Exploring the Promise of Generative Artificial Intelligence (AI) for Coding and Analysing Qualitative Patient Data: Results from a Pilot Study in Non-Hodgkin's Lymphoma

PCR94

Karen Bailey,[1] Anne Skalicky,[2] Carla Dias Barbosa,[3] Sonya Stanczyk,[2] Paul Cordero[4]

[1]Thermo Fisher Scientific, London, UK; [2]Thermo Fisher Scientific, Waltham, MA, USA; [3]Thermo Fisher Scientific, Lyon, France; [4]Sanofi, Reading, UK

Background

- Computer Assisted Qualitative Data Analysis Software (CAQDAS) programmes have integrated machine learning and artificial intelligence (AI) into their tools.
- These integrations aim to accelerate time-intensive processes related to qualitative data quality control, analysis, and reporting.
- The use of AI with sensitive patient experience data must comply with General Data Protection Regulation (GDPR) standards and documentation.

Objectives

- **Evaluate how AI performs compared with human tasks and processes.**
- **Provide recommendations for AI use-cases that can meet regulatory standards, using both primary and secondary analysis of qualitative data.**

Methods

Figure 1. Methods of primary data collection

Design	Non-interventional, cross-sectional, qualitative study.
Type of interviews	Concept elicitation (CE) and cognitive interviews (CI) to examine the relevance of PROs.
Population	US adult patients diagnosed with DLBCL or MCL.
Data collection	60-minute telephone or web-conference interviews using a semi structured interview guide, audio recorded and transcribed verbatim, conducted between June and September 2023.
PRO being assessed	EORTC QLQ-C30, EORTC QLQ-NHL-HG29, EORTC QLQ-NHL-LG20, FACT-Lym.
Objectives of the primary data collection study	• Identify relevant concepts to refine and validate existing CDMs. • Evaluate selected PROs for use as endpoints in NHL clinical trials, focusing on readability, comprehensibility, relevance, and patients' impressions of items, instructions, and recall periods.
Ethics	Approval granted by US WCG IRB. Participants provided informed consent for secondary use of their data.

Abbreviations: CDM = conceptual disease model; DLBCL = diffuse large b-cell lymphoma; EORTC = European Organisation for Research and Treatment of Cancer; FACT-Lym = Functional Assessment of Cancer Therapy–Lymphoma; HG29 = High Grade Module 29; IRB = institutional review board; LG20 = Low Grade Module 20; MCL = mantle cell lymphoma; NHL = Non-Hodgkin lymphoma; PRO = patient-reported outcome; QC = quality control; QLQ = Quality of Life Questionnaire; WCG IRB = Western Institutional Review Board Copernicus Group

- ATLAS.ti was the CAQDAS used for the case study (**Figure 2**).
- The human coding and the AI coding approaches were compared in terms of the:
  - Time and human resources required to finalise coding of one transcript and all the transcripts analysed
  - Coding issues identified in quality control (QC)
  - Number and type of codes applied to one selected transcript
  - Relevance of final analysis in relation to the study objectives

Figure 2. Data analysis approaches

	Human Analysis (ATLAS.ti v22)	AI Coding	Intentional AI Coding
No. transcripts analysed	30	30	16 (due to AI restrictions)
Coding	• CE: Deductive (interview guide and CDM) and inductive (participant-driven) codes. • CI: Deductive codes to encompass comprehension and relevance of the items, instrument instructions, and response options.	• AI coding function.	• Intentional AI coding function. • Researchers guided by entering a summary of the interview guide questions and two instruction statements.
QC	• Dual coding of first transcript. • Coding monitored by a qualitative data manager. • ICA of ≥80% was achieved for the first transcript.	• Human review not completed.	• Researchers reviewed the code dictionary and coding at a high level. • One transcript underwent detailed human QC.

Abbreviations: AI = artificial intelligence; CDM = conceptual disease model; CE = concept elicitation; CI = cognitive interview; ICA = intercoder agreement; QC = quality control

About ATLAS.ti (v25) AI function

- ATLAS.ti v25 (released in April 2024) has an AI function that uses OpenAI to perform qualitative analysis.
- If the AI coding function is not enabled, no data is uploaded to OpenAI.
- Before starting AI coding, ATLAS.ti asks for your consent before uploading data to ATLAS.ti servers, as well as OpenAI servers.
- The proprietary arrangement with OpenAI ensures that any data provided is not stored and will not be used to train the OpenAI large language model (LLM).
- **AI coding:** Codes inductively without human input to help categorise, interpret, and make sense of data as part of initial cycles of coding.
- **Intentional AI coding:** Allows human input of "intentions," which are akin to prompts and context about the research, to inform the automatic coding.

Results

Figure 3. Time and human resources[a] required to finalise coding of one transcript and all the transcripts analysed

Coding Approach	Human Coding (30 transcripts)	AI Coding (30 transcripts)	Intentional AI Coding (16 transcripts)
No. codes and categories	1,298 codes grouped into 40 categories	2,318 codes grouped into 578 categories	2,350 codes grouped into 27 categories
Time to code one transcript (4,340 words)	~120 minutes	50 seconds (no human QC)	Human QC of coding: ~180 mins
Time/human resources for initial coding	Draft code book: ~12 hours Coding of 30 transcripts: ~70 hours QC of coding and ICA: ~30 hours	AI function: ~10 minutes	Human development of intentions (prompts) to help generate code categories:~ 30 mins AI function: ~15 minutes
Time/human resources for QC and finalise coding	QC of coding and ICA: ~30 hours	Human review was not deemed feasible due to the excessive codes generated	Human review of all transcripts to remove codes only applied to interviewer questions (**Figure 5**): ~110 mins Estimated time to QC all transcripts: ~ 48 hours

[a] For human coding, the resources consisted of two senior and four junior qualitative researchers. For both AI approaches, the resources consisted of two senior qualitative researchers; Abbreviations: AI = artificial intelligence; ICA = intercoder agreement; QC = quality control

Results (cont.)

Figure 4. Coding issues identified in QC of AI Methods

Coding Approach	Human Coding	AI Coding	Intentional AI Coding
Coding issues	• Amount of text coded to a particular concept. • Differences between coders on the interpretation of inductive concepts related to emotional wellbeing e.g., being upbeat, positivity.	• Application of codes to interviewer questions without participant responses. • Application of irrelevant codes • General coding of symptoms "health concerns," "physical symptoms," "physical ailments." • Inability to discern between similar concepts at the category or code level.	• Application of multiple wrong codes to a quotation (**Figure 5**, Example 1). • Double coding of concept as a symptom and impact. • Inability to determine which PRO the participant was describing (**Figure 5**, Example 2). • Inability to discern between similar concepts at the category or code level, (e.g., "absence of impact" and "absence of impacts" or "bother rating" and "bothersomeness rating").

Abbreviations: AI = artificial intelligence; PRO = patient-reported outcome

Figure 5. Examples of issues with the Intentional AI Coding

Example 1. Application of multiple inappropriate codes to a quotation (transcript text to the left, codes applied to the right)

Interviewer: Anything in relation to kind of urinary issues at all?
Participant 001-002: No urinary issues, no. No.

Interviewer: Okay, and how about anything in relation to skin, so bruising or anything like that?
Participant 001-002: I don't think anything abnormal, no. No.

Interviewer: Anything respiratory at all?
Participant 001-002: Spiritual?

Interviewer: Respiratory. So any breathing difficulties or anything like that?

- ◆DLBCL Impacts: No Impacts
- ◆DLBCL Impacts: No Symptoms
- ◆DLBCL Impacts: Urinary Issues
- ◆EORTC QLQ NHL-LG20 Views: No Views
- ◆FACT-Lym Improvement: No Improvements
- ◆FACT-Lym Meaningful Improvement: No Suggestions
- ◆FACT-Lym Views: No Views
- ◆MCL Impacts: No Impacts
- **◆MCL Symptoms: No Symptoms**
- **◆MCL Symptoms: Urinary Issues**
- ◆QLQ-C30 Views: No Views
- ◆QLQ-NHL-HG29 Improvements: No Improvements
- ◆QLQ-NHL-HG29 Meaningful Impr…: No Suggestions
- ◆QLQ-NHL-HG29 Views: No Views

Example 2. Incorrectly coding statements associated with the PRO questionnaire (transcript text to the left, codes applied to the right)

Interviewer: So the next one I want to talk about is the one EORCTQLQ and ending with 2-9. So it has a long name. So when you completed that questionnaire, again, what was your overall impression? Any feedback that you have?

- **◆EORTC QLQ NHL-LG20 Views: Feedback**
- **◆EORTC QLQ NHL-LG20 Views: Overall Impression**

Participant 001-007: No, it was straightforward. I didn't have any problems at all filling it out. But pretty straightforward questions and pretty easy to answer.

- ◆QLQ-C30 Views: Ease of Use
- ◆QLQ-C30 Views: Simplicity

Abbreviations: DLBCL = diffuse large b-cell lymphoma; EORTC = European Organisation for Research and Treatment of Cancer; FACT-Lym = Functional Assessment of Cancer Therapy–Lymphoma; HG29 = High Grade Module 29; LG20 = Low Grade Module 20; QLQ = Quality of Life Questionnaire

Green text indicates correct codes. **Grey text** indicates incorrect codes.

Figure 6. Number and type of codes applied to one selected transcript 001-009

Coding Approach	Human Coding	Intentional AI Coding
No. of total codes applied	112	**Before QC: 101**; **After QC: 91** (76 codes removed, 41 newly created codes, 25 already existing codes applied, 5 codes renamed)
Strengths	• Accurate coding suitable in frequency counts. • Identification of symptom/impacts as spontaneous or probed. • Superior application of detailed deductive codes for CI. • Able to apply latent codes (beyond description).	• Suggestion of inductively driven relevant codes for symptoms and impacts not identified in human coding. • For example, coding of "fatigue" in a quotation was also double coded as "persistent tiredness" and "inadequate rest."
Weaknesses	• Lacked the identification of some inductive (participant-driven) codes in comparison to AI.	• Relied on substantial human input to correct AI codes (removal/renaming/creating codes). • Unable to identify the correct PRO questionnaire under discussion. • Most quotations required adjustment to incorporate participant response to interviewer question.

Abbreviations: AI = artificial intelligence; PRO = patient-reported outcome

Figure 7. Relevance of final analysis in relation to the study objectives

Human Coding	AI Coding
• Identification and frequency of signs and symptoms and revision of the CDMs. • Mapping of participants' reported concepts with the items covered by the PROs to draw conclusion on the relevance of PROs for each population. • Confirmation of content validity of the PROs.	• Recognition of relevant concepts not identified in the human coding, which can be useful for the creation of the initial code book. • However, due to the issues identified with the AI coding, the provision of concept frequency (symptoms and impacts) and evidence of content validity of PROs could not be provided.

Abbreviations: AI = artificial intelligence; CDM = conceptual disease model; PRO = patient-reported outcome

Conclusions

- **Given the large environmental impacts of AI, its adoption in qualitative research is justified only by the demonstration of time-saving efficiencies and quality improvements of outputs.**
- **The pilot results identified that the AI capabilities in ATLAS.ti (v25) were not suitable for the production of regulatory-grade qualitative research involving both concept elicitation and cognitive interviews and did not create time-saving efficiencies when compared with human coding.**
- **Major barriers included the inability to process more than 16 transcripts using the Intentional AI coding function and the lack of trustworthiness in the data, which required human input that was equivalent or more to the human input required to code without AI.**
- **The AI strengths lay in its ability to identify inductive (participant-driven) codes. Incorporating a hybrid approach combining the use of AI coding on 1–2 transcripts in the development of the human-driven codebook may be useful.**
- **Moving forward, the ability to train the LLM underlying the AI with the outputs of the human QC would enhance the quality and efficiency of the output.**
- **Future work should assess how the use of AI in qualitative research complies with GDPR standards and the response of regulatory and study sponsor guidelines and policies.**

Disclosures

KB, **AS**, **CDV**, and **SS** are employees of are employees of PPD™ Evidera™ Patient-Centered Research, Thermo Fisher Scientific. **PC** is employee of Sanofi and may hold stock and/or stock options in the company. This poster was funded by Thermo Fisher Scientific. Editorial and graphic design support were provided by Karissa Calara of Thermo Fisher Scientific.