Feasibility of GPT-4-based Content Extraction to Identify Eligible Titles and Abstracts in a Systematic Literature Review

MSR136

Cichewicz A,[1] Pande A,[1] Casañas i Comabella C,[2] Mittal L,[3] Slim M[4]

[1]Evidera Inc., a business unit of PPD, part of Thermo Fisher Scientific, Waltham, MA, USA; [2]Evidera Ltd., a business unit of PPD, part of Thermo Fisher Scientific, London, UK; [3]Evidera Ltd., a business unit of PPD, part of Thermo Fisher Scientific, Bengaluru, India; [4] Evidera Inc., a business unit of PPD, part of Thermo Fisher Scientific, Montréal, Canada

Background

- Screening publications for inclusion based on pre-specified eligibility criteria is a rigorous step in the systematic literature review (SLR) process. The growing body of published literature, particularly on well-researched topics, can result in large search yields that can be time consuming to screen, leading to increased resourcing needs and higher execution costs.
- Leveraging machine learning (ML)-based artificial intelligence (AI) capabilities to alleviate the burden of title/abstract screening in SLRs has been extensively explored.[1,2] However, two main challenges arise when using such models:
  - They must be pre-trained on a large set of human-screened records for optimal performance.[3]
  - They are unable to provide reasons for inclusion/exclusion.
- Conversely, large language models (LLMs), including Generative Pre-trained Transformer 4 (GPT-4), are designed for universal applicability and do not require specific pre-training.
- Smart Tag Recommendations (referred to as smart tags henceforth) is a GPT-4-based content extraction feature developed by Nested Knowledge to assist users in extracting study details and data. This feature produces non-generative AI recommendations based on relevant text excerpts from the underlying studies to provide responses to a series of user-specified questions.
- While smart tags were not specifically intended for screening purposes, the ability of the tool to respond to questions designed based on the eligibility criteria for an SLR may provide more granular rationale for inclusion/exclusion decisions without SLR-specific training.

Objectives

- To evaluate the feasibility of employing smart tags to provide detailed rationale for study eligibility to inform title and abstract screening decisions when conducting SLRs.

Methods

- A previously conducted SLR of trials in pre-treated ovarian cancers was used. The eligibility criteria are presented in **Table 1**. Nineteen records that met the inclusion criteria from the SLR were used to evaluate the performance of smart tags.
- In the context of AI, prompts are inputs or instructions given to an AI model to elicit a specific response or behavior. Question-based prompts (referred to as questions henceforth) were developed from pre-defined population, intervention, comparator, outcome, study design (PICOS) eligibility criteria (**Table 1**). Sixteen questions were formulated, including four related to the population, one for intervention/comparator, nine for outcomes, and two for study designs (**Figure 1**).

Table 1. PICOS Eligibility Criteria

Category	Inclusion Criteria	Exclusion Criteria
Population	Women with de novo locally advanced or metastatic OC or fallopian tube or primary peritoneal carcinomas who: • Have platinum-sensitive* disease • Have responded to a prior first-line platinum therapy	Women in the following categories: • Early OC (stage I) • Without previous platinum-based chemotherapy • Prior maintenance treatment • With central nervous system metastasis that remains untreated
Interventions	• Targeted treatments ○ PARP inhibitors ○ Monoclonal antibodies • Any of the above interventions	• Non-pharmacologic treatments, such as surgery or radiotherapy alone • Alternative doses, schedules, or formulations of the intervention as the only comparator arms
Comparators	• Chemotherapy (platinum and non-platinum-based) • Placebo or best supportive care	
Outcomes	• Efficacy: PFS, time on/to treatment, discontinuation, ORR, OS, and duration of response, time to progression to first treatment • Safety/tolerability: any AE, discontinuation due to AEs, tolerability for dose. • HRQoL and PROs, including symptom assessment	• Publications that do not report data on relevant outcomes • Publications that report only interim trial results
Study designs	• Systematic reviews and meta-analyses of RCTs^ • RCTs (phases II/III)	• Non-randomized, single-arm, or observational • Open-label extension phases of RCTs • Pre-clinical studies • Case reports, expert opinion articles, letters, narrative (non-systematic reviews)

* Defined as disease progression >6 months after completion of their penultimate platinum regimen (from last dose) ^Published SLRs will not be included in the results of this review, but will be used for citation-chasing purposes
Abbreviations: AE = adverse event; HRQoL = health-related quality of life; NA = not applicable; OC = ovarian cancer; ORR = overall response rate; OS = overall survival; PARP = poly adenosine diphosphate ribose polymerase; PFS = progression-free survival; PRO = patient-reported outcome; RCT = randomised controlled trial; RECIST = Response Evaluation Criteria In Solid Tumors

- The smart tags were applied only to the included abstracts to determine whether these records would be accurately included based on the specifics of the PICOS criteria. Upon applying the smart tags to each of the 19 abstracts, the eligibility of records was determined as follows:
  - Presence of a recommendation = the record met that criterion for inclusion, e.g., if AI extracted information on the population, records with smart tags fulfilling all PICOS criteria were deemed eligible for inclusion.
  - Absence of a recommendation = the record failed to meet that criterion, i.e., there was no information in the title/abstract to answer that question or the AI did not identify an appropriate text excerpt.
    - Since all PICOS criteria must be met for a record to be included, records with one or more missing recommendations were excluded.
- Two key metrics were evaluated: recall and accuracy. Recall measures the AI's ability to identify relevant studies. Accuracy evaluates whether the AI correctly identified and extracted the relevant information from the abstracts in relation to the presented PICOS question.
- To evaluate the completeness and accuracy of smart tags, records were assigned scores of 0, 0.5, and 1 for incorrect, partially correct, and correct smart tag, respectively, with a maximum possible score of 19.

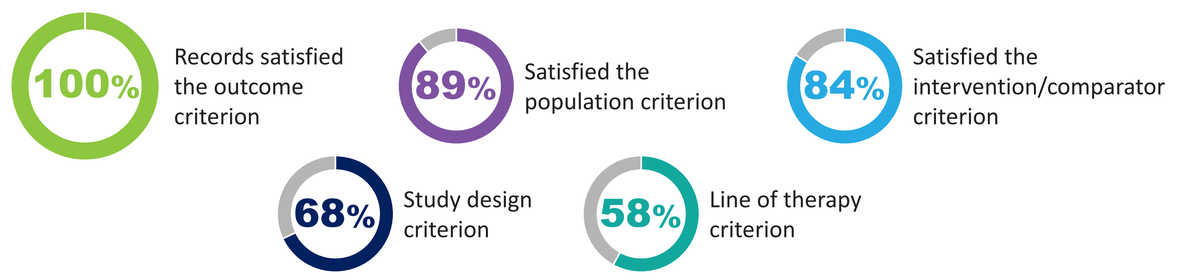
Results

Figure 1. Accuracy Rates at Title and Abstract Stage According to PICOS-based Questions

PICOS Category	Questions	Average Accuracy Score of Recommendations, When Available*	Average Score by PICOS Category
Population	Describe the population of the study	11	10
	Is the study on women with locally advanced or metastatic ovarian cancer?	16	
	Is the study on women with fallopian tube or primary peritoneal	2	
	Did patients receive previous treatment?	10	
Intervention/ Comparator	What treatments were patients randomized to?	15	15
Outcomes	Were any of the following efficacy outcomes reported? PFS using RECIST criteria, time on treatment, time to treatment discontinuation, ORR, OS, and duration of response, time to progression to second treatment, PFS on the subsequent line of treatment.	16	7
	Were response outcomes reported?	1.5	
	Does the study report time to progression outcomes?	12	
	Were survival outcomes reported?	4	
	Were any of the following safety outcomes reported? any adverse event, adverse events by grade, discontinuation due to adverse events, including tolerability for dose	13.5	
	Does the study report adverse events?	9	
	Does the study report on patient discontinuation rates?	3	
	Does the study report tolerability outcomes or dose adjustments?	3	
	Were health-related quality of life or patient-reported outcomes reported?	4	
Study Design	What type of study was conducted?	6.5	6
	Is this a randomized controlled trial?	6	

* Since the assessment was conducted on included trials, the maximum score for calculation was set at 19. This applied even in cases where no smart tags were extracted from the abstract or when the abstract did not report any of the PICOS criteria.

Recall



- Criteria related to outcomes, population and intervention/comparator had the highest recall rate while study design and line of therapy had the lowest. As a result, only 32% of records had smart tags for each PICOS criterion and were therefore included.
  - This low rate was driven in part by the treatment-related criteria (requiring information about line of therapy/previous therapy). When these criteria were not considered, 63% were deemed eligible

Accuracy

- Among the 19 trials, when the smart tags were available, the highest average score was recorded for questions related to intervention/comparator (15/19) and population (10/19), followed by outcomes (7/19). Questions about study design resulted in the lowest accuracy with an average score of 6/19 (**Figure 1**).
- Four questions focused on the population. The accuracy of the smart tags varied by question, with the lowest score (2) for the question "*Is the study on women with fallopian tube or primary peritoneal carcinomas?*" and the highest score (16) for the question "*Is the study on women with locally advanced or metastatic ovarian cancer?*"
- The question used for intervention/comparator—"*What treatments were patients randomized to?*"—achieved an accuracy score of 15.
- Nine questions were employed for the availability of outcomes, with accuracy scores ranging from 1.5 for "*Were response outcomes reported?*" to 16 for "*Were any of the following efficacy outcomes reported? [Progression-free survival] using [Response Evaluation Criteria in Solid Tumors] criteria, time on treatment, time to treatment discontinuation, [overall response rate], [overall survival], duration of response, time to progression to second treatment, and [progression-free survival] on the subsequent line of treatment.*"
- The questions used to inquire about study design showed an accuracy ranging from 6 for "*Is this a randomized controlled trial?*" to 6.5 for "*What type of study was conducted?*"

Discussion

- Overall, given the low recall rate and diminished accuracy of smart tags, particularly for questions related to outcomes and study designs, the smart tags feature of Nested Knowledge is not recommended to be used as an independent title/abstract screening tool.
- Broader questions incorporating multiple keywords or commonly used terms generally produced better recommendations. However, testing and refining comprehensive questions is a time-consuming process. This may hinder the efficient use of smart tags in practice, particularly if the set of questions needs to be created de novo, or if the screening team lacks AI literacy.
- The accuracy of smart tags is highly dependent on the choice of keywords. If critical keywords are omitted, the relevance of the smart tags may be diminished.
- Smart tags were not designed for screening purposes and are restricted to direct text excerpts. The effectiveness of smart tags for screening is hindered by inadequate, insufficient, or inconsistent reporting of study details reported within the underlying study – especially within the abstract which is often subject to word or character counts.
  - Assumptions are frequently made by humans during title/abstract screening in the absence of reporting of individual PICOS criteria. Specific details that are often unclear in the abstract (i.e., line of therapy, disease stage) may not be captured by the AI based on the question prompts, keywords, or abstract text.
- When deploying smart tags in this manner, it may be generally perceived as overly exclusive, which violates the main conception of an SLR which is to capture all relevant evidence.
- Since the assessment was conducted only on studies that ultimately meet all PICOS criteria (based on the full text), the accuracy estimation did not consider whether all PICOS criteria were reported in the abstract.

Conclusions

- Smart tags, initially designed for content extraction, may not be well-suited as a screening tool for SLRs. However, smart tags could accelerate human screening efforts by facilitating the categorization and prioritization of records that meet specific PICOS criteria.
- Our findings emphasize the importance of maintaining human oversight (i.e., human-in-the-loop) when using AI to aid the screening process so that critical nuances often overlooked by automated systems are appropriately captured. A less rigorous set of abstract-level specific PICOS criteria or hybrid approaches that combine LLMs with a human-in-the-loop component may lead to more effective AI integration within the SLR process.
- Further research should explore the optimization of question-based prompts and keywords to enhance the performance of smart tags in SLRs.

References

1. Cichewicz A, et al. Artificial Intelligence (AI)-Based Screening: Exploration of Differences in Two Health Technology Assessment (HTA)-Compliant Systematic Literature Reviews (SLRS). *Value Health*. 2023;26(11, S2).
2. Cichewicz A, et al. Application of Artificial Intelligence As a Decision Support Tool for Abstract Screening: Implications for Time and Cost Savings. *Value Health*. 2023;26(6, S2).
3. Cichewicz A, et al. Utility of Artificial Intelligence in Systematic Literature Reviews for Health Technology Assessment Submissions. *Value Health*. 2022;25(6, S1).

To view an electronic version of this poster, scan this Quick Response (QR) code.

Copies of this poster obtained through QR code are for personal use only and may not be reproduced without permission from the congress and the author of this poster.

Presented at the ISPOR Europe Conference ● 17–20 November 2024 ● Barcelona, Spain

Funding provided by Evidera Inc., a business unit of PPD, part of Thermo Fisher Scientific