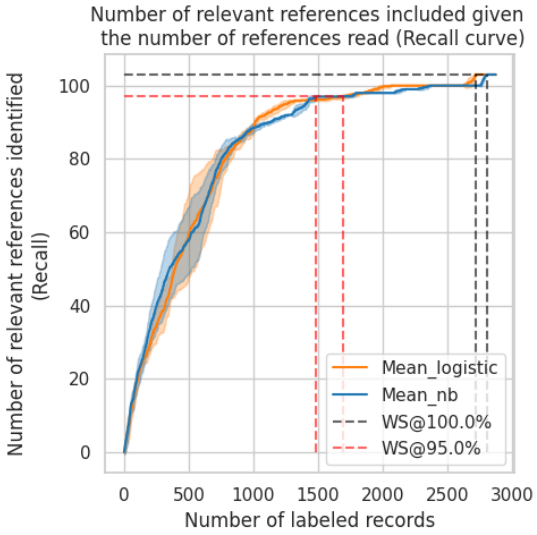

Figure 3 : Test_HPV, stratégie (Y||MN)

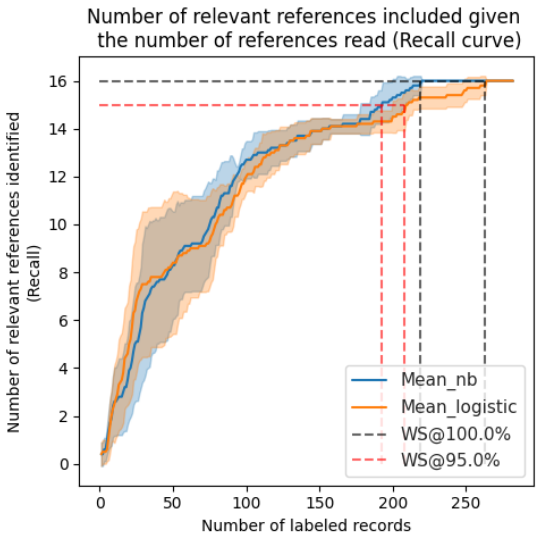

Figure 4 : Fluoroscopie, stratégie (Y||MN)

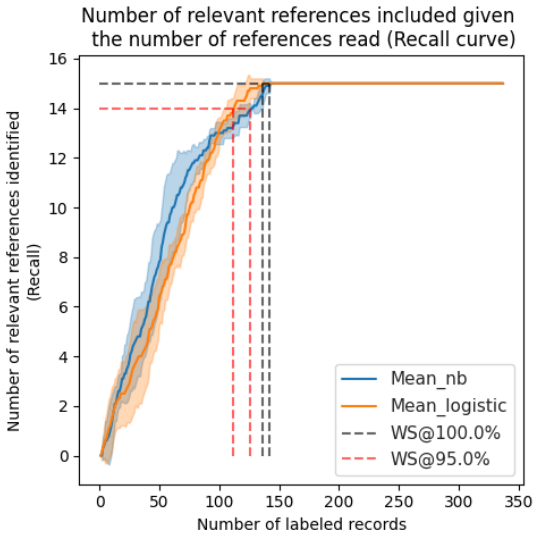

Figure 5 : Endomicroscopie, stratégie (Y||MN)

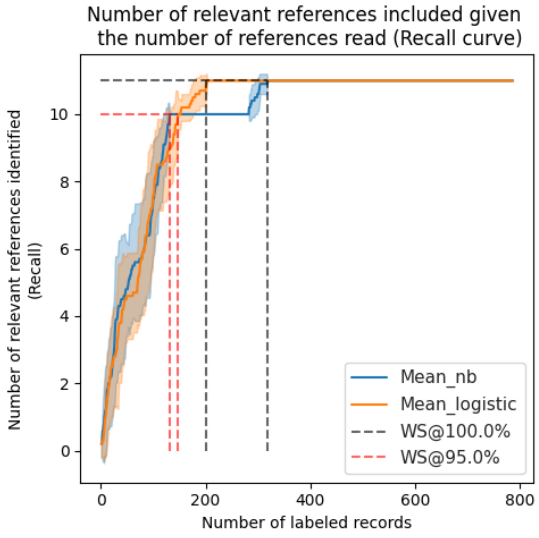

Figure 6 : TestsKSein, stratégie (Y||MN)

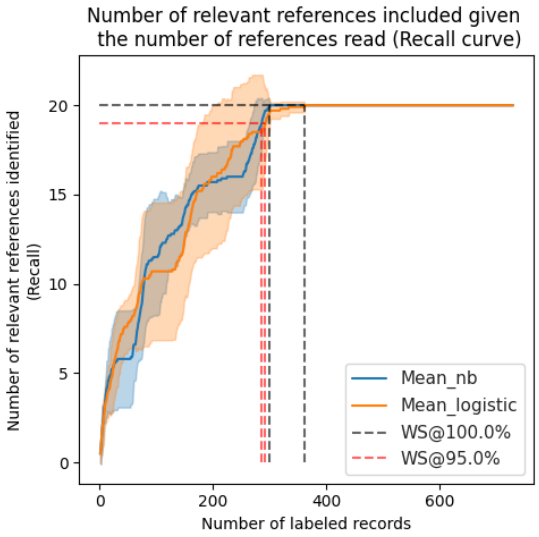

Figure 7 : TAVI, stratégie (Y||MN)

4. Discussion

4.1. Synthèse des apprentissages et des limites

Nous avons évalué l'efficacité du screening priorisé de façon rétrospective sur cinq revues de littérature de la HAS à l'aide d'un logiciel open source (ASReview).

Lors de la création des jeux de données labellisés, deux limites ne permettent d'évaluer qu'imparfaitement l'intérêt de l'outil pour des chefs de projet métiers. Premièrement, l'absence d'identifiant unique au cours du processus de revue complexifie grandement la création de ces jeux de données. En particulier, les changements dans les clefs de jointure (titre, auteur, année) entre les bases mères, filles et sélectives introduisent des erreurs dans les labels si l'on compare avec les références effectivement retenues dans les publications (cf. Partie 2.1). Deuxièmement, les bases bibliographiques utilisées pour le screening rétrospectif regroupent plusieurs questions de recherche, qui, en conditions réelles, conduiraient à des screenings distincts effectués par les chefs de projet. Cette situation a pu complexifier la tâche de screening unique réalisée rétrospectivement et met en évidence la nécessité de tester de façon prospective les performances de l'outil ASReview.

Les résultats obtenus en termes de WSS@95 sont comparables à ceux de la littérature (Van De Schoot et al. 2021; Ferdinands et al. 2023).

Le screening priorisé a un meilleur potentiel de détection des articles pertinents, pour la stratégie identifiant les références retenues sur texte complet (Y||MN) versus la stratégie les identifiant sur titre résumé (YM||N). Une première explication est le plus faible ratio d'inclusion à l'étape de sélection sur texte complet versus sélection sur titre/résumé (il y a moins de références pertinentes à identifier). Ce résultat suggère également que le screening priorisé est d'autant plus efficace pour identifier les revues pertinentes qu'on lui fournit des revues véritablement pertinentes (sur texte complet). En usage courant, cela signifie qu'il faut être exigeant dans le choix des références labellisées comme pertinentes lors du processus de revue. En considérant le processus actuel au sein de la HAS qui consiste en une première étape de sélection sur titre/résumé, il peut être recommandé à l'humain d'être exigeant dès cette première étape (en examinant le texte complet dès qu'il est accessible et en cas de doute) pour améliorer la performance de l'outil. Afin de refléter plus fidèlement les processus de revue à la HAS, il faudrait également évaluer la stratégie de labelling (Y||M) correspondant à la seconde étape après (YM||N).

Les gains de temps potentiels sont plus marqués pour les revues avec un grand nombre de références candidates et de faibles ratios d'inclusion. Ce résultat est valable pour les deux stratégies de screening. Il semblerait que les résultats soient également meilleurs lorsque la revue est ciblée sur un acte précis plutôt que sur un ensemble d'actes. Ce dernier résultat reste à confirmer sur d'autres exemples.

Les modèles régression logistique et naive Bayes ont des performances proches, malgré quelques différences sur le WSS@100 pour certaines revues.

4.2. Ouverture

Il serait judicieux de se doter d'identifiants uniques pour les références afin d'éviter les erreurs de labellisation provenant de changements dans les titres, auteurs ou année des références.

En outre, il est nécessaire de distinguer chaque question de recherche pour tirer le meilleur parti du screening priorisé. Il faudrait donc créer une base mère par question de recherche.

Un bénéfice attendu de l'outil pour un humain revoyant l'ensemble des références est un confort ressenti dans le processus de sélection, dans la mesure où les articles les plus pertinents sont priorisés par l'outil. Ce bénéfice reste à valider lors d'expérimentations prospectives.

Cependant le premier bénéfice attendu pour ce type d'outil est un gain de temps. Le véritable gain de temps devrait être étudié à l'aune d'un critère d'arrêt. En effet, la simulation de revue assistée étudiée dans cet article ne permet pas de conclure sur le gain de temps réel. Dans cette étude, aucun critère d'arrêt n'est fixé. La simulation est conduite comme si l'utilisateur revoyait toutes les références. En utilisation réelle, le screening priorisé permettrait de gagner en productivité uniquement si le chef de projet arrêtait le screening avant d'avoir revu toutes les publications. Un critère d'arrêt intuitif serait d'arrêter le screening lorsqu'aucune nouvelle référence n'est incluse après avoir revu successivement un certain nombre de références. Ce nombre de références successives non-pertinentes vues avant arrêt est un paramètre à fixer qui dépend du ratio de références pertinentes dans l'ensemble du corpus. Même si ce ratio est inconnu a priori, on peut le majorer par un ratio élevé parmi les cinq revues étudiées (ex. 0.2). Il serait intéressant d'étudier le temps réel gagné et le nombre de revues laissées de côté par le screening priorisé en fonction d'un critère d'arrêt simple terminant la revue après la lecture par exemple 10% de revues non pertinentes. Une revue des différents critères d'arrêt possibles est disponible sur le github d'ASReview[4].

Il serait également intéressant, dans la stratégie de sélection sur titre/résumé, de mesurer le rappel et le WSS pour un certain niveau de rappel sur le nombre de références pertinentes finalement retenues sur texte complet (« Y »), pour évaluer si le classifieur les a identifiées avant les références rejetées sur texte complet (« M »). Cela permettrait également de sécuriser l'utilisation d'un critère d'arrêt dans cette phase de sélection sur titre/résumé (puis que les « M » n'ont pas d'intérêt).

Enfin, il serait intéressant de comparer les résultats avec critères d'arrêt à ceux d'un grand modèle de langage. En effet, les grands modèles de langage ne nécessitent pas d'entraînement et pourrait s'appuyer sur les critères d'inclusion et d'exclusion. Il faudrait mettre en perspective ces résultats en termes de rappel et de spécificité avec le taux d'erreurs humaines, estimé par (Wang et al. 2020) sur 85 revues à 10% (fausses exclusions ou inclusions).

De nouvelles études prospectives avec l'utilisation d'ASReview en temps réel par des chefs de projet permettront d'évaluer plus finement les performances de l'outil et son intérêt.

[4] https://github.com/asreview/asreview/discussions/557

Table des annexes

Annexe 1. Création du jeu de données

La procédure de création du jeu de données à partir des bases mères, filles et sélectives est la suivante (décrite dans la Figure 8 : Processus de jointure en deux étapes entre la base mère et la base fille. Ce processus est répété pour joindre la base sélective.) :

- Déduplique sur les clés [*titre, premier auteur, année]* la base mère

- jointure 1 : joint la base fille à la base mère avec la clé (*titre* | *premier auteur* | *année)* en créant une nouvelle colonne contenant les identifiants des références de la base fille

- jointure 2 : joint les références non jointes de la base mère aux références non jointes de la base fille sur la clé permissive [*titre, année]*

- concatène les références de la jointure 1, de la jointure 2 et celles de la base fille sans correspondance identifiée

- recommence ce processus pour joindre la base obtenue avec la base sélective.

Les références non jointes de la base fille sont supprimées des jeux de données pour l'expérience de screening rétrospective. En effet, ces références ne sont pas disponibles dans l'extraction initiale de données. Pour refléter au mieux le process actuel, nous avons donc décider de les enlever.

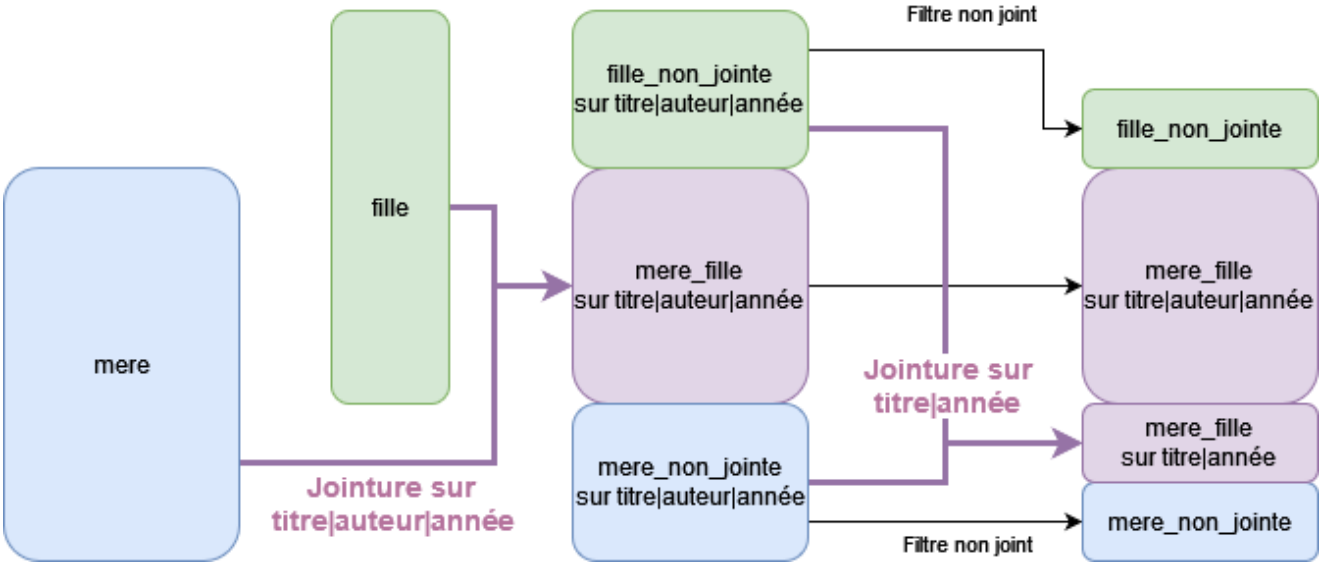

Figure 8 : Processus de jointure en deux étapes entre la base mère et la base fille. Ce processus est répété pour joindre la base sélective.

Annexe 2. Analyses supplémentaires des jeux de données des revue

Analyse des références incluses pour la revue Fluoroscopie

Les bases mères, filles et sélectives non dédoublonnées contiennent 345, 102 et 68 références. La base mère dédoublonnée contient 284 références dont 61 inclus sur titre/résumé et 32 incluses finalement.

Dans le corps de la publication, il est mentionné 345 références initiales, 57 références inclues sur titre/résumé et 24 références retenues finalement. Dans la publication, 33 références écartées sur lecture complète sont mentionnées en annexe de la publication et font donc partie de la base sélective.

De plus, 12 références mentionnées dans la publication et présentes dans la base sélective sont labellisées comme non pertinente par l'algorithme de création du jeu de données (ex : couturier_2017, NICE_2018, visser_2012). Ces erreurs sont dues à des clés de jointure défaillantes : l'auteur ou le titre, voire l'année qui change entre deux bases. Les références des sociétés savantes semblent particulièrement touchées par ce problème.

Analyse des références incluses pour la revue Endomicroscopie

La publication liée à la revue mentionne dans son corps de texte 298 références initiales, 124 inclues sur titre/résumé et 16 références retenues après lecture complète. De plus, le nombre de références dans la partie référence de la publication s'élève à 104. Or, la base de données que nous avons créée donne 340 références initiales, 116 inclues sur titre/résumé et 87 retenues parmi la base sélective.

Les 16 références retenues dans la publication sont bien présentes parmi les 87 retenues parmi la base sélective. La source de différence est expliquée par l'annexe 5 de la publication qui référence 71 publications écartées sur titre/résumé, soit exactement la différence entre les 87 obtenues à partir des bases du SDV et les 16 publications retenues.

Analyse des références incluses pour la revue TestsKSein

Seulement 3 des références mentionnées comme incluses dans la publication sont retrouvées dans la base sélective. Les trois manquantes (toutes associées à l'auteur Kalinsky) ne sont pas présentes dans la base mère et ont donc été écartées du processus de création du jeu de données.

Annexe 3. Résultats complets

Le Tableau 3 restitue le nombre d'articles évités pour atteindre un rappel de 95% ou 100% par en utilisant un screening priorisé par rapport à un screening aléatoire. Les résultats sont moyennés sur 10 seeds différentes et la déviation standard est reportée entre parenthèses. Le nombre total de références pour chaque revue est rappelé pour donner un ordre de grandeur de l'économie de travail réalisée.

Revue	Modèle	Stratégie	Nb réfs.	WSS@95 (std)	WSS@100 (std)
Test_HPV	logistic	YM_N	2877	502.1 (±21.7)	127.2 (±22.46)
Test_HPV	nb	YM_N	2877	330.4 (±11.75)	4.6 (±0.52)
Fluoroscopie	logistic	YM_N	284	19.7 (±7.83)	8.7 (±2.98)
Fluoroscopie	nb	YM_N	284	45.8 (±7.13)	19.6 (±5.02)
Endomicroscopie	logistic	YM_N	339	55.5 (±15.59)	2.0 (±2.58)
Endomicroscopie	nb	YM_N	339	60.8 (±7.89)	6.7 (±2.45)
TestsKSein	logistic	YM_N	788	352.1 (±13.14)	79.1 (±19.11)
TestsKSein	nb	YM_N	788	288.9 (±5.34)	262.6 (±7.4)
TAVI	logistic	YM_N	731	365.7 (±23.57)	143.0 (±11.67)
TAVI	nb	YM_N	731	318.1 (±18.91)	201.3 (±9.98)
Test_HPV	logistic	Y_MN	2877	1095.2 (±83.22)	166.1 (±28.1)
Test_HPV	nb	Y_MN	2877	1258.5 (±31.6)	68.8 (±6.44)
Fluoroscopie	logistic	Y_MN	284	63.1 (±31.15)	34.2 (±23.57)
Fluoroscopie	nb	Y_MN	284	75.5 (±29.44)	71.8 (±13.92)
Endomicroscopie	logistic	Y_MN	339	196.8 (±8.57)	204.8 (±9.24)
Endomicroscopie	nb	Y_MN	339	188.4 (±9.03)	190.3 (±3.97)
TestsKSein	logistic	Y_MN	788	546.8 (±16.35)	571.9 (±20.26)
TestsKSein	nb	Y_MN	788	561.5 (±15.16)	455.0 (±11.29)
TAVI	logistic	Y_MN	731	468.2 (±48.32)	426.2 (±43.65)
TAVI	nb	Y_MN	731	418.7 (±51.09)	440.4 (±30.7)

Tableau 3 : Résultats complets

Références bibliographiques

ASReview LAB. 2024. « ASReview LAB - A tool for AI-assisted systematic reviews ». Zenodo. https://doi.org/10.5281/zenodo.10912058.

Covidence. 2024. « Covidence - Better Systematic Review Management ». Covidence. 2024. https://www.covidence.org/.

DistillerSR. 2024. « Systematic Review and Literature Review Software ». DistillerSR. 2024. https://www.distillersr.com/.

Elliott, Julian H, Anneliese Synnot, Tari Turner, Mark Simmonds, Elie A Akl, Steve McDonald, Georgia Salanti, et al. 2017. « Living systematic review: 1. Introduction—the why, what, when, and how ». Journal of clinical epidemiology 91:23-30.

Elliott, Julian H, Tari Turner, Ornella Clavisi, James Thomas, Julian PT Higgins, Chris Mavergames, et Russell L Gruen. 2014. « Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap ». PLoS medicine 11 (2): e1001603.

EPPI. 2024. « EPPI-Reviewer ». 2024. https://eppi.ioe.ac.uk/.

Ferdinands, Gerbrich, Raoul Schram, Jonathan de Bruin, Ayoub Bagheri, Daniel L. Oberski, Lars Tummers, Jelle Jasper Teijema, et Rens van de Schoot. 2023. « Performance of Active Learning Models for Screening Prioritization in Systematic Reviews: A Simulation Study into the Average Time to Discover Relevant Records ». Systematic Reviews 12 (1): 100. https://doi.org/10.1186/s13643-023-02257-7.

Haute Autorité de Santé. 2019a. « Évaluation de la pertinence de l'acte de fluoroscopie de l'œil réalisé par l'ophtalmologue », 2019.

———. 2019b. « Évaluation de la recherche des papillomavirus humains (HPV) en dépistage primaire des lésions précancéreuses et cancéreuses du col de l'utérus et de la place du double immuno-marquage p16/Ki67 », 2019.

———. 2022. « Endomicroscopie confocale par aiguille de ponction pour la caractérisation des tumeurs kystiques pancréatiques », 2022.

———. 2023. « Actualisation 2023 : utilité clinique des signatures génomiques dans le cancer du sein RH+/HER2- de stade précoce », 2023.

———. 2024. « Critères d'éligibilité des centres implantant des TAVIs (évaluation de 2023) », 2024.

Higgins, JPT, Chandler J TJ, M Cumpston, T Li, MJ Page, et VA Welch. 2023. « Cochrane Handbook for Systematic Reviews of Interventions version 6.4(updated August 2023) ». www.training.cochrane.org/handbook.

Norman, Christopher. 2020. « Systematic review automation methods ».

Norman, Christopher, Mariska Leeflang, Pierre Zweigenbaum, et Aurélie Névéol. 2018. « Automating Document Discovery in the Systematic Review Process: How to Use Chaff to Extract Wheat. » In International Conference on Language Resources and Evaluation.

O'Mara-Eves, Alison, James Thomas, John McNaught, Makoto Miwa, et Sophia Ananiadou. 2015. « Using text mining for study identification in systematic reviews: a systematic review of current approaches ». Systematic reviews 4:1-22.

Ouzzani, Mourad, Hossam Hammady, Zbys Fedorowicz, et Ahmed Elmagarmid. 2016. « Rayyan—a web and mobile app for systematic reviews ». Systematic Reviews 5 (1): 210. https://doi.org/10.1186/s13643-016-0384-4.

Rayyan. 2022. « Rayyan – Intelligent Systematic Review - Rayyan ». 2022.

Settles, Burr. 2009. « Active learning literature survey ».

Van De Schoot, Rens, Jonathan De Bruin, Raoul Schram, Parisa Zahedi, Jan De Boer, Felix Weijdema, Bianca Kramer, et al. 2021. « An open source machine learning framework for efficient and transparent systematic reviews ». Nature machine intelligence 3 (2): 125-33.

Wang, Zhen, Tarek Nayfeh, Jennifer Tetzlaff, Peter O'Blenis, et Mohammad Hassan Murad. 2020. « Error rates of human reviewers during abstract screening in systematic reviews ». PloS one 15 (1): e0227742.

Abréviations et acronymes

HAS	Haute Autorité de santé
HPV	Papillomavirus humains
IA	Intelligence artificielle
M	Maybe (classe définie dans le processus de sélection de publications)
N	No (classe définie dans le processus de sélection de publications)
PICOTS	Population, Intervention, Comparator, Outcome, Timing, Setting
SBP	Service des Bonnes Pratiques
SDV	Service Documentation et Veille
SEAP	Service Evaluation des Actes Professionnels
SESPEV	Service Evaluation de Santé Publique et Evaluation des Vaccins
TAVI	Transcatheter aortic valve implantation
TF-IDF	Term frequency-inverse document frequency
WSS	Work saved over sampling
Y	Yes (classe définie dans le processus de sélection de publications)

Retrouvez tous nos travaux sur
www.has-sante.fr

HAUTE AUTORITÉ DE SANTÉ

Développer la qualité dans le champ sanitaire, social et médico-social