Accelerating Dynamic HTA Landscaping in Oncology Through Autonomous Generative AI–Driven Multilingual Data Extraction

Manuel Cossio,[1,2] Lilia Leisle,[3]
[1]Cytel, Inc., Geneva, Switzerland; [2]Universitat de Barcelona, Barcelona, Spain; [3]Cytel, Inc., Berlin, Germany.

Background

Health Technology Assessment (HTA) in the European Union (EU) has entered a new era with the introduction of the EU Joint Clinical Assessment (JCA) (1). The assessment scope will reflect the consolidated evidence requirements of all JCA Member States (2), defined using the Population, Intervention, Comparator, Outcomes (PICO) framework. Anticipating these PICO requests is key for JCA readiness and success and usually largely relies on an in-depth review of prior HTA decisions across the JCA Member States. However, HTA reports are often lengthy, heterogeneous, and published in multiple languages, making systematic data extraction both time-consuming and resource intensive.

Recent advances in large language models (LLMs) have enabled the development of autonomous agents capable of extracting structured information from unstructured and multilingual sources (3). While these approaches offer the potential to scale HTA evidence synthesis and support dynamic PICO simulations, challenges remain in ensuring accuracy, completeness, and robustness—particularly in the presence of nuanced, context-dependent information. This study investigates the use of LLM-based agents for automated HTA data extraction, with a focus on improving performance through prompt refinement and evaluating their applicability in regulatory-relevant settings.

Objectives

- To develop autonomous LLM-based agents for structured extraction of information from multilingual HTA reports to support EU JCA PICO simulation exercises or other HTA landscaping applications.
- To evaluate the performance of the developed LLM-based agents in terms of accuracy, completeness, and consistency across diverse HTA report formats and languages.

Methods

2.1 Data Sources and Data Extraction Framework

Publicly available HTA reports for osimertinib (Tagrisso) as a treatment of non-small cell lung cancer with an EGFR T790M mutation were selected from three European jurisdictions: Spain (4,678 words) (4), the Netherlands (2,512 words) (5), and France (9,876 words) (6). These documents were chosen to reflect variability in language, structure, and level of detail typical of HTA outputs across agencies.

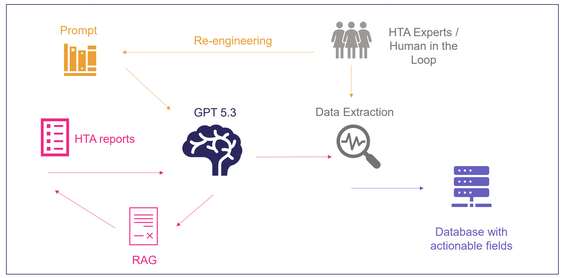
A structured data extraction framework was developed based an expert-generated questionnaire designed to capture both standard PICO elements and context-specific (CS) information, including methodological considerations, reasons for acceptance or rejection of evidence, and additional critique points relevant to HTA decision-making. The general categories are presented in Figure 2. This framework was designed to align with EU JCA requirements, ensuring relevance for downstream PICO simulations.

2.2 LLM-Based Agent Design and Prompting Strategy

Two consecutive LLM-based agents were developed to perform autonomous information extraction. Both agents were implemented using an OpenAI GPT-5.3–class model and integrated within a retrieval-augmented generation (RAG) framework, in which each HTA report was indexed and dynamically queried to provide relevant context for each question (Figure 1). Agent 1 employed a general prompting strategy, applying the same set of questions directly to each HTA report without additional contextual guidance. Agent 2 extended this approach by incorporating targeted clarification instructions within select questions, aimed at improving interpretation of nuanced or CS content. Both agents processed the full text of each report in original language and generated structured responses for each question in English, enabling systematic comparison of extraction performance across prompting strategies. The agents were implemented in a fully automated pipeline to ensure consistency in execution and eliminate variability introduced by manual intervention (Figure 1).

Methods (cont.)

Figure 1. Agentic framework for HTA data processing and extraction.



Abbreviations: HTA, health technology assessment, RAG, retrieval augmented generation, GPT, generative pre-trained transformer.

2.3 Performance Evaluation and Scoring Criteria

Agent performance was evaluated by a human expert using a custom scoring framework assessing accuracy, completeness, and the presence of hallucinations (defined as content not supported by the original document). Each agent-generated response was assigned up to two points: One point for factual correctness and one point for completeness relative to the source text. Half-points were scored whenever either of the aspects was not fully satisfied, according to human judgement. Responses containing hallucinations were punished with a total score of zero, independent of the scoring in accuracy or completeness. Aggregate performance scores were calculated per HTA report and agent across all responses, allowing comparison between agents and assessment of the impact of prompt refinement on data extraction quality.

Results

- Both agents successfully executed comprehensive data extractions across three HTA reports (Figure 2), utilizing human expert-generated frameworks. Agent 2, supplied with more granular prompting, demonstrated superior overall performance. Notably, only one instance of hallucination occurred (Spain, Agent 1, Category: pivotal trial population; Figure 2), indicating high reliability across multilingual sources.
- Across all reports, high-level questions—such as regulatory label, intervention name, and listing submitted clinical trials—were extracted with high accuracy and completeness, requiring no prompt adjustments (Figure 2). These elements were typically explicitly stated in the source documents, enabling reliable identification by both agents. Evidence that has not been accepted in HTA was captured surprisingly well too (including the reasoning), which may be attributed to the fact that these decisions were outlined in great detail across all reports.
- Low-performing categories requiring further Agent optimization (and human attention during the validation of AI-extracted data) are listed in Figure 2.
- The HAS opinion (France) seemed most amenable to AI-assisted extractions, likely due to its systematic structure and detailed descriptions. Agent 2's marginal score deficits in these extractions were primarily linked to response incompleteness rather than compromised accuracy (Figure 2).
- The AEMPS therapeutic positioning report (Spain) presented the greatest technical challenge but also exhibited the most significant gains following prompt refinement (Figure 2). Agent 1's error rate (10/19 responses lacking accuracy and/or completeness) was reduced by 50% in Agent 2 (5/19), with only a single instance of diminished accuracy (Figure 2). The hallucination detected for Agent 1 was abolished in Agent 2 (Figure 2).
- Of note, occasionally Agent 2 underperformed relative to Agent 1 (specifically, in posology extractions from Spanish and Dutch reports and in providing a complete list of accepted outcomes based on the Dutch report; Figure 2). Furthermore, increased prompt specificity failed to yield performance gains in a few categories (Figure 2), requiring further investigation.

Results (cont.)

Figure 2. Performance of Agent 1 and Agent 2 in extracting standard PICO elements and content-specific information from HTA reports.

CATEGORIES targeted by expert-generated questions		France (HAS)		Spain (AEMPS)		Netherlands (ZiN)	
		AGENT 1	AGENT 2	AGENT 1	AGENT 2	AGENT 1	AGENT 2
POPULATION	EMA label	2	2	2	2	2	2
	Population under HTA (definition + LoT)	2	2	1.5	1.5	2	2
	Reimbursed population	2	2	2	2	2	2
	Pivotal trial population	1.5	1.5	0*	1.5	1.5	1.5
	Subgroups assessed	2	2	1.5	2	1.5	2
INTERVENTION	Intervention under HTA	2	2	2	2	2	2
	Intervention posology	2	2	2	1.5	2	1.5
COMPARATORS	Comparators in pivotal trial	2	2	1.5	2	1.5	2
	Comparators accepted/suggested by HTA body	2	2	1.5	1.5	1.5	1.5
	Comparators not accepted (incl. reasons)	2	2	2	2	2	2
OUTCOMES	Efficacy endpoints in pivotal trial	2	2	1.5	1.5	2	2
	Safety endpoints in pivotal trial	0	2	0	2	0	2
	Outcomes accepted (incl. reasons)	1.5	2	1.5	2	2	1.5
	Outcomes not accepted (incl. reasons)	2	2	2	2	2	2
CLINICAL TRIALS ASSESSED	Name/ Ref ID per trial	2	2	2	2	2	2
	Design elements per trial	1.5	1.5	1	2	1.5	2
HTA DECISION SUMMARY	Decision drivers, critique (in detail)	1.5	1.5	1.5	2	2	2
	Recommendation for reimbursement (if applicable)	2	2	2	2	2	2
	HTA rating (if applicable)	2	2	2	2	2	2

* Total score was decreased to 0 due to a hallucination.

TOTAL SCORE

- 2 — accurate and complete
- 1.5 or 1 — inaccurate and/or incomplete
- 0.5 or 0 — inaccurate and incomplete

OVERALL PERFORMANCE: AGENT 1 vs AGENT 2

France		Spain		Netherlands	
AGENT 1	AGENT 2	AGENT 1	AGENT 2	AGENT 1	AGENT 2
14/19 replies	16/19 replies	9/19 replies	14/19 replies	13/19 replies	15/19 replies
4/19 replies	3/19 replies	8/19 replies	5/19 replies	5/19 replies	4/19 replies
1/19 replies	0/19 replies	2/19 replies	0/19 replies	1/19 replies	0/19 replies

LOW-PERFORMING CATEGORIES requiring further optimization:

- Pivotal trial population (missing relevant inclusion/ exclusion criteria such as age or ECOG PS)
- Pivotal trial design (missing relevant design elements such as number of trial arms or treatment cross-over)
- Comparators (e.g., listing platinum-based therapies without specifying the platinum compounds)
- Outcomes (e.g., listing PROMs without specifying the questionnaire names)

Conclusions

- **Expert-guided prompt refinement significantly improved autonomous extraction of both standard and content-specific information from multilingual HTA reports. Hence, LLM-based agents show promise for scalable HTA data extraction to support EU JCA PICO predictions or other HTA landscaping applications.**
- **Of note, inclusion of human experts in the loop is essential to control for hallucinations, verify completeness and accuracy of extracted data and thereby to ensure reliability of results.**
- **In future, further categories will be added for extraction. Moreover, Agent 2 will be tested on HTA reports from other European jurisdictions (in original languages) as well as on other therapies; further methodological improvement will be undertaken to optimize its performance.**

References

1. Regulation (EU) 2021/2282 on health technology assessment
2. Guidance on the scoping process (2024)
3. Ntinopoulos, Vasileios, et al. "Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation." BMJ health & care informatics 32.1 (2025): e101139.
4. Informe de Posicionamiento Terapéutico de osimertinib (Tagrisso®) en el cáncer de pulmón no microcítico con presencia de la mutación T790M (2018)
5. Osimertinib als behandeling van NSCLC met een EGFR-T790M-mutatie (2017)
6. TAGRISSO 40 mg, comprimé pelliculé B/30 (CIP: 34009 300 476 4 4), TAGRISSO 80 mg, comprimé pelliculé B/30 (CIP: 34009 300 476 5 1); Commission De La Transparence Avis 13 septembre 2017

Abbreviations: AEMPS, Agencia Española de Medicamentos y Productos Sanitarios; AI, artificial intelligence; EU JCA, European Union Joint Clinical Assessment; HAS, Haute Autorité de Santé; HTA; Health Technology Assessment; LLM; Large Language Model; EMA; European Medicines Agency; PICO; Population, Intervention, Comparators, Outcomes; Ref ID, reference identification number; ZiN, Zorginstituut Nederland

Disclosures and acknowledgements

The study was investigator initiated. All authors are employees of Cytel, Inc. MC is also a researcher at Universitat de Barcelona.