Article scientifique pour les exercices pratiques

Titre: Utilisation de l'intelligence artificielle pour l'extraction automatisée d'informations dans la littérature sur les maladies tropicales négligées en Afrique de l'Ouest

Résumé

Les maladies tropicales négligées (MTN) représentent un défi majeur de santé publique en Afrique de l'Ouest, mais l'exploitation efficace de la littérature scientifique sur ce sujet reste difficile en raison du volume croissant de publications. Cette étude présente une approche basée sur l'intelligence artificielle pour l'extraction automatisée d'informations à partir d'un corpus de 250 articles scientifiques sur cinq MTN prévalentes dans la région (schistosomiase, filariose lymphatique, onchocercose, trypanosomiase et ulcère de Buruli). Nous avons développé et évalué un système combinant des modèles de langage préentraînés adaptés au domaine biomédical avec des techniques d'apprentissage par transfert. Les résultats montrent une précision globale de 87,3% dans l'extraction des informations épidémiologiques, des interventions thérapeutiques et des facteurs de risque. L'analyse des données extraites révèle des disparités significatives dans la couverture géographique et thématique de la recherche, avec une sous-représentation des études sur les déterminants sociaux et environnementaux. Cette approche offre un potentiel considérable pour accélérer la synthèse des connaissances et identifier les lacunes dans la recherche sur les MTN en Afrique de l'Ouest.

1. Introduction

Les maladies tropicales négligées (MTN) affectent plus d'un milliard de personnes dans le monde, principalement dans les régions tropicales et subtropicales où la pauvreté est endémique (WHO, 2021). En Afrique de l'Ouest, ces maladies constituent un fardeau sanitaire et économique considérable, entravant le développement socioéconomique et perpétuant le cycle de la pauvreté (Hotez et al., 2019). Malgré l'importance de ces

maladies, les efforts de recherche et d'intervention restent fragmentés et insuffisamment coordonnés.

La littérature scientifique sur les MTN en Afrique de l'Ouest s'est considérablement développée au cours des deux dernières décennies, avec une augmentation annuelle moyenne de 8,7% du nombre de publications (Deribe et al., 2022). Cette prolifération d'informations, bien que bénéfique, pose des défis significatifs pour les chercheurs, les professionnels de la santé et les décideurs politiques qui doivent synthétiser ces connaissances pour orienter leurs actions. L'extraction manuelle d'informations à partir de cette vaste littérature est chronophage, sujette aux erreurs et difficilement évolutive.

Les récentes avancées en intelligence artificielle (IA), particulièrement dans le domaine du traitement du langage naturel (NLP), offrent des opportunités prometteuses pour automatiser l'extraction d'informations à partir de textes scientifiques (Wang et al., 2020). Des modèles comme BioBERT (Lee et al., 2020) et SciBERT (Beltagy et al., 2019) ont démontré des performances impressionnantes dans la compréhension de textes biomédicaux. Cependant, leur application à la littérature sur les MTN en Afrique de l'Ouest présente des défis spécifiques, notamment en raison des particularités épidémiologiques, géographiques et socioculturelles de la région.

Cette étude vise à développer et évaluer un système d'extraction automatisée d'informations adapté à la littérature sur les MTN en Afrique de l'Ouest. Nos objectifs spécifiques sont de : (1) concevoir une architecture d'IA capable d'extraire avec précision les données épidémiologiques, les interventions thérapeutiques et les facteurs de risque ; (2) évaluer les performances du système par rapport à l'extraction manuelle par des experts ; et (3) analyser les tendances et lacunes dans la recherche sur les MTN en Afrique de l'Ouest à partir des données extraites.

2. Méthodologie

2.1 Constitution du corpus

Nous avons constitué un corpus de 250 articles scientifiques publiés entre 2000 et 2022, traitant de cinq MTN prévalentes en Afrique de l'Ouest : schistosomiase (n=65), filariose lymphatique (n=52), onchocercose (n=58), trypanosomiase humaine africaine (n=42) et ulcère de Buruli (n=33). Les articles ont été sélectionnés à partir des bases de données PubMed, Scopus et African Journals Online (AJOL) selon les critères d'inclusion suivants : (1) études originales portant sur au moins une des cinq MTN ciblées ; (2) recherche menée dans au moins un pays d'Afrique de l'Ouest ; (3) texte intégral disponible en anglais ou en français ; et (4) inclusion de données épidémiologiques, thérapeutiques ou sur les facteurs de risque.

2.2 Prétraitement des documents

Les articles au format PDF ont été convertis en texte brut à l'aide de la bibliothèque PyMuPDF. Un prétraitement spécifique a été appliqué pour normaliser le texte, segmenter les sections (introduction, méthodes, résultats, discussion), et identifier les tableaux et figures. Pour les articles en français (n=47), une traduction automatique a été réalisée avec MarianMT, suivie d'une vérification manuelle des termes techniques par un locuteur bilingue.

2.3 Architecture du système d'extraction

Notre système d'extraction repose sur une architecture hybride combinant :

1. **Un modèle de base préentraîné** : Nous avons utilisé BioBERT-Large, préentraîné sur PubMed et PMC, comme modèle de base pour la compréhension du texte biomédical.

2. **Fine-tuning spécifique au domaine** : Le modèle a été affiné sur un sous-ensemble de 50 articles annotés manuellement par trois experts en MTN, avec un accord inter-annotateurs de $\kappa=0.82$.

3. **Module d'extraction structurée** : Un système de règles et d'expressions régulières a été développé pour extraire des informations spécifiques comme les données démographiques, les prévalences, les protocoles thérapeutiques et les coordonnées géographiques.

4. **Module de résolution d'entités** : Les entités extraites (noms de maladies, médicaments, lieux) ont été normalisées et liées à des ontologies standard comme MeSH, DrugBank et GeoNames.

Le système a été conçu pour extraire trois catégories principales d'informations :

- **Données épidémiologiques** : prévalence, incidence, distribution géographique, caractéristiques démographiques des populations affectées.
- **Interventions thérapeutiques** : médicaments, dosages, durées de traitement, efficacité, effets secondaires.
- **Facteurs de risque** : déterminants environnementaux, comportementaux, socioéconomiques et biologiques.

2.4 Évaluation des performances

Les performances du système ont été évaluées sur un ensemble de test de 50 articles annotés manuellement, distincts de l'ensemble d'entraînement. Nous avons calculé la précision, le rappel et le F1-score pour chaque catégorie d'information extraite. Une

analyse d'erreur qualitative a également été réalisée pour identifier les types d'informations les plus difficiles à extraire correctement.

2.5 Analyse des tendances et lacunes

Les informations extraites ont été analysées pour identifier les tendances temporelles, la distribution géographique des études, les thématiques dominantes et les lacunes dans la recherche. Une analyse de réseau a été réalisée pour visualiser les connexions entre pays, maladies et approches thérapeutiques.

3. Résultats

3.1 Performances du système d'extraction

Le système a démontré une performance globale satisfaisante avec une précision moyenne de 87,3%, un rappel de 82,1% et un F1-score de 84,6% sur l'ensemble des catégories d'information (Tableau 1). Les performances variaient selon les catégories, avec les meilleures performances pour l'extraction des données épidémiologiques (F1=88,2%) et les moins bonnes pour les facteurs de risque socioéconomiques (F1=76,5%).

Tableau 1. Performances du système d'extraction par catégorie d'information

Catégorie d'information	Précision (%)	Rappel (%)	F1-score (%)
Données épidémiologiques	90,4	86,1	88,2
Caractéristiques démographiques	89,7	84,3	86,9
Distribution géographique	92,3	88,5	90,3
Interventions thérapeutiques	86,8	83,2	84,9
Médicaments et dosages	91,5	87,3	89,3
Efficacité des traitements	84,2	80,6	82,3
Effets secondaires	84,7	81,7	83,2
Facteurs de risque	84,7	77,1	80,7
Environnementaux	87,3	79,8	83,4
Comportementaux	85,1	78,2	81,5

Catégorie d'information	Précision (%)	Rappel (%)	F1-score (%)
Socioéconomiques	81,6	72,1	76,5
Biologiques	84,9	78,2	81,4
Moyenne globale	**87,3**	**82,1**	**84,6**

L'analyse d'erreur a révélé que les principales difficultés concernaient l'extraction d'informations implicites, la résolution des coréférences complexes et l'interprétation correcte des résultats négatifs ou des associations non significatives.

3.2 Analyse des tendances dans la recherche sur les MTN

L'analyse des 250 articles a révélé plusieurs tendances notables :

1. **Distribution temporelle** : Une augmentation significative des publications sur les MTN en Afrique de l'Ouest a été observée, passant de 5,2 articles par an en 2000-2005 à 21,7 articles par an en 2018-2022 ($p<0,001$).

2. **Couverture géographique** : Une distribution inégale des études a été constatée, avec une forte concentration au Nigeria (28,4%), au Ghana (18,2%) et au Sénégal (14,6%), tandis que des pays comme la Guinée-Bissau (1,2%) et le Liberia (2,0%) étaient sous-représentés (Figure 1).

3. **Focus thématique** : Les études épidémiologiques descriptives dominaient (42,8%), suivies par les évaluations d'interventions (31,6%), les études sur les facteurs de risque (18,4%) et les recherches fondamentales (7,2%).

4. **Approches thérapeutiques** : La chimiothérapie préventive était l'intervention la plus étudiée (52,3% des articles sur les interventions), suivie par la gestion des cas (27,8%), la lutte antivectorielle (12,5%) et les approches intégrées (7,4%).

3.3 Lacunes identifiées dans la recherche

L'analyse systématique des informations extraites a permis d'identifier plusieurs lacunes importantes dans la recherche sur les MTN en Afrique de l'Ouest :

1. **Déséquilibre géographique** : Certains pays fortement touchés par les MTN, comme la Sierra Leone et la Guinée, sont sous-représentés dans la littérature scientifique.

2. **Sous-représentation des déterminants sociaux** : Seulement 18,4% des études abordaient les facteurs socioéconomiques et comportementaux, malgré leur importance reconnue dans la transmission et le contrôle des MTN.

3. **Manque d'études longitudinales** : La majorité des recherches (76,8%) étaient transversales, avec peu d'études suivant l'évolution des interventions sur le long terme.

4. **Faible intégration des approches One Health** : Seulement 5,2% des études adoptaient une perspective One Health intégrant santé humaine, animale et environnementale.

5. **Limites méthodologiques** : 34,8% des études présentaient des échantillons de taille insuffisante ou des méthodes statistiques inadéquates pour les conclusions avancées.

4. Discussion

Notre étude démontre l'utilité et l'efficacité des approches basées sur l'IA pour l'extraction automatisée d'informations à partir de la littérature scientifique sur les MTN en Afrique de l'Ouest. Le système développé a atteint des performances satisfaisantes, particulièrement pour l'extraction des données épidémiologiques et des informations sur les interventions thérapeutiques.

Les performances légèrement inférieures pour l'extraction des facteurs de risque socioéconomiques reflètent la complexité inhérente à ces informations, souvent présentées de manière nuancée et contextuelle dans les textes scientifiques. Cette observation rejoint les conclusions de Johnson et al. (2021), qui ont souligné les défis spécifiques de l'extraction automatisée d'informations sociocontextuelles dans la littérature biomédicale.

L'analyse des tendances et lacunes dans la recherche sur les MTN en Afrique de l'Ouest révèle des déséquilibres importants qui méritent l'attention de la communauté scientifique et des bailleurs de fonds. La concentration des études dans certains pays, principalement anglophones, soulève des questions d'équité dans la production et l'utilisation des connaissances scientifiques. Comme l'ont souligné Okorie et al. (2020), cette disparité peut conduire à des interventions mal adaptées aux contextes locaux spécifiques.

La sous-représentation des déterminants sociaux et environnementaux dans la recherche sur les MTN constitue une lacune préoccupante. Plusieurs études ont démontré l'importance cruciale de ces facteurs dans la persistance des MTN malgré les interventions biomédicales (Bardosh, 2018). Nos résultats suggèrent la nécessité d'une approche plus holistique intégrant les dimensions biomédicales, environnementales et sociales dans la recherche sur les MTN.

Les limites méthodologiques identifiées dans une proportion significative d'études soulignent l'importance d'un renforcement des capacités en méthodologie de recherche dans la région. Cette observation rejoint les recommandations de Mitra et al. (2021) sur la nécessité d'investir dans la formation et le mentorat des chercheurs locaux.

Notre étude présente plusieurs limitations. Premièrement, malgré les performances satisfaisantes du système, l'extraction automatisée ne peut pas encore égaler la nuance et la profondeur de l'analyse humaine, particulièrement pour les informations contextuelles complexes. Deuxièmement, notre corpus était limité aux publications en anglais et français, excluant potentiellement des recherches pertinentes publiées dans d'autres langues. Troisièmement, nous nous sommes concentrés sur cinq MTN spécifiques, ce qui ne représente pas l'ensemble du spectre des maladies négligées affectant la région.

5. Conclusion

Cette étude démontre le potentiel des approches basées sur l'IA pour faciliter l'extraction et la synthèse d'informations à partir de la littérature scientifique sur les MTN en Afrique de l'Ouest. Le système développé offre un outil précieux pour les chercheurs, les professionnels de la santé et les décideurs politiques confrontés au défi de rester informés dans un contexte de prolifération rapide des publications scientifiques.

L'analyse des informations extraites a permis d'identifier des tendances et des lacunes importantes dans la recherche sur les MTN, notamment un déséquilibre géographique, une sous-représentation des déterminants sociaux et environnementaux, et des limitations méthodologiques dans une proportion significative d'études.

Ces résultats soulignent la nécessité d'une approche plus équilibrée et holistique de la recherche sur les MTN en Afrique de l'Ouest, intégrant les dimensions biomédicales, environnementales et sociales. Ils appellent également à un renforcement des capacités de recherche dans les pays sous-représentés et à une meilleure coordination des efforts de recherche à l'échelle régionale.

Les développements futurs de notre système incluront l'intégration de capacités multilingues plus avancées, l'amélioration de l'extraction des informations sociocontextuelles complexes, et l'extension à d'autres MTN affectant la région. Ces améliorations contribueront à une utilisation plus efficace et équitable des connaissances scientifiques pour lutter contre le fardeau des MTN en Afrique de l'Ouest.

Références

1. Bardosh, K. (2018). Towards a science of global health delivery: A socio-anthropological framework to improve the effectiveness of neglected tropical disease interventions. PLoS Neglected Tropical Diseases, 12(7), e0006537.

2. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (pp. 3615-3620).

3. Deribe, K., Simpson, H., Pullan, R. L., Bosco, M. J., Wanji, S., Weaver, N. D., ... & Cano, J. (2022). Mapping the global distribution of podoconiosis: Applying an evidence consensus approach. PLoS Neglected Tropical Diseases, 16(1), e0010036.

4. Hotez, P. J., Fenwick, A., & Molyneux, D. H. (2019). The new WHO neglected tropical diseases roadmap for 2021–2030. PLoS Neglected Tropical Diseases, 13(5), e0007408.

5. Johnson, A. E., Bulgarelli, L., & Pollard, T. J. (2021). Extracting socioeconomic information from biomedical literature: Challenges and opportunities. Journal of Biomedical Informatics, 117, 103771.

6. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

7. Mitra, A. K., Mawson, A. R., Biswas, A., & Reine, A. (2021). Strengthening research capacity in Africa: A case study of the Africa Center of Excellence in Public Health and Herbal Medicine. Global Public Health, 16(1), 120-136.

8. Okorie, P. N., Bockarie, M. J., Molyneux, D. H., & Kelly-Hope, L. A. (2020). Neglected tropical diseases mapping in Nigeria: Historical perspective and systematic review. PLoS Neglected Tropical Diseases, 14(10), e0008565.

9. Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., ... & Kohlmeier, S. (2020). CORD-19: The Covid-19 Open Research Dataset. ArXiv, abs/2004.10706.

10. World Health Organization. (2021). Ending the neglect to attain the Sustainable Development Goals: A road map for neglected tropical diseases 2021–2030. Geneva: World Health Organization.