Accelerating Evidence Synthesis in Observational Studies: A Living NLP-Assisted Intelligent Systematic Literature Review System

Jingcheng Du, Dong Wang

Submitted to: JMIR Medical Informatics
on: November 17, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Accelerating Evidence Synthesis in Observational Studies: A Living NLP-Assisted Intelligent Systematic Literature Review System

Jingcheng Du[1] PhD; Dong Wang[1]

Corresponding Author:
Dong Wang

Abstract

Background: Systematic literature review (SLR), a robust method to identify and summarize evidence from published sources, is considered as a complex, time-consuming, labor-intensive and expensive task.

Objective: To present a solution based on Natural Language Processing (NLP) that accelerates and streamlines the SLR process for observational studies using real world data.

Methods: We followed an agile software development and iterative software engineering methodology to build a customized intelligent end-to-end living NLP-assisted solution for observational SLR tasks. Multiple machine learning-based NLP algorithms were adopted to automate article screening and data element extraction processes. The NLP prediction results can be further reviewed and verified by domain experts, following the human-in-the-loop design. The system integrates Explainable AI (XAI) to provide evidence to NLP algorithms and add transparency to extracted literature data elements. The system was developed based on three existing SLR projects of observational studies, including the epidemiology studies of human papillomavirus-associated diseases, the disease burden of pneumococcal diseases, and cost-effectiveness studies of pneumococcal vaccines.

Results: Our Intelligent SLR Platform, covers major SLR steps, including study protocol setting, literature retrieval, abstract screening, full-text screening, data element extraction from full-text articles, results summary, and data visualization. The NLP algorithms have achieved 0.86 to 0.90 accuracy scores on article screening tasks (framed as text classification tasks) and 0.57 to 0.89 macro-average F1 scores on data element extraction tasks (framed as named entity recognition tasks).

Conclusions: Cutting-edge NLP algorithms expedite SLR for observational studies, thus allowing scientists to have more time to focus on the quality of data and the synthesis of evidence in observational studies. Aligning the living systematic literature review concept, the system has the potential to update literature data and enable scientists to easily stay current with the literature related to observational studies prospectively and continuously.

(JMIR Preprints 17/11/2023:54653)
DOI: https://doi.org/10.2196/preprints.54653

Original Manuscript

Accelerating Evidence Synthesis in Observational Studies: A Living NLP-Assisted Intelligent Systematic Literature Review System

Frank J. Manion, PhD[1], Jingcheng Du, PhD[1], Dong Wang, PhD[2*], Long He, MS[1], Bin Lin, MS[1], Jingqi Wang, PhD[1], Siwei Wang, MS[1], David Eckels[2], Jan Cervenka[2], Peter C. Fiduccia, PhD[2], MBA[2], Nicole Cossrow, PhD[2], Lixia Yao, Ph.D[2]

[1]Intelligent Medical Objects, Houston, TX, USA

[2]Merck & Co., Inc, Rahway, NJ, USA

*Corresponding author

Name: Dong Wang, PhD

Email: dong.wang10@merck.com

Keywords: machine learning, deep learning, natural language processing, systematic literature review

Abstract

Background

Systematic literature review (SLR), a robust method to identify and summarize evidence from published sources, is considered as a complex, time-consuming, labor-intensive and expensive task.

Objective

To present a solution based on Natural Language Processing (NLP) that accelerates and streamlines the SLR process for observational studies using real world data.

Methods

We followed an agile software development and iterative software engineering methodology to build a customized intelligent end-to-end living NLP-assisted solution for observational SLR tasks. Multiple machine learning-based NLP algorithms were adopted to automate article screening and data element extraction processes. The NLP prediction results can be further reviewed and verified by domain experts, following the human-in-the-loop design. The system integrates Explainable AI (XAI) to provide evidence to NLP algorithms and add transparency to extracted literature data elements. The system was developed based on three existing SLR projects of observational studies, including the epidemiology studies of human papillomavirus-associated diseases, the disease burden of pneumococcal diseases, and cost-effectiveness studies of pneumococcal vaccines.

Results

Our Intelligent SLR Platform, covers major SLR steps, including study protocol setting, literature retrieval, abstract screening, full-text screening, data element extraction from full-text articles, results summary, and data visualization. The NLP algorithms have achieved 0.86 to 0.90

accuracy scores on article screening tasks (framed as text classification tasks) and 0.57 to 0.89 macro-average F1 scores on data element extraction tasks (framed as named entity recognition tasks).

Conclusion

Cutting-edge NLP algorithms expedite SLR for observational studies, thus allowing scientists to have more time to focus on the quality of data and the synthesis of evidence in observational studies. Aligning the living systematic literature review concept, the system has the potential to update literature data and enable scientists to easily stay current with the literature related to observational studies prospectively and continuously.

INTRODUCTION

Systematic literature reviews (SLRs) are widely recognized as a robust method to identify and summarize evidence from published sources. [1] However, conducting an SLR can be a complex, time-consuming, labor-intensive and expensive task, depending on the breadth of the topic, level of granularity or resolution of the review needed.[2,3] One recent study estimated the time and cost required to conduct an SLR can be as high as 1.72 person-years of scientist effort and approximately $140,000 per review.[4] Because SLRs are so resource intensive, it is difficult to stay up to date and once an SLR is complete and new literature is published, the SLR may become incomplete and obsolete as time goes by.

Natural Language Processing (NLP) refers to Artificial Intelligence technologies that can extract structured information from textual documents such as medical charts, lab results and many other types of unstructured text. NLP has significantly advanced a variety of biomedical applications in recent years. There is considerable community interest in using AI such as machine learning and NLP to improve automation in aspects of literature reviews.[2,5–7] For example, Thomas et al used NLP to identify randomized controlled trials for Cochrane reviews, and Wallace et al developed methods to extract sentences from literature related to clinical trial reports. There are also some SLR management software, such as Raynan.ai[8], which leverage NLP to expedite certain SLR steps, including article screening.

Despite these existing efforts, there is a lack of systematic and integrated NLP solutions for SLR to cover its full aspects, preventing the wide adoption of such tools in SLR projects.

Thus, in this study, we evaluated an intelligent SLR system (hereinafter referred as ISLR) for observational SLR tasks. The use of NLP improves efficiency while the human-in-the-loop approach improves accuracy and reduces errors. The system uses cutting-edge NLP tools that employ machine learning (ML) and deep learning (DL) approaches to expedite the time-consuming processes involved in an SLR by making a series of learned recommendations to the end user. The purpose of this study is to evaluate an AI tool that accelerates and streamlines the SLR process and to demonstrate the validity of this tool in three use cases.

METHODS

Workflow and System Architecture

ISLR has two major views that target two types of users in the observational studies in an SLR lifecycle: 1) An intelligent SLR workbench for literature reviewers who conduct routine literature reviews, 2) A living literature data dashboard for researchers and analysts who focus on analyzing SLR data and keep up to date on new evidence. Figure 1 shows the overview architecture including the two major views and data flow of the SLR system. ISLR integrates AI technologies and an SLR workflow management system to support literature collection, screening, and data extraction. The living literature dashboard continuously searches and updates the SLR allowing users to interactively navigate the updated literature and develop new insights.

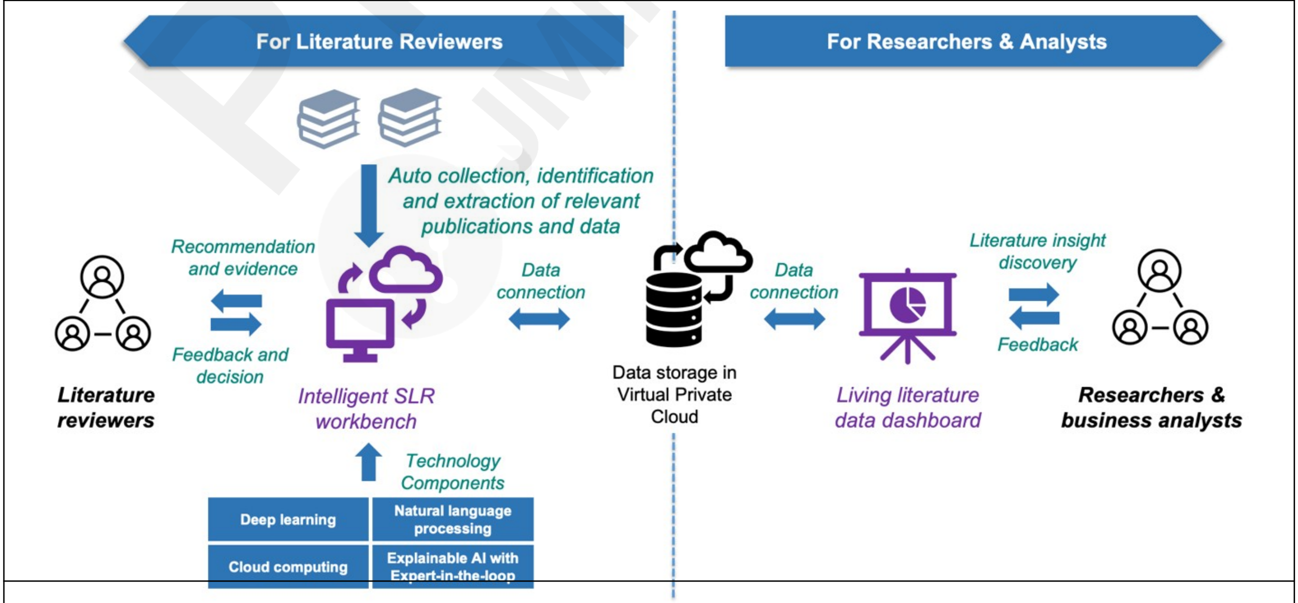

For Literature Reviewers
For Researchers & Analysts
Auto collection, identification and extraction of relevant publications and data
Recommendation and evidence
Feedback and decision
Literature reviewers
Intelligent SLR workbench
Data connection
Data storage in Virtual Private Cloud
Data connection
Living literature data dashboard
Literature insight discovery
Feedback
Researchers & business analysts
Technology Components
Deep learning
Natural language processing
Cloud computing
Explainable AI with Expert-in-the-loop

Figure 1: Overall data flow architecture of ISLR demonstrating the two major views.

Reliable NLP systems depend heavily on the development of a reasonable workflow, user interfaces, and high-performance NLP algorithms. To develop the system and define the system workflow and user interfaces, we collaborated with end users who are experts in SLR using an iterative approach that employed industry-standard agile methodology. The team identified six major functional areas that were essential for the application: 1) protocol specification assistance, 2) literature search and indexing, 3) abstract screening with NLP assistance, 4) support for full-text searching, uploading, and screening, 5) full-text data element extraction using NLP assistance to identify and extract relevant data elements from full-text and embedded tables, and 6) literature data visualization to enable users to assess the SLR results and perform data discovery. Figure 2 shows the system workflow and the embedded NLP services to expedite two of the most time-consuming steps which are article screening and data element extraction.

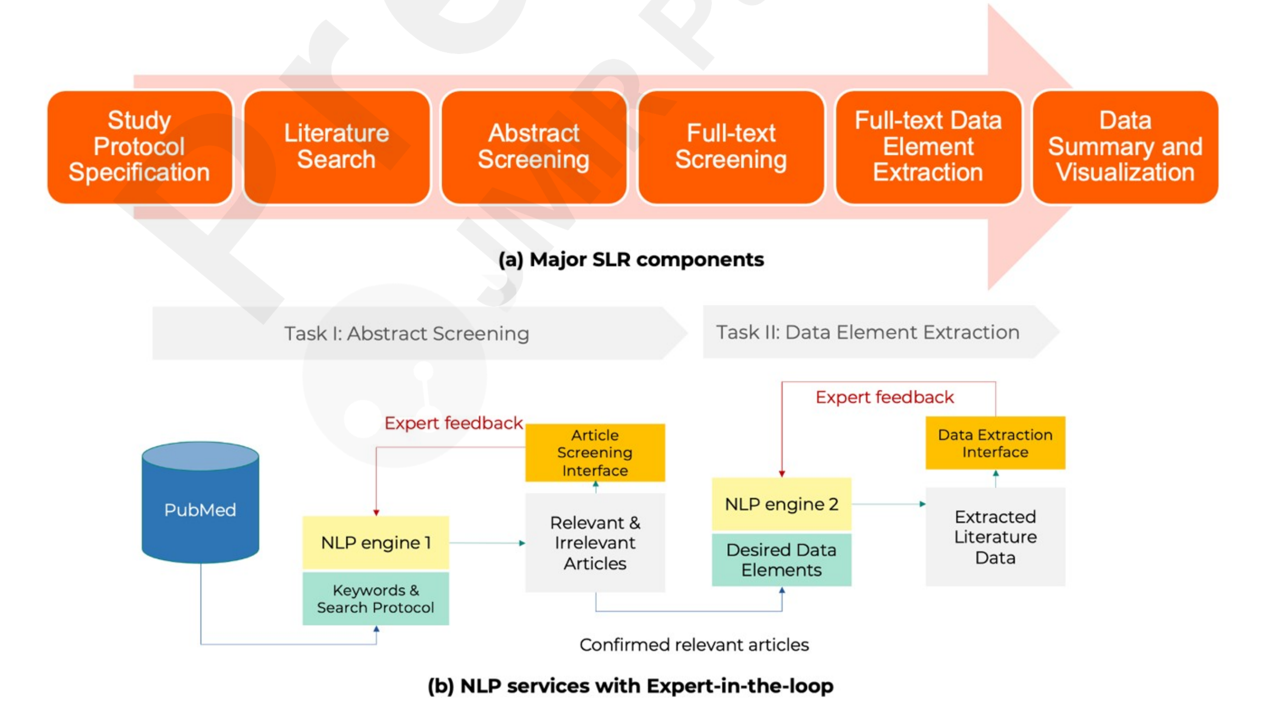

Study Protocol Specification
Literature Search
Abstract Screening
Full-text Screening
Full-text Data Element Extraction
Data Summary and Visualization
(a) Major SLR components
Task I: Abstract Screening
Task II: Data Element Extraction
PubMed
Expert feedback
Article Screening Interface
NLP engine 1
Keywords & Search Protocol
Relevant & Irrelevant Articles
Confirmed relevant articles
Expert feedback
Data Extraction Interface
NLP engine 2
Desired Data Elements
Extracted Literature Data
(b) NLP services with Expert-in-the-loop

Figure 2: ISLR workflow and embedded NLP engines.

NLP Algorithms Development and Validation

As mentioned earlier, two sets of NLP algorithms are required for a specific SLR project, including abstract screening and full-text data element extraction. Figure 3 outlines the NLP algorithm development process for these two steps separately. For abstract screening, the first step is to annotate and build a corpus that includes the abstract text, citation metadata, and inclusion/exclusion status. Once the corpus is prepared, NLP algorithms training, evaluation, and selection can be performed, and the best-performing algorithms will be chosen for deployment.

Similar to abstract screening, the NLP algorithm for the full-text data element extraction also requires a complete NLP development lifecycle. Unlike abstract screening, where labeled corpora may be available from previous SLR projects, data annotation is required to curate a labeled data set for training and evaluating NLP algorithms. The best-performing algorithms will be selected for deployment after evaluation. The following figure describes details on NLP algorithms development and validation process for SLR projects.

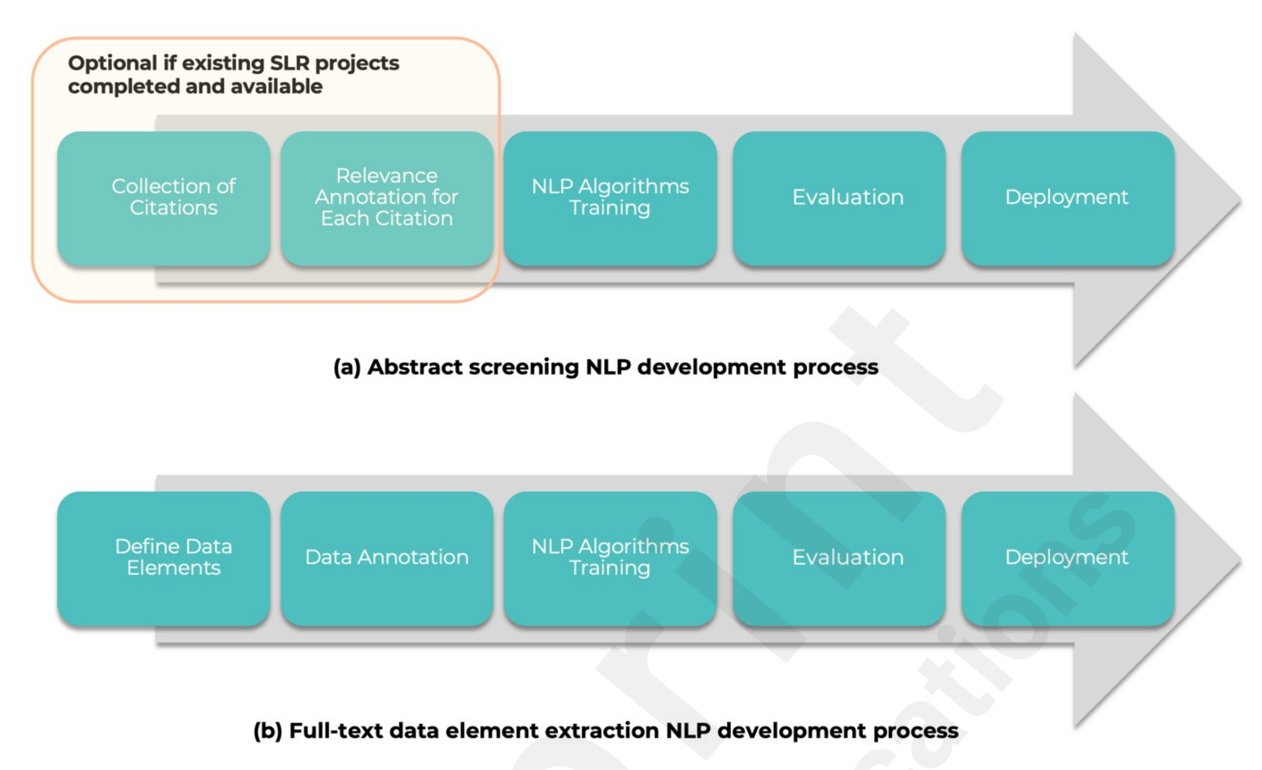

Optional if existing SLR projects completed and available
Collection of Citations
Relevance Annotation for Each Citation
NLP Algorithms Training
Evaluation
Deployment
(a) Abstract screening NLP development process
Define Data Elements
Data Annotation
NLP Algorithms Training
Evaluation
Deployment
(b) Full-text data element extraction NLP development process

Figure 3. SLR NLP algorithms development steps.

Three previously completed SLRs were used to guide and validate NLP development. These three projects included: 1) the prevalence of human papillomavirus (HPV) detected in head and neck squamous cell carcinomas (referred to as *HPV Prevalence*); 2) the epidemiology of the pneumococcal disease (referred to as *Pneumococcal Epidemiology*), and 3) the economic burden of pneumococcal disease (referred to as *Pneumococcal Economic Burden*). The inclusion and exclusion criteria for these three SLRs can be found in Table S1.

Developing the Abstract Screening Corpora: Abstract screening was treated as a binary document classification task, i.e., inclusion or exclusion of the article based on the abstract. Consequently, it was necessary to select and train NLP models for the task that demonstrated adequate performance and that had a reasonable computational time. The annotated screening literature sets from the three previous SLRs were used as the gold standard to train and validate models, including 1,697, 207, and 421 articles for *HPV Epidemiology*, *Pneumococcal*

Epidemiology, and *Pneumococcal Economic Burden* respectively. The corpora contained citation metadata including title, authors, Medical Subject Heading (MeSH) terms [9], and the text of the corresponding abstracts.

Developing the Full-text Data Element Extraction Corpora: We selected 190, 25, and 24 full-text articles for *HPV Prevalence*, *Pneumococcal Epidemiology,* and *Pneumococcal Economic Burden* for annotation, respectively. Based on the key outcome variables defined in the three SLRs, we annotated 12 types of data elements, covering information related to general observational studies, such as *Study Population,* to disease-specific information such as *HPV Lab Technique* and *Pneumococcal Disease Type*.

Abstract screening NLP algorithms: For abstract screening, the NLP model classifies each article for its relevance based on its title, abstract and other citation meta data. To build the abstract screening module, we evaluated four traditional ML-based document classification algorithms, XGBoost,[10] Support Vector Machines (SVM),[11] Logistic regression (LR),[12] and Random Forest[13] on the binary inclusion/exclusion classification task for abstract screening. The abstract screening corpora were used to evaluate NLP models by calculating standard metric of *precision (fraction of relevant instances among the retrieved instances, also called positive predictive value), recall (fraction of relevant instances that were retrieved, also called sensitivity), accuracy,* and *F1-scores* (the harmonic mean of precision and recall). The full features include Title, abstract, authors, keywords, journal, MeSH term, and publication types. we concatenated all features and extracted the TF-IDF (term frequency-inverse document frequency) vector as feature representation.

Data element extraction NLP algorithms: To construct the module for data element extraction, we treated the problem of data element recognition and extraction as a Named Entity

Recognition (NER) problem, which aims to recognize the mentions of entities from the text.[14] We evaluated a series of NLP algorithms consisting of ML and DL algorithms to recognize and extract data elements from full-text, including 1) Conditional Random Fields, a classic statistical sequence modeling algorithm that has been widely applied to NER tasks;[15,16], 2) Long Short-term Memory (LSTM), a variation of Recurrent Neural Networks (RNNs) that has achieved remarkable success in NER tasks;[17,18] and 3) "Clinical BERT (Bidirectional Encoder Representations from Transformers)"[19], a novel Transformer-based deep learning model. Standard metrics, including *precision*, *recall*, *accuracy*, and *F1-scores*, were calculated.

NLP Results

Here we report the results of the construction of the annotation corpora, the results of the NLP algorithm for abstract screening and data element extraction respectively.

Abstract screening corpora description: The *HPV Prevalence* corpus we constructed from the existing SLR project contained 1,697 total citations, of which 538 were included, and 1,159 were excluded due to study criteria. The constructed *Pneumococcal Epidemiology* contained 207 citations, of which 85 were included, and 122 were excluded. The constructed *Pneumococcal Economic Burden* corpus contained 421 citations, of which 79 were included, and 342 were excluded.

Abstract screening NLP evaluation results: Extensive studies have shown the superiority of transformer-based deep learning (DL) models for many NLP tasks.[20–23] Based on our experiments, however, adding features to the pre-trained language models did not significantly boost their performance. The performance comparison results for each task are shown in Table 1. XGBoost achieved the highest accuracy on *HPV Prevalence* and *Pneumococcal Economic Burden* tasks, while a Support Vector Machine achieved the highest accuracy on *Pneumococcal*

Epidemiology task. XGBoost was ultimately chosen for deployment due to its better generalizability.

Table 1. Comparison of article screening NLP model performance.

Task	Algorithm	F1 score	Precision	Recall	Accuracy
HPV Prevalence (n=1,697)	XGBoost	0.808	0.769	0.851	0.888
	Support vector machine	0.727	0.781	0.681	0.859
	Logistics regression	0.684	0.897	0.553	0.859
	Random forest	0.523	0.944	0.362	0.818
Pneumococcal Economic Burden (n=421)	XGBoost	0.750	0.857	0.667	0.907
	Support vector machine	0.533	0.667	0.444	0.667
	Logistics regression	0.333	0.667	0.222	0.831
	Random forest	0.429	0.600	0.333	0.814
Pneumococcal Epidemiology (n=207)	XGBoost	0.667	0.533	0.889	0.619
	Support vector machine	0.667	0.667	0.667	0.861
	Logistics regression	0.429	0.600	0.333	0.619
	Random forest	0.615	1.000	0.444	0.762

Full-text data element extraction corpora description: The human annotators annotated 190, 25, and 24 full-text articles for the HPV Prevalence, Pneumococcal Epidemiology, and Pneumococcal Economic Burden tasks respectively. Among these full-text articles, 4,498, 579, and 252 entity mentions were annotated for three projects respectively. However, distribution of annotated entities is highly imbalanced. For example, data elements like *HPV Lab Technique* and *HPV Sample Type* were very prevalent, but data elements like *Maximum/Minimum Age in Study Cohort* were rarely annotated in the corpora.

Results of Full text screening and data element extraction NLP methods: Table 2 and 3 show the comparison of NLP performance among Conditional Random Fields (CRF), LSTM, and BERT on the three corpora. For each of the three corpora used to train the NLP models, LSTM demonstrated superiority over the conventional machine learning algorithm (i.e., CRF) on entity recognition. Among DL models, we did not observe significant improvement in F-1 scores by use of the BERT model. The BERT model achieved similar or worse performance on most data elements. The performance across different tasks varies, primally due to availability of annotated data. For example, on average, models' performance on *HPV Prevalence* is higher than Pneumococcal Epidemiology and Pneumococcal Economic Burden, as *HPV Prevalence* has the largest annotated data. Due to the highly imbalanced distribution of annotated entities, we observe a variation in performance across different data elements for the same model. For example, in Pneumococcal Epidemiology Task, the LSTM model has achieved 0.412 in the identification of *Study Cohort* and 0.768 in the identification of *Pneumococcal Disease Type*.

Table 2. Overall performance comparison for the NER recognition task in the three NLP training corpora. Scores averaged across all 12 extracted data elements. Measured in lenient F-1 score.

	HPV Prevalence			**Pneumococcal Epidemiology**			**Pneumococcal Economic Burden**		
Measure	**CRF**	**LSTM**	**Clinical BERT**	**CRF**	**LSTM**	**Clinical BERT**	**CRF**	**LSTM**	**Clinical BERT**
Micro-average (global average that uses the	0.856	0.890	0.782	0.571	0.646	0.444	0.609	0.615	0.478

total number of true positives, false positives, and false negatives)									
Macro-average score (arithmetic mean of all the per-entity type scores)	0.522	0.674	0.685	0.270	0.295	0.227	0.216	0.238	0.231

Table 3. Performance comparison for the NER recognition task on selected data elements. Measured in lenient F-1 score.

	HPV Prevalence			**Pneumococcal Epidemiology**			**Pneumococcal Economic Burden**		
Measure	**CRF**	**LSTM**	**Clinical BERT**	**CRF**	**LSTM**	**Clinical BERT**	**CRF**	**LSTM**	**Clinical BERT**
Study Cohort	0.482	0.695	0.727	-	0.412	0.278	-	-	-
Study Location	0.434	0.520	0.574	0.514	0.508	0.546	0.586	0.484	0.497
Study Type	0.733	0.760	0.753	0.364	0.525	-	-	0.328	0.299
Pneumococcal Disease Type	-	-	-	0.725	0.768	0.526	0.644	0.715	0.523
Incidence or Prevalence	0.986	0.983	0.924	-	-	-	-	-	-
Study Time	0.714	0.888	0.930	0.222	0.636	0.328	-	-	-

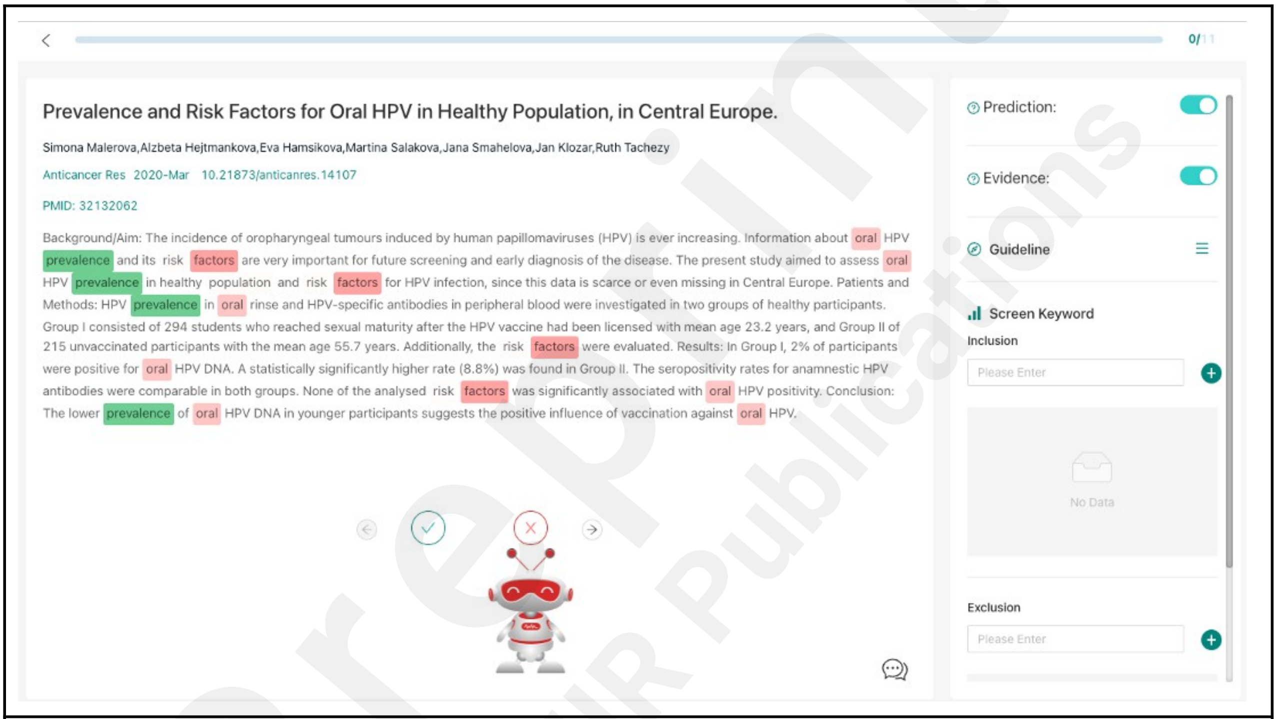
Final NLP algorithm selection: NLP algorithms were needed for the two tasks, abstract screening, and data element extraction, in the ISLR system. The abstract screening was treated as a classification task. Based on our experimental results, XGBoost was selected for this task due to good performance on our document classification experiments and less computational complexity than DL-based models. For the data element extraction task, LSTM was selected over CRF and BERT for same reasons.

ISLR System Components

Study Protocol Specification: Study protocol specification is one of the first steps in an SLR project. Users can upload a PDF document to the system that describes the SLR study protocol for reference. The SLR system has a default list of data elements with their descriptions and answer types (e.g., free text, multiple choice, checkbox), which will be extracted from full-text PDFs of articles. The system also allows users to create and modify the list. At the end of the project, all the extracted data elements can be exported in a structured format.

Literature Search: The ISLR system is integrated with the PubMed E-utilities Application Programming Interface (API), which enables users to perform direct searches on PubMed. Citation metadata such as abstracts, titles, journals, and authors can be retrieved from PubMed and indexed in the system for further screening and data element extraction. Additionally, the system provides an option for users to retrieve this citation metadata by uploading a list of individual PubMed IDs.

Abstract Screening: The purpose of abstract screening is to review collected articles' relevance based on their title, abstract, and other relevant metadata, such as journal names, article types, keywords. The relevant articles will be included for the following full-text screening and data elements extraction steps. NLP services are provided at this step to make recommendations on whether a particular article should be included for full-text review. The supporting information (e.g., salient words that are impactful to inclusion and exclusion) for the NLP recommendation will also be shown to provide explainable evidence. Human experts can further review the predictions for each article and decide on abstract screening status (keep or exclude). Figure 4 shows the abstract screening interface demonstrating prediction results and relevant terms discovered by the NLP algorithms.



Prevalence and Risk Factors for Oral HPV in Healthy Population, in Central Europe.

Simona Malerova,Alzbeta Hejtmankova,Eva Hamsikova,Martina Salakova,Jana Smahelova,Jan Klozar,Ruth Tachezy

Anticancer Res 2020-Mar 10.21873/anticanres.14107

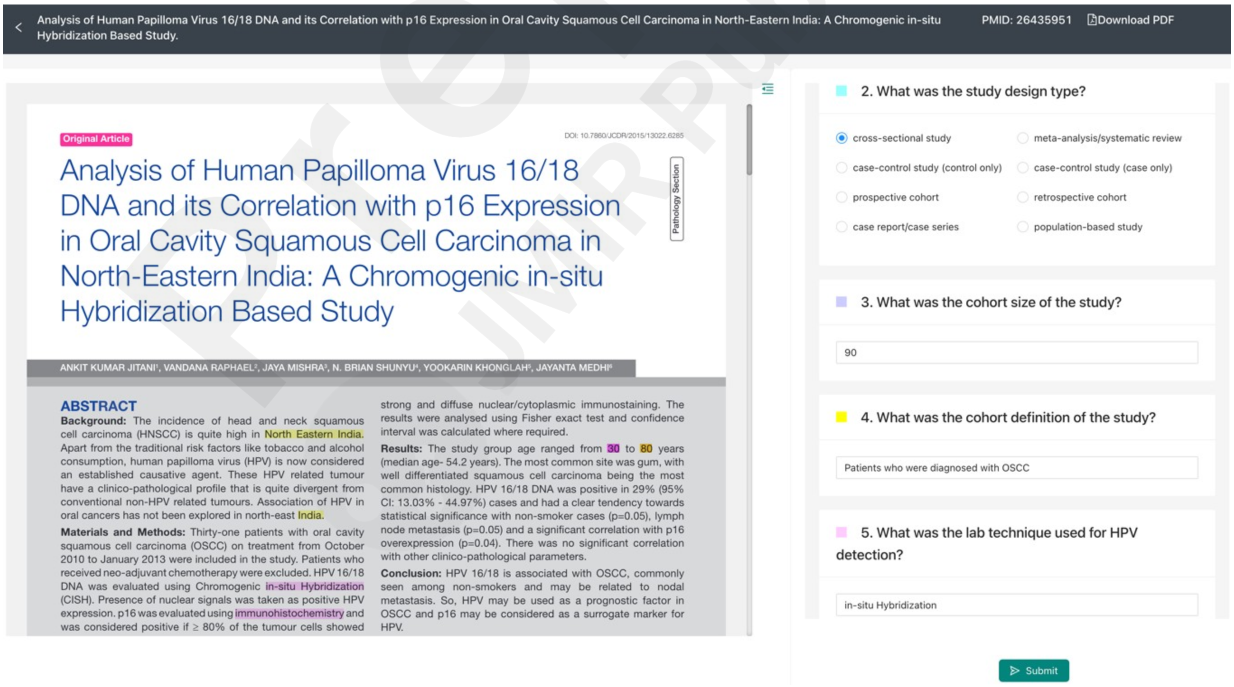
PMID: 32132062

Background/Aim: The incidence of oropharyngeal tumours induced by human papillomaviruses (HPV) is ever increasing. Information about oral HPV prevalence and its risk factors are very important for future screening and early diagnosis of the disease. The present study aimed to assess oral HPV prevalence in healthy population and risk factors for HPV infection, since this data is scarce or even missing in Central Europe. Patients and Methods: HPV prevalence in oral rinse and HPV-specific antibodies in peripheral blood were investigated in two groups of healthy participants. Group I consisted of 294 students who reached sexual maturity after the HPV vaccine had been licensed with mean age 23.2 years, and Group II of 215 unvaccinated participants with the mean age 55.7 years. Additionally, the risk factors were evaluated. Results: In Group I, 2% of participants were positive for oral HPV DNA. A statistically significantly higher rate (8.8%) was found in Group II. The seropositivity rates for anamnestic HPV antibodies were comparable in both groups. None of the analysed risk factors was significantly associated with oral HPV positivity. Conclusion: The lower prevalence of oral HPV DNA in younger participants suggests the positive influence of vaccination against oral HPV.

Figure 4. Abstract screening interface. Terms that support inclusion in the finalized cohort of relevant articles are shown in green, while terms that detract from inclusion are shown in red. The scale of the colors shows how significantly one term can impact prediction decisions (e.g., darker color indicates higher impact).

__Full-text Searching, Uploading and Screening:__ This step aims to identify full-text PDF documents for each included article and further screen their relevance based on the SLR study protocol. Only the articles that are deemed relevant after this stage will be included in the final full-text data element extraction step. The process of locating full-text PDF documents for each article can be time-consuming. The ISLR system integrates with PubMed Central to automatically find and collect full-text PDFs if they are publicly available. However, for articles whose full-text PDFs are not publicly available, users need to manually locate the articles through publishers and upload the corresponding PDFs to the system though the provided user interface.

__Full-text Data Element Extraction:__ Extracting full-text data elements is a time-consuming process in SLR projects. It requires reviewing the full-text article and extracting multiple relevant pieces of information defined in the study protocol. These data elements are often found in various sections of an article, including tables. The ISLR system uses Amazon Textract[24] for Optical Character Recognition (OCR) to extract text and tables from PDF files, followed by NLP services to further extract information from both text and tables. The NLP services can recommend potential answers for each data element, and human experts can review, select and modify the extracted information. Figure 5 shows a screenshot of the user interface for this step.



Analysis of Human Papilloma Virus 16/18 DNA and its Correlation with p16 Expression in Oral Cavity Squamous Cell Carcinoma in North-Eastern India: A Chromogenic in-situ Hybridization Based Study.

Original Article

DOI: 10.7860/JCDR/2015/13022.6285

Pathology Section

Analysis of Human Papilloma Virus 16/18 DNA and its Correlation with p16 Expression in Oral Cavity Squamous Cell Carcinoma in North-Eastern India: A Chromogenic in-situ Hybridization Based Study

ANKIT KUMAR JITANI^1, VANDANA RAPHAEL^2, JAYA MISHRA^3, N. BRIAN SHUNYU^4, YOOKARIN KHONGLAH^5, JAYANTA MEDHI^6

ABSTRACT

Background: The incidence of head and neck squamous cell carcinoma (HNSCC) is quite high in North Eastern India. Apart from the traditional risk factors like tobacco and alcohol consumption, human papilloma virus (HPV) is now considered an established causative agent. These HPV related tumour have a clinico-pathological profile that is quite divergent from conventional non-HPV related tumours. Association of HPV in oral cancers has not been explored in north-east India.

Materials and Methods: Thirty-one patients with oral cavity squamous cell carcinoma (OSCC) on treatment from October 2010 to January 2013 were included in the study. Patients who received neo-adjuvant chemotherapy were excluded. HPV 16/18 DNA was evaluated using Chromogenic in-situ Hybridization (CISH). Presence of nuclear signals was taken as positive HPV expression. p16 was evaluated using immunohistochemistry and was considered positive if ≥ 80% of the tumour cells showed strong and diffuse nuclear/cytoplasmic immunostaining. The results were analysed using Fisher exact test and confidence interval was calculated where required.

Results: The study group age ranged from 30 to 80 years (median age- 54.2 years). The most common site was gum, with well differentiated squamous cell carcinoma being the most common histology. HPV 16/18 DNA was positive in 29% (95% CI: 13.03% - 44.97%) cases and had a clear tendency towards statistical significance with non-smoker cases (p=0.05), lymph node metastasis (p=0.05) and a significant correlation with p16 overexpression (p=0.04). There was no significant correlation with other clinico-pathological parameters.

Conclusion: HPV 16/18 is associated with OSCC, commonly seen among non-smokers and may be related to nodal metastasis. So, HPV may be used as a prognostic factor in OSCC and p16 may be considered as a surrogate marker for HPV.

🟦 2. What was the study design type?

● cross-sectional study
○ meta-analysis/systematic review
○ case-control study (control only)
○ case-control study (case only)
○ prospective cohort
○ retrospective cohort
○ case report/case series
○ population-based study

🟪 3. What was the cohort size of the study?

90

🟨 4. What was the cohort definition of the study?

Patients who were diagnosed with OSCC

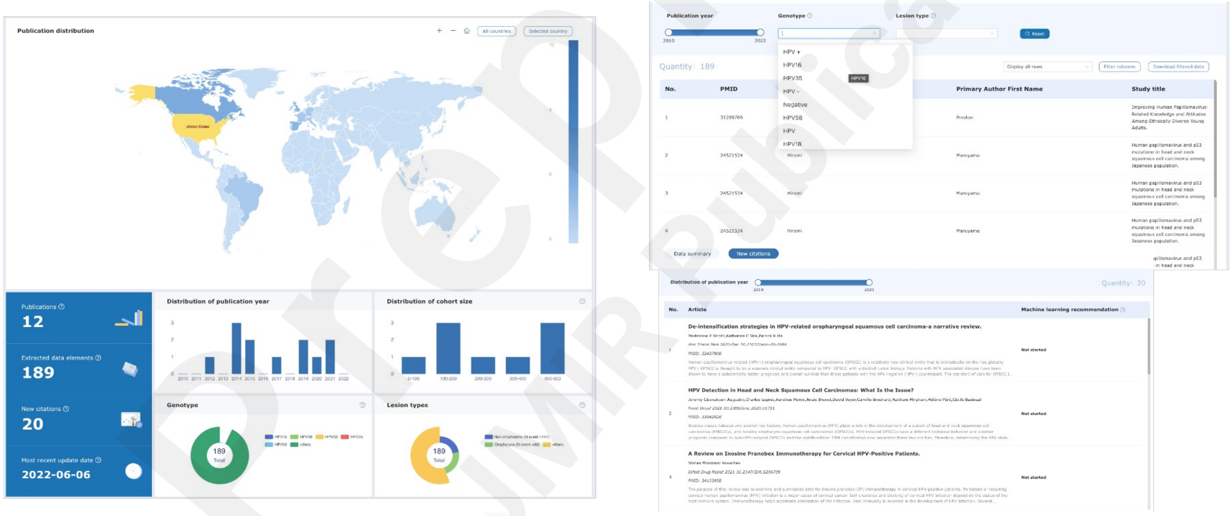
🟪 5. What was the lab technique used for HPV detection?

in-situ Hybridization

Figure 5. Full-text data elements extraction user interface. Data elements from the article extracted by the NLP algorithms are color coded and highlighted in the PDF. Highlight colors in the PDF text are linked to the data elements as shown in the right-hand frame. For the data element list on the right side, all the extracted data elements can pop up as

candidates for the users to choose from.

Data Summary & Visualization: The ISLR system offers interactive dashboards to end users, such as researchers, for exploring the SLR results and data. These dashboards allow users to apply data filters, such as study location and cohort size, to refine their search results. For each data element extracted from full-text articles, users can click on the element to navigate to the corresponding article, ensuring traceability and appropriate references to source documents in the SLR project. Additionally, the dashboards recommend recent relevant articles and suggest articles that may require full-text screening. Figure 6 displays the major functions and screenshots of the dashboard.



No.	PMID		Primary Author First Name	Study title
1	31289786		Preston	Improving Human Papillomavirus-Related Knowledge and Attitudes Among Ethnically Diverse Young Adults.
2	24521534	Hiromi	Maruyama	Human papillomavirus and p53 mutations in head and neck squamous cell carcinoma among Japanese population.
3	24521534	Hiromi	Maruyama	Human papillomavirus and p53 mutations in head and neck squamous cell carcinoma among Japanese population.
4	24521534	Hiromi	Maruyama	Human papillomavirus and p53 mutations in head and neck squamous cell carcinoma among Japanese population.

No.	Article	Machine learning recommendation
1	**De-intensification strategies in HPV-related oropharyngeal squamous cell carcinoma-a narrative review.** Madeleine P Strohl,Katherine C Wai,Patrick K Ha *Ann Transl Med 2020-Dec 10.21037/atm-20-2984* *PMID: 33437800* Human papillomavirus-related (HPV+) oropharyngeal squamous cell carcinoma (OPSCC) is a relatively new clinical entity that is dramatically on the rise globally. HPV+ OPSCC is thought to be a separate clinical entity compared to HPV- OPSCC with a distinct tumor biology. Patients with HPV associated disease have been shown to have a substantially better prognosis and overall survival than those patients with the HPV negative (HPV-) counterpart. The standard of care for OPSCC i...	Not started
2	**HPV Detection in Head and Neck Squamous Cell Carcinomas: What Is the Issue?** Jeremy Ghenakoun Augustin,Charles Lepine,Aurelien Morin,Anais Brunet,David Veyer,Camille Brochard,Haitham Mirghani,Hélène Péré,Cécile Badoual *Front Oncol 2020 10.3389/fonc.2020.01751* *PMID: 33042820* Besides classic tobacco and alcohol risk factors, human papillomavirus (HPV) plays a role in the development of a subset of head and neck squamous cell carcinomas (HNSCCs), and notably oropharynx squamous cell carcinomas (OPSCCs). HPV-induced OPSCCs have a different biological behavior and a better prognosis compared to non-HPV-induced OPSCCs and the eighth-edition TNM classification now separates these two entities. Therefore, determining the HPV statu...	Not started
3	**A Review on Inosine Pranobex Immunotherapy for Cervical HPV-Positive Patients.** Stefan Miladinov Kovachev *Infect Drug Resist 2021 10.2147/IDR.S296709* *PMID: 34103950* The purpose of this review was to examine and summarize data for inosine pranobex (IP) immunotherapy in cervical HPV-positive patients. Persistent or recurring cervical human papillomavirus (HPV) infection is a major cause of cervical cancer. Self-clearance and blocking of cervical HPV infection depend on the status of the host immune system. Immunotherapy helps accelerate elimination of the infection. Host immunity is involved in the development of HPV infection. Several...	Not started

Figure 6. Interactive visualization of existing SLR data, lists of relevant publications, and data exportation control.

DISCUSSION

As described in the introduction, conducting an SLR is complex and expensive. There is also a rapid growth of the available number of publications and other data such as clinical trial reports used in the article search and screening processes, with an average annual growth rate for the life sciences of around 5%.[40] Consequently, there is considerable community interest in applying various types of automation, including AI, DL, and NLP to the multiple tasks required for producing an accurate SLR.[2,5–7]

An important consideration for using the results of an SLR is how often the SLR is updated and hence how timely and complete these data are with respect to the real-world evidence. “Living” ISLR system addresses the difficulty of updating an SLR by providing an automated workflow including review tools to detect when new data are available and to trigger at least a semi-automated update process for the expediated review. The system is also expandable to cover additional data elements of interests by updating existing NLP pipelines.

The major accomplishments of this ISLR system include improving the time, efficiency, cost, completeness of evidence, and error avoidance through techniques to assist researchers with decision-making (so-called human-in-the-loop). The ISLR system is aligned with the living SLR concept, as it supports a rapid update of existing literature data. Additionally, since the classification and data element extraction tasks are maintained by the system, results can be used for retraining the classification and NLP algorithms on a routine basis. Consequently, the performance of the system should improve over time.

The focus of this work was to evaluate an intelligent system that includes all major steps of an SLR with human in the loop. The corpora evaluated in this study mostly focus on health economics and outcomes research in specific therapeutical areas. The generalizability of the learning algorithms to another domain will benefit from further formal examination. Since we have not yet conducted a time analysis of an SLR study conducted both manually and with this tool, we are unable to precisely quantify the time savings from the ISLR system. In addition, our NLP technologies limit to the extraction of relevant information directly from the text but are not able to conduct reasoning with long context to support complex data elements extraction, such as GRADE (Grading of Recommendations, Assessment, Development and Evaluation) or ROB2 (Risk of Bias 2). The recent advances in Large Language Models (LLMs), such as Generative Pre-trained Transformer 4 (GPT-4), bring NLP technologies expert-level performance on various professional and academic benchmarks. Given its high performance, and generalizability and reasoning capacity, it would be interesting to further assess the efficacy and accuracy of LLMs in various SLR tasks and complex data elements extraction.

As an early and innovative attempt to automate SLR lifestyle through NLP technologies, ISLR does not fully support PRISMA reporting yet. We plan to continuously iterate ISLR to cover PRISMA checklist and report generation in the future. In addition, we have not yet conducted formal usability studies of the user interface, although agile methods involving iterative refinement of the interface through input from domain experts in SLR were employed throughout the software development process.

CONCLUSION

Our ISLR system is a user-centered, end-to-end intelligent solution to automate and accelerate the SLR process and supports “living” SLRs with human in the loop. The system integrates cutting-edge ML and DL-based NLP algorithms to make recommendations on article screening and data element extraction, which allow the system to prospectively and continuously update relevant literature in a timely fashion. This allows scientists to have more time to focus on the quality of data and the synthesis of evidence, and to stay current with literature related to observational studies.

REFERENCES

1 Munn Z, Stern C, Aromataris E, *et al.* What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC Med Res Methodol*. 2018;18:5.

2 Systematic review automation technologies | Systematic Reviews. https://link.springer.com/article/10.1186/2046-4053-3-74 (accessed 23 April 2024)

3 Cochrane Handbook for Systematic Reviews of Interventions. https://training.cochrane.org/handbook/current (accessed 7 August 2022)

4 Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. *Contemporary Clinical Trials Communications*. 2019;16:100443.

5 Michelson M, Ross M, Minton S. AI2 LEVERAGING MACHINE-ASSISTANCE TO REPLICATE A SYSTEMATIC REVIEW. *Value in Health*. 2019;22:S34.

6 Fiol GD, Michelson M, Iorio A, *et al.* A Deep Learning Method to Automatically Identify Reports of Scientifically Rigorous Clinical Research from the Biomedical Literature: Comparative Analytic Study. *Journal of Medical Internet Research*. 2018;20:e10281.

7 Elliott JH, Turner T, Clavisi O, *et al.* Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. *PLOS Medicine*. 2014;11:e1001603.

8 Rayyan – Intelligent Systematic Review - Rayyan. 2021. https://www.rayyan.ai/ (accessed 23 April 2024)

9 Medical Subject Headings - Home Page. https://www.nlm.nih.gov/mesh/meshhome.html (accessed 30 May 2022)

10 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery 2016:785–94. https://doi.org/10.1145/2939672.2939785

11 Noble WS. What is a support vector machine? *Nat Biotechnol*. 2006;24:1565–7. doi: 10.1038/nbt1206-1565

12 *Logistic Regression*. https://link.springer.com/book/10.1007/978-1-4419-1742-3 (accessed 30 May 2022)

13 Random forest classifier for remote sensing classification: International Journal of Remote Sensing: Vol 26, No 1. https://www.tandfonline.com/doi/abs/10.1080/01431160412331269698 (accessed 30 May 2022)

14 Nadeau D, Sekine S. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*. 2007;30:3–26.

15 Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Departmental Papers (CIS)*. Published Online First: 28 June 2001.

16 Lin S, Ng J-P, Pradhan S, *et al.* Extracting Formulaic and Free Text Clinical Research Articles Metadata using Conditional Random Fields. *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*. Los Angeles, California, USA: Association for Computational Linguistics 2010:90–5. https://aclanthology.org/W10-1114 (accessed 7 August 2022)

17 Chiu JPC, Nichols E. Named Entity Recognition with Bidirectional LSTM-CNNs. *arXiv:151108308 [cs]*. Published Online First: 19 July 2016.

18 Lample G, Ballesteros M, Subramanian S, *et al.* Neural Architectures for Named Entity Recognition. *arXiv:160301360 [cs]*. Published Online First: 7 April 2016.

19 Alsentzer E, Murphy JR, Boag W, *et al.* Publicly Available Clinical BERT Embeddings. 2019. https://doi.org/10.48550/arXiv.1904.03323

20 Devlin J, Chang M-W, Lee K, *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 2019. https://doi.org/10.48550/arXiv.1810.04805

21 BioBERT: a pre-trained biomedical language representation model for biomedical text mining | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/36/4/1234/5566506 (accessed 3 June 2020)

22 Gu Y, Tinn R, Cheng H, *et al.* Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans Comput Healthcare*. 2021;3:2:1-2:23.

23 Chen Q, Du J, Allot A, *et al.* LitMC-BERT: transformer-based multi-label classification of biomedical literature with an application on COVID-19 literature curation. arXiv 2022. https://doi.org/10.48550/arXiv.2204.08649

24 Intelligently Extract Text & Data with OCR - Amazon Textract - Amazon Web Services. Amazon Web Services, Inc. https://aws.amazon.com/textract/ (accessed 8 August 2022)

ACKNOWLEDGMENT

This research was supported by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA

DATA AVAILABILITY

The annotated corpora underlying this article are available at https://github.com/Merck/NLP-SLR-corpora.

CONTRIBUTION

Study concept and design: JD and LY

Corpus preparation: DW, JD and LY

Experiments: JD and BL

Draft of the manuscript: FJM, JD, DW, NC and LY

Acquisition, analysis, or interpretation of data: JD, DW, NC and LY

Critical revision of the manuscript for important intellectual content: All authors

Study supervision: JD, LY, and NC

CONFLICT OF INTEREST STATEMENT

DW, JC, DE, NC, PCF, and LY are employees of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA. JD, BL, SW, XW, LH, JW, and FJM are employees of Melax Tech.

DISCLAIMER

The content is the sole responsibility of the authors and does not necessarily represent the official views of Merck & Co., Inc., Rahway, NJ, USA or Melax Tech

Supplementary Files

Multimedia Appendixes

Untitled.
URL: http://asset.jmir.pub/assets/f05377e735240e6dc7608557cdc2384b.docx