



Vers l'automatisation des examens systématiques sur la vaccination au moyen d'un système d'extraction avancé fondé sur le traitement du langage naturel

David Begert¹, Justin Granek¹, Brian Irwin¹, Chris Brogley^{2*}

Résumé

Les décisions fondées sur des données probantes reposent sur le principe selon lequel tous les renseignements sur un sujet sont recueillis et analysés. Les examens systématiques permettent l'évaluation rigoureuse de différentes études selon les principes de PICO (population, intervention, contrôle, résultats). Toutefois, le fait de réaliser une révision est un processus généralement lent qui impose un fardeau important sur les ressources. Le problème fondamental est qu'il est impossible d'élargir l'approche actuelle à la réalisation d'un examen systématique pour faire face aux difficultés découlant d'un corpus important de données non structurées. Pour cette raison, l'Agence de la santé publique du Canada envisage l'automatisation de différentes étapes de synthèse des données visant à accroître les gains d'efficacité.

Dans le présent article, les auteurs présentent le résumé d'une version préliminaire d'un nouveau système d'apprentissage automatique fondé sur des avancements récents quant au traitement du langage naturel (TLN), comme BioBERT, où d'autres optimisations seront réalisées par l'entremise d'une nouvelle base de données de documents portant sur la vaccination. Le modèle de TLN optimisé obtenu et qui est au cœur de ce système peut déceler et extraire les champs relatifs aux principes de PICO des publications sur la vaccination avec une exactitude moyenne s'élevant à 88 % dans cinq classes de texte. La fonctionnalité est rendue possible par une interface Web directe.

Citation proposée : Begert D, Granek J, Irwin B, Brogley C. Vers l'automatisation des examens systématiques sur la vaccination au moyen d'un système d'extraction avancé fondé sur le traitement du langage naturel. *Relevé des maladies transmissibles au Canada* 2020;46(6):197– 203. <https://doi.org/10.14745/ccdr.v46i06a04f>

Mots-clés : automatisation, traitement du langage naturel, TLN, extraction de données, révisions systématiques, apprentissage automatique

Introduction

La médecine fondée sur des données probantes s'appuie sur des examens systématiques à titre de sources principales d'information sur une variété de sujets (1) parce qu'elles fournissent les évaluations et les analyses de données provenant de différentes études. La publication rapide d'examens systématiques pertinents et de qualité supérieure est idéale, mais en pratique, le processus de publication est généralement lent (1,2). Cette lenteur est largement attribuable à la quantité considérable de données non structurées qui doivent être filtrées. La synthèse des données clés provenant de plusieurs articles visant à effectuer un examen systématique demande beaucoup de temps pour les experts. Le temps requis pour

effectuer une publication dépasse souvent un an (3), et les coûts s'élèvent à des centaines de milliers de dollars (4).

Des méthodes d'apprentissage automatique pour l'automatisation des examens systématiques ont été définies dans le passé (1,5). Elles permettent l'élaboration de systèmes logiciels capables de déterminer automatiquement des types distincts de données textuelles, pourvu que le nombre d'exemples à partir desquels le système peut apprendre soit suffisant. Des méthodes de traitement du langage naturel (TLN) ont aussi été définies pour l'automatisation des examens systématiques (1,5) puisque ces méthodes permettent d'analyser

Cette œuvre est mise à la disposition selon les termes de la licence internationale Creative Commons Attribution 4.0



Affiliations

¹ Xtract AI, Vancouver, BC

² Centre de l'immunisation et des maladies respiratoires infectieuses, Agence de la santé publique du Canada, Ottawa, ON

*Correspondance :

christopher.brogley@canada.ca



des textes écrits au moyen d'approches statistiques ou fondées sur des connaissances, et permettent ainsi la définition d'éléments et de modèles clés.

Contexte

L'Agence de la santé publique du Canada envisage l'automatisation des aspects de la synthèse des données en fonction des principes de PICO (population, intervention, contrôle, résultats) afin d'éliminer certaines des barrières à l'obtention des résultats d'un examen systématique, notamment la mobilisation directe d'experts, le temps et les coûts. Pour ce faire, l'Agence a collaboré avec Xtract AI (Vancouver, Colombie-Britannique) afin d'élaborer un système qui s'appuie sur l'apprentissage automatique et le TLN de pointe pour cibler l'extraction des principes de PICO à partir d'articles portant sur la vaccination. La fonctionnalité de notre système est basé sur son apprentissage pour l'examen d'articles contenus dans une base de données formée de 249 articles portant sur la vaccination et qui renferment des éléments PICO marqués manuellement. Lorsqu'il est démontré que l'exactitude du système à extraire des données textuelles pertinentes relatives au PICO dans des articles jusque-là inconnus est élevée, il est possible de s'appuyer là-dessus pour réaliser des travaux qui seraient normalement réalisés manuellement. L'objectif consistait à élaborer le système permettant d'examiner de façon manuelle beaucoup moins d'articles.

Dans le présent article, les termes « modèle de TLN » et « système » sont utilisés. À proprement parler, le « modèle de TLN » renvoie à la collection de méthodes d'apprentissage automatique et de TLN employées pour traiter des documents relatifs à la vaccination. Le « système » renvoie au modèle de TLN doté d'une interface utilisateur sur le Web qui permet la convivialité du modèle (voir la **figure 1**). Le modèle de TLN est fondamental au système et réalise des tâches d'extraction et de prédiction, mais le terme « système » est utilisé plus souvent dans le présent article en raison des liens de dépendance entre le modèle de TLN et les composantes du système.

Figure 1 : Capture d'écran de la page d'accueil de l'interface Web du système d'extraction fondé sur la TLN montrant une interrogation de recherche

The screenshot shows the TLN web interface with the following search parameters:

- Population:** Human papillomavirus OR HPV
- Intervention:** Live
- Outcome:** Safety

At the bottom, it says "Almost 11 Results" and has a "RESEARCH" button.

Abréviation : TLN, traitement du langage naturel
(Figure en anglais seulement)

Le système a été conçu pour extraire automatiquement 27 catégories de texte fondées sur le principe de PICO à parti

d'articles inconnus portant sur la vaccination. Pour mettre à l'essai cette version initiale du système, le rendement a été mesuré en fonction de la capacité du système à repérer le texte portant sur le vaccin principal, le type d'étude, l'état de santé de la population et le résultat, ainsi que le texte descriptif du résultat dans 40 articles initialement inconnus portant sur la vaccination. Nous avons considéré ces cinq catégories comme des moyens appropriés de mesurer le rendement initial de l'extraction de textes relatifs au principe de PICO parce qu'elles englobent une vaste gamme de textes, et avons présumé que les catégories étaient plus variées que bon nombre des autres catégories de texte que le système peut extraire. Le **tableau 1** montre les cinq catégories principales utilisées pour mesurer le rendement du système.

Tableau 1 : Catégories d'extraction de texte principales, avec exemples et score d'exactitude

Catégorie de texte	Description	Exemple extrait	Résultat d'exactitude, en pourcentage (%)
Vaccin principal	De quel vaccin cet article traite-t-il?	Vaccin quadrivalent contre le virus du papillome humain, ou vaccin heptavalent antipneumococcique	90
Type d'étude	De quel type d'étude s'agit-il?	Essai randomisé, contrôlé par placebo	92,5
État de santé de la population	État de santé de la population	VIH positif	85
Résultat – effet indésirable	Un effet indésirable décrit dans l'article	[traduction] « L'effet indésirable le plus courant était la douleur; les autres effets indésirables courants étaient de nature neurologique, gastro-intestinale ou épidermique. »	85
Résultat – phrase descriptive	Toutes les phrases importantes relatives aux résultats	Résultat concernant l'innocuité et l'immunogénicité, e.g. « L'EV était de 93,0 % (85,1–97,3) dans l'ECV-PE (Tableau S1). »	87,5

Au moment de la rédaction, l'exactitude moyenne du système dans ces cinq catégories de texte principales était de 88 %. Le tableau 1 présente le résumé des résultats de l'exactitude. Afin d'atteindre ce niveau d'exactitude, le système a appris de 209 exemples sur les 249 qui se trouvent dans notre base de données de documents. La tâche d'automatisation a été mise à l'essai au moyen des 40 autres exemples se trouvant dans la base de données de documents. Deux versions de ces 40 documents d'essai existaient; la première version a été



marquée par un expert qui a analysé les documents pour des cas des 27 catégories de texte, et l'autre n'était pas marquée. Afin de mettre à l'essai la capacité d'extraire des textes relativement aux principes de PICO du système, le système a traité les versions non marquées des documents d'essai. Les extraits automatisés du système ont ensuite été comparés au texte des documents marqués de l'expert. L'extraction par le système d'un texte de qualité comparable à celle de l'un des textes marqués de l'expert était comptée comme un succès ayant contribué au score d'exactitude moyen de 88 %.

Nous présumons qu'il est possible d'obtenir des résultats semblables à la mise à l'essai de la tâche d'extraction de textes relatifs aux principes de PICO sur de nombreux autres documents à mesure que les travaux d'élaboration se poursuivent et que d'autres articles portant sur la vaccination marqués par des experts s'ajoutent à la base de données de documents.

Le présent article décrit l'approche technique à l'égard de l'élaboration du modèle de TLN et le processus par lequel le modèle de TLN a appris à réaliser cette tâche. Il présente ensuite une analyse plus détaillée de l'exactitude au moyen de plusieurs mesures du rendement.

Approche technique

Le modèle de TLN a été conçu en tant que modèle d'extraction de séquences à plusieurs classes. Un modèle d'extraction de séquences à plusieurs classes fonctionne en traitant le texte intégral d'un document auparavant inconnu, puis en extrayant les séquences de texte qui correspondent à chacune des catégories de texte qu'il a appris à extraire. Dans ce cas, selon les domaines marqués par des experts dans 209 documents portant sur la vaccination des 249 documents disponibles, le système extrait jusqu'à 27 catégories de texte. Chaque catégorie peut être comprise en double.

BioBERT, la variante de langue biomédicale de Représentations des codeurs bidirectionnels des transformateurs (BERT) pour l'exploration de texte biomédicale a été employée comme fondement du modèle de TLN afin d'accroître le rendement (6,7). En tant qu'avancée récente dans le TLN (7), BERT est essentiellement un modèle qui a traité un corpus de texte énorme et qui a appris de ce corpus. BERT, et les variantes comme BioBERT, sont de plus en plus utilisés comme base des nouveaux systèmes logiciels d'apprentissage automatique et de TLN. Leur utilisation a entraîné une augmentation importante de l'exactitude de ces systèmes. L'utilisation de BioBERT, en tant que variante de langue biomédicale du modèle original BERT, était considérée comme la plus appropriée dans le présent travail.

Création d'un ensemble de données

Les données d'apprentissage de départ employées pour le modèle de TLN du système étaient le corpus de Traitement automatique du langage naturel sur les données probantes à des fins d'extraction relative aux principes de PICO. Le corpus de TLN sur les données probantes renferme 5 000 résumés annotés d'articles médicaux décrivant des essais cliniques randomisés contrôlés (8). Les annotations du corpus de TLN sur les données probantes marquent les passages principaux de ces résumés, comme la description des participants (e.g. groupe d'âge, état), le type d'intervention (e.g. pharmacologique) et les résultats (e.g. douleur, effets indésirables, ou mortalité). Ces champs ou catégories de texte ont été déterminés par les personnes responsables d'élaborer le corpus de TLN sur les données probantes. Une description détaillée de la méthodologie d'annotation employée pour le corpus de TLN sur les données probantes se trouve au <https://ebm-nlp.herokuapp.com/annotations> (en anglais seulement).

Le corpus TLN sur les données probantes a été employé pour « enseigner » à notre système à travailler avec ces champs, parce qu'il fallait être en mesure d'extraire les mêmes champs ou catégories de texte. Toutefois, puisque le corpus de TLN sur les données probantes annote un petit nombre de champs de texte qui ne portent pas sur la vaccination, la base de données contenant les 249 documents portant sur la vaccination a été générée par l'annotation de données sur place. Le temps exact qu'il a fallu pour intégrer et marquer chaque article dans la base de données de documents n'a pas été consigné, mais le temps qu'il faut pour ajouter et marquer un document équivaudrait généralement au temps requis pour examiner manuellement un article qui pourrait faire l'objet d'un examen systématique. Toutefois, les efforts à cette étape signifieront au bout du compte moins de participation humaine au processus général d'examen systématique, puisque l'exactitude du système devrait montrer qu'il est capable de réaliser la tâche lui-même avec fiabilité.

À l'heure actuelle, le modèle traite tous les types de documents, sans arrêt, mais ne rend aucun résultat pour ceux qui ne contiennent aucun texte reconnaissable semblable aux exemples d'apprentissage. Nous prévoyons que les versions ultérieures du système sont en mesure de repérer les documents qui ne portent pas sur la vaccination par manque de résultat. Au moyen de l'outil d'annotation BRAT (9), les sections relatives au titre, au résumé, aux méthodes et aux résultats des 249 articles de la nouvelle base de données ont été annotées. Les articles ont été obtenus auprès de PubMed Central au moyen d'une recherche par mot-clé. Les champs qui n'étaient pas compris dans le corpus de TLN sur les données probantes, c'est-à-dire, les équivalents anglais de « type d'études », « vaccin principal », « résultats – effets indésirables », « résultat – phrase descriptive » et « état de santé de la population », ont été annotés. Le modèle de TLN comprend ces catégories de texte parce qu'elles ont été



marquées manuellement dans la nouvelle base de données de documents portant sur la vaccination. Des exemples de ces catégories sont énumérés au tableau 1.

Processus d'apprentissage et essai de l'exactitude du système

Le modèle de TLN fondé sur BioBERT a été lancé à partir du modèle de BioBERT original, et l'apprentissage ultérieur a eu lieu au moyen du corpus de TLN sur les données probantes. Le modèle a ensuite poursuivi son apprentissage au moyen de la base de données de documents portant sur la vaccination des présents chercheurs. Alors que 209 échantillons de la base de données de documents ont été utilisés au cours de cette dernière étape d'apprentissage, 40 articles marqués ont été exclus pour mettre le rendement du système à l'essai.

Nous avons ciblé la mise à l'essai de cinq catégories de texte principales dans cet article, mais le système n'a besoin d'extraire et de produire qu'une petite quantité de textes simples pour mettre à l'essai plusieurs des catégories de texte qu'il est possible d'extraire (voir le **tableau 2** pour des exemples). En ce qui concerne les autres catégories, le système ne produit pas le texte extrait lorsqu'il le trouve, mais attribue plutôt l'indicateur « vrai » (true) ou « faux » (false) au contenu de l'article de recherche. Le corpus de TLN sur les données probantes utilisé dans la formation initiale du modèle de TLN contient des exemples de quelques-unes de ces catégories. Il était prévu que ces exemples permettent au modèle de TLN de détecter n'importe quoi qui se trouvait dans le corpus de TLN sur les données probantes avec exactitude.

Tableau 2 : Catégories d'extraction de texte supplémentaires avec exemples^a

Catégorie de texte ^b	Exemple extrait
Innocuité	Vrai
Efficacité	Vrai
Pharmacologique ^b	Vaccin quadrivalent contre le virus du papillome humain (types 6, 11, 16 et 18) recombinant intramusculaire de 0,5 mL
État	VIH positif
Pays	Mali
Âge	Adultes âgés d'au moins 27 ans
Taille de l'échantillon	535
Sexe	Femmes

^a Le système extrait aussi les équivalents anglais de « cible de l'agent pathogène du vaccin », « immunogénicité », « immunodéprimé », « en santé », « non vivant », « sans adjvant », « vivant », « sexe », « grossesse », « identifiant numérique d'objet », « score », « résumé », « méthodes », « résultats », « version_là », et « mot-clé », mais ils ne sont pas énumérés à des fins de concision.

^b Pour certaines catégories de texte, e.g. pharmacologique, plusieurs milliers d'exemples d'apprentissage ont été montrés au système par l'entremise d'un corpus de traitement du langage naturel de médecine fondée sur des données probantes

Évaluation

Afin d'évaluer le rendement de la prédiction des cinq catégories de texte, des mesures de rendement courantes de systèmes d'apprentissage automatique et de TLN ont été calculées. Elles comprennent la précision, le rappel et le score F1 (une autre mesure d'exactitude fondée sur la précision et le rappel). Le nombre de réussites et d'erreurs ainsi que le pourcentage d'exactitude général ont été calculés. Toutes les mesures de rendement sont définies au **tableau 3**. Les mesures énumérées s'appliquent aux documents dans leur intégralité. Le nombre de « vrais positifs », de « vrais négatifs », de « faux positifs » et de « faux négatifs » est utilisé pour calculer le nombre de réussites et d'erreurs. Le nombre de vrais positifs, de vrais négatifs, de faux positifs et de faux négatifs est le résultat du niveau d'exactitude de l'extraction de textes effectuée par le système. Les pourcentages d'exactitude généraux présentés dans le tableau 1 sont calculés en fonction des réussites et des erreurs.

Tableau 3 : Définition des mesures de rendement du modèle

Mesure de l'exactitude	Méthodes ou formule de calcul	Signification des résultats
VP	Nombre de documents	Documents correctement repérés Les résultats élevés sont préférables
VN	Nombre de documents	Documents correctement repérés Les résultats élevés sont préférables
FP	Nombre de documents	Documents incorrectement repérés Les faibles résultats sont préférables
FN	Nombre de documents	Documents incorrectement repérés Les faibles résultats sont préférables
Réussites	VP + VN	Somme des documents correctement repérés Les résultats élevés sont préférables
Erreurs	FP + FN	Somme des documents incorrectement repérés Les faibles résultats sont préférables
Pourcentage d'exactitude général	$\frac{(VP + VN)}{(VP + VN + FP + FN)} * 100$	Exactitude générale allant de 0 % à 100 %
Précision	$\frac{VP}{(VP + FP)}$	Mesure de la recherche documentaire Résultats allant de 0,0 à 1,0; plus un résultat est élevé, meilleur il est

Tableau 3 : Définition des mesures de rendement du modèle (suite)

Mesure de l'exactitude	Méthodes ou formule de calcul	Signification des résultats
Rappel	$\frac{VP}{(FN + VP)}$	Mesure de la recherche documentaire Résultats allant de 0,0 à 1,0; plus un résultat est élevé, meilleur il est
Score [F1]	$2*(P*R) / (P+R)$	Résultats allant de 0,0 à 1,0; plus un résultat est élevé, meilleur il est.

Abréviations : FN, faux négatif; FP, faux positif; P, précision; R, rappel; VN, vrai négatif; VP, vrai positif

Il est important de noter que le système extrait du texte de format libre. Dans ce cas, la longueur et le contenu d'une prédition extraite peuvent varier grandement de la séquence correcte et marquée du texte d'un document d'essai (pour connaître des exemples, consulter la **figure 2**). Puisque c'est le cas, il est très important de définir clairement ce qui constitue une réussite (vrai positif ou vrai négatif). Par exemple, si la prédition extraite est « vaccin antipneumococcique » alors que la bonne réponse est « vaccin heptavalent antipneumococcique », le résultat pourrait être catégorisé comme une erreur si le système juge « heptavalent » comme étant trop important pour l'exclure du texte sur le « vaccin principal » de ce document.

Figure 2 : Capture d'écran de l'interface Web du modèle de TLN présentant une prédition extraite^a

Age	Vaccine	Safety	EXTRACT
19-65	HPV-16/18 AS04-adenovirus vaccine	true	
Condition	Came from results	Immunogenicity	EXPORT
healthy Chinese women	Safety	The safety outcomes are presented in Table 3. A total of 126 (2.1%) women reported at least one SAE in the current follow-up (Table 3). One SAE (epigastric pain) was considered by the investigator as possibly related to vaccination (the group assignment is not reported to avoid unblinding). The most common SAE was appendicitis, reported by 16 (0.3%) women. Five women experienced at least one SAE leading to premature study discontinuation, but none of these SAEs were considered by the investigator as possibly related to vaccination. One fatal SAEs (suicide) assessed by the investigator as not related to vaccination was reported during the study (the group assignment is not reported to avoid unblinding). A total of 361 women (6.0%) reported at least one medically significant AE (Table 3), the most common was pelvic inflammatory disease (22 women, 4%). Overall, NOCDs and NOADs were rarely reported (0.3% and 0.06% of women, respectively) (Table 3). The most common NOCD was allergic dermatitis, reported by three women (0.1%) from the vaccine group and four women (0.1%) from the control group. Among the four NOADs reported, psoriasis was reported by two women (one in each group), and hyperthyroidism and VIIIth nerve paralysis were reported in one woman each (blinded data). None of the NOADs reported were assessed by the investigator as being related to vaccination.	
Health Status			
Healthy			
Sample Size	6081 women were enrolled in the study. 6051 were included in the TVC (3026 in the vaccine group and 3025 in the control group), and 5972 (N = 6051)	false	
		<p>The most common SAE was appendicitis, reported by 16 (0.3%) women.</p>	

Abréviation : TLN, traitement du langage naturel

La prédition extraite est soulignée dans l'interface Web et est encerclée dans le panneau des résultats principaux de l'interface Web (Figure en anglais seulement)

Lorsque le problème associé aux définitions n'est pas réglé, les mesures de rendement sont dépourvues de contexte. À cette première étape, les prédictions extraites sont inspectées manuellement pour évaluer l'exactitude en fonction des critères définis. L'application de critères d'exactitude aux prédictions extraites n'avait pas été effectuée au moment de la rédaction, mais il est prévu qu'elle soit automatisée dans les versions

ultérieures du système. Ces critères d'exactitude ont été imposés aux cinq catégories de texte principales afin qu'une prédition extraite compte comme une valeur prédictive positive (VPP). Puisqu'un document peut contenir de nombreuses prédictions extraites, il faut un grand nombre de VPP correctes pour que la tâche de catégorisation du document soit comptée comme une réussite entière, ou un vrai positif. La façon la plus facile de penser à une VPP par rapport à un vrai positif est que si VPP se trouve au niveau textuel, alors qu'un vrai ou un faux positif est attribué à un document entier en fonction du nombre de VPP. Le **tableau 4** en définit les critères généraux.

Tableau 4 : Critères établis sur mesure pour les vrais positifs prévus et les formules utilisées pour calculer les vrais positifs et les vrais négatifs

Critères 1 de VPP ^a	Critères 2 de VPP ^a	Exigence pour obtenir une réussite (vrai positif) ^{b,c}	Exigence pour obtenir une réussite (vrai négatif) ^{b,c}
La PE comprend au moins une des réponses marquées OU bonnes réponses non marquées, y compris toutes les données importantes	La PE ne peut pas comprendre trop de données inutiles	$\frac{VPP}{(VPP + FPP + FNP)} > 0,8$ $> 0,5 \text{ pour les catégories de texte « Résultat » attribuable à la subjectivité}$	$VPP + FPP + FNP = 0$

Abréviations : FNP, faux négatif prévu; FPP, faux positif prévu; PE, prédictions prévues; VPP, vrai positif prévu

⁸ Les VPP sont déterminés en fonction des prédictions extraites effectuées par le système à partir du texte du document après une évaluation par un spécialiste d'examen. De nombreuses prédictions extraites par catégories de texte pourraient compter comme VPP ou autre.

^b Les EPP et les FNP sont tout aussi importants, mais leur nombre varie selon l'article.

• Les FPP et les FNP sont tout aussi importants, mais leur nombre varie selon l'article
• Si les exigences d'un vrai positif ou d'un vrai négatif ne sont pas respectées, un faux positif peut être attribué lorsqu'un $FPP > FNP$, ou un faux négatif peut être attribué lorsqu'un $FPP < FNP$

Tous les résultats de l'essai d'exactitude sur 40 documents sont présentés dans le **tableau 5**.

Le système a continuellement donné un bon rendement. Le taux de réussite était élevé et le taux d'erreur était faible, ce qui démontre l'efficacité générale de la tâche d'extraction en fonction des principes de PICO. Il était impossible d'assurer l'équilibre entre les exemples d'essai négatifs et positifs pour chaque catégorie en raison des données limitées, quoiqu'un équilibre ne reflète pas nécessairement le rendement réel.

Par exemple, le nombre de vrais négatifs pour la catégorie « population - état de santé » était beaucoup plus élevé parce que les articles de cette catégorie ne contenaient aucun texte qui aurait pu être extrait. Néanmoins, ce déséquilibre a entraîné un problème; les résultats d'exactitude de ces catégories de texte peuvent être biaisés en faveur du groupe (soit positif ou négatif) qui présente le plus d'exemples d'essai. Toutefois, il est prévu que ces résultats demeurent élevés puisque ce problème se règle par l'élargissement de la base de données de documents portant sur la vaccination.



Tableau 5 : Tous les résultats de rendement des cinq catégories principales

Mesure du rendement	Catégories principales de texte				
	Vaccin principal	Type d'étude	Résultat – effet indésirable	Résultat – phrase descriptive ^a	Population – État de santé
Résultat F1	0,8824	0,947	0,727	0,9315	0,75
Précision ^c	1	0,964	1	0,9444	0,9
Rappel ^d	0,7895	0,931	0,571	0,9189	0,643
VP ^e	15	27	4	34	9
VN ^e	21	10	33	1	25
FP ^f	0	1	0	2	1
FN ^f	4	2	3	3	5
Réussites (VP ou VN) ^g	36	37	37	35	34
Erreurs (FP ou FN) ^h	4	3	3	5	6
Pourcentage d'exactitude de la catégorie, % ⁱ	90	92,5	92,5	87,5	85

Abréviations : FN, faux négatif; FP, faux positif; P, précision; R, rappel; VN, vrai négatif; VP, vrai positif

^a Les catégories de texte présentaient des exemples de positifs et de négatifs visiblement déséquilibrés. L'exactitude générale peut être faussée en faveur du groupe qui comprend le plus d'exemples. Toutefois, un déséquilibre entre des exemples peut survenir au sein des données réelles

^b Une autre mesure d'exactitude fondée sur la précision et le rappel. Les résultats varient entre 0,0 et 1,0, plus un résultat est élevé, meilleur il est

^c Les résultats varient entre 0,0 et 1,0, plus un résultat est élevé, meilleur il est

^d Les résultats varient entre 0,0 et 1,0, plus un résultat est élevé, meilleur il est

^e Mesure des documents correctement repérés. Plus un résultat est élevé, meilleur il est

^f Mesure des documents incorrectement repérés. Les faibles résultats sont préférables

^g Mesure de la somme des documents correctement repérés. Les résultats élevés sont préférables

^h Somme des documents incorrectement repérés. Les faibles résultats sont préférables

ⁱ Exactitude générale. Les résultats varient de 0 % à 100 %

Comme le montre la figure 1, les résultats d'extraction relatifs aux principes de PICO sont accessibles au moyen d'une interface Web conviviale. La **figure 3** présente l'exemple d'une recherche effectuée affichant les résultats d'un grand nombre des catégories de texte.

Limites du système

Comme indiqué précédemment, il a été impossible d'obtenir des groupes d'exemple d'essai positifs et négatifs équilibrés parce que les données étaient limitées. Ce problème pourrait biaiser les résultats d'exactitude en faveur du groupe qui comprend le plus d'exemples d'essai. Toutefois, il est important de noter que ces exemples de positifs et de négatifs peuvent présenter un déséquilibre dans les documents inconnus dans les situations réelles.

L'élaboration de la nouvelle base de données de documents portant sur la vaccination a nécessité une certaine participation d'experts, c'est-à-dire qu'elle n'était pas automatisée. Il a aussi fallu faire des travaux manuellement pour examiner les prédictions extraites du texte du document pour apporter des corrections. Ces travaux effectués manuellement au départ

Figure 3 : Exemples de résultats d'extraction associés au VPH après l'envoi des termes de recherche

Title	Countries	Age	Vaccine	Safety	EXTRACT
Efficacy, immunogenicity, and safety of the HPV-16/18 AS04 adjuvanted vaccine in Chinese women aged 18–25 years: event-triggered analysis of a randomized controlled trial	China	19-65	HPV-16/18 AS04-adjuvanted vaccine	true	EXPORT
StudyType	Condition	healthy Chinese	HPV-16/18 vaccine or Al(OH)3	Immunogenicity	PUBLISHED PAPER
Gender	Gender	women	Adjuvanted	Efficacy	Score
Health Status	Non-Adjuvanted	Healthy	false	Main Outcomes	122.97265625
Authors	Sample Size	Feng-cai Zhu Shang-Ying Hu Ying Hong Yue-Mei Hu Xun Zhang Yi-Ju Zhang Qin-Jing Pan Wen-Hua Zhang Fang-Hui Zhao	6081 women were enrolled in the study; 6051 were included in the TVC (3026 in the vaccine group and 3025 in the control group), and 5972	In the TVC-E, VE was 83.2% (24.7–98.2) (Table S1). The HPV-16/18 baseline serostatus	
Live	Non-Live	false	false	Outcome Safety	The most common SAE was appendicitis, reported by 16 (0.3%) women.

(Figure en anglais seulement)

sont en fin de compte requis pour permettre l'automatisation ultérieurement.

Prochaines étapes

Au moment de la rédaction, le système était toujours en cours d'élaboration. Les prochains travaux consisteront entre autres à augmenter le nombre de documents marqués dans la nouvelle base de données de documents portant sur la vaccination afin d'améliorer l'apprentissage du système. De plus, l'interface Web continuera d'être peaufinée. Idéalement, le système repérera les documents qui ne traitent pas de la vaccination, et arrêtera immédiatement de les traiter pour empêcher même le court délai qu'il faut pour analyser un texte à l'heure actuelle. De plus, un système connexe, conçu pour englober tous les documents de nature biomédicale (fondés sur la même technologie dont il est question dans le présent article), est en cours d'élaboration.

Enfin, il faudra mettre à l'essai l'efficacité d'un système plus exhaustif en consultation avec les responsables des décisions en matière de santé publique.

Conclusion

Les auteurs ont décrit un système fondé sur l'apprentissage automatique et les méthodes de TLN qui vise à automatiser les travaux manuels répétitifs de l'analyse de documents assujettis au processus d'examen systématique. Ce système cible les documents portant sur la vaccination seulement. Les résultats prometteurs quant au rendement dans le cadre des présents travaux préliminaires montrent qu'il est possible de s'éloigner des approches manuelles et laborieuses à l'égard des examens systématiques et d'avancer vers des systèmes automatisés dans le but d'éliminer éventuellement (ou réduire de façon considérable) la participation de spécialiste aux tâches répétitives du processus.



La conception générale du système présente une manière prometteuse pour les responsables de décisions en matière de santé publique d'utiliser des données non structurées plus rapidement et de façon économique lorsqu'ils prennent des décisions et appliquent les principes de la médecine fondée sur des données probantes. La contribution unique des présents chercheurs à ce domaine est la convivialité du système au moyen d'une interface Web directe jumelée au rendement attribuable à l'application de méthodes d'apprentissage automatique et de TLN de pointe à notre nouvelle base de données de documents portant sur la vaccination.

Déclaration des auteurs

D. B. — Conception et mise en œuvre du système, mise à l'essai, rédaction, examen, révision
J. G. — Conception et mise en œuvre du système, mise à l'essai, rédaction, examen, révision
B. I. — Conception et mise en œuvre du système, mise à l'essai, rédaction, examen, révision
C. B. — Exigences en matière de mise à l'essai, rédaction, examen, révision

Conflits d'intérêts

Aucun.

Remerciements

Les auteurs souhaitent souligner la contribution de Xtract AI et des représentants suivants de l'Agence de la santé publique du Canada : O. Baclic, M. Tunis, M. Laplante, K. Young, H. Swerdfeger, C. Doan.

Financement

Le travail a reçu l'appui de l'Agence de la santé publique du Canada.

Références

1. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. *Syst Rev* 2014;3(74):1-15. [DOI](#) [PubMed](#)
2. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. *Syst Rev* 2015;4(78):1-16. [DOI](#) [PubMed](#)
3. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* 2017;7(2):e012545. [DOI](#) [PubMed](#)
4. Lau J. Editorial: systematic review automation thematic series. *Syst Rev* 2019;8(70):1-2. [DOI](#) [PubMed](#)
5. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *BMC Syst Rev* 2019;8(163):1-10. [DOI](#) [PubMed](#)
6. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020 Feb;36(4):1234-40. [DOI](#) [PubMed](#)
7. Devlin J, Chang MW, Lee K, Toutanova K. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); BERT: pre-training of deep bidirectional transformers for language understanding; 2019-06. NAACL-HLT. Minneapolis, Minnesota (US): Association for Computational linguistics; 2019;4171-86. [DOI](#)
8. Nye B, Li JJ, Patel R, Yang Y, Marshall I, Nenkova A, Wallace B. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. Melbourne (AUS): Association for Computational linguistics; July 15-20, 2018. [DOI](#)
9. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. BRAT: A web-based tool for NLP-assisted text annotation. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012 April. Avignon (FR): EACL.