ACR Open Rheumatology
Vol. 6, No. 6, June 2024, pp 375–379
DOI 10.1002/acr2.11662

American College of Rheumatology
Empowering Rheumatology Professionals

BRIEF REPORT

A Novel Approach for Mixed-Methods Research Using Large Language Models: A Report Using Patients' Perspectives on Barriers to Arthroplasty

Insa Mannstadt,[1] Susan M. Goodman,[2] Mangala Rajan,[1,3] Sarah R. Young,[4] Fei Wang,[1,5] Iris Navarro-Millán,[2] and Bella Mehta[2]

Objective. Mixed-methods research is valuable in health care to gain insights into patient perceptions. However, analyzing textual data from interviews can be time-consuming and require multiple analysts for investigator triangulation. This study aims to explore a novel approach to investigator triangulation in mixed-methods research by employing a large language model (LLM) for analyzing data from patient interviews.

Methods. This study compared the thematic analysis and survey generation performed by human investigators and ChatGPT-4, which uses GPT-4 as its backbone model, using data from an existing study that explored patient perceptions of barriers to arthroplasty. The human- and ChatGPT-4–generated themes and surveys were compared and evaluated based on their representation of salient themes from a predetermined topic guide.

Results. ChatGPT-4 generated analogous dominant themes and a comprehensive corresponding survey as the human investigators but in significantly less time. The survey questions generated by ChatGPT-4 were less precise than those developed by human investigators. The mixed-methods flowchart proposes integrating LLMs and human investigators as a supplementary tool for the preliminary thematic analysis of qualitative data and survey generation.

Conclusion. By utilizing a combination of LLMs and human investigators through investigator triangulation, researchers may be able to conduct more efficient mixed-methods research to better understand patient perspectives. Ethical and qualitative implications of using LLMs should be considered.

INTRODUCTION

Qualitative research techniques provide valuable insights into patient experiences and perceptions. An in-depth understanding of patient perspectives, coupled with the quantitative evaluation of data allows researchers to better understand health care gaps. Mixed-method studies, therefore, relies on a thorough analysis of the qualitative data.[1]

However, analyzing qualitative data, such as text from patient interviews, can be a time-consuming and challenging process that requires specialized expertise. Moreover, the credibility of mixed-method research is typically contigent on investigator triangulation, which entails multiple researchers analyzing the same data to mitigate biases and reinforce validity.[2] Recent advancements in large language models (LLMs) such as Generative Pretrained Transformer (GPT) and Bidirectional Encoder Representations from Transformers can aid in identifying patterns and themes in textual data and therefore serve as an investigator in investigator triangulation to offer an efficient and objective analysis of qualitative data.[3]

The objective of this study is to evaluate the application of LLMs, specifically ChatGPT-4, in mixed-methods research. This study will specifically compare the thematic analysis of interview text and the subsequent survey generation conducted by ChatGPT-4 with that performed by human investigators. The

[1]Insa Mannstadt, BS, BA, Mangala Rajan, MPH, Fei Wang, PhD: Hospital for Special Surgery, New York, NY; [2]Susan M. Goodman, MD, Iris Navarro-Millán, MD, Bella Mehta, MBBS, MS, MD: Weill Cornell Medicine and Hospital for Special Surgery, New York, NY; [3]Mangala Rajan, MPH: Department of Medicine, Weill Cornell Medicine, New York, NY; [4]Sarah R. Young, MSW: Binghamton University, Binghamton, NY; [5]Fei Wang, PhD: Department of Population Health Sciences, Weill Cornell Medical College, Cornell University, New York, NY.

Additional supplementary information cited in this article can be found online in the Supporting Information section (http://onlinelibrary.wiley.com/doi/10.1002/acr2.11662).

Author disclosures are available at https://onlinelibrary.wiley.com/doi/10.1002/acr2.11662.

Address correspondence via email to Bella Mehta, MBBS, MS, MD, at mehtab@hss.edu.

Submitted for publication April 27, 2023; accepted in revised form February 9, 2024.

comparison will focus on evaluating the quality of the content and assessing the time taken for these tasks. We seek to determine the potential of LLMs in enhancing mixed-method research in health care.

MATERIALS AND METHODS

This study is designed as a case report to investigate the effectiveness of using ChatGPT-4 for a mixed-methods study, including thematic analysis of qualitative data and survey generation. Study components were approved by the ethics committee of the Hospital for Special Surgery Institutional Review Board (IRB#2023-2439). All participants provided implicit consent to being in the study by completing the survey, and the study was undertaken in accordance with the Declaration of Helsinki.

Thematic analysis. As a part of a previous study (Protocol #1807019476), we conducted six semistructured interviews and one focus group of patients with arthritis needing total joint replacement to identify the barriers preventing them from undergoing surgery. The topic guide for these interviews is provided as Supplementary Table 1.

The qualitative data was analyzed and used to generate survey questions following a mixed-methods research approach. Figure 1 delineates our process, showing our collection of qualitative data, analysis of dominant themes, and generation of a survey, and shows stages of the workflow in which ChatGPT-4 performed analysis independently of the human investigators. Dominant themes regarding barriers for total joint replacement were identified by human investigators and ChatGPT-4 independently and subsequently compared.

Independent investigators. Two independent investigators (INM, SRY) used Nvivo software[4] that allows investigators to assign codes to textual data and explore themes and patterns in the data. Investigators coded a subset of the data and met to discuss the emerging codebook, reconciling themes when there was disagreement. Investigators then recoded the transcripts with the consensus codebook and compared levels of agreement between the coders in Nvivo, achieving consensus of 0.7 or higher for each theme, suggesting commonly accepted levels of agreement and consistency in applying the codebook to the data. One investigator (SRY) used the consensus codebook to code the remainder of the data, and the second investigator (INM) reviewed the coded data set, again reconciling any disagreements. After conducting six interviews and a focus group, investigators determined that thematic saturation had been reached. The human investigator approach to thematic analysis involved primarily deductive methods but also incorporated elements of inductive analysis. Human investigators used a consensus codebook and Nvivo software to identify dominant themes through deductive reasoning, and inductive analysis was used for the iterative interpretation and theme reconciliation in cases of investigator disagreement.

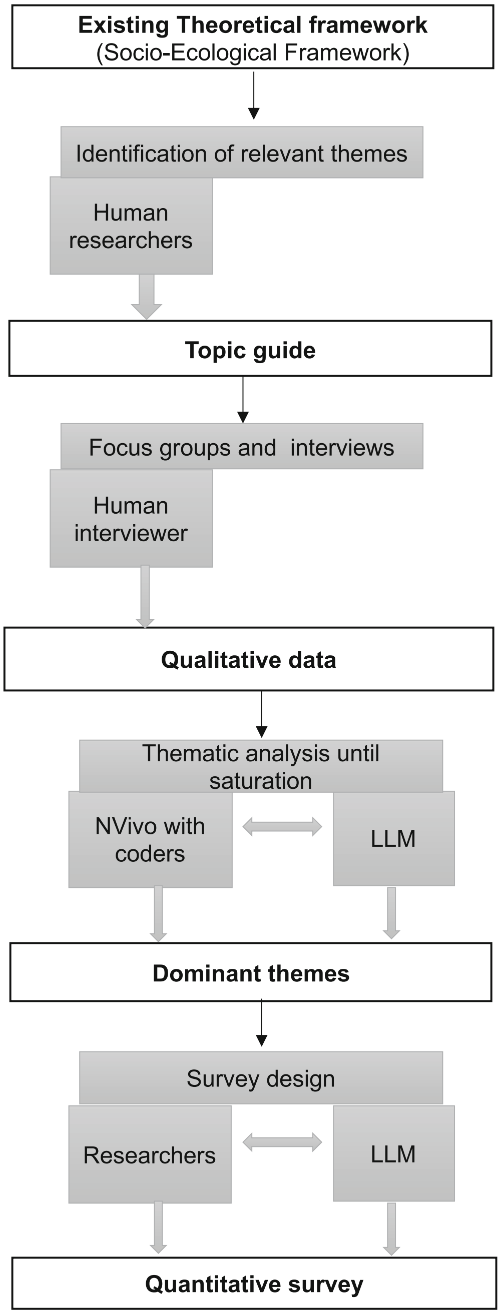
Existing Theoretical framework
(Socio-Ecological Framework)
Identification of relevant themes
Human researchers
Topic guide
Focus groups and interviews
Human interviewer
Qualitative data
Thematic analysis until saturation
NVivo with coders
LLM
Dominant themes
Survey design
Researchers
LLM
Quantitative survey

Figure 1. Workflow for gathering and examining qualitative data to create surveys; this workflow encompasses the combined efforts of Large Language Models (LLMs) and human researchers.

LLM. For each interview, a distinct chat was created, and the interview transcriptions were fed into ChatGPT-4 in segments of one thousand words due to its input limitations. The transcripts were analyzed for themes in one thousand word segments, then the next one thousand word segment was assessed. For each interview, ChatGPT-4 was prompted to generate dominant themes related to the participant's barriers to receiving arthroplasty care: "Analyze the interview transcription to identify the dominant themes related to Participant's barriers to receiving arthroplasty care."

After the themes were generated from all the interview transcripts, a list containing themes from all interviews were reentered into a new chat, and ChatGPT-4 was asked to identify the prominent themes to summarize themes and avoid repetition. The prompt given to ChatGPT-4 was, "From the following list of barriers to arthroplasty, please identify the prominent themes." In the LLM approach, ChatGPT-4 generated themes based on learned patterns and associations rather than explicit reasoning, making it a nonexclusive combination of inductive and deductive methods.

Survey generation. The dominant themes were used to create a survey that aimed to collect quantitative data on the barriers to total joint replacement identified through patient interviews. Surveys were generated by human investigators and ChatGPT-4 independently and subsequently compared.

Independent investigators. Based on the thematic analysis and quotes from patients, two independent researchers (INM, SRY) generated a list of potential survey questions that aligned with the identified themes and topic guide. The researchers pulled direct quotes from patients from the interview text to illustrate the specific experiences, perspectives, and opinions expressed by the interviewed participants. These quotes were rephrased to fit into the questions included in the final survey. The survey questions formulated by humans were created to elicit responses on a five-level Likert scale, ranging from 0 = "Not very important" to 4 = "Extremely important."

LLM. To maintain comparability, ChatGPT-4 was directed to generate survey questions that also used a five-level Likert scale. ChatGPT-4 was prompted to "Generate a Likert-scale survey based on these dominant themes to better understand and quantify patient's barriers to arthroplasty care."

Comparison. The comparative assessment of methods considered both the time required to complete the tasks, which were approximated by the investigators, and the evaluation of theme content and corresponding survey questions. Comparison of thematic analysis was done by visually examining the identified themes from both sources and noting any similarities or overlaps until group consensus by all authors was reached.

RESULTS

The human-led thematic analysis produced six themes and 30 survey questions. ChatGPT-4 generated six themes and 16 survey questions. The human analysis was completed in 3,100 minutes, and the ChatGPT-4 analysis was completed in less than 45 minutes.

Thematic analysis. Four of the six (67%) dominant themes generated by ChatGPT-4 (financial barriers, lack of trust in the health care system, fear and uncertainty about the procedure, challenges during recovery) exhibited high similarity to the human-generated themes (cost, physician, procedure-specific concerns, recovery). Two out of the six themes (33%) generated by ChatGPT-4 (personal responsibilities and social isolation, lack of information) showed more substantial differences with the human generated themes (trust/pride, timing) (Table 1).

Survey questions. Overall, the survey questions generated by ChatGPT-4 addressed similar topics as those generated by human investigators but were shorter and less specific. Ten of the 16 (63%) survey questions generated by ChatGPT-4 exhibited high similarity to the human generated themes. Six of the 16 themes (37%) generated by ChatGPT showed more substantial differences with the human-generated questions. (Table 2).

For theme 1, human- and ChatGPT-4–generated survey questions differed slightly in phrasing but addressed identical content. For theme 2, the surveys gauged participants' trust in surgeons' expertise. Theme 3 had more substantial differences, with humans using "trust/pride" and ChatGPT generating "personal responsibility and social isolation." However, there was overlap in survey questions assessing comfort levels with seeking help during recovery. Theme 4 had the most differences, with humans focusing on "timing" and the LLM emphasizing a "lack of information." Although both surveys evaluated the extent of missing information about arthroplasty surgery, the human version also addressed comorbidities, age, pain, and prioritization of health concerns, which were areas not covered by the LLM. For theme 5, both surveys assessed fears about potential arthroplasty surgery complications and the procedure's benefits for pain management. For theme 6, both

Table 1. Comparison of human- and ChatGPT-4–generated themes regarding barriers to arthroplasty*

Theme	Human-generated themes and survey questions	ChatGPT-4–generated themes and survey questions
Theme 1	✓ Cost	✓ Financial barriers
Theme 2	✓ Physician	✓ Lack of trust in healthcare system
Theme 3	○ Trust/pride	○ Personal responsibility and social isolation
Theme 4	○ Timing	○ Lack of information
Theme 5	✓ Procedure-specific concerns	✓ Fear and uncertainty about procedure
Theme 6	✓ Recovery	✓ Challenges during recovery

*Checkmarks indicate matched themes between groups.

Table 2. Comparison of human- and ChatGPT-4–generated survey questions regarding barriers to arthroplasty organized by theme

Theme	Human-generated themes and survey questions	ChatGPT-4–generated themes and survey questions
Theme 1	✓ Cost of a joint replacement ✓ Cost of the co-pay for a joint replacement ✓ Cost of a co-pay for physical therapy after joint replacement ○ Insurance status	✓ How concerned are you about the cost of arthroplasty surgery? ✓ How concerned are you about copays for specialist visits related to arthroplasty surgery?
Theme 2	✓ Finding a surgeon I trust ○ Figuring out how to find a qualified and experienced surgeon ○ Finding a surgeon who understands what I need	✓ How much trust do you have in your surgeon's expertise and success rate?
Theme 3	○ Not having trust in any doctors or hospitals ○ Having someone I know have a bad result from a joint replacement ○ Already had a bad experience with a joint replacement ○ Already having bad experiences with other surgery or medical procedures ○ Having people to see me using a cane or walker	○ How much assistance and support do you think you will need during recovery from arthroplasty? ○ How much caregiving responsibility do you have for family members? ○ How difficult do you find it to accept help during recovery from arthroplasty surgery? ○ How satisfied are you with the social support you currently receive?
Theme 4	✓ Not having enough information to decide about having a joint replacement ○ Having a joint replacement is the last resort, and I think I should wait longer ○ Having many medical problems and having a joint replacement is not a priority now ○ Being too young to undergo joint replacement surgery ○ Being too old to undergo joint replacement surgery ○ Not doing everything I can do (like lose weight) to avoid having a joint replacement ○ Not having bad enough joint pain to have a joint replacement	✓ How much information do you have about the specific procedures involved in arthroplasty surgery?
Theme 5	✓ Fear that I will need another joint replacement after the first one because I am young ✓ Fear that a joint replacement will not help me walk and function better ✓ Fear that the joint replacement will not improve my pain ○ Fear of needles	✓ How uncertain are you about the potential complications of arthroplasty surgery? ✓ How fearful are you about undergoing arthroplasty surgery? ✓ How skeptical are you about the long-term benefits of arthroplasty surgery for managing pain? ○ How satisfied are you with current treatments for managing arthritis?
Theme 6	✓ Availability of someone to help me recover from a joint replacement ✓ Availability to take care of my family/friends while I undergo joint replacement surgery ✓ Accessing transportation to get to physical therapy appointments ✓ Finding good physical therapy centers in my community ✓ Concerns about how hard the recovery after a joint replacement will be ○ Concern of being healthy enough to undergo joint replacement surgery ○ Taking care of myself after a joint replacement because my building doesn't have an elevator	✓ How concerned are you about the healing process after arthroplasty surgery? ✓ How accessible do you think comprehensive physical therapy services are for arthroplasty patients? ✓ How concerned are you about the need for multiple surgeries every 20 years? ○ How concerned are you about the duration of recovery after arthroplasty surgery?

*Checkmarks indicate matched questions between groups.

surveys addressed concerns regarding the healing process, the need for comprehensive physical therapy, and the possibility of revision surgeries.

DISCUSSION

The utility of LLMs as an instrument for aiding in investigator triangulation was evident in both the qualitative aspect of mixed-methods research by assisting with thematic analysis, and in the quantitative aspect by facilitating survey development. In this case study, ChatGPT-4 analyzed qualitative data and generated dominant themes and corresponding survey questions to explore patients' perspectives on barriers to arthroplasty, which were comparable to those generated by humans, demonstrating that LLMs can serve as a supplement to human analysis.

Although the overall themes were similar between groups, there were notable differences in the wording and level of detail between themes and survey questions identified by humans and those generated by the LLM. The observed differences in themes and survey questions generated by human investigators versus the LLM may reflect variations in the underlying perspectives and biases of humans and the LLM. For the thematic analysis, human investigators are influenced by personal perspectives, expertise, and understanding of the research context,[5] which an LLM lacks. Additionally, the human investigators used predetermined theoretical frameworks that guide their thematic analysis whereas the LLM did not. ChatGPT-4 uses patterns and associations that it learned from a vast amount of text data humans are not trained on. LLMs may prioritize patterns or associations based on the training data.

In order to generate survey questions, human analysts may have drawn on expertise not explicitly mentioned during interviews or created questions based on salient quotes. The survey questions generated by the LLM were more succinct, resulting in shorter length, less specificity, or the potential omission of certain important details. It is important to take into account any variations in wording if the LLM is to be utilized for survey generation, as wording can influence how participants perceive and respond to the questions. The differences in question wording between human-generated and LLM-generated responses should be analyzed in a controlled setting to quantify any discrepancies in participant responses.

The limitations of our study's approach to comparing outputs include missing subtle differences between the themes and not providing a quantitative analysis of their similarities. A more robust assessment could include additional methods such as frequency analysis or kappa's-alpha. The segmentation of interviews into 1,000-word segments may have affected ChatGPT-4's ability to capture the full context and nuances of the data. The segmentation process may impact the way ChatGPT-4 identified and/or summarized themes.

To assess the utility and limitations of LLM in mixed-methods research, several further steps are necessary. Further studies should evaluate LLM's ability to perform complex tasks, such as inductive and deductive reasoning and understanding abstract concepts. Refining the commands for ChatGPT-4 and exploring the fine-tuning of LLMs for specific research queries or contexts should be assessed. Understanding the mechanisms behind LLM-generated results and how it was trained could enhance trustworthiness and acceptance for their utility in health care research. The ethical considerations of employing LLMs in mixed-methods research, including data privacy and informed consent, require continued assessment.[6,7]

Although LLMs like ChatGPT-4 show potential use in analyzing qualitative data in this qualitative–quantitative mixed-methods case study, other qualitative research approaches may present complexities that may make LLMs less amenable to analysis. For example, qualitative methodologies, such as grounded theory, phenomenology, narrative research, ethnography, case studies, and discourse analysis, may require an understanding of nuanced aspects of human experiences, emotional intuition, cultural references, interpretation of nonverbal cues, and implicit knowledge. These complexities may surpass the current capabilities of LLMs.

Large language models like ChatGPT-4 can be useful in mixed-method data analysis, particularly when there are disagreements among investigators or when dealing with large datasets. This study demonstrated the potential of ChatGPT-4 in identifying themes in qualitative data efficiently, which suggests that it can be a helpful complement to human analysis. Although LLMs have the potential to excel at tasks such as identifying recurring patterns in data, their capacity for inductive and deductive reasoning is quite limited when it comes to understanding the significance of the data. This is because they are currently incapable of engaging in the higher-level cognition, abstract thinking, and interpretive judgment required to generate results that are not explicitly present in the data. As a result, human investigators' active involvement in qualitative analysis remains a crucial factor in mixed-methods research.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr. Mehta had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Mannstadt, Goodman, Navarro-Millán, Mehta.

Acquisition of data. Mannstadt, Goodman, Rajan, Young, Wang, Navarro-Millán, Mehta.

Analysis and interpretation of data. Mannstadt, Goodman, Rajan, Young, Wang, Navarro-Millán, Mehta.

REFERENCES

1. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE guide no. 131. Med Teach 2020;42(8):846–854.
2. Bhandari P. Triangulation in research | guide, types, examples. Scribbr. January 3, 2022. Accessed April 4, 2023. https://www.scribbr.com/methodology/triangulation
3. Introducing ChatGPT. OpenAI. Accessed March 30, 2023. https://openai.com/blog/chatgpt
4. NVivo - Lumivero. Lumivero. Accessed April 14, 2023. https://lumivero.com/products/nvivo/
5. Flanagin A, Bibbins-Domingo K, Berkwits M, et al. Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. JAMA 2023;329(8):637–639.
6. Chenail RJ. Interviewing the investigator: strategies for addressing instrumentation and researcher bias concerns in qualitative research. Qual Rep 2011;16(1):255–262.
7. Sanderson K. GPT-4 is here: what scientists think. Nature. 2023; 615(7954):773.