Leveraging Artificial Intelligence for Thematic Analysis of Qualitative Transcripts: A Feasibility Study in Insurance Payer Interviews

Quinn Levin, Alexa C. Klimchak, Katherine L. Gooch

Sarepta Therapeutics, Inc., Cambridge, MA, USA

Key Findings

In a thematic analyses of semi-structured interviews, Copilot generated broad topics rather than meaningful themes, highlighting the continued importance of human oversight for in-depth contextual understanding.

Background

- In qualitative research, manually reviewing, coding, interpreting themes, and extracting insights are labor-intensive, expensive, time-consuming, and susceptible to human error and biases[1-3]
- Recent advancements in artificial intelligence (AI) offer a promising way to improve the speed, rigor, transparency, and/or consistency of interview analysis while also potentially reducing bias[4,5]
- ChatGPT-based, AI-augmented reflexive thematic analyses have reported alignment with human analyses while also highlighting limitations in contextual understanding[5-8]
- There are limited data addressing the application of AI in qualitative interview analysis as well as validation of AI-generated results compared with human analysis

Objective

- To assess the accuracy and robustness of themes generated using Microsoft Copilot compared with traditional human-generated themes, using semi-structured insurance payer interview transcripts as a case study

Methods

- An independent investigation on the accuracy of historical drug-spending predictions, using interferon-free direct-acting antiviral (DAA) therapies approved to treat hepatitis C virus infection as well as cell and gene therapies, was used as a case study (**see presentations here at ISPOR 2025** by Kulkarni N, et al. Posters HPR122 and HPR87)
- Semi-structured interviews with 10 payers were conducted to understand the drivers of discrepancies between national drug spending predictions and actual spending
- Anonymized transcripts of the interviews were captured in Microsoft Word, one document per interview
- Key themes were identified by human researchers; Copilot was also leveraged for thematic analysis by two researchers (**Figure 1**)
- Quality and accuracy of Copilot-generated themes were compared with human-identified themes

Methods (cont.)

Figure 1 *Copilot Approach*

01 Platform Selection
- The Copilot AI model was selected based on existing enterprise-level integration and its advanced data security features

02 Data Acquisition
- Semi-structured payer interviews related to predictive budget modeling were used
- Experienced human researchers summarized key themes traditionally

03 Prompt Development
- A targeted literature review informed initial standardized AI prompts that were refined iteratively
- Best practices included articulating the source for analysis, defining the goals, specifying context, using polite phrasing, and avoiding abbreviations
- To minimize bias, one of the two analysts conducting the Copilot AI assessment was unfamiliar with the study

04 Copilot Transcript Analysis
- Unedited, anonymized interview transcripts were loaded into Copilot
- A series of prompts were applied sequentially that included describing documents with clear direction, requesting details on themes identified iteratively, and creating a codebook for future research
- Two human analysts used the same prompts and captured responses repeating the process in three unique chats to assess reproducibility

05 Analysis
- The quality and accuracy of themes generated by Copilot were compared with those generated by experienced human researchers
- The depth and richness of insights as well as response variability between human and Copilot approaches to identical prompts across repeated analyses were assessed
- Hallucinations were documented to assess the proportion of inaccurate responses by Copilot

- Prompt engineering (**Figure 2**) utilized insights from previously published ChatGPT studies and Copilot training/best practices, with an emphasis on politeness, directness, and detail
- Prompt structure focused on the following key elements when phrasing:
  - Goal or desired response
  - Context for the task
  - Data sources for analysis
  - Expectations
- Both researchers initially adhered to an identical set of structured prompts to evaluate reproducibility; prompts were subsequently dynamically modified based on interactions with Copilot to assess quality of results
- In each phase, different prompts were used and the thematic analysis was done inductively, no preexisting coding framework or examples were given to the model, and all the codes and the themes were entirely based on Copilot's "interpretation" of the prompt

Figure 2 *Prompt Engineering and Response Workflow*

Results

- Copilot identified broad topics but had **difficulty synthesizing meaningful themes** compared with human assessments (**Figure 3**)
  - For example, "Payers confirmed that plans spent less on DAA therapies vs initial predictions" vs "Accuracy of National Drug Spending Predictions"

Figure 3 *Comparison of Key Themes Identified by Human Researchers vs Copilot*

	Human Researchers		Copilot (Sample)[a]
1	Payers were highly focused on total spend / budget implications for hepatitis C DAA therapies	1	Accuracy of National Drug Spending Predictions
2	Payers confirmed that plans spent less on DAA therapies vs initial predictions	2	Challenges in Budget Impact Prediction
3	Payers acknowledged that "lessons learned" in HCV may be applicable to other treatment areas, but with limits	3	Evolving Treatment Guidelines and Practices
4	Payers are actively forecasting / predicting plan spend for single-administration CGTs	4	Impact of Competitive Market Dynamics
5	Payers rely on a combination of internal analyses and external inputs / research to inform budget predictions	5	Role of Value-Based Contracts and Warranties
6	Generally, payers have experienced lower costs for single-administration CGTs compared with predictions	6	Patient Access and Equity Considerations
7	Payers acknowledge the difficulty in predicting total spend across a diverse set of single-administration CGTs		

[a]Representative results after prompt refinements and multiple iterations of Copilot analyses.
CGT, cell and gene therapy; DAA, direct-acting antiviral; HCV, hepatitis C virus.

- While both broad and targeted prompts showed some alignment between human and Copilot analyses, notably Copilot:
  - **Missed key insights** (eg, "Payers rely on a combination of internal analyses and external inputs/research to inform budget predictions") (**Figure 3**)
  - **Overemphasized certain concepts** despite minimal mention from interviewees (eg, listing "Role of Value-Based Contracts and Warranties" as a theme) (**Table 1**)
  - **Linked concepts that interviewees did not connect** (eg, "clinical guidelines evolved…as prices decreased") (**Table 1**)
  - **Overlooked context** (eg, despite directive prompting that enhanced Copilot's focus on prevalence predictions, the context provided by payers regarding prevalence data and analyses was overlooked)
  - Was **unable to reproduce insights** across conversations and researchers

Table 1 *Copilot Thematic Analysis From Broad, Non-directive Prompts*

Agreement With Human Researchers	Key Themes Identified		Description of Key Themes From Copilot
High	1	Impact of Competitive Market Dynamics	The competitive dynamics of the market significantly influenced drug pricing and access. Respondents mentioned: - The introduction of competing therapies led to significant price reductions - Manufacturers' strategies, such as offering substantial rebates, played a crucial role in making therapies more accessible
Medium / High: Captured some, but not all, key themes identified by human researchers	2	Accuracy of National Drug Spending Predictions	Several respondents highlighted the discrepancies between predicted and actual US spending on hepatitis C DAA treatments and single-administration CGTs. They noted that initial predictions often overestimated the spending due to factors such as: - Overestimated patient populations - Assumptions of universal uptake - Lack of consideration for discounts and rebates
Medium / High: Redundant to "Accuracy of National Drug Spending Predictions"	3	Challenges in Budget Impact Predictions	Respondents discussed the challenges in predicting the budget impact of new therapies. Key points included: - Difficulty in estimating the number of eligible patients - Variability in patient response and adherence to treatment - The impact of new therapies on existing treatment paradigms
Medium: Copilot elevated a less relevant theme and incorrectly linked (clinical) guidelines and prices	4	Evolving Treatment Guidelines and Practices	The evolution of treatment guidelines and practices was a recurring theme. Respondents noted that: - Initial guidelines were often restrictive, focusing on the sickest patients - Over time, guidelines evolved to include a broader patient population as more data became available and as prices decreased
Low: Copilot elevated a less relevant theme and did not specify this was mentioned in the context of only CGTs	5	Role of Value-Based Contracts and Warranties	Several respondents emphasized the importance of value-based contracts and warranties in managing the financial risk associated with high-cost therapies. They highlighted: - The need for outcomes-based agreements to ensure that payers receive value for their expenditures - The role of warranties in providing financial protection if therapies do not deliver the expected outcomes
Low: Copilot hallucinated theme; no mention of equitable access or underserved populations	6	Patient Access and Equity Considerations	Patient access and equity were also significant themes. Respondents discussed: - The importance of ensuring equitable access to new therapies across different patient populations - The challenges faced by underserved populations in accessing high-cost therapies

CGT, cell and gene therapy; DAA, direct-acting antiviral.

- Copilot sometimes **inaccurately extracted quotes** that were misattributed to the wrong respondent or irrelevant to the topic and sometimes generated incorrect information
  - Between 11% and 25% of quotes were "**hallucinations**," defined as AI chatbot–generated fictional, erroneous, or unsubstantiated information in response to queries[9]
  - The proportion of "hallucinations" increased with more targeted prompts
- Copilot-based **analyses were unbalanced** and did not seem to reflect all transcripts (eg, Copilot often selected quotes from only the first few uploaded interviews, leading to biased results). Additionally, Copilot struggled to accurately count the number of payer interview transcript files uploaded despite targeted prompting
- An **overemphasis on responses to initial prompts** was observed (eg, prompts for new quotes to additionally support themes were met with the same initial quotes being provided)
- There was a **lack of contextual understanding**, with Copilot struggling to grasp the context of who was making a comment and the significance of their perspective (eg, an interviewee with commercial vs Medicaid payer experience), which is crucial for accurate interpretation and understanding

Conclusions

- Across varying prompts for thematic analyses of semi-structured interview transcripts, Copilot identified only superficial topics that lacked deeper insights into underlying patterns/nuances, emphasizing the continued importance of human oversight for in-depth, contextual understanding
- Copilot-based analyses contained inaccuracies including hallucinations (eg, erroneous information), overlooking of respondents' perspectives (eg, experience with private or public insurance), and failure to analyze across all available data
- Exclusive usage of Copilot is a notable limitation of this study; further research is necessary to evaluate other AI platforms
- While AI has the potential to conduct qualitative analysis rapidly with increased objectivity and consistency, additional studies are needed to fully realize its capability to support these assessments
- Future studies should explore using Application Programming Interfaces (enabling the use of Retrieval-Augmented Generation, facilitating workflow integration and scalability, and modifying source code for parameter adjustments to reduce hallucinations) and carefully consider prompt engineering

Scan the QR code

The QR code is intended to provide scientific information for individual reference, and the information should not be altered or reproduced in any way.

Presented at International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Annual Meeting; May 13-16, 2025; Montreal, QC, Canada.

Acknowledgments & Disclosures

This study was sponsored and funded by Sarepta Therapeutics, Inc., Cambridge, MA, USA.

The authors would like to thank JP Maguzzu, Thomas Basso, and Scott Pierce for support with Copilot-related queries. Medical writing and editorial support were provided by Srividya Venkitachalam, PhD (HCG), in accordance with Good Publication Practice (GPP) 2022 guidelines (https://www.ismpp.org/gpp-2022) with funding from Sarepta Therapeutics, Inc., Cambridge, MA, USA.

QL, **ACK**, and **KLG** are employees of Sarepta Therapeutics, Inc., and may own stock and/or stock options in the company.

References

1. Bachiochi PD, Weiner SP. Qualitative data collection and analysis. In: Rogelberg SG, ed. *Handbook of Research Methods in Industrial and Organizational Psychology.* Blackwell Publishing; 2004:161-183. 2. Pope C, et al. *BMJ.* 2000;320(7227):114-116. 3. Leeson W, et al. *Int J Qual Methods.* 2019;18. doi:10.1177/1609406919887021. 4. Raiaan MAK, et al. *IEEE Access.* 2024;12:26839-26874. 5. Wachinger J, et al. *Qual Health Res.* 2024:10497323241244669. 6. De Paoli S. *Soc Sci Comput Rev.* 2024;42(4):997-1019. 7. Hamilton L, et al. *Int J Qual Methods.* 2023;22. doi:10.1177/16094069231201504. 8. Morgan DL. *Int J Qual Methods.* 2023;22. doi:10.1177/16094069231211248. 9. Kumar M, et al. *Cureus.* 2023;15(8):e43313.