Bigger, Better Studies – Prioritizing Literature Review Using an Approach Based on Large Language Models

SA17

Polly Field, Christian Eichinger, Marta Radwan, Kim Wager

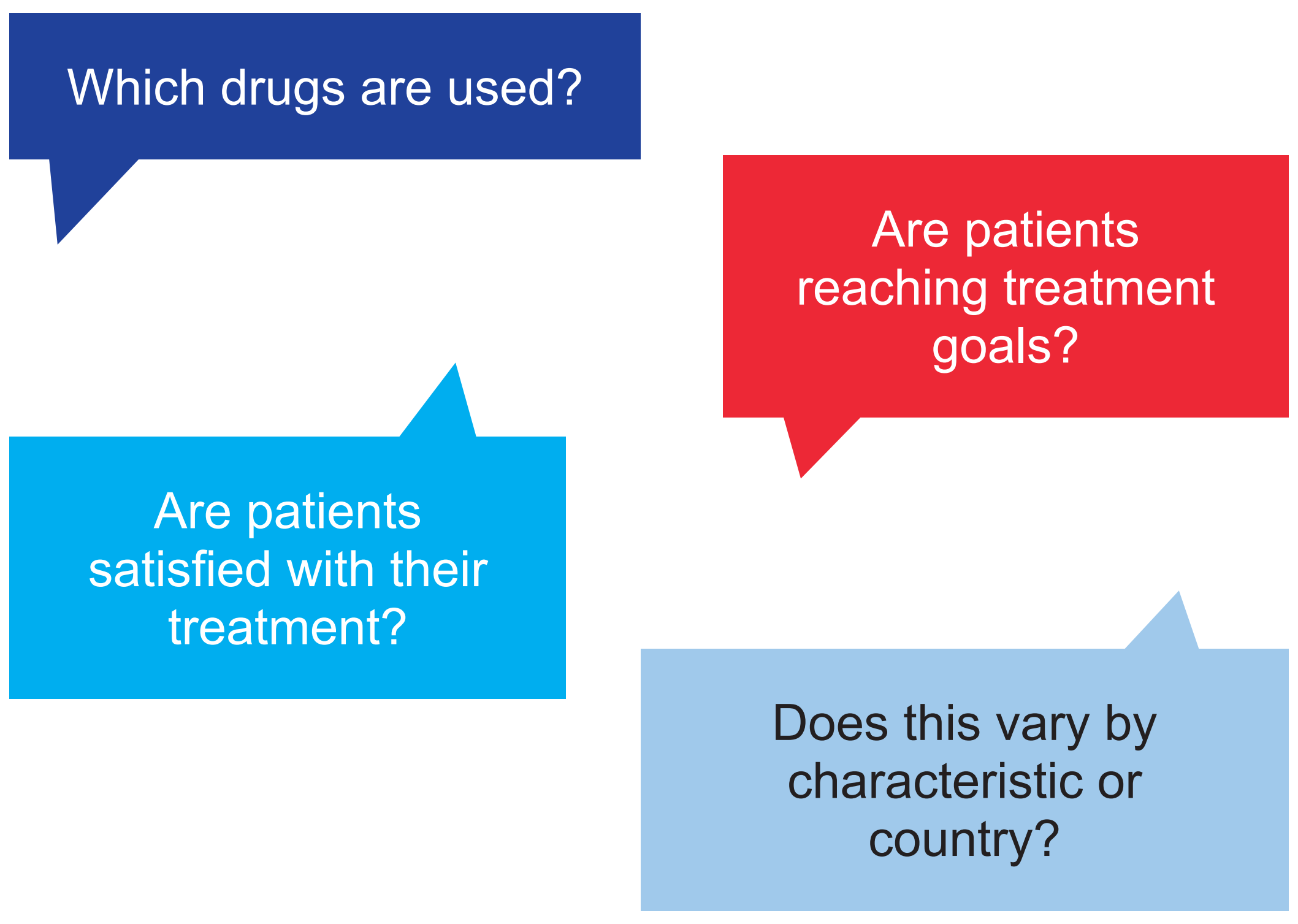
Oxford PharmaGenesis, Oxford, UK

Scan here to download the poster

ISPOR main topic/taxonomy: Study Approaches
ISPOR subtopics: Artificial Intelligence, Machine Learning, Literature Review & Synthesis

INTRODUCTION

- Sometimes in literature review we are not starting with one very specific question – we want answers across multiple questions.



Which drugs are used?

Are patients reaching treatment goals?

Are patients satisfied with their treatment?

Does this vary by characteristic or country?

- We also want the answers quickly – we want to go straight to the most relevant studies for us. These could be:
  - 'bigger' studies, of the populations that we are interested in
  - 'better' studies, because their main focus is the question/s that we are interested in.
- These criteria are difficult to include comprehensively in a Boolean search strategy, so we often have large numbers of studies to screen.
- **Here, we describe how we have added an extra characterization stage in the literature review process, using artificial intelligence (AI) for simple data extraction, allowing us to rank and prioritize studies at title/abstract screening stage. This helps us to find the bigger, better studies hidden in a large volume of literature quickly.**

OBJECTIVES

- To develop an AI-supported approach to review large bodies of literature rapidly and reproducibly:
  - across multiple questions
  - adjusting sensitivity to prioritize the bigger and better studies, according to factors that cannot easily be specified in search strings (e.g. study size, reporting multiple outcomes of interest)
  - enabling subject matter experts (SMEs) to make adaptive decisions based on the evidence available.

METHODS

- Our approach included an approach with human oversight of large language models (LLMs) (**Figure 1**).

Searches

- Our approach included a sensitive, reproducible, Boolean search strategy and searched bibliographic databases via OVID, exporting results to Excel.

Title/abstract screening

- AI and SMEs co-developed the prompt-engineering strategy for initial screening.
- An automated system implemented in Python cleaned and processed the data before iterative screening by a LLM (GPT-4o) based on terms for the patient populations of interest.
- Abstracts included in the first phase were passed to an initial data extraction phase, giving a ranking of publications by question and data relevance, with the bigger, better studies ranked highest.
  - This initial data extraction included study size, type, location, outcomes, patient characteristics and yes/no classification of whether specific data items were reported.
- SMEs reviewed the results over multiple rounds of data extraction and prompting. They iteratively refined the screening sensitivity based on the evidence available for each research question.
  - For example, SMEs included smaller studies if they found limited evidence for a particular question, but prioritized larger studies if they found more evidence. Data extraction is required for this prioritization, so this usually occurs far later in the review process.

Full-text review and reporting

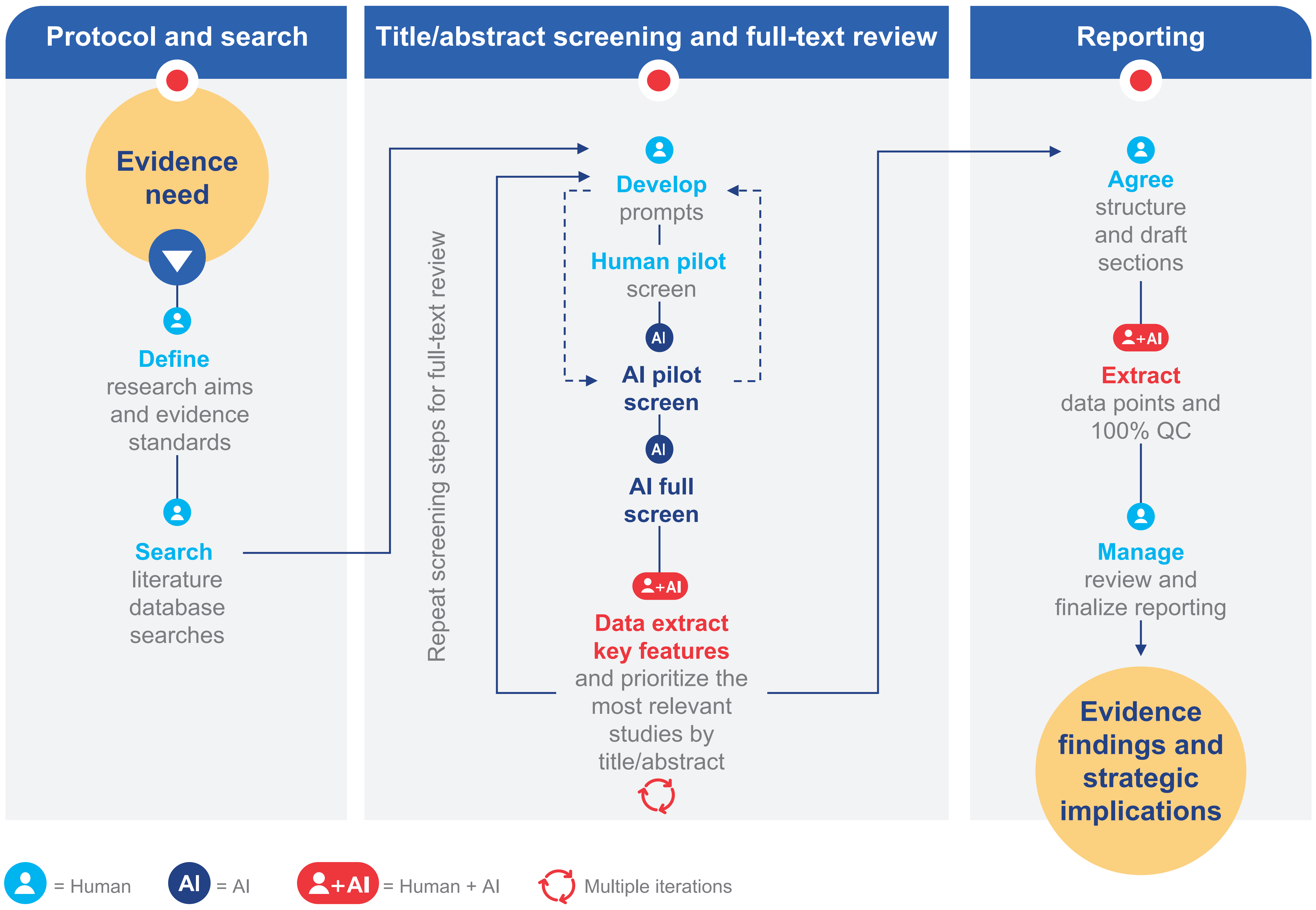
- SMEs then reviewed the full-text PDFs of the prioritized articles and extracted data from the final set of included articles, supported by the LLM tool, Elicit. SMEs reported the findings.

Quality checks

- SMEs refined prompts for title/abstract screening based on the initial responses, checking the AI versus human decision.
- SMEs spot-checked the data extraction used for prioritization, checking against the original abstract that data were extracted fully and correctly.
- SMEs checked every data point in the full data extraction, ensuring all data were reported correctly.
- We also compared machine run time with the theoretically equivalent manual time.

RESULTS

- A test project identified a large volume of results: 23 997 publications (after deduplication) by OVID searching.
- The use of LLM screening was iterative.
  - The initial title/abstract screen reduced this number to 9532 publications. Approximately 10 rounds of data extraction and characterization led to 250 prioritized publications for SME review, or 1% of those originally identified, with 120 reported in full.
  - Sensitivity was dialled up for some study types (e.g. for questions related to treatment satisfaction, few studies were identified so SMEs reviewed all studies).
  - Sensitivity was dialled down for other study types (e.g. for questions related to treatment goals, many studies were identified so SMEs set a threshold of 10 000 patients and raised this for countries with multiple studies [the biggest included > 4 million patients]).
- We achieved more with AI than we could have achieved manually.
  - The extra characterization stage, using AI for simple data extraction, allowed us to prioritize approximately 1% of the 23 997 publications at title/abstract screen, deprioritizing the remaining 99%.
    - We then removed 50% at full-text review, extracting and reporting 0.5% of the bigger, better and most relevant studies from those originally identified.
    - In contrast, we usually expect that 10–20% of studies are included at title/abstract screening and 10–20% again after full-text review, with 1–4% of the original studies included for data extraction and reporting – often the 0.5% most relevant studies are the focus of the report.
  - Therefore, this extra characterization stage allowed us to review a far higher number of publications in the time available than when reviewing without the use of AI or without the use of AI-assisted prioritization.
  - AI-assisted screening was faster than manual screening, even with time for prompting and prioritizing based on the simple data extractions.
  - Screening required time for prompt set-up, data processing, prompt iterations and SME review; this was less than 20% of the time we would theoretically require if the screening was fully manual and if all 23 997 publications were screened manually.
  - The total machine run time was approximately 50 hours.



Protocol and search

Evidence need

Define research aims and evidence standards

Search literature database searches

Title/abstract screening and full-text review

Repeat screening steps for full-text review

Develop prompts

Human pilot screen

AI pilot screen

AI full screen

Data extract key features and prioritize the most relevant studies by title/abstract

Reporting

Agree structure and draft sections

Extract data points and 100% QC

Manage review and finalize reporting

Evidence findings and strategic implications

Figure 1. Our literature review process with the extra characterization phase at title/abstract screening.
AI, artificial intelligence; QC, quality control.

Conclusions

- This approach, supported by LLMs, allows rapid reviews of large bodies of literature based on sensitive, reproducible Boolean searches.
- The iterative approach to screening and data extraction relies on SMEs developing prompts based on their expert understanding of the questions we are asking.
- We can adjust sensitivity and rapidly prioritize the most relevant studies, resulting in reports that include the bigger, better studies and that rapidly give answers and inform the next steps.
- This opens up new opportunities in literature review because we review quickly and expertly across high volumes of citations.

Presented at International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Europe 2025, 9–12 November 2025, Glasgow, UK