MSR42

Automating PICOS Criteria Assessment in Systematic Reviews with LLMs: Insights From Two Case Studies

Theodora Oikonomidi[1], Lenon Mendes [2], Ketevan Rtveladze[3], Inês Guerra[3]

[1]IQVIA, Athens, Greece; [2]IQVIA, Chicago, US; [3]IQVIA, London, United Kingdom

OBJECTIVES

- Study selection for inclusion in systematic literature reviews (SLRs), based on prespecified population, intervention, comparison, outcome, and study design (PICOS) criteria is fundamental to evidence synthesis. However, this process is labour-intensive and time-consuming.
- This proof-of-concept (POC) study evaluates the performance of a large language model (LLM) in assessing whether scientific article abstracts meet PICOS criteria.

Table 1. Key aspects of SLR test cases

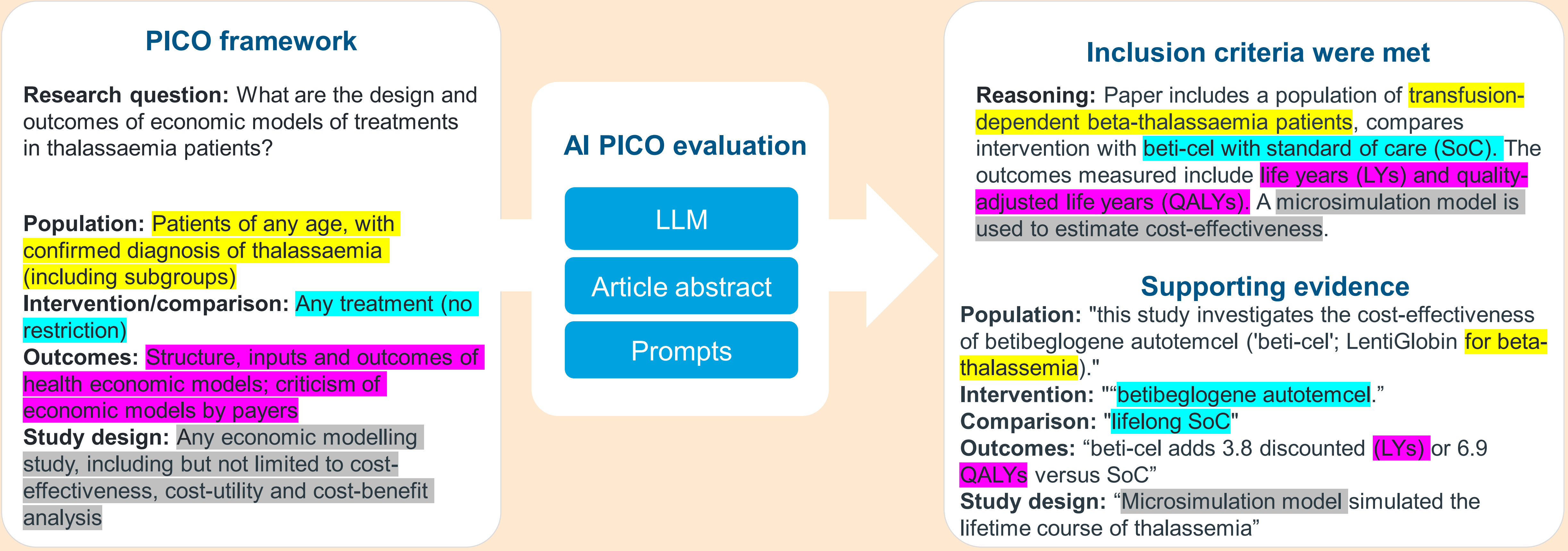
	Test case 1	Test case 2
Disease area	Thalassaemia	Muscle-invasive bladder cancer
Eligible studies	Economic modelling studies, relevant SLRs	Economic modelling studies, cost and health care resource use (HCRU) studies, relevant SLRs
Abstracts (N)	833*	1,715*

*Abstracts eligible for screening by the PICOS screener; excludes search results where only the title was available.

METHODS

- A prompt was developed to identify PICOS elements in titles and abstracts **(Figure 1).**
- Subject matter experts (SMEs) tested how accurately this prompt categorised abstracts for inclusion/exclusion compared to a manual gold standard (i.e., decisions by two humans). Two completed economic SLRs were used as test cases **(Table 1).**
- Sensitivity (i.e. the ability of the LLM to correctly classify included studies), specificity (i.e. the ability of the LLM to correctly classify excluded studies), precision (i.e. proportion of classified includes that were correct) and accuracy (i.e. proportion of total correctly classified abstracts) were calculated. LLM/human decision discrepancies were analysed by a SME.

Figure 1. PICO screener input, process and output in the assessment of abstracts for case 1



PICO framework

Research question: What are the design and outcomes of economic models of treatments in thalassaemia patients?

Population: Patients of any age, with confirmed diagnosis of thalassaemia (including subgroups)
Intervention/comparison: Any treatment (no restriction)
Outcomes: Structure, inputs and outcomes of health economic models; criticism of economic models by payers
Study design: Any economic modelling study, including but not limited to cost-effectiveness, cost-utility and cost-benefit analysis

AI PICO evaluation

LLM

Article abstract

Prompts

Inclusion criteria were met

Reasoning: Paper includes a population of transfusion-dependent beta-thalassaemia patients, compares intervention with beti-cel with standard of care (SoC). The outcomes measured include life years (LYs) and quality-adjusted life years (QALYs). A microsimulation model is used to estimate cost-effectiveness.

Supporting evidence

Population: "this study investigates the cost-effectiveness of betibeglogene autotemcel ('beti-cel'; LentiGlobin for beta-thalassemia)."
Intervention: "“betibeglogene autotemcel.”
Comparison: "lifelong SoC"
Outcomes: “beti-cel adds 3.8 discounted (LYs) or 6.9 QALYs versus SoC”
Study design: “Microsimulation model simulated the lifetime course of thalassemia”

RESULTS

- Across the two cases, sensitivity ranged from 67% to 80% and specificity from 97% to 98%.
- Precision ranged from 73% to 78% and accuracy from 92% to 97%.
- In test case 1, the LLM correctly included 80% of the 46 abstracts included by humans **(Table 2);** nine abstracts were misclassified as excluded by the LLM. This error had minimal impact on final inclusions, as none of the nine abstracts were included after full-text review.
- In test case 2, the LLM correctly included 67% of the 258 abstracts included by humans **(Table 3);** 85 abstracts were misclassified as excluded by the LLM. Of the 85 abstracts, 19 were later included after full-text review.
- Analysis of the LLM’s exclusion rationale **(Figure 2)** revealed that in most cases, incorrect exclusions were due to misclassification of the study design or the population.
- The findings suggest that refining the prompt PICOS could enhance performance (e.g. by providing definitions of economic/HCRU study designs, or providing precise instructions for the handling of ‘grey area’ studies with mixed populations)

Table 2. PICOS screener accuracy – test case 1

		PICOS screener decision		
		Include	Exclude	
Human decision	Include	37	9	80% sensitivity
	Exclude	14	773	98% specificity

Table 3. PICOS screener accuracy – test case 2

		PICOS screener decision		
		Include	Exclude	
Human decision	Include	173	85	67% sensitivity
	Exclude	49	1,408	97% specificity

Figure 2. Analysis of abstracts excluded by the PICOS screener, but included by humans

Multiple diseases
Population not recognised in studies enrolling multiple populations, including thalassaemia

Implicit diagnosis
Population not recognized where patients were implicitly described as thalassaemic, e.g., "patients with transfusional iron overload"

HCRU in clinical studies
Outcomes such as hospitalisation rate or length of hospital stay, reported in clinical studies, were not recognised as HCRU outcomes

Broad SLRs
Study design not recognized in broad SLRs not focused on economic models, e.g., SLRs of economic burden, retained by humans for reference-checking

Multiple cancer stages
Population not recognised in studies enrolling patients with multiple stages of the same cancer, including muscle-invasive cancer as a subgroup

CONCLUSIONS

- LLM performance in identifying PICOS elements and making inclusion decisions at the abstract level is promising. Importantly, missed inclusions might be minimised by further elaborating PICOS criteria (e.g., by providing definitions of the study designs of interest, instructions for ‘grey area’ study selection such as handling studies with population subgroups) and/or providing examples of relevant studies.
- Future work should focus on developing and prospectively validating integrated workflows that incorporate LLMs alongside human reviewers in the SLR process.

Abbreviations: GenAI: Generative Artificial Intelligence; HCRU: Health care resource use; LLM: Large language model; LY, Life year; PICOS: Population, intervention, comparison, outcome, study design; POC: Proof-of-concept; QALY: Quality-adjusted life year; SLR: Systematic literature review; SME: Subject matter expert; SoC: Standard of care; UK: United Kingdom.

Correspondence: Inês Guerra (Ines.Guerra@iqvia.com)

International Society for Pharmacoeconomics and Outcomes Research Europe 2025, November 10, 2025, Glasgow, UK