www.nature.com/scientificreports

scientific reports

OPEN

Use of deep learning-based NLP models for full-text data elements extraction for systematic literature review tasks

Jingcheng Du[1], Dong Wang[2]✉, Bin Lin[1], Long He[1], Liang-Chin Huang[1], Jingqi Wang[1], Frank J. Manion[1], Yeran Li[2], Nicole Cossrow[2] & Lixia Yao[2]

Systematic literature review (SLR) is an important tool for Health Economics and Outcomes Research (HEOR) evidence synthesis. SLRs involve the identification and selection of pertinent publications and extraction of relevant data elements from full-text articles, which can be a manually intensive procedure. Previously we developed machine learning models to automatically identify relevant publications based on pre-specified inclusion and exclusion criteria. This study investigates the feasibility of applying Natural Language Processing (NLP) approaches to automatically extract data elements from the relevant scientific literature. First, 239 full-text articles were collected and annotated for 12 important variables including study cohort, lab technique, and disease type, for proper SLR summary of Human papillomavirus (HPV) Prevalence, Pneumococcal Epidemiology, and Pneumococcal Economic Burden. The three resulting annotated corpora are shared publicly at [https://github.com/Merck/NLP-SLR-corpora], to provide training data and a benchmark baseline for the NLP community to further research this challenging task. We then compared three classic Named Entity Recognition (NER) algorithms, namely Conditional Random Fields (CRF), Long Short-Term Memory (LSTM), and the Bidirectional Encoder Representations from Transformers (BERT) models, to assess performance on the data element extraction task. The annotation corpora contain 4,498, 579, and 252 annotated entity mentions for HPV Prevalence, Pneumococcal Epidemiology, and Pneumococcal Economic Burden tasks respectively. Deep learning algorithms achieved superior performance in recognizing the targeted SLR data elements, compared to conventional machine learning algorithms. LSTM models have achieved 0.890, 0.646 and 0.615 micro-averaged F1 scores for three tasks respectively. CRF models could not provide comparable performance on most of the elements of interest. Although BERT-based models are known to generally achieve superior performance on many NLP tasks, we did not observe improvement in our three tasks. Deep learning algorithms have achieved superior performance compared with machine learning models on multiple SLR data element extraction tasks. LSTM model, in particular, is more preferable for deployment in supporting HEOR SLR data element extraction, due to its better performance, generalizability, and scalability as it's cost-effective in our SLR benchmark datasets.

Systematic literature review (SLR) is a robust, systematized method to identify and synthesize evidence from literature and to integrate and present findings on a research question or the efficacy and effectiveness of an intervention[1]. SLR is a major methodological tool in many areas of the health sciences for both scientists and biopharmaceutical companies. In the field of health economics and outcomes research (HEOR), SLRs are routinely conducted to understand the burden of disease, and the research landscape, synthesize evidence around unmet medical needs, compare the values of various treatment options, and assess the issues related to efficiency, effectiveness, and value of resources in health sciences[2].

Unfortunately, conducting an SLR is known to be manually intensive and expensive – a recent study found the time and cost required to conduct an SLR has been estimated to require approximately 1.72 person-years of clinical or research scientist effort and cost approximately an average of $140,000 per review[3]. The top ten pharmaceutical companies were found to publish an average of 41.71 SLRs per year for an estimated cost of $5.8 million per company[3]. The time and cost required for completing an SLR negatively impact the

[1]Intelligent Medical Objects, Houston, TX, USA. [2]Merck & Co., Inc., Rahway, NJ, USA. ✉email: dong.wang10@merck.com

timeliness of SLR results[3]. The major steps required to conduct an SLR include defining the research questions, screening articles, extracting data elements, assessing the risk of bias and quality of the work, analyzing data, etc[4]. Consequently, automation of any of these steps presents an opportunity for improvement. Previously we developed machine learning models to automatically identify relevant publications based on given inclusion and exclusion criteria[5].

In this paper, we demonstrate the use of natural language processing (NLP) to facilitate data element extraction to reduce the time and cost required to complete or update an SLR. NLP is a branch of artificial intelligence and used here refers to any technology (e.g., novel machine learning and deep learning algorithms) that can extract structured information from textual documents. NLP has been widely used in automatic mining and knowledge extraction from biomedical literature, such as recognition of biomedical concepts from the text (e.g., Chemical, Gene, Disease), relation extraction (e.g., drug-drug interaction)[6], yet there has been limited effort focused on automating the data element extraction phase of SLR, despite being one of the most time-consuming steps[7]. The majority of existing efforts have been focusing on sentence-level information extraction from the article abstract, e.g., recognizing PICO (Population, Intervention, Comparison, Outcomes) sentences from the abstract[8,9]. There are a few studies focusing on data element extraction from full text, but their efforts have been limited to eligibility criteria[10,11]. Jonnalagadda et al. concluded that NLP technologies have not been fully utilized to automate the data element extraction step[7].

The main objective of this study is to develop and present valuable resources for advancing the field of automated data extraction in observational Systematic Literature Review (SLR) projects. Our primary contributions are twofold: (1) Performance Benchmark: We have conducted a comprehensive evaluation of multiple machine learning and deep learning algorithms for recognizing data elements crucial to SLR projects. This evaluation establishes a robust performance benchmark that will serve as a valuable reference point for future studies in this field. (2) Annotated Corpora: To support our benchmark and facilitate further research, we have created and are making publicly available a set of annotated corpora. These corpora cover 12 types of data elements across three distinct SLR projects: HPV Prevalence, Pneumococcal Epidemiology, and Pneumococcal Economic Burden. This resource allows the broader research community to reproduce our benchmark results, test new algorithms, and develop innovative approaches to automated data extraction.

Materials and methods

SLR corpora annotation

In this study, we built upon three SLR projects that were conducted at Merck & Co., Inc., Rahway, NJ, USA in the past. The first was detecting human papillomavirus (HPV) prevalence in head and neck squamous cell carcinomas (referred to as *HPV Prevalence*), which aims to identify the available peer-reviewed evidence on the prevalence of HPV detected in head and neck squamous cell carcinomas (HNSCCs). It specifically focused on the detection of overall HPV, and where reported, high-risk HPV, low-risk HPV, and HPV genotypes 6, 11, 16, 18, 31, 33, 45, 52, and/or 58, The second SLR was epidemiology of the pneumococcal disease (both invasive and noninvasive, caused by Streptococcus pneumonia) in children and adults in 13 countries of interests, including United States, Canada, and Japan (referred to as *Pneumococcal Epidemiology*). The third SLR was the economic burden of pneumococcal disease with the same scope (referred to as *Pneumococcal Economic Burden*). The scope and inclusion criteria for these three SLRs are summarized in Table 1.

We collected 239 full-text articles for annotation, including 190 articles, 25 articles, and 24 articles for *HPV Prevalence*, *Pneumococcal Epidemiology*, and *Pneumococcal Economic Burden*, respectively. These articles were all selected for final data analysis after both abstract screening and full-text screening steps. Based on the key outcome elements defined in the three SLR protocols, we annotated 12 types of data elements, covering general HEOR-related information such as *Study Population* and *Lab Technique*, disease-specific information such as *HPV Lab Technique*, and *Pneumococcal Disease Type*. A detailed description of each data element can be found in Table 2.

Text preprocessing

We leveraged Amazon Textract[13] to extract printed text from full-text articles in PDF format. From all the text extracted, we further selected three major sections for data annotation, including the Title, Abstract/Summary, and Methodology/Material sections. The data elements we aimed to extract are generally mentioned in these

	HPV prevalence	Pneumococcal epidemiology	Pneumococcal economic burden
Purpose of SLR	To identify the available peer-reviewed evidence on the prevalence of HPV detected in head and neck squamous cell carcinomas (HNSCCs).	To understand the epidemiology of the pneumococcal disease (both invasive and noninvasive, caused by Streptococcus pneumonia) in children and adults in 13 countries of interests	To understand the economic burden of the pneumococcal disease (both invasive and noninvasive, caused by Streptococcus pneumonia) in children and adults in 13 countries of interests
Populations	adults (age > = 13) with histologically confirmed invasive HNSCCs (oral cavity, oropharynx, larynx, hypopharynx)	Populations with Invasive and noninvasive pneumococcal disease (Excluding children < 5 years and adults > 18 years)	Populations with Invasive and noninvasive pneumococcal disease (Excluding children < 5 years and adults > 18 years)
Primary outcomes	overall HPV prevalence and/or type distribution (types 16 and 18 and at least one of the new Gardasil 9 vaccine types [i.e., types 31, 33, 45, 52 or 58])	Epidemiology data (incidence, prevalence, etc.) of invasive or noninvasive pneumococcal disease	Direct resource use or costs by health state; Indirect or other resource use or costs of interest; quality-of-life data
Time of publications	Publication date from 2015 to 2020	Publication date from 2012 to 2017	Publication date from 2012 to 2017

Table 1. SLR scope and primary inclusion criteria.

Data element	Definition	Example	Associated corpora
Study time	The time when the study was conducted	From *1977 till 2008*, a total of 54 patients aged < 40 years with newly diagnosed, previously untreated HNSCC were identified in the Netherlands Cancer Institute Database	HPV Prevalence; Pneumococcal Epidemiology; Pneumococcal Economic Burden
Study location	The location where the study population was recruited.	Frequency and genotype distribution of multiple human papillomavirus infections in cancer of the head and neck in a *Mexican* population	HPV Prevalence; Pneumococcal Epidemiology; Pneumococcal Economic Burden
Study cohort	Description of the study cohort, including disease type, cohort size, etc.	*Patients included in this study were diagnosed with SCC of the head and neck, and treated with chemora- diotherapy* at the Yorkshire Cancer Centre between 2003 and 2006	HPV Prevalence; Pneumococcal Epidemiology; Pneumococcal Economic Burden
Maximum age in study cohort	Higher age-cut-off for study population	Median age was 56.3 years (range, 34–*76* years)	HPV Prevalence; Pneumococcal Epidemiology; Pneumococcal Economic Burden
Minimum age in study cohort	Lower age-cut-off for study population	Median age was 56.3 years (range, *34*–76 years)	HPV Prevalence; Pneumococcal Epidemiology; Pneumococcal Economic Burden
Study type	Study type descriptor (e.g., meta-analysis, population-based study, etc.)	A *retrospective study* was performed analyzing 114 patients with oral SCC treated from 1970 to 2006	HPV Prevalence
Incidence or prevalence	Specifies whether the study reports prevalence or incidence	The aim of this study was to analyze the *prevalence* of HPV and EBV in oral and oropharyngeal squamous cell carcinoma in south-eastern Poland	HPV Prevalence
HPV lab technique	HPV amplification and typing assays used	HPV DNA detection was performed using *SPF10-DEIALiPA25* version 1 system (Labo BiomedicalProducts, Netherlands)	HPV Prevalence
HPV sample collection method	Describes how samples were collected; i.e., biopsy, brush, spatula, etc.	For each specimen, routine histological examination on hematoxylin-eosin stained slides prepared from a formalin-fixed, paraffin-embedded (FFPE) *biopsy*, confirmed invasive diagnosis	HPV Prevalence
HPV Sample Type	Specifies type of sample (i.e., tissue specimen, cervical swab, etc.)	Additionally, we evaluated the quality of our *specimens* by means of cellular tubulin detection and subsequently excluded the HPV DNA negative and tubulin negative cases from the statistical analyses.	HPV Prevalence
Pneumococcal disease type	The sub-type of pneumococcal disease	To determine changes in mortality among adults with *invasive pneumococcal disease (IPD)* after introducing pneumococcal conjugate vaccines (PCVs) in children	Pneumococcal Epidemiology; Pneumococcal Economic Burden
Study purpose	The objective of the study	*To evaluate the cost-effectiveness of introducing universal vaccination of adults aged 60 years with the 23-valent pneumococcal polysaccharide vaccine (PPV23) into the National Immunization Program (NIP) in Brazil*	Pneumococcal Epidemiology; Pneumococcal Economic Burden

Table 2. The list of data elements in the annotation corpora. In this study, three annotators with bachelor's degrees or above were recruited for this annotation task, under another lead annotator with MD/PhD degrees. To mitigate potential inconsistency, each article was annotated at least twice. A team-based NLP annotation software that offers statistical quality control support – LANN[12] – was used as the tool for annotation. The resulting annotated corpora are publicly available at [https://github.com/Merck/NLP-SLR-corpora*]*.

sections. We used the NLP tool CLAMP (Clinical Language Annotation, Modeling, and Processing)[14] to perform preprocessing steps including tokenization, sentence boundary detection, and Part-of-Speech (POS) tagging.

Named entity recognition (NER) algorithms for data element extraction

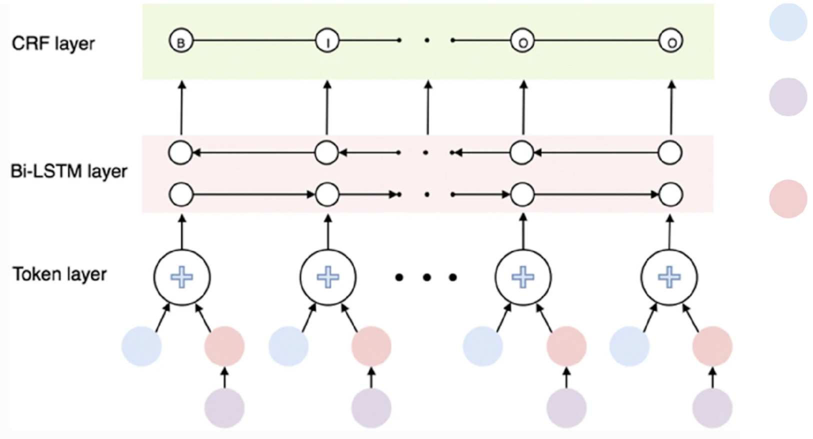
We framed the extraction of SLR data elements from the full text as NER tasks. NER is one of the most fundamental NLP tasks to locate and classify named entities in text into predefined classes[15]. We converted annotated entities into the BIO format, where "B" represents the word at the beginning of the entity, "I" represents the word inside of the entity, and "O" represents the word outside of the entity. We evaluated a variety of machine learning and deep learning-based NER models, as follows, to predict the BIO labels for words in sentences.

Conditional random fields (CRF)

CRF is a classic statistical sequence modeling algorithm that has been widely applied to NER tasks before the rise of deep learning algorithms. We again leveraged CLAMP[14] to implement the CRF algorithm. We employed an extensive set of features including lexical features, syntactic features, context features, distributional representation of words, and domain knowledge features.

Long Short-Term memory (LSTM)

LSTM is a variation of Recurrent Neural Networks (RNN) that has achieved remarkable success in NER tasks[16,17]. We used an LSTM-based deep learning framework for the NER tasks (See Fig. 1 for model architecture). Specifically, token embedding mapped each token in the text to a token vector. Considering the issues of out-of-vocabulary words or misspellings, we added a character embedding layer to map the characters in each token to character vectors. The character Bi-LSTM received the character embedding output as input at each time step, and output the summary of each character for each token. The token embedding input was concatenated with the character LSTM output and then fed to the token Bi-LSTM layer. In addition to token Bi-LSTM, a sequential Conditional Random Fields (CRFs) was added to jointly decode labels for each word. The pre-trained clinical embedding (dimension: 200[18]) was used to initiate the word embedding layer. The maximum number of epochs was set at 50.



Token embedding mapped each token in the text to a token vector.

Character embedding mapped the characters in each token to character vectors.

Character Bi-LSTM took the character embedding at each time step as the input and outputs the summary of each character for each token.

The token embedding was concatenated with the character LSTM output and then fed to the *Token Bi-LSTM* layer together.
On top of token Bi-LSTM, we applied a *Sequential CRF* to jointly decode labels for the whole text

Fig. 1. The architecture of LSTM-based deep learning model for NER tasks.

Bidirectional encoder representations from Transformers (BERT)
To address the issue of resource-deficient datasets, pre-trained language models were proposed to first pre-trained on large volumes of unannotated datasets (i.e., pre-training) and then further adapted to guide other tasks (i.e., fine-tune). Pre-trained language models have recently achieved superior performance in many general and biomedical NLP tasks[19–22]. We evaluated four variants of BERT models, including Clinical BERT[21], BioBERT[23], BLUEBERT[24] and PubMedBERT[25], for full-text data element extraction. For all BERT models, we used base we set maximum sequence length at 64, learning rate at 2e-5, batch size at 4, and maximum number of training epochs at 50 with early stop.

Evaluation

For each dataset, the annotated corpus was divided into train, validation, and test sets with a proportion of 6:2:2. Hyperparameters tuning were conducted and the values for major hyperparameters are included in Supplementary Table S19. Standard metrics, including precision, recall, and F1-measure, were calculated, based on lenient match (overlap in entity boundary) as described below:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Where *True Positive* means that there is an overlap between a predicted entity with an annotated entity (gold standard entity): *False Positive* means that a predicted entity is not overlapped with any gold standard entities; *False Negative* means that an annotated entity is not overlapped with any predicted entities.

For every model evaluation, we repeated the experiments five times and reported the average scores.

Results

Annotation statistics

Table 3 shows the statistics for the annotated corpora. As can be observed, the 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.

NLP results comparison

Tables 4, 5 and 6 show the comparison of NLP performance among CRF, LSTM, and BERT on *HPV Prevalence*, *Pneumococcal Epidemiology*, and *Pneumococcal Economic Burden* tasks respectively. Due to the highly imbalanced distribution of annotated entities, we observed a significant variation in performance across different data elements. For example, for the *HPV Prevalence* task, data elements like *Maximum Age in Study Cohort*, *Minimum Age in Study Cohort*, and *Study Location* had relatively low F-1 scores, while data elements like *HPV Sample Type* and *HPV Lab Technique* had much higher F-1 scores. Similar variations were also shown for the *Pneumococcal Epidemiology*, and *Pneumococcal Economic Burden* tasks.

In general, the deep learning models (LSTM and BERT) demonstrated superiority over the conventional machine learning algorithm (i.e., CRF) on entity recognition for all 3 tasks. For some data elements, the improvements were quite remarkable. For example, for *Study Type* in the *HPV Prevalence* task, the LSTM model improved the F-1 score from 0 in the CRF model to 0.683; for *Study Time* in the *Pneumococcal Epidemiology*

Data element	Number of annotated entities in HPV prevalence corpus (no. of articles: 190)	Number of annotated entities in pneumococcal epidemiology corpus (no. of articles: 25)	Number of annotated entities in pneumococcal economic burden corpus (no. of articles: 24)
Study time	206	58	25
Study location	322	112	126
Study cohort	365	32	13
Maximum age in study cohort	27	28	7
Minimum age in study cohort	29	11	22
Study type	119	38	38
Incidence or prevalence	414	–	–
HPV lab technique	1349	–	–
HPV sample collection method	69	–	–
HPV sample type	1598	–	–
Pneumococcal disease type	-	287	
Study purpose	-	13	21

Table 3. SLR corpora annotation statistics.

Data element	CRF	LSTM	Clinical-BERT	Bio-BERT	BLUE-BERT	PubMed-BERT
Maximum age in study cohort	0.182	0.571	0.529	0.667*	0.630	0.667*
Minimum age in study cohort	0.308	0.480	0.514	0.713	0.791	0.889*
Study location	0.434	0.520	0.574	0.708*	0.628	0.578
Incidence or prevalence	0.986*	0.983	0.924	0.978	0.924	0.921
HPV lab technique	0.905	0.939*	0.656	0.895	0.684	0.702
HPV sample collection method	–	0.222	0.338	0.515*	0.374	0.433
HPV sample type	0.942	0.951	0.903	0.946*	0.890	0.886
Study cohort	0.482	0.695	0.727	0.749	0.748	0.752*
Study type	0.733	0.760	0.753	0.785	0.710	0.794*
Study time	0.714	0.888	0.930	0.966	0.954	0.980*
Micro-average score	0.856	0.890*	0.782	0.888	0.790	0.795
Macro-average score	0.620	0.741	0.838*	0.809	0.754	0.775

Table 4. Performance comparison on HPV prevalence task. Measured in lenient F-1 score. (Note: * denotes the highest value for a specific data element)

Data element	CRF	LSTM	Clinical-BERT	Bio-BERT	BLUE-BERT	PubMed-BERT
Maximum age in study cohort	0.333	0.571*	0.138	0.205	0.323	0.148
Minimum age in study cohort	–	–	–	–	–	–
Study location	0.514	0.508	0.546	0.748*	0.387	0.471
Pneumococcal disease type	0.725	0.768*	0.526	0.683	0.605	0.586
Study cohort	–	0.412	0.278	0.524*	0.387	0.353
Study purpose	–	0.261	–	0.625*	0.571	–
Study type	0.364	0.525*	–	0.500	0.083	0.043
Study time	0.222	0.636*	0.328	0.317	0.255	0.349
Micro-average score	0.571	0.646*	0.444	0.620	0.480	0.469
Macro-average score	0.405	0.588*	0.385	0.537	0.409	0.400

Table 5. Performance comparison on Pneumococcal epidemiology. Measured in lenient F-1 score. (Note: * denotes high value).

task, the LSTM model improved the F-1 score from 0 in CRF model to 0.625. Among deep learning models, we did not observe significant improvement in F-1 scores by use of the BERT model on these NER tasks. The BERT model achieved similar or worse performance on most data elements.

Table S1-18 from supplement material contains all lenient and exact precision, recall F-1 scores for all three tasks for reference.

Data element	CRF	LSTM	Clinical-BERT	Bio-BERT	BLUE-BERT	PubMed-BERT
Maximum age in study cohort	–	–	–	–	–	–
Minimum age in study cohort	–	0.235	–	0.462‡	0.071	0.148
Study location	0.586	0.484	0.497	0.697‡	0.542	0.524
Pneumococcal disease type	0.644	0.715	0.523	0.716‡	0.533	0.622
Study cohort	–	–	–	–	–	–
Study purpose	0.500	0.143	0.529	0.571‡	0.571‡	0.529
Study type	–	0.328	0.299	0.4‡	0.319	0.328
Study time	–	–	–	0.080	0.240	0.267‡
Micro-average score	0.609	0.615	0.478	0.648‡	0.497	0.559
Macro-average score	0.606‡	0.366	0.511	0.514	0.423	0.466

Table 6. Performance comparison on Pneumococcal economic burden. Measured in lenient F-1 score. ‡ denotes the best performance in each data element across all models.

Discussions and conclusions

Principal findings

Some prior work exist that applied NLP to improve efficiency of the SLR process. For example, Thomas et al. used NLP to identify randomized controlled trial for Cochrane reviews[26]. Wallace et al. developed methods to extract sentences from literature related to clinical trial reports[27]. Despite these existing efforts, there is a lack of robust NLP solutions for to extract detailed data elements from the full-texts of the articles, which is addressed in our study.

As one of the first efforts targeting multiple therapeutic areas, this study evaluated the use of deep learning algorithms to extract 12 types of data elements from full-text articles for three SLR projects. The data elements cover general HEOR-related information as well as some disease-specific information. We've further made study corpora publicly available for the broad research communities as benchmark corpora for advanced algorithm development and evaluation.

Not surprisingly, deep learning algorithms have achieved significant improvement in the recognition of targeted SLR data elements. CRF models, which were widely-adopted classic machine learning models before the rise of deep learning, were not able to provide comparable performance on a majority of SLR data elements. In addition, although BERT-based models have achieved improved performance on many NLP tasks compared with conventional deep learning models (e.g., LSTM), we've not observed improvement on our three tasks. Considering the much lower requirements for computation resources, in our case LSTM models will be preferable for deployment.

In this study, we evaluated the used of deep learning models to support data element extraction on three separate SLR tasks and on a dozen of data elements related to observational studies. This proposed methodology framework is both cost-effective and scalable, and can be potentially embedded in SLR software to support end-to-end "living systematic literature review"[28].

Limitations and future work

Some limitations remain for this study. First, although we're targeting a variety of SLR data elements, there are still many important SLR data elements that we did not cover. Many of these data elements appear in various tables in the articles, which are not in the scope of this study. However, the NLP development process we adopted in this study can be used to cover additional data elements when needed. Second, the annotation of these SLR corpora may be biased as well. Although each article was annotated and reviewed by two annotators, however, these two annotators were not annotating the corpora independently and their understanding of the contents may also be limited due to their training and education background difference. Third, from our NLP algorithms comparison, we can observe that some data elements suffer from low performance, mainly due to the limited size of annotated entities. To further improve the performance in these minority classes we will need more annotation in the future. Data element-specific optimization may also help.

Additionally, we evaluated the performance of NLP models based on standard NER evaluation metrics at sentence level, however, there is a risk of identifying a correct entity in a sentence which may not be the ground truth at document level (e.g., hallucination). Further studies are needed to quantitively assess these types of errors. Finally, we evaluated in this study mainly focus on health economics and outcomes research. The generalizability of machine learning algorithms to other domains including other therapeutic areas will benefit from additional experiments.

The recent advances in large language models (LLMs), such as ChatGPT and GPT-4, show promising results in information extraction for scientific articles. However, leveraging LLMs for SLR tasks also face additional challenges like cost, latency, and risks of hallucination. We will explore the use of LLMs to expedite SLR in the future by conducting comprehensive and quantitative analysis on LLMs.

Conclusions

Data element extraction is a critical but very time-consuming step for an SLR project as it needs expert reviewers to examine the full-text articles carefully to extract relevant study data. Our study finds that Deep learning algorithms have achieved superior performance compared with machine learning models on multiple SLR data element extraction tasks. LSTM model, in particular, is more preferable for deployment in supporting SLR data element extraction, due to its robust performance, generalizability, and scalability as it's cost-effective. As one of the early efforts in evaluating deep learning-based NLP approaches to extract data elements for SLR tasks, this study built the basis to leverage NLP approaches to facilitate automating SLR studies in the future. One particular application could be a living systematic literature review system, which enables scientist to prospectively and continuously reviewing comprehensive and up-to-date literature for scientific discovery. Scientists could spend more time focusing on the quality of data and synthesis of information, rather than the labor-consuming SLR process.

Data availability

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

Glossary

Artificial intelligence (AI)	A set of technologies that enable computers or machines to perform tasks that typically require human intelligence.
Machine learning	A set of artificial intelligence methods that can learn from data and make a prediction without being explicitly programmed. Conventional machine learning refers to non-deep learning algorithms, such as logistic regression, support vector machines and Conditional Random Fields (CRFs).
Deep learning	A branch of machine learning algorithms based on deep artificial neural networks algorithms. Typical deep learning algorithms include multi-layer perceptron, convolutional neural networks, recurrent neural networks, graph neural networks, and Transformer.
Natural language processing (NLP)	A subfield of Artificial intelligence that uses machine learning to enable computers to understand and communicate with human language.
Named Entity Recognition (NER)	A natural language processing (NLP) technique that identifies and categorizes important information in text. NER is used to extract key information from unstructured text, such as sentences, paragraphs, or documents. It's also known as entity extraction or chunking.

Received: 15 May 2024; Accepted: 23 May 2025
Published online: 03 June 2025

References

1. Pati, D. & Lorusso, L. N. How to write a systematic review of the literature. *HERD* **11**, 15–30. https://doi.org/10.1177/1937586717747384 (2018).
2. Mishra, D. & Nair, S. R. Systematic literature review to evaluate and characterize the health economics and outcomes research studies in India. *Perspect. Clin. Res.* **6**, 20–33. https://doi.org/10.4103/2229-3485.148802 (2015).
3. 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. *Contemp. Clin. Trials Commun.* **16**, 100443. https://doi.org/10.1016/j.conctc.2019.100443 (2019).
4. Cochrane Handbook for Systematic Reviews of Interventions. Aug (2022). https://training.cochrane.org/handbook/current (accessed 7.
5. Jingcheng Du, E. et al. Yao. Machine learning models for abstract screening task - A systematic literature review application for health economics and outcome research. *Health Data Sci. (Under Review)* (2022).
6. Recent advances in. biomedical literature mining | Briefings in Bioinformatics | Oxford Academic. May (2022). https://academic.oup.com/bib/article/22/3/bbaa057/5838460?login=true (accessed 30.
7. Automating data extraction in. systematic reviews: a systematic review | Systematic Reviews | Full Text. https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-015-0066-7 (accessed 7 Aug 2022).
8. Huang, K-C. et al. Classification of PICO elements by text features systematically extracted from PubMed abstracts. In: *2011 IEEE International Conference on Granular Computing*. 279–83. (2011). https://doi.org/10.1109/GRC.2011.6122608
9. Combining classifiers for robust PICO element detection | BMC Medical Informatics and Decision Making | Full Text. https://bmcmedinformdecismak.biomedcentral.com/articles/https://doi.org/10.1186/1472-6947-10-29 (accessed 7 Aug 2022).
10. Kiritchenko, S. et al. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Med. Inf. Decis. Mak.* **10**, 56. https://doi.org/10.1186/1472-6947-10-56 (2010).
11. Lin, S. et al. Extracting Formulaic and Free Text Clinical Research Articles Metadata using Conditional Random Fields. In: *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 Aug 2022).
12. LANN | HOME. Aug (2022). https://lann.melaxtech.com/ (accessed 8.
13. Intelligently Extract Text & Data with OCR - Amazon Textract. - Amazon Web Services. Amazon Web Services, Inc. https://aws.amazon.com/textract/ (accessed 8 Aug 2022).
14. Soysal, E. et al. CLAMP – a toolkit for efficiently Building customized clinical natural Language processing pipelines. *J. Am. Med. Inf. Assoc.* **25**, 331–336. https://doi.org/10.1093/jamia/ocx132 (2018).
15. Nadeau, D. & Sekine, S. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*. **30**, 3–26. https://doi.org/10.1075/li.30.1.03nad (2007).
16. Chiu, J. P. C. & Nichols, E. Named Entity Recognition with Bidirectional LSTM-CNNs. *arXiv:151108308 [cs]* Published Online First: 19 July 2016.http://arxiv.org/abs/1511.08308 (accessed 11 Jun 2020).
17. Lample, G. et al. Jun. Neural Architectures for Named Entity Recognition. *arXiv:160301360 [cs]* Published Online First: 7 April 2016. (2020). http://arxiv.org/abs/1603.01360 (accessed 3.

18. Zhang, Y. et al. BioWordVec, improving biomedical word embeddings with subword information and mesh. *Sci. Data*. **6**, 52. https://doi.org/10.1038/s41597-019-0055-0 (2019).
19. Devlin, J. et al. BERT: Pre-training of deep bidirectional Transformers for Language Understanding. *ArXiv* https://doi.org/10.48550/ArXiv.1810.04805 (2019).
20. 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 Jun 2020).
21. Alsentzer, E. et al. *Publicly Available Clin. BERT Embeddings* doi:https://doi.org/10.48550/arXiv.1904.03323 (2019).
22. Chen, Q. et al. LitMC-BERT: transformer-based multi-label classification of biomedical literature with an application on COVID-19 literature curation. *ArXiv* https://doi.org/10.48550/ArXiv.2204.08649 (2022).
23. Lee, J. et al. BioBERT: a pre-trained biomedical Language representation model for biomedical text mining. *Bioinformatics* **36** (4), 1234–1240 (2020).
24. Peng, Y., Yan, S. & Lu, Z. Transfer learning in biomedical natural Language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. ArXiv Preprint ArXiv:1906.05474. Jun 13. (2019).
25. Gu, Y. et al. Domain-specific Language model pretraining for biomedical natural Language processing. *ACM Trans. Comput. Healthc. (HEALTH)*. **3** (1), 1–23 (2021).
26. Thomas, J. et al. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane reviews. *J. Clin. Epidemiol.* **133**, 140–151 (2021).
27. Wallace, B. C., Kuiper, J., Sharma, A., Zhu, M. & Marshall, I. J. Extracting PICO sentences from clinical trial reports using supervised distant supervision. *J. Mach. Learn. Res.* **17**, 4572–4596 (2016).
28. Du, J. et al. EPH153 A natural Language processing solution for health economics and outcomes research systematic literature review. *Value Health*. **26** (6), S192 (2023).

Acknowledgements

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

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 Intelligent Medical Objects.

Author contributions

Study concept and design: JD and LYCorpus preparation: DW, YL and LYExperiments: JD and BLDraft of the manuscript: JD, DW, FJM and LYAcquisition, analysis, or interpretation of data: JD, DW and LYCritical revision of the manuscript for important intellectual content: All authorsStudy supervision: LY.

Declarations

Competing interests

DW, YL, NC, and LY are employees of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA. JD, ES, DW, LH, JW, and FJM are employees of Intelligent Medical Objects. All the remaining authors declare no conflict of interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-03979-5.

Correspondence and requests for materials should be addressed to D.W.