EVALUATING THE USE OF LARGE LANGUAGE MODELS (LLM) IN SUMMARISING TREATMENT GUIDELINES: A PILOT STUDY

MSR97

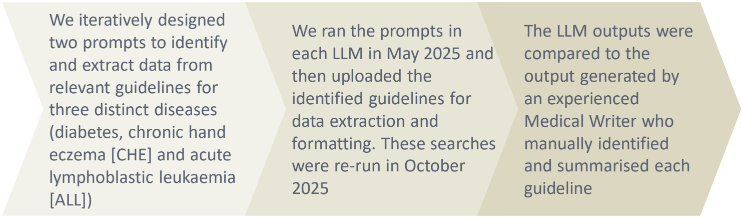
Karakusevic, A[1], Heminsley CE[1], Hopwood, NP[2]
[1]RJW&partners, Filleigh, England, UK; [2]Axia MedComms, Manchester, England, UK.

RJW&partners

Introduction

- Under the EU's Health Technology Assessment (HTA) Regulation, any submission to the Joint Clinical Assessment (JCA) made after 12 January 2025 must describe current clinical management, including the care pathway and variations across European-level clinical guidelines[1]
- This study therefore evaluated the accuracy, readability and time-efficiency of LLM-generated treatment guideline summaries for inclusion in submission dossiers, compared to manual compilation, to support the development of dossiers

Methods



We iteratively designed two prompts to identify and extract data from relevant guidelines for three distinct diseases (diabetes, chronic hand eczema [CHE] and acute lymphoblastic leukaemia [ALL])

We ran the prompts in each LLM in May 2025 and then uploaded the identified guidelines for data extraction and formatting. These searches were re-run in October 2025

The LLM outputs were compared to the output generated by an experienced Medical Writer who manually identified and summarised each guideline

Final prompts used in this study

Guideline identification: *"I am a Medical Writer working on a Global Value Dossier for a product to treat [disease]. I want to identify publicly-available guidelines across Europe and the USA for [disease]. Identify all publicly-available guidelines for [disease] in Adults in Europe and the USA and provide references. The guidelines and references should be peer-reviewed and not from a website page. They should be the most up to date version of the guidelines. The guidelines should not focus on paediatric patients. Provide the available guidelines in a bullet point list format."*

Guideline data extraction: *"I am a Medical Writer working on a Global Value Dossier for a product to treat [disease]. I want to extract and summarise data from guidelines across Europe and the USA for the treatment [disease] in adults. Extract and summarise the guidelines in a structured table format. The data I wish to be included in the table from each guideline are:*
- The country/ region of origin for the guideline
- The name of the guideline, year of publication, and reference
- Recommended treatments across different treatment phases
- Include treatment options in each treatment phase for all relevant patient sub-populations
Only extract and summarise data from the attached guidelines. The table should be written in British English, using a professional tone. The extracted data should be presented using bullet points and should be concise. Please define any abbreviations used in the table below the table as a footnote."

Results

Table 1: Visual summary of the results

	ChatGPT-4o			Gemini 2.5 Flash			Microsoft 365 CoPilot		
	Diabetes	CHE	ALL	Diabetes	CHE	ALL	Diabetes	CHE	ALL
Guideline identification	🟩	🟢	🟢	🟩	🟢	🟢	🟢	🟢	🟢
Completeness	🟩	🟢	🟧	🟢	🟢	🟧	🟢	🟢	🟧
Error rate	No errors	Low	High	No errors	Medium	High	No errors	Low	High
Clarity and readability	🟩	🟩	🟧	🟧	🟧	🟧	🟩	🟩	🟧

Green indicates almost perfect/ no changes required. Light green indicates minor changes needed (>75% complete). Orange indicates moderate changes needed (>50% complete). In the error rate row, "low" indicates ≤5 mistakes, "medium" indicates >5 to ≤15 mistakes and "high" indicates >15 mistakes versus human identification and extraction.

Guideline identification:

- All LLMs successfully identified the EU guidelines for ALL (European Society of Medical Oncology), CHE (Thyssen 2020), and diabetes (European Association for the Study of Diabetes) **(Table 1)**
  - All LLMs successfully identified the US guidelines for ALL (National Comprehensive Cancer Care Network) and diabetes (American Diabetes Association)
- ChatGPT-4o identified >75% of the guidelines overall, including some country-specific guidelines (ALL, Onkopedia; CHE, Bauer 2023, Silvestre Salvador 2020; Diabetes, NICE)
- Country-specific consensus-based guidance was identified less frequently, likely due to the prompt not being tailored for that purpose
  - Subsequent prompting often led to the identification of country-specific guidelines

Completeness:

- Diabetes guideline outputs were the most complete, with ChatGPT-4o successfully differentiating treatments by patient characteristics such as age and pregnancy
- For CHE, Microsoft 365 Copilot identified all treatments, while ChatGPT-4o and Gemini 2.5 Flash missed several later-line therapies
- The ALL guidelines had the lowest completeness across all LLMs, with ChatGPT-4o failing to follow treatment pathways or differentiate by age in NCCN outputs, although extraction of the ESMO guidelines was mostly accurate. Gemini 2.5 Flash struggled to separate treatment stages, and Microsoft 365 Copilot missed multiple treatments and patient-specific variations (e.g. patient characteristics [age, comorbidities], prior treatments received, disease subtype, or measurable residual disease status) for ALL treatment
- A follow-up prompt referencing specific guideline figures in the ALL guidelines, identified by the human extraction, did not improve the output generated by the LLMs
- LLM outputs often had lower word counts than those of medical writers, typically due to missing treatments, patient subgroups, or gaps in understanding treatment logic and flow

Error rate, including hallucinations:

- Errors were observed in both the CHE (n=12) and ALL (n=80) outputs, while no errors were identified versus human extraction for the diabetes guidelines
- Overall, the most common errors were omitted treatments (n=62) and lack of differentiation for treatment lines by patient subgroups (e.g. patient characteristics, prior treatments received, disease subtype, n=16) **(Figure 1)**
- Hallucinations were also observed across all LLMs (n=14), with LLMs often populating treatment stages in the absence of specific guidance **(Figure 1)**
  - The majority of the hallucinations (n=10) were observed across the LLM outputs for ALL
  - Two hallucinations were observed in the ChatGPT-4o output and one in the Gemini 2.5 Flash output for CHE; these concerned additional first-line treatments for severe CHE and treatment potency descriptions not found in source materials

Clarity and readability:

- ChatGPT-4o and Microsoft 365 Copilot produced readable outputs with well-formatted tables and consistent use of language for diabetes and CHE
- Inconsistencies in treatment descriptions and pathway steps appeared across all LLM outputs from ALL treatment guidelines
- ChatGPT-4o showed logical reasoning by referencing prior content instead of duplicating text within the output generated
- Gemini 2.5 Flash failed to follow formatting instructions, producing unstructured outputs with random characters

Time savings:

- LLMs showed modest time savings (~10–15%) compared to manual summarisation, due to the time needed to iteratively develop and refine a usable prompt
- Creation of the prompts required multiple (>10) iterations to refine the output and ensure these were aligned with the human searches/extraction

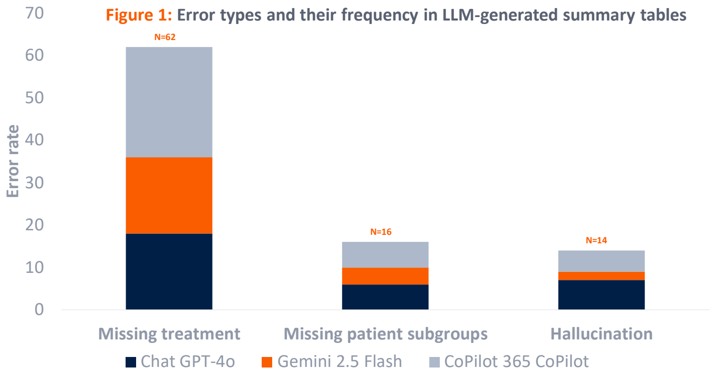


Figure 1: Error types and their frequency in LLM-generated summary tables

Discussion and conclusions

- This pilot study had several limitations: guideline documents were manually uploaded rather than identified and extracted by the models; multiple prompt revisions were required; LLMs used in this study could only analyse publicly-available data that was not behind a paywall or that required an account to access data. LLM knowledge may also be outdated due to time lags in data updates, and the prompt wording did not explicitly request country-level guideline identification. Additionally, ChatGPT-4o used a premium feature, which may have contributed to better performance compared to the free versions of Gemini 2.5 Flash and Microsoft 365 CoPilot
- LLMs provide a foundation for identifying and summarising treatment guidelines to support dossier development. However, a Medical Writer's review is essential, given the human input required to ensure an output suitable for inclusion in HTA and value dossiers. Learnings from this pilot will reduce prompt design iterations, saving time in the future
- Greater time savings may be observed for diseases with simple treatment pathways, while complex conditions, with multiple patient subgroups and lines of treatment would still likely require thorough content review by a medical writer

Abbreviations: ADA, American Diabetes Association; ALL, acute lymphoblastic leukaemia; CHE, chronic hand eczema; EASD, European Association for the Study of Diabetes; ESMO, European Society for Medical Oncology; HTA, health technology assessment; JCA, Joint Clinical Assessment; LLM, large language model; NCCN, national Comprehensive Cancer Network; NICE, National Institute of Healthcare and Excellence.

References: **1.** European Commission (2025). Joint Clinical Assessments. https://health.ec.europa.eu/health-technology-assessment/implementation-regulation-health-technology-assessment/joint-clinical-assessments_en