Comparing GPT-4 and Human Researchers in Qualitative Analysis of Healthcare Data: Qualitative Description Study

Kevin Danis Li, Adrian M Fernandez, Rachel Schwartz, Natalie Rios, Marvin Nathaniel Carlisle, Gregory M Amend, Hiren V Patel, Benjamin N Breyer

Submitted to: Journal of Medical Internet Research
on: January 17, 2024

Table of Contents

Comparing GPT-4 and Human Researchers in Qualitative Analysis of Healthcare Data: Qualitative Description Study

Kevin Danis Li[1, 2] BS; Adrian M Fernandez[1] MD; Rachel Schwartz[1, 3] PhD; Natalie Rios[1] BS; Marvin Nathaniel Carlisle[1] BS; Gregory M Amend[4] MD; Hiren V Patel[1] PhD, MD; Benjamin N Breyer[1, 2] MD, MAS, FACS

[1]Department of Urology University of California San Francisco San Francisco US
[2]Department of Epidemiology and Biostatistics University of California San Francisco San Francisco US
[3]Department of Anesthesia and Perioperative Care University of California San Francisco San Francisco US
[4]Department of Urology Icahn School of Medicine at Mount Sinai New York US

Corresponding Author:
Kevin Danis Li BS
Department of Urology
University of California San Francisco
400 Parnassus Avenue
San Francisco
US

Abstract

Background: Large language models like GPT-4 have opened new avenues in healthcare and qualitative research. Traditional qualitative methods are time-consuming and require expertise to capture nuance. Although large language models have demonstrated enhanced contextual understanding and inferencing compared to traditional natural language processing, their performance in qualitative analysis versus that of humans remains unexplored.

Objective: We evaluated the effectiveness of GPT-4 versus human researchers in qualitative analysis of interviews from patients with adult-acquired buried penis (AABP).

Methods: Qualitative data were obtained from semi-structured interviews with 20 AABP patients. Human analysis involved a structured thematic process in three stages: initial observations, line-by-line coding, and consensus discussions to refine themes. In contrast, artificial intelligence (AI) analysis with GPT-4 underwent two phases: a naïve phase where GPT-4 outputs were independently evaluated by a blinded reviewer to identify themes/subthemes, and a comparison phase where AI-generated themes were compared with human-identified themes to assess agreement.

Results: The study population (n=20) comprised predominantly white (85%), married (60%), heterosexual (95%) men, with a mean age of 58.8 years and BMI of 41.1 kg/m2. Human thematic analysis identified "urinary issues" in 95% and GPT-4 in 75% of interviews, with the subtheme "spray/stream" noted in 60% and 35%, respectively. "Sexual issues" were prominent (95% humans vs. 80% GPT-4), though humans identified a wider range of subthemes, including "pain with sex or masturbation" (35%) and "difficulty with sex or masturbation" (20%). Both analyses similarly highlighted "mental health issues" (55% humans vs. 44% GPT-4), although humans coded "depression" more frequently (50% humans vs. 20% GPT-4). Humans frequently cited "issues using public restrooms" (60%) as impacting social life, whereas GPT-4 emphasized "struggles with romantic relationships" (45%). "Hygiene issues" were consistently recognized (70% humans vs. 65% GPT-4). Humans uniquely identified "contributing factors" as a theme in all interviews. There was moderate agreement between human and GPT-4 coding (Cohen's Kappa = 0.401). Reliability assessments of GPT-4's analyses showed consistent coding for themes like "Body image struggles" and "Chronic pain" (100%), and "Depression" (90%). Other themes like "Motivation for surgery" and "Weight challenges" were reliably coded (80%), while less frequent themes were variably identified across multiple iterations.

Conclusions: Large language models like GPT-4 can effectively identify key themes in analyzing qualitative healthcare data, showing moderate agreement with human analysis. While human analysis provided a richer diversity of subthemes, the consistency of AI suggests its utility as a complementary tool in qualitative research. With AI rapidly advancing, future studies should iterate analyses and circumvent token limitations by segmenting data, furthering the breadth and depth of large language model-driven qualitative analyses.

(JMIR Preprints 17/01/2024:56500)

DOI: https://doi.org/10.2196/preprints.56500

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Comparing GPT-4 and Human Researchers in Qualitative Analysis of Healthcare Data: Qualitative Description Study

Abstract

Background: Large language models like GPT-4 have opened new avenues in healthcare and qualitative research. Traditional qualitative methods are time-consuming and require expertise to capture nuance. Although large language models have demonstrated enhanced contextual understanding and inferencing compared to traditional natural language processing, their performance in qualitative analysis versus that of humans remains unexplored.

Objective: We evaluated the effectiveness of GPT-4 versus human researchers in qualitative analysis of interviews from patients with adult-acquired buried penis (AABP).

Methods: Qualitative data were obtained from semi-structured interviews with 20 AABP patients. Human analysis involved a structured process in three stages: initial observations, line-by-line coding, and consensus discussions to refine themes. In contrast, artificial intelligence (AI) analysis with GPT-4 underwent two phases: a naïve phase where GPT-4 outputs were independently evaluated by a blinded reviewer to identify themes/subthemes, and a comparison phase where AI-generated themes were compared with human-identified themes to assess agreement. We employed a general qualitative description approach.

Results: The study population (n=20) comprised predominantly white (85%), married (60%), heterosexual (95%) men, with a mean age of 58.8 years and BMI of 41.1 kg/m2. Human qualitative analysis identified "urinary issues" in 95% and GPT-4 in 75% of interviews, with the subtheme "spray/stream" noted in 60% and 35%, respectively. "Sexual issues" were prominent (95% humans vs. 80% GPT-4), though humans identified a wider range of subthemes, including "pain with sex or masturbation" (35%) and "difficulty with sex or masturbation" (20%). Both analyses similarly highlighted "mental health issues" (55% humans vs. 44% GPT-4), although humans coded "depression" more frequently (50% humans vs. 20% GPT-4). Humans frequently cited "issues using public restrooms" (60%) as impacting social life, whereas GPT-4 emphasized "struggles with romantic relationships" (45%). "Hygiene issues" were consistently recognized (70% humans vs. 65% GPT-4). Humans uniquely identified "contributing factors" as a theme in all interviews. There was moderate agreement between human and GPT-4 coding (Cohen's Kappa = 0.401). Reliability assessments of GPT-4’s analyses showed consistent coding for themes like "Body image struggles" and "Chronic pain" (100%), and "Depression" (90%). Other themes like "Motivation for surgery" and "Weight challenges" were reliably coded (80%), while less frequent themes were variably identified across multiple iterations.

Conclusions: Large language models like GPT-4 can effectively identify key themes in analyzing qualitative healthcare data, showing moderate agreement with human analysis. While human analysis provided a richer diversity of subthemes, the consistency of AI suggests its utility as a complementary tool in qualitative research. With AI rapidly advancing, future studies should iterate analyses and circumvent token limitations by segmenting data, furthering the breadth and depth of large language model-driven qualitative analyses.

Keywords: Artificial Intelligence; ChatGPT; large language models; qualitative analysis; content

analysis; buried penis; qualitative interviews; qualitative description; urology

Introduction

Recent advancements in artificial intelligence (AI), particularly in large language models, have significantly expanded their applications in healthcare and academic research. These developments raise critical questions about their potential and ethical use.[1–3] GPT-4, developed by OpenAI, is a large language model that uses deep learning algorithms, specifically the Generative Pre-trained Transformer (GPT), to process and generate human-like text.[4] Its training on diverse internet text sources through unsupervised learning enables it to interpret complex language data, making it a potentially invaluable tool for qualitative research.[5] This is especially important in areas where traditional qualitative data analysis is labor-intensive and requires expertise to understand subtle nuances.[6] Furthermore, it is unknown how AI-driven qualitative analysis may differ from human-driven analysis in research contexts.

Despite its potential, the application of AI and large language models to qualitative data remains underexplored.[7,8] Prior studies in the realm of qualitative data analysis have employed traditional Natural Language Processing (NLP) models, which often require benchmark-specific training and hand-engineering, leading to a more constrained contextual understanding and inferencing abilities. For example, Lennon et al. combined human coding with an NLP system trained on internal data, significantly reducing coding time,[9] while Cheligeer et al. used a model based on Bidirectional Encoder Representations from Transformers for faster keyword analysis.[10] However, such models fall short of the advanced contextual and inferencing abilities exhibited by widely-trained large language models like GPT-4, which has been shown to outperform traditional systems on standard NLP benchmarks.[11] Although the field is rapidly evolving, there remains a limited number of studies that directly compare AI-driven qualitative analysis to human-driven approaches.[12–17]

In this study, we used GPT-4 to re-examine qualitative data from a previously published study of 20 patients with adult-acquired buried penis (AABP)—a urological condition with significant psychosocial consequences—and compare its performance with that of human researchers.[18] Evaluating GPT-4 for qualitative analysis in this patient population is particularly important due to the unique and profound psychosocial distress associated with AABP, including issues related to body image, sexual function, and mental health. Understanding patients' experiences through qualitative analysis can provide increased understanding of their lived experiences. To accomplish these objectives, we created a series of generalizable prompts that allow the application of GPT-4 to qualitative analysis without requiring specialized knowledge or skills.[19] Finally, we evaluated the validity of our approach by measuring agreement between GPT-4 and human analysis and reliability by assessing if prompts consistently elicited similar outputs from the same data.

Methods

Data Source

Qualitative data were from a convenience sample of 20 patients who presented to urology clinics participating in TURNS (Trauma and Urologic Reconstructive Network of Surgeons), a multi-institutional collaborative research group focused on urologic trauma and reconstruction.[18] We conducted semi-structured interviews focusing on the impact of AABP on personal relationships, social life, mental health, and physical health. Participants were interviewed for 15-30 minutes and audio was transcribed electronically using Otter transcription software.[20] Interviews were

conducted over Zoom live video conferencing.[21] For both human and GPT-4 qualitative analyses, only de-identified text transcripts were used, ensuring that the qualitative data were interpreted solely from text, providing a comparable basis for both human and AI-driven analyses.

The study was approved by the University of California San Francisco (UCSF) Institutional Review Board (#20-32062), and consent was obtained from all participants. In addition to the original study's IRB approval, we obtained an exemption from our institution's IRB for the secondary analysis using GPT-4, as the data were de-identified.

Human Analysis

Our human-driven analysis employed a general qualitative description approach which differs from other qualitative methods in that the analytic process stays close to the data, describing informants' experiences using their own language.[22–24] The research team initially reviewed interview transcripts, taking notes to capture observations and ideas and facilitate a comprehensive understanding of the overall content. This preparatory work informed the subsequent structured coding process. To ensure consistency and reliability, the team convened at three key stages: 1. prior to coding, to share initial text impressions and establish a standardized coding protocol; 2. after initiating line-by-line coding, to discuss applied codes and refine categorization strategies; and 3. to assess coder inter-rater reliability using weighted Fleiss' kappa coefficients.[25] Codes with a kappa value below 0.75 were discussed among all authors until a coding consensus was reached. This approach enabled the identification and categorization of relevant sub-themes and themes.

Artificial Intelligence Analysis

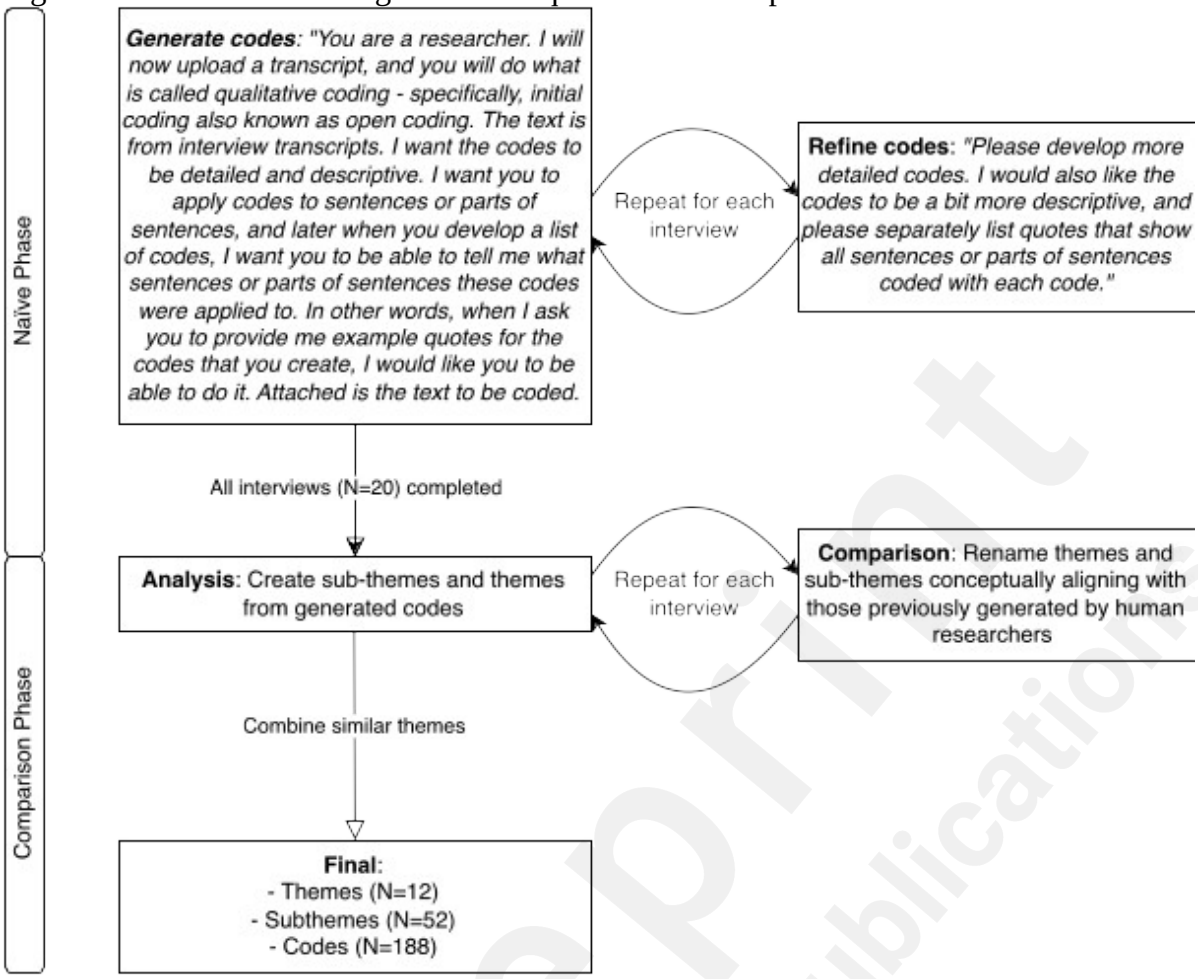
Prior to analysis, all transcripts were reviewed to ensure they contained no protected health information or identifiable data to maintain participant confidentiality. We utilized a private instance of GPT-4, known as Versa, which operates independently of OpenAI's commercial model and does not retain or learn from the data inputted.[26] This instance was used to develop our AI qualitative analysis methodology. For subsequent analyses, all data were confirmed to be thoroughly de-identified before using the commercial version of GPT-4. Each de-identified transcript underwent text formatting removal before analysis by GPT-4 using a standardized prompt set (Figure 1).[27] The analysis of the GPT-4-generated output was conducted in two phases: the naïve phase and the comparison phase.

In the naïve phase, GPT-4's outputs for each interview were examined to extract relevant codes and quotes. These were then combined into subthemes, with groupings based on conceptual coherence and content relevance, following a standard qualitative description process.[24] Subsequently, similar subthemes were grouped to form overarching themes. Multiple iterations were conducted to refine the subthemes before synthesizing generalizations that held true across the data. Memo-writing was integral to this process, capturing the evolving understanding of the data. Importantly, no discussions with the human-analyst team were conducted during this phase to avoid biasing the process. All interactions with and evaluations of GPT-4's analyses were conducted by a blinded reviewer (KDL) who was not involved in the initial human-driven analysis and kept naïve to its outcomes.

In the comparison phase, AI-identified subthemes and themes were compared against those previously identified through human-driven analysis. This phase focused on identifying parallels and alignments between the two analyses to provide a direct comparison.

Interview data were collected in 2021, and human analyses were completed by 2022. All GPT-4 analyses were processed in separate instances on December 1, 2023, using the latest model of GPT-4 available at that time.

Figure 1. Procedure for using GPT-4 for qualitative description



Measures to Ensure Rigor

The analytic team included KDL: a medical and data science master's student, NR: a clinical research coordinator with extensive experience in managing and coordinating clinical studies in healthcare settings, and GMA: a urology fellow and practicing physician/clinician specializing in urologic conditions, including adult acquired buried penis. Additionally, we consulted BNB, an expert in urologic reconstruction who frequently treats patients with buried penis, to provide in-depth clinical insights and ensure the medical accuracy of our interpretations, and RS, a health services researcher and communication scientist with expertise in qualitative methods, to guide us on appropriate methodologies and ensure the rigor of our analyses.

To ensure rigor, we implemented several strategies addressing credibility, transferability, dependability, and confirmability.[28] For credibility, we built patient rapport through prolonged engagement, as most patients had existing longitudinal relationships at the urology clinics where they received care, allowing for deeper insights into their experiences. For transferability, we reported clinical characteristics of the study participants to inform the applicability of our findings to other populations with AABP and employed a multi-institutional sampling strategy to account for potential geographic or local institutional characteristics, ensuring broader applicability of our results.

Dependability was ensured through methodological documentation, where all codes, subthemes, and themes were documented at each step to provide transparency and replicability of our coding decisions. We also maintained detailed audit trails of raw outputs from GPT-4, processed outputs,

and the subsequent organization into subthemes and themes, which the team reviewed to ensure consistency and reliability. Confirmability was achieved by having BNB, an expert in urologic reconstruction, review the study findings and provide critical insights during the design phase, and RS, who provided qualitative methodological support. Additionally, data were shared with the entire research team, and feedback from all co-authors was incorporated into subsequent interpretation and analysis.

Comparison of Analyses

Qualitative analyses, including themes and subthemes, were summarized using descriptive statistics, including frequencies and proportions. To visually represent agreement between human and AI-identified themes (validity), an agreement matrix was constructed. We measured interrater reliability using Cohen's Kappa coefficient. A separate analysis was performed 10 times on the same interview transcript to assess the reliability of GPT-4's analysis. Themes identified exclusively by GPT-4 were highlighted with exemplar quotes that best represented each theme. All analyses were performed using R statistical software (Version 4.3.1).

Results

Study Population

Participant characteristics are summarized in Table 1. Participants' mean age and body mass index were 58.8 years (standard deviation [SD] = 13.9) and 41.1 kg/m^2 (SD = 9.4). Most participants were white (17/20, 85%), married (60%), heterosexual (95%) men residing in the Western region of the United States (50%). 55% of participants underwent surgical correction of their AABP, with interviews conducted an average of 497 days (SD = 666) post-operation.

Table 1. Participant demographics and characteristics.

Characteristic	n (%)
Mean ±SD yrs age	58.8 ±13.9
Mean ±SD kg/m2 body mass index	41.1 ±9.4
No. self-identified race (%):	
White/Caucasian	17 (85)
Black/African American	1 (5)
Other	2 (10)
No. Hispanic/Latinx ethnicity (%)	3 (15)
No. relationship status (%):	
Married	12 (60)
Single	6 (30)
In a relationship	2 (10)
No. sexual orientation (%):	
Heterosexual	19 (95)
Homosexual	1 (5)
No. region (%):	
West	10 (50)
Northeast	7 (35)
Midwest	2 (10)
South	1 (5)
No. pts who underwent AABP surgical correction (%):	11 (55)
Escutcheonectomy	9 (45)

Excision of penile skin with split-thickness skin graft	6 (30)
Ventral slit scrotal flap	5 (25)

Qualitative Description

Table 2 presents a comparative analysis of themes and subthemes identified by human researchers versus GPT-4. "Urinary issues" were common in interviews analyzed by human researchers (95%) and GPT-4 (75%). Issues with "spray/stream" was a notable subtheme (60% humans vs. 35% GPT-4). "Sexual issues" were prominently coded as well, present in 95% of human-analyzed interviews and 80% by GPT-4, with "inability to perform intercourse" coded as a subtheme more frequently by human researchers (60% vs. 30%). Humans coded a broader array of sexual function issues, such as "pain with sex or masturbation" (35%) and "difficulty with sex or masturbation" (20%). "Mental health issues" were similarly recognized by both humans and GPT-4 (55% vs. 44%, respectively), with "depression" more frequently coded by humans compared to GPT-4 (50% vs. 20%, respectively). “Impact on social life" was an additional significant theme, with humans coding "issues using public restrooms" (60%), while GPT-4 emphasized "struggles with romantic relationships" (45%). Both methods identified "hygiene issues" (70% humans vs. 65% GPT-4), highlighting difficulties in maintaining cleanliness. Human researchers uniquely identified "contributing factors" as a theme in all interviews.

Table 2. Human researchers versus GPT-4 qualitative analysis.

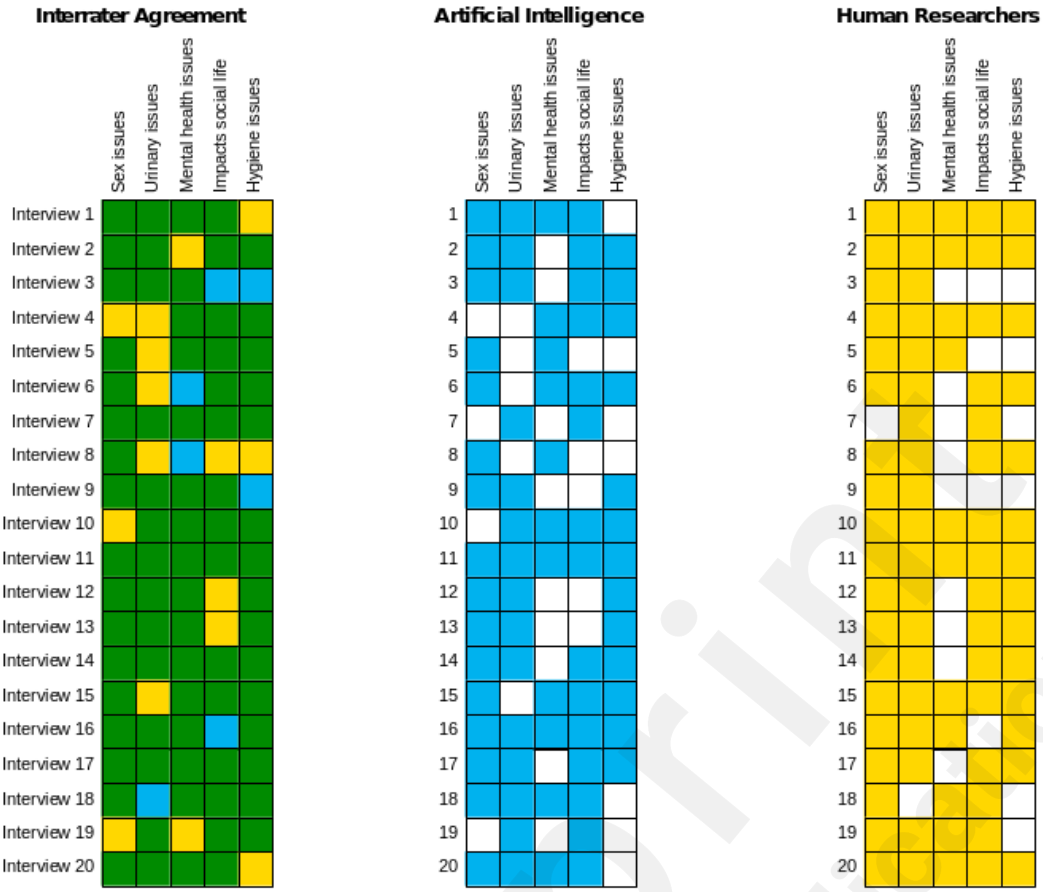
Human Researchers			GPT-4		
Themes and subthemes	No.	%	Themes and subthemes	No.	%
Urinary issues	**19**	**95**	**Urinary issues**	**15**	**75**
Spray/stream	12	60	Spray/stream	7	35
Hovers over toilet	8	40	Sits to urinate	4	20
Pain with urination	7	35	Incontinence	3	15
History of urethral stricture disease	3	15	Pain with urination	3	15
Incontinence	3	15	Incomplete bladder emptying	2	10
Incomplete bladder emptying	3	15	Frequent urination	2	10
Sits to urinate	2	10	Getting up at night to urinate	1	5
Smelly urine	1	5	Smelly urine	1	5
Trouble with catheter	1	5			
Uses shower/tub to urinate	1	5			
Sex issues	**19**	**95**	**Sex issues**	**16**	**80**
Unable to perform intercourse	12	60	Unable to perform intercourse	6	30
Unable to get erection	9	45	Avoids sex	4	20
Pain with sex or masturbation	7	35	Unable to get erection	3	15
Difficulty with sex or masturbation	4	20	Adaptive masturbation techniques	2	10
Painful erection	4	20	Poor cosmetic appearance	2	10
Unable to maintain erection	3	15	Painful erection	2	10
Avoids sex	2	10	Brittle skin	1	5
Unable to orgasm	2	10	Unable to use condom	1	5
Reduced genital sensation	1	5	Overuse of pornography	1	5
Takes longer to orgasm	1	5			
Pain with ejaculation	1	5			
Intercourse not enjoyable	1	5			
Mental health issues	**11**	**55**	**Mental health issues**	**11**	**44**
Depression	10	50	Depression	6	20
Feels like less of a man	7	35	Feels like less of a man	4	10
Anxiety	4	20	Emotional turmoil	2	10
Decreased self-esteem	3	15	Anxiety	2	10
Stress	1	5	Stress	1	5

			Loss of confidence	1	5
			Guilt	1	5
Impacts social life	**16**	**80**	**Impacts social life**	**15**	**75**
Issues using public restrooms	12	60	Struggles with romantic relationships	9	45
Avoids travel	6	30	Issues using public restroom	8	40
			Mobility impairment	6	30
			Spousal support	3	15
			Avoids hobbies	1	5
			Avoids social activities	1	5
			Negative impact on career	1	5
Hygiene issues	**14**	**70**	**Hygiene issues**	**13**	**65**
Hard/effort to clean	11	55	Hard/effort to clean	11	55
Skin tearing	7	35	Infections	6	30
Penile bleeding	6	30	Penile bleeding	2	10
Contributing factors	**20**	**100**			
Worse after wt gain	14	70			
Worse after multiple surgeries	8	40			
Worse after wt loss	4	20			
Improvement after wt loss	0	0			

Validity and Reliability of GPT-4 Analysis

To further assess validity of GPT-4 analysis, we generated an agreement matrix comparing themes coded by human researchers and GPT-4 per interview (Figure 2). There were 63 instances where both human and GPT-4 analyses agreed on the presence of a theme, and 14 instances of agreement on a theme being absent. There was disagreement in 23 cases—16 where humans identified a theme that GPT-4 did not, and 7 where GPT-4 identified a theme that humans did not (Table 3). The overall Cohen's Kappa coefficient was 0.401, indicating moderate agreement.

Figure 2. Themes identified per interview by GPT-4 versus human researchers.[a]



[a]Boxes depict interview theme analysis: Blue (AI) and yellow (humans) squares indicate presence; green reflects agreement on presence or absence.

Table 3. Codes and exemplar quotes identified exclusively by GPT-4

Interview No.	GPT-4 Code Applied: Exemplar Quote	Theme
3	Impact on Marital Relationship: "I am married? And you know it's it is... strained or? I wasn't meeting her needs."	Impacts social life
3	Hygiene Management Efforts: "I try to keep myself pretty clean... I really tried to wash my genitals really well."	Hygiene issues
6	Mental Health Impact and Resilience: "Yes in some ways it did affect me but other ways I don't really don't think it did."	Mental health issues
8	Mental Health and Self-Image Concerns: "the preconceived notion you know but the man's function is supposed to be."	Mental health issues
9	Improved Hygiene Post-Surgery: "I actually feel that hygiene became a lot easier simply because I didn't have to dig my finger in and run around the shaft to try and wash everything out."	Hygiene issues

16	Day-to-Day Discontent and Social Withdrawal: "It's just I just I would hate for other candidates that going forward thinking there is nothing that can be done need to be here they need to have options on the table."	Impacts social life
18	Urinary Dysfunction and Social Anxiety: "I would say they're abnormal for somebody my age a lot of times it's needing the needing to push… And that can cause anxiety in a public sort of restroom atmosphere."	Urinary issues

We assessed reliability by analyzing the same interview transcript 10 times with the same prompt set (Table 4). There was consistent identification of "Body image struggles/disfigurement" and "Chronic pain and discomfort," both appearing in all iterations (100%). "Depression" was also frequently coded, appearing in 90% of analyses. High reliability was observed for "Motivated to have surgery," "Uses shower/tub to urinate," and "Weight challenges," each occurring in 80% of the analyses. Other codes such as "Issues using public restrooms," "Unable to perform intercourse," and "Negative healthcare experiences" were present in 70% of iterations. Codes for "Hard/effort to clean," "Decreased self-esteem," and "Necrotizing fasciitis diagnosis" were identified 60% of the time. Less frequent were codes for "Urinary tract infections" (30%), "Sits to urinate" (20%), and a cluster of codes that included "Dependency on others for care," "Social isolation and loneliness," "High frequency of urination," "Anxiety," "Loss of physical autonomy," "Financial burden," and "Hematuria," each appearing once (10%).

Table 4. Reliability of GPT-4 generated codes.[a]

Code	No.	%
Body image struggles/disfigurement	10	100
Chronic pain and discomfort	10	100
Depression	9	90
Motivated to have surgery	8	80
Uses shower/tub to urinate	8	80
Weight challenges	8	80
Issues using public restrooms	7	70
Unable to perform intercourse	7	70
Negative healthcare experiences	7	70
Hard/effort to clean	6	60
Decreased self-esteem	6	60
Necrotizing fasciitis diagnosis	6	60
Urinary tract infections	3	30
Sits to urinate	2	20
Dependency on others for care	1	10
Social isolation and loneliness	1	10
High frequency of urination	1	10
Anxiety	1	10
Loss of physical autonomy	1	10
Financial burden	1	10
Hematuria	1	10

[a]Presence of codes from the same interview analyzed 10 times by GPT-4. Each code was counted only once per analysis, indicating whether it was identified (present) or not (absent) during each

separate analysis.

Discussion

Principal Results

In this investigation, we directly compared the performance of artificial intelligence (GPT-4) to human researchers in conducting qualitative analysis of interviews with patients affected by AABP. Our study is the first of its kind, to our knowledge, to perform such a direct comparison, highlighting the potential utility of AI in qualitative research. By employing generalized prompts, our method allows researchers without specialized NLP knowledge to utilize GPT-4 for rigorous qualitative analysis, significantly reducing the time investment required.

Our results showed moderate alignment between GPT-4 and human analyses in identifying key themes, including urinary challenges, sexual health issues, and mental health impacts. Human analysis identified more subthemes, capturing the data's complexities more thoroughly than GPT-4. This difference may stem from GPT-4’s token size limitations, which restrict its ability to perform comprehensive analyses as the input length increases.[29] The reliability tests revealed that while GPT-4 consistently recognized key codes, its identification of subtler codes was more variable. This suggests that implementing repeated analysis cycles, similar to the human multi-rater approach, could refine AI's analytical reliability. Overall, our findings underscore a complementary role for AI and human collaboration in qualitative research, where each can augment the strengths of the other.

The question of how to evaluate the accuracy and reliability of AI-driven analysis is crucial for future research. We adopted a quantitative approach to directly compare the presence of themes and subthemes in both human and AI analyses. By calculating Cohen's Kappa, a statistic that measures inter-rater reliability by considering the agreement occurring by chance, we provided an objective assessment of the consistency of themes identified by GPT-4 compared to human analysis, presupposing human analysis as the "gold standard." Additionally, to ensure consistency in GPT-4's outputs, we conducted multiple iterations of the same interview transcript analysis, analogous to traditional qualitative research methods where multiple analysts and iterative coding processes are employed to standardize analyses and minimize biases. It is important to note that while these quantitative metrics offer a clear criterion for comparison, they may not fully capture the depth and richness of qualitative insights. GPT-4 has demonstrated the ability to detect subtle nuances and emotional contexts from text data, suggesting that incorporating more qualitative approaches in AI analysis evaluation could enhance the understanding of its analytical capabilities.[30,31]

Limitations

A primary limitation of this study arises from the comparison phase, where themes and subthemes generated by GPT-4 were aligned with those identified by human researchers. Although a blinded reviewer was employed to mitigate potential bias, the subjective nature of qualitative analysis means that a degree of bias is likely to remain. This is a common challenge in qualitative research, where analysts' subjective interpretations inherently influence their analysis. However, it can be argued that the use of a large language model such as GPT-4 may present a more objective method of analysis compared to the potential variability inherent between different human researchers' analyses, due to the large language model's consistent application of its transformer model.

We deliberately chose qualitative description as our analytic approach, favoring accuracy to source

material over depth of analysis. Qualitative description involves the systematic categorization and interpretation of qualitative data to uncover patterns and insights while staying close to the original data.[22–24] A more context-based approach, such as thematic analysis, could generate richer themes and subthemes but poses challenges for comparability. More interpretative methods may introduce subjectivity, reducing reproducibility. While our methodological choice ensures that our study remains accessible as a framework for others to build on and develop more interpretative techniques, the need for comparison limited our depth of insights.

Qualitative methods have inherent limitations, such as potential bias and limited generalizability due to smaller, non-random samples, and aim to produce in-depth insights and understanding rather than population inferences.[32,33] Consequently, our findings may not capture the full diversity of patient experiences, potentially limiting the generalizability of our results. Nevertheless, our study primarily aims to provide a comparative analysis, focusing on GPT-4 as a suitable tool for qualitative research applications.

As GPT-4 and other large language models advance, their analytical capabilities are expected to become more sophisticated, which may alter their proficiency in qualitative analysis. For example, while GPT-3.5 scored in the bottom 10% on a simulated bar exam, GPT-4 has demonstrated a significant improvement, placing within the top 10% of test-takers.[11] The study's findings are therefore a snapshot of GPT-4’s capabilities at a specific point in time and may not fully represent its future potential in qualitative analysis. Despite this limitation, the current trajectory of AI indicates that the use of GPT-4 and similar large language models in qualitative research is likely to become increasingly robust and refined.

Comparison with Prior Work

While studies applying GPT-4 or other large language models to qualitative research are limited, a growing body of work has compared the performance of OpenAI's Generative Pre-trained Transformer models, including GPT-3, 3.5, and 4, to that of humans in academic research and medical education.[12–15] Wang et al. found that while ChatGPT can generate accurate and relevant information, it is not without gaps when compared to official sources, indicating a need for supplementary validation from reliable references.[34] Other studies have shown that ChatGPT can mimic the style of human-written research abstracts, albeit with limitations in quality and accuracy as indicated by the ability of blinded reviewers to distinguish AI-generated content.[35] In the field of medical education, ChatGPT has been shown to outperform medical students on examinations, suggesting valuable applications in examination preparation.[36] Similarly, ChatGPT's performance on USMLE further showcases the potential utility of AI in medical education, where it achieved scores near the passing threshold without specialized training.[37] These findings emphasize that while advanced large language models like GPT-4 are becoming increasingly competent in complex tasks, their current role remains complementary to human expertise.

The application of GPT-4 and other large language models to healthcare is a burgeoning field with substantial promise, resting on the fundamental ability of AI to process qualitative data efficiently. In patient care, large language models can enhance communication by translating complex medical language into more accessible terms for healthcare providers and patients.[38] The performance of large language models on medical licensing exams also indicates their potential utility in supporting clinical decision-making.[39] In administrative contexts, large language models are particularly valuable for generating concise clinical summaries and synthesizing extensive electronic medical record documentation, tasks that typically consume considerable time for healthcare professionals. The integration of large language models into administrative workflows may increase efficiency and

allow clinicians to allocate more time to direct patient care. Healthcare companies are already beginning to integrate large language models into electronic health records (EHRs), such as Epic's recent partnership with Microsoft to embed Azure OpenAI Service into its own EHR systems.[40]

Despite its promise, integration of large language models in healthcare raises several ethical concerns that warrant careful consideration.[41] Foremost among these is data privacy, particularly regarding the handling of sensitive patient information, necessitating robust safeguards against data breaches. The opacity of these models, due to the unavailability of public training datasets and model weights, poses another concern, as it obscures the understanding of their decision-making processes and challenges their trustworthiness in clinical applications.[42] Additionally, the commercialization of large language models by major corporations such as OpenAI, Microsoft, Meta, and Google brings into question the potential influence of commercial interests on model development and deployment, possibly overshadowing patient welfare. A crucial concern is the risk of patient harm arising from incorrect or biased models, emphasizing the need for rigorous testing and validation of large language models to ensure their reliability and prevent adverse clinical outcomes.[43]

Conclusions

Our research demonstrates that large language models like GPT-4 can discern key themes from qualitative healthcare data when utilized with standardized prompts. This 'out-of-the-box' approach aligns moderately well with qualitative description analysis by human analysts. Future work should employ more context-based prompts for deeper, richer themes. As this may introduce greater subjectivity, researchers should also explore iterative analyses, such as synthesizing output from multiple iterations, to improve large language model output reliability. Additionally, researchers should assess the qualitative analytic abilities of other popular models like Gemini, LLaMa, and Claude, and develop methods to circumvent the token limitations inherent in models like GPT-4 by segmenting qualitative data inputs, enriching the depth and breadth of qualitative analyses.

Conflicts of Interest

None declared.

Abbreviations

AABP: adult-acquired buried penis
AI: artificial intelligence
HER: electronic health record
GPT: generative pre-trained transformer
NLP: natural language processing
SD: standard deviation
TURNS: Trauma and Urologic Reconstructive Network of Surgeons
UCSF: University of California San Francisco

References

1. Liu Y, Han T, Ma S, et al. Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiol*. 2023;1(2):100017. doi:10.1016/j.metrad.2023.100017
2. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in

medicine. *Commun Med*. 2023;3(1):1-8. doi:10.1038/s43856-023-00370-1
3. Meyer JG, Urbanowicz RJ, Martin PCN, et al. ChatGPT and large language models in academia: opportunities and challenges. *BioData Min*. 2023;16(1):20. doi:10.1186/s13040-023-00339-9
4. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys Syst*. 2023;3:121-154. doi:10.1016/j.iotcps.2023.04.003
5. Schopow N, Osterhoff G, Baur D. Applications of the Natural Language Processing Tool ChatGPT in Clinical Practice: Comparative Study and Augmented Systematic Review. *JMIR Med Inform*. 2023;11:e48933. doi:10.2196/48933
6. Queirós A, Faria D, Almeida F. STRENGTHS AND LIMITATIONS OF QUALITATIVE AND QUANTITATIVE RESEARCH METHODS. *Eur J Educ Stud*. 2017;3(9). doi:10.5281/zenodo.887089
7. Kantor J. Best practices for implementing ChatGPT, large language models, and artificial intelligence in qualitative and survey-based research. *JAAD Int*. 2023;14:22-23. doi:10.1016/j.jdin.2023.10.001
8. Hitch D. Artificial Intelligence Augmented Qualitative Analysis: The Way of the Future? *Qual Health Res*. Published online December 8, 2023:10497323231217392. doi:10.1177/10497323231217392
9. Lennon RP, Fraleigh R, Scoy LJV, et al. Developing and testing an automated qualitative assistant (AQUA) to support qualitative analysis. *Fam Med Community Health*. 2021;9(Suppl 1):e001287. doi:10.1136/fmch-2021-001287
10. Cheligeer C, Yang L, Nandi T, et al. Natural language processing (NLP) aided qualitative method in health research. *J Integr Des Process Sci*. 2023;27(1):41-58. doi:10.3233/JID-220013
11. OpenAI, Achiam J, Adler S, et al. GPT-4 Technical Report. Published online December 18, 2023. Accessed December 27, 2023. http://arxiv.org/abs/2303.08774
12. Zhang H, Wu C, Xie J, Lyu Y, Cai J, Carroll JM. Redefining Qualitative Analysis in the AI Era: Utilizing ChatGPT for Efficient Thematic Analysis. Published online May 27, 2024. doi:10.48550/arXiv.2309.10771
13. Hamilton L, Elliott D, Quick A, Smith S, Choplin V. Exploring the Use of AI in Qualitative Analysis: A Comparative Study of Guaranteed Income Data. *Int J Qual Methods*. 2023;22:16094069231201504. doi:10.1177/16094069231201504
14. Morgan DL. Exploring the Use of Artificial Intelligence for Qualitative Data Analysis: The Case of ChatGPT. *Int J Qual Methods*. 2023;22:16094069231211248. doi:10.1177/16094069231211248
15. Wachinger J, Bärnighausen K, Schäfer LN, Scott K, McMahon SA. Prompts, Pearls, Imperfections: Comparing ChatGPT and a Human Researcher in Qualitative Data Analysis. *Qual Health Res*. Published online May 22, 2024:10497323241244669. doi:10.1177/10497323241244669
16. A. Fuller K, Morbitzer KA, Zeeman JM, M. Persky A, C. Savage A, McLaughlin JE. Exploring the use of ChatGPT to analyze student course evaluation comments. *BMC Med Educ*. 2024;24(1):423. doi:10.1186/s12909-024-05316-2
17. Amirova A, Fteropoulli T, Ahmed N, Cowie MR, Leibo JZ. Framework-based qualitative analysis of free responses of Large Language Models: Algorithmic fidelity. *PloS One*. 2024;19(3):e0300024. doi:10.1371/journal.pone.0300024
18. Amend GM, Holler JT, Sadighian MJ, et al. The Lived Experience of Patients with Adult Acquired Buried Penis. *J Urol*. 2022;208(2):396-405. doi:10.1097/JU.0000000000002667
19. Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. *J Med Internet Res*. 2023;25:e50638. doi:10.2196/50638

20. Otter.ai - AI Meeting Note Taker & Real-time AI Transcription. Accessed January 17, 2024. https://otter.ai/
21. One platform to connect. Zoom. Accessed January 17, 2024. https://zoom.us/
22. Sandelowski M. Whatever happened to qualitative description? *Res Nurs Health*. 2000;23(4):334-340. doi:10.1002/1098-240X(200008)23:4<334::AID-NUR9>3.0.CO;2-G
23. Sandelowski M. What's in a name? Qualitative description revisited. *Res Nurs Health*. 2010;33(1):77-84. doi:10.1002/nur.20362
24. Neergaard MA, Olesen F, Andersen RS, Sondergaard J. Qualitative description – the poor cousin of health research? *BMC Med Res Methodol*. 2009;9(1):52. doi:10.1186/1471-2288-9-52
25. Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Med Res Methodol*. 2016;16(1):93. doi:10.1186/s12874-016-0200-9
26. Now Available: Versa, UCSF Generative AI Platform | Office of the Chancellor. Accessed May 30, 2024. https://chancellor.ucsf.edu/news/now-available-versa-ucsf-generative-ai-platform
27. *Thematic Analysis with ChatGPT | PART 1- Coding Qualitative Data with ChatGPT*.; 2023. Accessed December 26, 2023. https://www.youtube.com/watch?v=8dTs7D42ge0
28. Ahmed SK. The pillars of trustworthiness in qualitative research. *J Med Surg Public Health*. 2024;2:100051. doi:10.1016/j.glmedi.2024.100051
29. Kohn R. Mastering Token Limits and Memory in ChatGPT and other Large Language Models. Medium. Published March 21, 2023. Accessed December 27, 2023. https://medium.com/@russkohn/mastering-ai-token-limits-and-memory-ce920630349a
30. Baktash JA, Dawodi M. Gpt-4: A Review on Advancements and Opportunities in Natural Language Processing. Published online May 4, 2023. doi:10.48550/arXiv.2305.03195
31. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. *Front Psychol*. 2023;14. doi:10.3389/fpsyg.2023.1199058
32. Borgstede M, Scholz M. Quantitative and Qualitative Approaches to Generalization and Replication-A Representationalist View. *Front Psychol*. 2021;12:605191. doi:10.3389/fpsyg.2021.605191
33. Tenny S, Brannan JM, Brannan GD. Qualitative Study. In: *StatPearls*. StatPearls Publishing; 2024. Accessed July 6, 2024. http://www.ncbi.nlm.nih.gov/books/NBK470395/
34. Wang G, Gao K, Liu Q, et al. Potential and Limitations of ChatGPT 3.5 and 4.0 as a Source of COVID-19 Information: Comprehensive Comparative Analysis of Generative and Authoritative Information. *J Med Internet Res*. 2023;25:e49771. doi:10.2196/49771
35. Cheng SL, Tsai SJ, Bai YM, et al. Comparisons of Quality, Correctness, and Similarity Between ChatGPT-Generated and Human-Written Abstracts for Basic Research: Cross-Sectional Study. *J Med Internet Res*. 2023;25:e51229. doi:10.2196/51229
36. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial Intelligence in Medical Education: Comparative Analysis of ChatGPT, Bing, and Medical Students in Germany. *JMIR Med Educ*. 2023;9:e46482. doi:10.2196/46482
37. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
38. Decker H, Trang K, Ramirez J, et al. Large Language Model-Based Chatbot vs Surgeon-Generated Informed Consent Documentation for Common Procedures. *JAMA Netw Open*. 2023;6(10):e2336997. doi:10.1001/jamanetworkopen.2023.36997
39. Benary M, Wang XD, Schmidt M, et al. Leveraging Large Language Models for Decision Support in Personalized Oncology. *JAMA Netw Open*. 2023;6(11):e2343689. doi:10.1001/jamanetworkopen.2023.43689

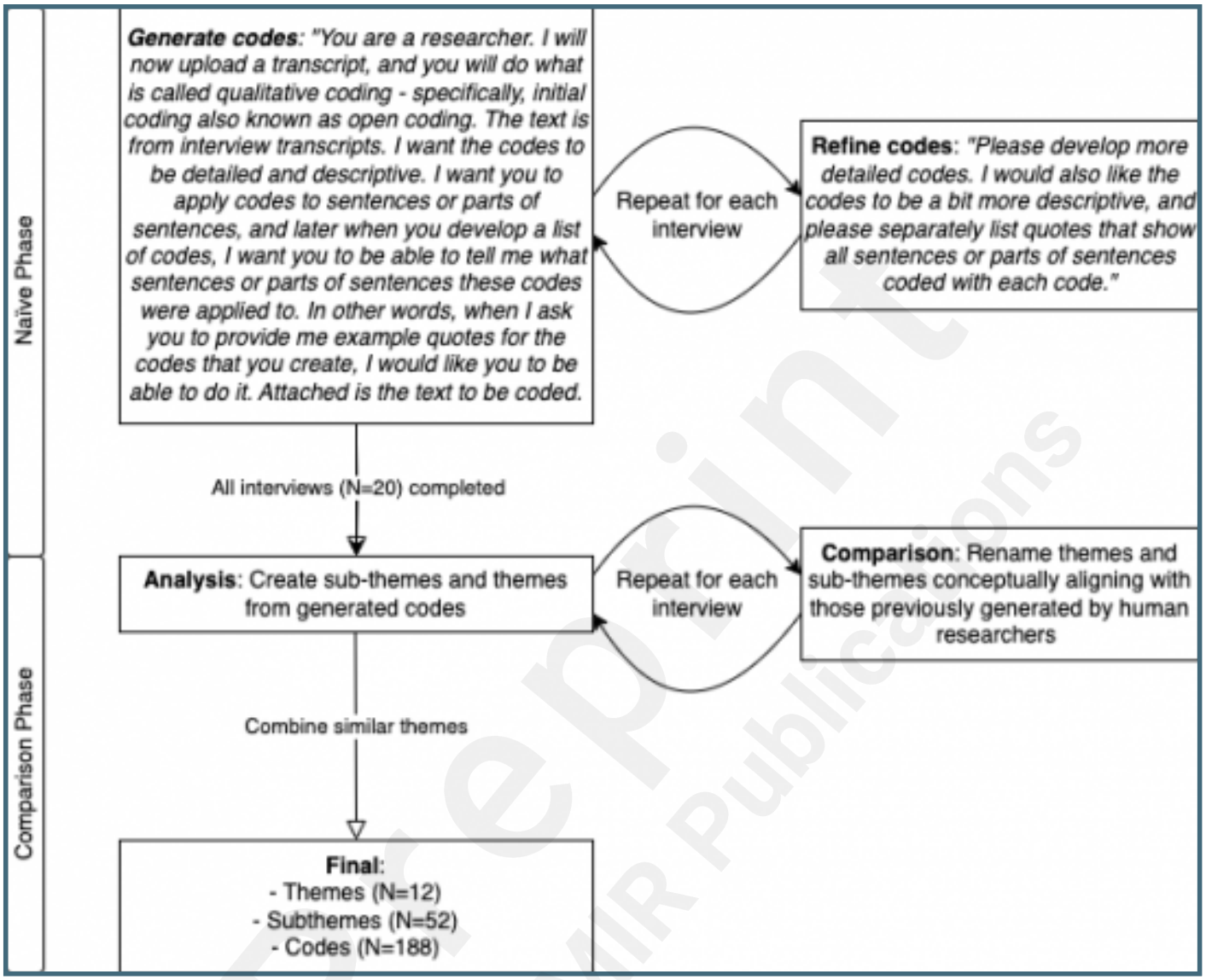
40. Center MN. Microsoft and Epic expand strategic collaboration with integration of Azure OpenAI Service. Stories. Published April 17, 2023. Accessed December 27, 2023. https://news.microsoft.com/2023/04/17/microsoft-and-epic-expand-strategic-collaboration-with-integration-of-azure-openai-service/
41. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *Npj Digit Med*. 2023;6(1):1-6. doi:10.1038/s41746-023-00873-0
42. Sanderson K. GPT-4 is here: what scientists think. *Nature*. 2023;615(7954):773-773. doi:10.1038/d41586-023-00816-5
43. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health*. 2024;6(1):e12-e22. doi:10.1016/S2589-7500(23)00225-X

Supplementary Files

Literature review of previously published literature using large language models for qualitative analysis.
URL: http://asset.jmir.pub/assets/4c9a590311dfda87aab0338354941478.xlsx

Figures

Procedure for using GPT-4 for qualitative description.



Themes identified per interview by GPT-4 versus human researchers.