Performance Evaluation of AI-Assisted Systematic Literature Review for Studies on Burden of HPV-associated Head and Neck Cancers

Dong Wang, Ph.D. [1], Surabhi Datta, Ph.D.[2], Yi-Ling Huang, Ph.D.[1], Kyeryoung Lee, Ph.D.[2], Chris Liston, PharmD, MBA[2], Yi Zheng, Ph.D.[1], Jun Zhang, MSPH, MD[3] , Majid Rastegar-Mojarad, Ph.D.[2], Nicole Cossrow, MPH, Ph.D.[1]

[1] Merck & Co., Inc., Rahway, NJ, USA; [2] IMO Health, Rosemont, IL, United States; [3] MSD R&D (China) Co., Ltd., Beijing, China

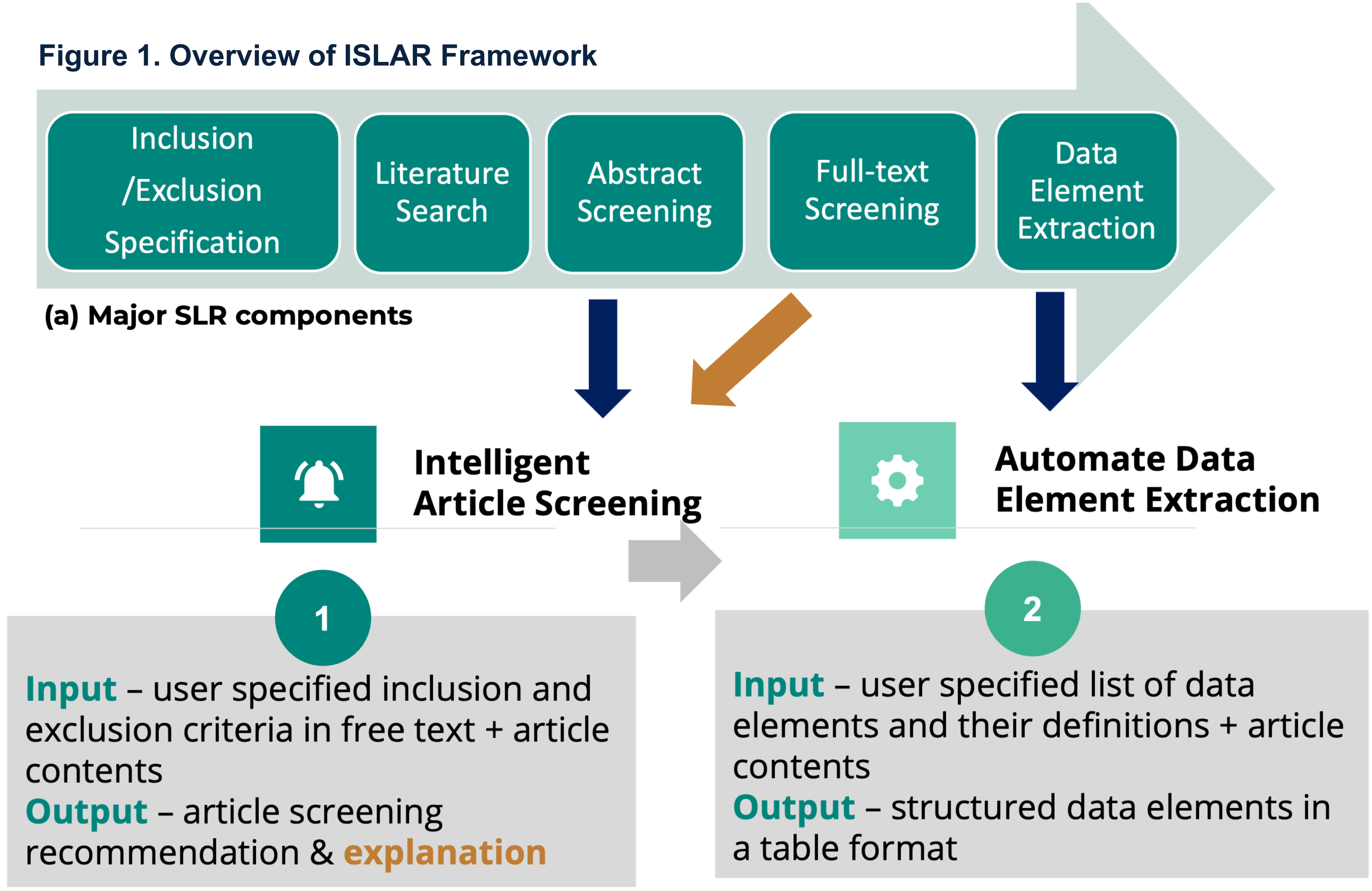
Introduction

- Systematic literature reviews (SLRs) of clinical research provide high quality evidence essential for informing policy decisions, research planning, and clinical practice.[1]
- Traditional SLR approaches are time-consuming and resource-intensive[2], often demanding an extensive review of large volume of studies.
- Natural language processing techniques, and, more recently, artificial intelligence (AI) hold potential to efficiently address this challenge by automating or semi-automating the major components of an SLR process.[3,4]
- We have developed a large language model (LLM)-based platform to assist SLR development – Intelligent Systematic LiterAture Review (ISLAR).
- Specifically, we focused on evaluating ISLAR in performing SLR in the domain of burden of HPV-associated head and neck cancers.

Methodology

- A protocol for burden of HPV-associated head and neck cancers was created by domain experts. After identifying related literatures, a gold standard corpus was constructed by randomly selecting and annotating 103 of the articles.
- ISLAR system includes major SLR components as shown in **Figure 1**. Among those, three most time-consuming components - abstract screening, full-text screening, and data extraction are powered by LLM to automate the process with human oversight. The SLR on burden of HPV detected in head and neck squamous cell carcinomas was run in ISLAR and the outputs from ISLAR were compared with the gold standard.
- The abstract and full-text screening components recommend eligibility decision based on the inclusion and exclusion criteria outlined in the study protocol.
- Data element extraction component identifies a predefined set of elements such as HPV sample type and collection method, HPV laboratory testing technique, sample size, cancer type, HPV genotype, HPV biomarker status, HPV prevalence, and HPV incidence along with study cohort and value information.
- We evaluated ISLAR's performance for the three components. Accuracy, sensitivity, and F1 scores were calculated for each component using human review as the gold standard.

Figure 1. Overview of ISLAR Framework


Inclusion /Exclusion Specification
Literature Search
Abstract Screening
Full-text Screening
Data Element Extraction
Intelligent Article Screening
Automate Data Element Extraction
1
2

(a) Major SLR components

Input – user specified inclusion and exclusion criteria in free text + article contents
Output – article screening recommendation & **explanation**

Input – user specified list of data elements and their definitions + article contents
Output – structured data elements in a table format

(b) Two NLP engines

Results

- For abstract screening component, ISLAR included 46 of 51 gold standard-included articles and excluded 30 of 52 gold standard-excluded articles, achieving an F1 score of 77.3%.
- For full-text screening, ISLAR included 33 of 36 articles included by the expert and excluded 3 of 6 articles excluded by the expert, achieving an F1 score of 91.7%.
- The confusion matrices for the screening components are shown in Tables 1 and 2 and the detailed performance measures of ISLAR on the three components including the data element extraction are presented in **Table 3**.

Table 1: Confusion matrix for abstract screening.

Gold standard	ISLAR		
	Include	Exclude	Total
Include	46	5	51
Exclude	22	30	52
Total	68	35	103

Table 2: Confusion matrix for full-text screening.

Gold standard	ISLAR		
	Include	Exclude	Total
Include	33	3	36
Exclude	3	3	6
Total	36	6	42

Table 3: Performance measures of the ISLAR system on the three components including screening and data element extraction.

Component	Accuracy (%)	Sensitivity (%)	F1 score
Abstract screening	73.80	90.10	77.30
Full-text screening	85.70	91.70	91.70
Data element extraction	80.80	94.60	86.60

Discussion

- ISLAR demonstrated high sensitivity (over 90%) across all SLR components with its data extraction component achieving a 94.60% sensitivity score denoting the system's capability to capture most of the defined data elements well.
- Although the abstract screening component obtained a moderate accuracy, it still demonstrated high sensitivity, indicating the system's priority to include articles that are good candidates for full-text review, and to ensure no relevant articles get excluded due to limited details in abstracts.
- A qualitative analysis of the errors made by the system revealed the following two main categories:
  - Inclusion of studies of non-eligible design - ISLAR erroneously included 9 abstracts that had the desired outcome information but were not epidemiological. A further 6 abstracts of narrative or systematic reviews were also erroneously included.
  - Cohort misidentification - ISLAR sometimes mistakenly attributed data elements to the wrong study cohort, typically doing so for multiple related elements.
- Future work will focus on:
  - further improvement of the SLR pipeline with refinement in prompt designs to efficiently incorporate expert feedback and domain-specific knowledge (e.g., providing instructions regarding the context of epidemiological studies)
  - evaluating the performance on a larger dataset and across different domains
  - investigating more advanced large language models

Conclusions

- ISLAR system exhibited excellent performance in full-text screening and data element extraction.
- The results demonstrated substantial potential for ISLAR's role in AI-assisted SLR development.
- Nevertheless, the occurrence of moderate performance in some instances underscores the underlying principle in ISLAR's design, which is that it should assist, not replace, human reviewers in carrying out SLR tasks. In line with this, ISLAR's design includes the reporting of reasons for article inclusion and exclusion to assist human decision-making, an approach that is transparent and consistent with high quality evidence generation.[2]
- Future research should evaluate ISLAR across a range of subject areas, supporting further development of this important research tool.

References

1. Bolaños F, Salatino A, Osborne F, Motta E. Artificial intelligence for literature reviews: opportunities and challenges. Artificial Intelligence Review 2024;57(259)doi:https://doi.org/10.1007/s10462-024-10902-3
2. van Dijk SHB, Brusse-Keizer MGJ, Bucsan CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open. Jul 7 2023;13(7):e072254. doi:10.1136/bmjopen-2023-072254
3. Blaizot A, Veettil SK, Saidoung P, et al. Using artificial intelligence methods for systematic review in health sciences: A systematic review. Res Synth Methods. May 2022;13(3):353-362. doi:10.1002/jrsm.1553
4. Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. Syst Rev. Aug 21 2024;13(1):219. doi:10.1186/s13643-024-02609-x

Acknowledgments

The authors thank Ingrid Peterson, PhD, in collaboration with ScribCo for medical writing assistance.

Disclosures

Dong Wang, Yi-Ling Huang, Yi Zheng, Jun Zhang, and Nicole Cossrow are employees of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA and may hold stock or stock options in Merck & Co., Inc., Rahway, NJ, USA.

Datta, Lee, Liston and Rastegar-Mojarad are employees of IMO Health who received funding from Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA for their participation in the study.

Contact information

Please contact corresponding author Dong Wang via email: dong.wang10@merck.com for any questions or suggestions.

https://bit.ly/3ILXa1c

Presented at ISPOR; Glasgow, Scotland, UK, Europe; November 9-12, 2025.