

Accuracy and Efficiency of Artificial Intelligence Tools for Transcription in Qualitative Research

Monique Hennink¹

Rollins School of Public Health, Emory University, Atlanta, GA, USA

Bahaa Kazzi

Emory University School of Medicine, Atlanta, GA, USA

Erin Yoon

Emory University College of Arts and Sciences, Atlanta, GA, USA

Nael Alami

Modern University for Business and Science, Beirut, Lebanon

Ghada Farhat

Emory University School of Medicine, Atlanta, GA, USA

ABSTRACT

Verbatim transcription in qualitative research is hugely time consuming, yet critical for data validity. The rise in artificial intelligence tools for transcription provides new opportunities for researchers. We assessed the accuracy and efficiency of two widely accessible AI tools and human transcription to provide empirical guidance for researchers. We compared 10 verbatim transcripts generated via three transcription modes: human, Otter.ai and MAXQDA. We compared transcription time, cost, errors (volume and type), and correction time. Results show that the AI tools were highly time/cost efficient but had 2-3 times more errors than human transcription. Accounting for average correction time, AI tools generated a transcript in 1.25 - 2.5 hours for USD31-38 per transcript; thus 3-6 times faster and 3 times cheaper than human transcription. However, human review of AI transcription remains critical for data validity; thus, AI tools can assist in the transcription process versus generating the final product. We develop recommendations for each phase of the research process to maximize the benefits of using AI transcription tools.

¹Corresponding Author: Professor, Hubert Department of Global Health, Rollins School of Public Health, Emory University, 1518 Clifton Road NE Atlanta, GA 30322. 404-7274882, mhennin@emory.edu
Copyright © 2026 by Author(s) and Licensed by CECS Publications, United States. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

KEYWORDS: Qualitative, Verbatim Transcription, Artificial Intelligence, Accuracy, Automated Speech Recognition

Verbatim transcription of recorded interviews is a central task in much qualitative research. The accuracy of verbatim transcription is paramount, since it influences data validity (Eftekhari, 2024; Johnson et al., 2020; King & Horrocks, 2010; Poland, 1995; Tracy, 2010). Typically, qualitative researchers conduct transcription manually by listening to interview recordings and transcribing these verbatim, sometimes using software to adjust the recording speed. Although this facilitates active engagement with data, it is hugely time consuming because speech is up to 10 times faster than typing (Wald, 2006). For example, 1 hour of recording can take 5–10 hours to transcribe verbatim (Eftekhari, 2024; Hennink et al., 2020; McMullin, 2023; Sponholtz et al., 2025; Tolle et al., 2024), which can create long delays between data collection and analysis. The recent rise in automated speech recognition (ASR) technology and artificial intelligence (AI) tools have opened new opportunities for verbatim transcription. AI technology for verbatim transcription is available through web-based AI programs for transcription (e.g., Otter.ai; Rev) or AI tools within qualitative data analysis software (e.g., MAXQDA; NVivo). Although AI-generated transcription offers a unique opportunity to reduce the manual burden of transcription, empirical research is needed to assess transcription quality and guide researchers on which transcription tools to use. Our study aims to contribute to the literature on the quality of AI-generated transcription in widely available programs. Many types of qualitative data analysis (e.g., grounded theory, narrative analysis) use verbatim transcription of interviews, which retains natural speech rather than a “cleaned up” version with grammar corrections. Verbatim transcription is important for these types of analyses because it reflects natural speech which conveys much contextual meaning (Denham & Onwuegbuzie, 2013; Green & Thorogood, 2018; Kvale and Brinkman, 2009; Oliver et al., 2005). However, making a verbatim transcription of a recorded interview is not an objective clerical task. Transcription is an interpretive act (Hammersley, 2010; Jaffe, 2007; Miles & Huberman, 1994; Ochs, 1979; Oliver et al., 2005; Poland, 1995; Van Maanen, 1979) that involves reconstruction from one medium (sound) to another (text). In this process the transcriber continually interprets speech – transcribing colloquial or slang words, pauses, verbal stutters and false starts, and adding grammar and speech filler sounds; therefore, they are continually interpreting what is said and (re)presenting this in text form (Lapadat, 2000; Tolle et al., 2024). As such, human-generated transcription is itself already an interpretation, and thus one version of an audio recording (Bucholtz, 2000; Oliver et al., 2005). The introduction of AI tools for verbatim transcription, provides another interpretation of an audio recording which may be close to or different from a human generated transcript. We are beginning to understand how AI tools “interpret” human speech, how they manage colloquial words and speech fillers, how they “decide” on grammar/punctuation, and the extent to which AI tools literally transcribe sounds into words versus “understand” and construct a meaningful sentence. Despite the mode of transcription used (human or AI), manual checking of verbatim transcription against the audio recording remains paramount, thus there will always be a human element in verbatim transcription. What is needed is continued empirical research to determine the extent to which various AI tools can reduce the burden of manual transcription and provide an effective verbatim account of a recorded interview.

Few published papers report how transcription was conducted. This sidelines transcription as an objective task that warrants no detailed description. For example, Nascimento and Steinbruch (2019) reviewed qualitative research articles in the top three management journals and found that

none described how data were transcribed, stating only that “interviews were transcribed” (p. 423). McMullin (2023) found that 41% of articles reviewed had no mention of transcription, and 11% mentioned transcription but not the process. This lack of transparency on transcription in qualitative research has long been an issue (McMullin, 2023; Point & Baruch, 2023; Tilley & Powick, 2002). However, with the availability of AI tools for transcription, the need for transparency is greater because how transcription is done (i.e., which AI tool is used, parameters set, quality checks, etc.) reflects the validity of data, thus transcription becomes an important data quality criterion. Concerns about transcription accuracy and transparency are warranted, particularly with the growing interest in using AI tools for transcription. For example, a recent survey found a high level of willingness (4.4 on 5-point scale) amongst researchers to use AI tools for transcription in qualitative research (Marshall & Naff, 2024).

Previous Studies

Voice recognition software (VRS) is continually evolving and its recent integration with AI technology has accelerated its appeal for transcribing recordings of qualitative interviews. We briefly review previous research which we present in four phases with different automated transcription techniques in each. The first phase of studies assessed the effectiveness of a “listen and repeat” transcription technique aided by voice recognition software (VRS). Park and Zeanah (2005) and Matheson (2007) used VRS to transcribe qualitative interviews, by listening to a segment of a recorded interview, then repeating it verbally into a digital recorder or voice editing software to generate a written transcript. Both studies reported that this technique took the same time as a typist, but more time was needed to edit transcripts, although no comparative data were provided. This technique also necessitated specific equipment, technological competence, and training. Johnson (2011) directly compared the listen a repeat technique with a human transcriber. Results showed that transcribing a single 17-minute recording took longer using VRS than human transcription (96 vs. 90 minutes respectively), had similar accuracy (96% vs. 98% accuracy respectively), but error correction took longer in a VRS versus human generated transcript (28 vs. 11 minutes), thus it cancelled out the time saved by using the software. Typical errors using VRS were homonyms and proper nouns versus minor typing errors for human transcription.

The second phase of studies assessed the accuracy of fully automated ASR technology for transcription (without “listen and repeat” step of earlier studies). Stolcke and Droppo (2017) assessed the differences between transcriptions of conversations from switchboard and home call audio transcribed by humans versus ASR technology. They found that both modes of transcription had a similar error rate (6% errors on switchboard and 13% on home call data) and types of errors; however, the ASR mode had more errors due to misrecognition of filler words (uh, um) and acknowledgement sounds (uh-huh, mhm) than human transcripts. Bokhove and Downey (2018) used automated captioning tools to generate transcripts of audio and video data, then originality checking software (Turnitin) to determine the percentage match between the automated and human transcripts. Results of auto generated data showed 91-92% similarity to the human transcripts. Errors included incorrect transcription of names, similar sounding words, interpretation of word meanings, and word contractions. A significant limitation is the use of text comparison software that is based on proprietary algorithms to determine a match, close match, or no match, making it unclear how a match / mismatch is determined, and what transcription accuracy actually means. Although authors claim the majority of errors in AI transcripts could easily be identified and rectified via editing, they don’t tabulate errors or provide any data on the volume of errors per

transcript, nor correction time. Authors also claim that using AI tools for transcription can reduce transcription time and cost but provide no empirical data on this.

The third phase involves assessing dedicated programs (Otter.ai, Sonix, Teams, YouTube) that have AI technology embedded with ASR for transcription of voice recordings. Louw (2021) assessed the accuracy of two AI transcription programs (Otter and Sonix) compared with human transcription. They used a text comparison program (Text Compare) to compared three minutes of audio from five documents to determine whether AI-generated transcription was a match, close match or no match to human transcription. Results from both AI programs were similar and showed a match or close match to the human transcript for 40-80% of the text segments compared. However, close match segments still contained errors, such as missing/added/incorrect words, homophone errors, missing dialogue and incorrect speaker designation. Transcription accuracy was most influenced by audio quality and number of speakers (80% match/near match for a monologue vs. 25% for three speakers) but not by a speaker's accent. This study shares similar limitations to Bokhove and Downey (i.e. use of a text comparison tool, lack of error tabulation, no data on transcription time /cost). Furthermore, only 15 minutes of audio were compared, providing no indication of the volume of errors in a whole transcript to determine the practical effectiveness of AI transcription in a full project, with many interviews. Eftekhari (2024) used the ASR transcription function in Microsoft Teams to simultaneously transcribe 41 interviews online. The main benefits reported were the immediacy of transcripts and cost saving. However, editing transcripts took 1.5-3.5 hours per transcript, and accuracy was limited by speaker accents/pronunciation, tempo, and grammar (e.g., incorrectly transcribed homophones, speech fillers and punctuation). This study did not tabulate errors, providing only examples of errors. Both Kuhn (2023) and Rao (2023) assess the performance of ASR for transcribing audio in YouTube videos. Kuhn compared multiple ASR services for transcribing 3-minute segments of audio in video lectures for higher education for the deaf and hard of hearing community. Results showed that transcription accuracy varied significantly between ASR services and across videos within a service, concluding that ASR services lack consistent accuracy of transcription. Similarly, Rao (2023) assessed Whisper ASR to transcribe brief segments of 25 educational videos and compared it with a baseline transcript generated by YouTube. They reported no strong conclusions on transcription accuracy, but show how ASR manages inaudible speech with predicted text that ranges in accuracy. Kuhn (2023) acknowledges a key limitation of assessing ASR transcription using educational videos is that they typically have pre-prepared content that may not be comparable to spontaneous conversational speech such as in research interviews and may have a bias towards White male voices.

The fourth and most recent phase of studies focus on using ASR to transcribe offline. Cloud-based transcription platforms typically require uploading audio files to a server, with limited information on where servers are located, how long files remain and whether data are used for training AI, all of which raise concerns for data security (Eftekhari, 2024; Samuel & Wassenaar, 2024). In response to data security concerns, recent studies have piloted technologies that use ASR transcription offline, thus harnessing time saving benefits of ASR technology while maintaining data security. Da Silva (2021) describes an offline semi-automated transcription method using Google Recorder. It was tested on five interviews and showed errors were 31% higher than human transcription (based on a paragraph of text) but it reduced the ratio of transcription to recording time from 7:1 to 3:1 for manual vs automated transcription respectively. However, it requires technical knowledge, specific hardware that increases costs, and a quiet environment for clear and confidential audio playback. Battaglia (2024) used a similar technique using Whisper technology

by OpenAI to allow fully offline transcription, with audio playback done internally within the computer device. They estimate an 83% time saving compared with human transcription and state it is “very accurate” (Battaglia 2024, p.7), but no time or error data are presented. They highlight that technical knowledge and competence with Whisper is needed which can be a limitation (Battaglia, 2024). Tolle et al. (2024) developed Vink, by adapting Whisper technology for qualitative transcription to allow transcription on a user’s computer rather than a cloud-based platform. It was tested on 14 audio files of 3-6 minutes in different languages. Results showed the word-error-rate (WER) varied widely by language and correction time varied between 1.7-16 times the recording time. It produced mostly de-naturalised speech (removing speech fillers, hesitations etc) which may be important to some types of data analysis and struggled with cross-talk and speaker recognition. Sponholtz et al. (2025) describe a method of transcribing audiofiles offline using Adobe Premier Pro CC. It was tested with 12 interviews and showed an average transcription and correction time of 2 hours 19 minutes per transcript in optimal sound quality conditions and with standard languages (without dialects or accents) and a WER of 17%. There was no comparison data. Meanwell et al (2024) went beyond the studies described above to develop an institution-based serverless cloud option for automated transcription based at Indiana University, called the Automated Transcription Service (ATS). Testing showed it performed best with high quality audio, individual speakers and English language. It is limited to transcribing 15 minutes of audio and is restricted to Indiana University users; nonetheless, it shows potential for other institutions with similar technologies and experience.

Even though there is a growing body of research broadly assessing the quality of AI-based transcription, there is significant variation in ASR technologies assessed (listen and repeat, ASR, AI-assisted programs, etc), data sources used (audio, video), assessment metrics (text comparison tools, WER) and amount of data used to identify errors (a paragraph versus whole documents). This makes comparisons across the literature challenging. Our study, situated within and mostly comparable to studies in phase three, focused on assessing widely accessible transcription programs with AI technology embedded in ASR. Currently, the pilot studies of offline transcription technologies (in phase four) are not widely accessible; therefore, researchers still need to use existing AI transcription programs, such as those we evaluate. Our study extends existing research by further assessing the accuracy and efficiency of widely accessible AI-based transcription tools, providing pragmatic evidence to researchers who are considering their use.

Study Aims

Our study aims to assess the accuracy (number/type of errors) and efficiency (time/cost of transcription and error correction) of verbatim transcription using three modes of transcription: two AI tools and human transcription. We sought to answer the following research questions:

1. How accurate are AI tools (MAXQDA, Otter.ai) and human transcribers for verbatim transcription in qualitative research?
 - a. What types of errors are present in AI-generated and human transcription?
 - b. What causes errors in AI-generated and human transcription and how can these be reduced?
2. How efficient are AI tools (MAXQDA, Otter.ai) and human transcribers for verbatim transcription in qualitative research?

3. What guidelines can be derived for using AI transcription tools in qualitative research?

Methods

Data

Data used in our analysis comprised audio recordings of 10 in-depth interviews from a study on *"Barriers to Non-Communicable Disease (NCD) care among Syrian refugees in Lebanon"* conducted in 2023 (under review). Study participants included healthcare providers from governmental, non-governmental, and intergovernmental organizations involved in healthcare delivery for refugees in Lebanon. Most interviewees had Arabic as their first language. All interviews were conducted in English, in-person or via Zoom, and digitally recorded. Interviews ranged from 45–120 minutes, generating transcripts between 8,000–13,000 words. Interviews were conducted by public health graduate students, trained in qualitative research. All interviews were completed a year prior to developing this study on AI transcription; therefore, interviewers could not have been influenced by this analysis while conducting the interviews. The parent study was approved by the Institutional Review Boards of Emory University, USA (STUDY00006226) and Modern University for Business and Science, Lebanon (MU-20230612-42)

Generating Transcripts

For this study, we generated verbatim transcripts of the 10 interview recordings via three modes: human transcription, web-based AI transcription using Otter.ai (v.2024), and software-based AI transcription using MAXQDA v.24 (VERBI Software). This generated 30 transcripts that we used for our comparative analysis. Human Transcription was done by two English-speaking analysts trained in verbatim transcription and using Express Scribe software (ACH Software). It involved listening to each interview recording via Express Scribe, which enables transcribers to pause, rewind or reduce the speed of the recording to enable continuous transcription. Transcribers followed a transcription protocol for consistency that included: using speaker identifiers (e.g., I for interviewer and P for participant) and following a transcription key of symbols to use for pauses, crosstalk, or inaudible speech. Human transcription did not transcribe speech fillers (e.g., um, er), and colloquial speech was transcribed exactly as spoken, including incorrect grammar and false starts. Where the recording used Arabic words, these were transcribed in Arabic. Web-based AI Transcription was conducted using Otter.ai, a web-based transcription platform available online (both free and fee-based) which uses advanced ASR software and machine learning technology to transcribe audio into written text. We used Otter.ai to transcribe the same 10 interview recordings. We needed to use the fee-based (Pro version) transcription option because the number of recording minutes exceeded the allowance for the free option. We purchased a 2-month subscription for Otter.ai, which allowed 1,200 minutes of recording and 10 file uploads. The transcription time was noted for each recording and included a brief period of file upload time (often only seconds). Each recording was uploaded and transcribed individually via Otter.ai to enable the transcription time to be recorded for each interview. Software-Based AI Transcription was conducted using the MAXQDA qualitative data analysis program, which has an embedded AI based transcription function. We purchased a transcription package (15 hours of transcription time) in MAXQDA to

transcribe the same 10 interview recordings. The transcription time was noted for each recording and defined as the time between pressing “Start Transcription” to when “Complete” was shown, which included a brief period of file download time (usually only seconds). Prior to using the AI transcription programs, we piloted the process with one recording on each platform (Otter.ai and MAXQDA) to identify available settings and ensure these were similar for both (e.g., excluding time stamps).

Developing a Codebook

Transcription accuracy was assessed by categorizing and counting errors within and across transcription modes. A transcription error was defined as any deviation from the digital recording. We adapted an error matrix, originally developed by Hennink and Weber (2013) for focus group transcripts. We piloted the matrix on our data and adapted it for interview data by removing items related to dialogue between group participants and multiple speaker identifiers and adding error items that are unique to AI-generated transcription (e.g., homophone errors, cleaned false starts, inaudible speech transcribed, acronym errors, incorrect spelling/word conjugation, Arabic word errors). Our final matrix of errors (Table 1) included four categories of transcription errors. The first category was inaudible speech, whereby speech was inaudible on the recording and either a) not transcribed (i.e., words are missing) or b) transcribed despite being inaudible (i.e., transliteration). The second category were omissions, whereby a word, phrase, speaker identifier, sound/pause was omitted but clearly heard on the recording. Omissions were further categorized by whether the word or phrase omitted was related to the study topic (e.g., topical words such as refugee, clinic) or generic non-topical words (e.g., the, and, for, etc.). This category also included omitted acknowledgement words (e.g., OK, I see). The third category were grammatical errors, such as incorrect word conjugation, grammar correction, false start correction and spelling errors. It also included homophone errors and acronyms that were incorrect or transcribed as words. The final category included errors and additions, such as incorrect words or phrases, additional words and speaker identifier errors (e.g., incorrect, unknown, or additional speaker). This category also included spoken Arabic words that were omitted or transcribed phonetically into an English word (transliteration). We did not include an error code for punctuation issues (e.g., commas, full stops, capitalization) because the nature of recorded speech makes the determination of punctuation unclear and therefore too subjective to code as an error.

Table 1
Transcription Error Codebook

Type of Error	Description of Error
I. Inaudible Speech	
Inaudible not transcribed	Speech is inaudible on recording and not transcribed nor marked in transcript as inaudible (i.e., speech is missing).
Inaudible transcribed	Speech is inaudible on recording, but some words are transcribed.
II. Omission of:	
Participant acknowledgement	Omission of acknowledgement words from participant. (e.g., OK, I see, yes, etc.) <u>Not</u> for sounds (a-ha, mm-hm).

Table 1

Continued

Type of Error	Description of Error
Interviewer acknowledgement	Omission of acknowledgement words from interviewer. (e.g., OK, I see, yes, etc.) (e.g., P: It's a good program [I: Ok good], but there are issues"). <u>Not</u> for sounds (a-ha, mm-hm).
Topical word(s)	Omission of up to 3 consecutive words on the study topic (e.g., refugee, clinic, physician).
Non-topical word(s)	Omission of up to 3 consecutive words <u>not</u> on study topic. (e.g., the, and, so, for, etc.).
Topical phrase	Omission of a whole sentence/ phrase (more than 3 consecutive words) on the study topic. (e.g., "It's not NCD care but primary health care that needs support").
Non-topical phrase	Omission of a whole sentence/ phrase (more than 3 consecutive words) not on the study topic. (e.g., " Ok let's see , we discussed the availability of drugs").
Speaker identifier	Speaker identifier is omitted but can be determined on recording.
Verbal sounds	Omission of clear verbal sounds by interviewer or participant (e.g., laughter). Also, where a long pause (5+ seconds) is not marked with (pause).
III. Grammar	
Incorrect word conjugation	Incorrect conjugations of words, such as missing/added prefixes, (e.g., unfunded , thinking). This differs from incorrect word errors where a different word is transcribed (e.g., into vs to; are vs a).
Grammar correction	Grammar is corrected versus transcribing colloquial speech. (e.g., "I am going to" instead of "I'm gonna"). This is different from additional words in the transcript – use "word added" code for this.
Fase starts cleaned	When participant stumbles or repeats a word/phrase, and this is transcribed with false start text removed. (e.g., spoken words are "I can, I think, I can't think of any more issues" but is transcribed as "I can't think of any more issues" or "rec-rec-recognized" is transcribed as "recognized." This is different from grammar correction.
Spelling	Any spelling errors, including use of UK instead of US spelling (e.g., centre vs center; programme vs program).
Homophone	Word(s) transcribed are homophones of the correct word or sound like the correct word on recording. (e.g., "for" instead of "four"; "injury" instead of "engineer", "find the nice" instead of "finalize."

Table 1*Continued*

Type of Error	Description of Error
Acronyms	Where acronyms are transcribed as words. (e.g., who vs WHO) or incorrect acronym is used (e.g., WFP vs WFPS).
IV. Errors & Additions	
Word added	Addition of 1-3 consecutive words, not heard on recording. (e.g., “The Ministry of <i>Public</i> Health has”)
Arabic word	A word or phrase spoken in participants own language (e.g., Arabic) and transcribed as such or transcribed as an English word (phonetically not as a translation) or is heard but omitted.
Incorrect word	Incorrect word(s) used (for 3 or less consecutive words). Including nouns, names, place names that are incorrect (e.g., “standards” instead of “centers”). Does not refer to acronyms.
Incorrect phrase	Incorrect phrase (more than 3 consecutive words) is transcribed. (e.g., “The upcoming record” instead of “they are coming regularly”).
Incorrect speaker identifier	Incorrect speaker is marked on transcript, but clearly identifiable from recording. (e.g., I: instead of P:).
Unknown speaker	An unknown speaker is identified on transcript (e.g., “unknown speaker” or “UU”) but speaker is clearly identifiable from recording.
Additional speaker	An additional speaker is marked on transcript, which is different from I: or P: (e.g., P3, S1).

Note. Table adapted from Hennink & Weber (2013). In examples, bold text denotes missing text on transcript, italics denotes addition of words.

Coding Transcripts

All 30 transcripts were entered into MAXQDA v.24 to be coded for errors. To avoid the possibility of bias, the mode of transcription was blinded from analysts during coding, and no analyst coded a transcript they had transcribed themselves. Blinding transcripts was done by using consistent formatting (font type/size) of transcripts across the three modes, standardizing the format of speakers to I (for interviewer) and P (for participant), labeling each transcript with a random number and removing any identifying text from the AI modes (e.g., “Transcribed by Otter.ai”). Coding transcripts for errors was done in MAXQDA and involved listening to each digital recording and coding each instance where the transcript deviated from the recording by applying the relevant type of error code from the codebook. We piloted the coding process on one transcript by conducting an inter-coder agreement to assess coding consistency between coders. Coding discrepancies were discussed, and refinements made to the codebook before all transcripts were coded for errors.

Data Analysis

To assess transcription efficiency, we recorded the transcription time and cost for each transcription mode. Transcription time for both AI modes was recorded for each transcript and calculated as the number of minutes from initiating the AI transcription to its completion. An average time per transcription mode was then calculated. The transcription time for human transcription included both transcription and review time and was averaged across all transcripts. All AI-based transcriptions were completed at the same institution and time of day to provide consistency on internet connectivity/speed in the results. Transcription cost was estimated as an average per transcript for comparability across modes of transcription. The cost of AI transcription was estimated as the package or subscription price divided by the number of transcripts. The cost of human transcription was calculated as the hourly rate of transcribers multiplied by the time taken and averaged per transcript.

To assess transcription accuracy, we first tabulated the frequency of errors by the mode of transcription and compared the number and type of errors by mode. We then conducted further analysis by examining each type of error individually to understand the context and potential causes of each. This involved listening to segments of the recordings for each type of error, categorizing potential causes, and comparing these across transcription modes. Once we understood the context and causes of transcription errors, we re-grouped errors into four categories which had similar errors or causes as shown in Table 3. We also assessed whether the length of the recordings influenced transcription errors but found no patterns. Finally, we assessed errors by individual transcripts for each mode, to determine whether transcription errors are influenced by characteristics of the recording (e.g., background noise, unclear voices, speaker accent, etc.). For example, we wanted to examine whether an unclear recording would generate high errors regardless of the mode of transcription. However, we found no patterns. This underscores that the errors identified in our analysis are influenced by the mode of transcription rather than other underlying factors.

To assess the average time needed for error correction, we selected one transcript from each mode that represented the average number of errors for that mode. To do this, we divided the total number of errors for the mode by 10 interviews in that mode, then identified the transcript that had the closest number of errors to this average. We then calculated the correction time for each transcript by recording the time needed to correct all the errors in the transcript. This involved listening to the recording, correcting errors, and timing this process. We then calculated the total correction time per hour of audio, to account for differences in the length of the three audio recordings. Our results thus represent an average correction time for our data. We excluded formatting, deidentification and punctuation edits from error correction time and focused only on correcting errors shown in Table 1.

Results

Transcription Efficiency

Transcription efficiency was assessed by comparing the transcription time and cost across modes (Table 2). Human transcription was the most time-consuming, taking an average of 368.5 minutes (or 6 hours) per transcript (Range: 160 to 461 minutes). We found little correlation between the length of the recording and transcription time, whereby the longest recording did not take the

longest time to transcribe. Instead, transcribers reported that transcription time was influenced by the clarity and speed of speech, the amount of crosstalk, the extent of interviewer/participant acknowledgement words (e.g., OK, I see) and the audio quality, all of which influence the extent to which they had to rewind the recording during transcription. However, the shortest recording took the shortest amount of time to transcribe. The average transcription time for Otter.ai was 7.1 minutes per transcript (range: 1 min. 43 sec. - 12 min. 19 sec.) The longest recording took the longest time to transcribe, but the shortest recording was not the fastest to transcribe. Transcription time using MAXQDA was the fastest at 4.4 minutes per transcript (Range: 1 min. 26 sec.- 14 min 18 sec.) and the longest recording was the fastest to transcribe; however, the longest (114 min.) and shortest (30 min.) recordings took almost the same amount of time to transcribe - 1 minute 26 seconds and 1 minute 56 seconds, respectively. The cost of transcription is shown in Table 2, with human transcription being significantly more expensive per transcript than either of the AI modes. This is due to the time intensive nature of human transcription. However, there was a large difference in the cost per transcript between the two AI modes, with MAXQDA costing almost four times that of Otter.ai.

Table 2
Time and Cost of Transcription by Mode

Indicator	Mode of Transcription		
	Human ^a	Otter.ai ^b	MAXQDA – AI ^c
Transcription Time (Averaged per transcript)	368.5 minutes	7.1 minutes	4.4 minutes
Transcription Cost in USD (Averaged per transcript)	\$92.1	\$3.40	\$13.0

Note. ^aCost is based on the current hourly rate of \$15 for a graduate student at the time of this study. ^bA 2-month subscription was needed at \$17 per month, costing a total of \$34. ^cTwo transcription bundles were needed – 10 hours at \$80 plus an additional 5 hours at \$50, costing a total of \$130. These were the costs at the time each platform was used and are likely to change over time.

Transcription Accuracy

Transcription accuracy was assessed by the number of errors for each mode of transcription. Table 3 shows that all modes of transcription had a range of errors. Human transcription had the lowest total number of errors compared with either AI mode, with 3 times less errors than Otter.ai and 1.9 times less than MAXQDA. Otter.ai had the highest number of errors compared with other modes. The two AI modes differed considerably, with Otter.ai having 1.5 times more errors than MAXQDA.

Table 3
Type of Errors by Mode of Transcription

Type of Error	Mode of Transcription		
	Human	MAXQDA- AI	Otter.ai
I. Speaker Errors			
Omitted Speaker Identifier	21 (1.0%)	743 (19.3%)	589 (9.6%)
Incorrect Speaker Identifier	12 (0.5%)	402 (10.4%)	90 (1.4%)
Unknown Speaker	0 (0.0%)	132 (3.4%)	186 (3.0%)
Additional Speaker	0 (0.0%)	153 (3.9%)	0 (0.0%)
Omitted Interviewer Acknowledgement	84 (4.1%)	132(3.4%)	239 (3.8%)
Omitted Participant Acknowledgement	25 (1.2%)	42 (1.0%)	76 (1.2%)
<i>Subtotal</i>	142 (6.8%)	1604 (41.4 %)	1180 (19.0%)
II. Word Errors			
Omitted Topical Word(s)	77 (3.8%)	110 (2.8%)	142 (2.3%)
Omitted Topical Phrase	8 (0.3%)	14 (0.3%)	17 (0.2%)
Omitted Non-Topical Word(s)	579 (28.6%)	214 (5.5%)	557 (9.0%)
Omitted Non-Topical Phrase	9 (0.4%)	15 (0.3%)	35 (0.5%)
Incorrect Word	288 (14.2%)	454 (11.8%)	1836 (29.9%)
Incorrect Phrase	1 (0.0%)	14 (0.3%)	99 (1.6%)
Word Added	151(7.4%)	146 (3.8%)	216 (3.5%)
Homophone	43 (2.1%)	54 (1.4%)	165 (2.6%)
<i>Subtotal</i>	1156 (56.8%)	1021 (26.2%)	3067 (49.6%)
III. Grammar Errors			
Incorrect Word Conjugation	155 (7.6%)	107 (2.7%)	285 (4.6%)
Spelling	38 (1.8%)	25 (0.6%)	9 (0.1%)
Grammar Correction	88 (4.3%)	29 (0.7%)	46 (0.7%)
False Start Cleaned	283 (14.0%)	162 (4.2%)	360 (5.8%)
Acronym Error	30 (1.4%)	492 (12.8%)	653 (10.6%)
Verbal Sound	26 (1.2%)	38 (0.9%)	32 (0.5%)
Arabic Word Error	39 (1.9%)	264 (6.8%)	243 (3.9%)
<i>Subtotal</i>	659 (32.3%)	1117 (28.7%)	1628 (26.2%)

Table 3*Continued*

Type of Error	Mode of Transcription		
IV Inaudible Speech Errors			
Inaudible not Transcribed	17 (0.8%)	19 (0.4%)	52 (0.8%)
Inaudible Transcribed	44 (2.1%)	78 (2.0%)	207 (3.3%)
<i>Subtotal</i>	61 (2.9%)	97 (2.4%)	232 (4.1%)
Total Errors by Mode	2018	3839	6134

Speaker Errors

Six types of errors were related to issues with the identification of speakers (See Table 3). These errors together accounted for 41.4% of errors in MAXQDA, 19.0% in Otter.ai and 6.8% in human transcription. We assessed data to determine causes of these errors. We found that speaker errors had similar causes across all modes of transcription, including speaker interjections, crosstalk, and speaker designation. Speaker interjections caused errors when one speaker interrupts or acknowledges the other with one to three words (e.g., “I understand,” “I see,” “Yeah,” “OK”) and the change of speakers is omitted, incorrectly identified or transcribed as an additional speaker. Crosstalk caused errors when speakers talk at the same time, but this is transcribed as a single speaker. E.g., “So, it will include things like availability, accessibility, affordability among others, cancer, yeah. exactly. Yeah. Um, so um next few questions...” instead of “I: So, it will include things like availability, accessibility. P: Yeah. I: Affordability, among others. P: Cancer. I: Yeah. Exactly. P: Yeah. I: Um, so um, next few questions....” Speaker designation caused errors because AI modes labelled the first speaker on the recording as “speaker 1” and the next as “speaker 2” retaining this association throughout the transcript. Thus, if the interviewer speaks first on one recording and the participant speaks first on another recording the speaker identifiers will be inconsistent across the data set. Also, if an unknown or additional speaker is identified, subsequent speakers were sometimes incorrect. Although these issues were typically found in specific transcripts, it caused a high number of errors. For example, in MAXQDA speakers were switched in three transcripts which accounted for 87% of the incorrect speaker identifier errors for this mode of transcription. There were other issues with the transcription of speakers in the AI modes that could not be explained. For example, some sentences were broken up and transcribed as separate paragraphs, sometimes assigned to a different or unknown speaker; yet we heard nothing on the recording that might cause this.

Word Errors

Eight types of word errors were identified (see Table 3), which accounted for 26.2% of errors in MAXQDA, 56.8% and 49.6% in human and Otter.ai transcription, respectively. The omission of topical words or phrases was found across all modes of transcription at a similar level. These errors reduced the accuracy of statements particularly when the omitted word qualified a statement. For example, (omitted topical words in brackets):

- “It was (not) lack of commitment from staff ...” or “So we have this (potential) continuity of care...” (transcribed by human)
- “A patient needs (enough) information and education...” (transcribed by MAXQDA)
- “It's lack of affordability to buy (good) food.” (transcribed by Otter.ai).
- Other omitted words changed the specificity of a statement, for example:
- “Many health care providers are not available (in the country)” (transcribed by Otter.ai)
- “People are coming from a middle (income) country to...” (transcribed by human)
- “I didn’t get exactly what you mean by (safety of) the provider...” (transcribed by MAXQDA).

These types of word omissions compromise the meaning of a statement and therefore data validity. The omission of non-topical words and phrases were found in all modes, with the highest in human transcription (29% of all errors). Typically, these were omissions of short words of 2–3 letters (e.g., we, if, so, the), or speech filler words (e.g., “yeah,” “you know”). Mostly the omission of non-topical words had little effect on the sentence. For example (omitted non-topical word in brackets), “As we mentioned [before]...” (Otter.ai), or “They are following the trends, [the] markets...” (human), or “For people with disability and [for] elderly.” (MAXQDA). Errors due to an incorrect word or phrase was highest in Otter.ai (30% of all errors). These types of errors had similar causes for both AI modes. They sounded similar (but not a homophone), for example in MAXQDA, “cancer” was transcribed as “counsellor,” and “do not, so far” transcribed as “do not suffer”. Some incorrect words were proper nouns, for example, city and region names which were recognized as such but in the wrong country. For example, regions in Lebanon - Bekaa and Akaar - were transcribed as Dhaka (in Bangladesh) and Jakarta (in Indonesia) in MAXQDA. In other cases, the transcribed words made no sense, for example, “the same nurse is the director” was incorrectly transcribed as “capacitor seminars is a direct departure” (Otter.ai). In human transcription, incorrect word or phrase errors had no clear cause (e.g., “would” instead of “with” or “ad hoc” instead of “at home”) and there were no proper noun errors. The word added errors were found across all modes of transcription and included the addition of speech filler words, such as “like,” “you know,” or other words such as “and” or “the” which are not heard on the recording. Homophone errors were seen across all modes of transcription at similarly low percentages. Some homophones were difficult to distinguish on the recordings (e.g., to/two, weather/whether, patience/patients) but could be corrected based on the sentence. However, most homophone errors generated erroneous words particularly by Otter.ai, for example (correct word in brackets) - attacks (a tax), mitigation (medication), impairment (empowerment), prices (crisis), education (medication). These errors created nonsensical sentences that clearly reduced data validity.

Grammar Errors

Seven types of grammar errors were identified (see Table 3). These errors accounted for 32.3% of all errors in human transcription, 28.7% in MAXQDA, and 26.2% in Otter.ai. There were few spelling errors across all modes. In MAXQDA some errors were due to British vs. US spelling (e.g., “centre” vs. “center,” “programme” vs. “program”). In human transcription, spelling errors were often typos. Similar word conjugation errors were found across transcription modes. Human transcription had missing pluralization (e.g., “intervention” instead of “interventions”), and all

modes had prefix errors (e.g., “productive” instead of “reproductive,” “power” instead of “empower”) or suffix errors (e.g., “referred” instead of “referral,” “minister” instead of “ministry”). Grammar correction errors included words spoken as contractions that were transcribed in full, or vice versa (e.g., “cause” transcribed as “because,” “gonna” transcribed as “going to”), as well as incorrectly spoken grammar that was transcribed corrected (e.g., “they will check (the) medical files” – the word “the” was added or “in partner with” corrected to “in partnership with”). Several types of false start errors were found in all modes of transcription. These included: word repetition (e.g., “for for” or “they they”) where the repeated word is omitted; speaker’s self-correction (e.g., “how do...how much do”) where only the corrected speech is transcribed; and word stutters (e.g., “disa-disability” or “admin-administrative”) where the stutter is omitted. Two additional false start errors were found only in AI modes: word stutters transcribed as a different word (e.g. “primary health care” transcribed as “private primary health care” and “social worker per-per center” transcribed as “social worker births per center”) and a speaker’s mid-sentence word change cleaned up (e.g. “rega-concerning” as in regarding-concerning transcribed as “concerning”). A range of acronym errors were also found across all transcription modes, but higher in AI modes. These included incomplete or incorrect acronyms (e.g. INGO transcribed as NGO, PHC as PhD, MOPH as MPH); and acronyms transcribed into words or phrases (e.g. YMCA transcribed as “what am I saying”; PHCC as “purchases” or “pieces”). Arabic-word errors were higher in AI modes. Otter.ai most often omitted an Arabic word, while MAXQDA more often transcribed it as an English word (e.g., “Yaani” transcribed as “Yeah” or “many”) and human transcription included a mix of these. Finally, there were few verbal sound errors across all modes of transcription. AI modes did not transcribe laughter or long pauses heard on the recording; however, these were omitted at a similar level in human transcription.

Inaudible Speech Errors

We counted errors where the recording was inaudible, but some words were still transcribed (inaudible transcribed). This error occurred at low percentages across all modes of transcription (3.3% in Otter.ai, 2.0% in MAXQDA and 2.1% in human transcription), whereby inaudible speech was still transcribed, most commonly as a transliteration of the sounds. Where speech was inaudible and not transcribed, it was often marked as such in human transcription but not indicated and not transcribed in the AI modes.

Error Correction

Table 4 shows the average error correction time needed for each mode of transcription, and the average correction time per hour of recording. These results show that the average correction time per minute of recording was highest in Otter.ai which took twice as long to correct than transcripts generated by MAXQDA and human modes, which were similar per minute. Based on these results we calculated the average cost of error correction per transcript and added this to the initial transcription cost for each mode (from Table 2) to identify a total transcription cost per transcript. This was then multiplied by the 10 transcripts in the study to calculate the total cost of transcription and average error correction for the whole study. Overall, both AI transcription modes were more cost-effective than human transcription, despite the higher number of errors and longer correction time needed for these modes. The total cost of AI transcription for our project was on average 2.8 to 3.5 times less than human transcription (average of 3 times less).

Table 4

Average Error Correction Time and Cost per Mode of Transcription

Mode of Transcription	Average Error Correction Time (per transcript)	Average Error Correction Time (minutes per hour of recording) ^a	Cost of Transcription in USD and Average Error Correction (per transcript) ^b	Total Cost of Transcription in USD and Average Error Correction (per study) ^c
Human	81 minutes	66 min/hour	\$92.10 + \$16.50 (\$108.60)	\$1,086 /10 interviews
Otter.ai	169 minutes	140 min /hour	\$3.40 + \$34.95 (\$38.35)	\$383.50 /10 interviews
MAXQDA	141 minutes	72 min/hour	\$13.00 + \$18.00 (\$31)	\$310 /10 interviews

Note. ^a Number of minutes to correct 1 minute / 1 hour of recording (e.g., 1.1 minutes to correct 1 minute of recording). ^b Cost of transcription is the average for each mode from Table 2. Cost of correction is based on the average correction time for a 60-minute interview at \$15 hourly rate for graduate students at time of study. ^c Cost is based on the average correction time for a 60-minute interview at \$15/hour by the 10 interviews in the study.

Discussion

Our study contributes to the methodological literature in qualitative research by assessing the quality of three modes of transcription: human transcription and two commonly accessible AI transcription tools. Distinct from previous studies, we compared all modes of transcription to the audio recordings (rather than to human transcription), thereby acknowledging that all transcription is interpretive including human transcription. We assessed 10 interview transcripts in their entirety and provide a comprehensive assessment of the total number, type and nature of transcription errors within a full transcript, thereby providing a more complete and realistic assessment of the extent of errors across an entire study than other studies. We also provide data on the average time and cost of error correction by mode of transcription, which is inconsistent in previous studies. Finally, we empirically assess the causes of different types of errors and use these to develop guidelines for reducing each type of error, which is absent from previous literature. Where possible below we compare our results to similar previous studies.

Our results show that the AI transcription tools we assessed are highly efficient by generating an initial verbatim transcript within 4-7 minutes compared with 6 hours for a human transcriber, and at a cost of USD3-13 compared with USD92 per transcript for a human transcriber. However, this efficiency masks the reduced accuracy of these AI tools, which have two to three times more errors than human transcription. Manually reviewing AI generated transcripts to correct errors somewhat reduces their efficiency. Accounting for an average manual correction time, the time and cost estimates for AI transcription increase to 1.25-2.5 hours and USD31-38 per transcript, respectively. This is still three to six times faster and three times cheaper than human transcription, making these AI transcription tools an efficient option for qualitative research. However, our results emphasize the need to include error correction time in the study budget. Our results are similar to Eftekhari (2024) that showed editing AI generated transcripts from Teams took between 1.5-3.5 hours per transcript, with cost savings of GBP3,772 on transcription costs.

Our results on transcription accuracy show that all modes of transcription (human and AI) had errors, but the AI transcription tools had 2-3 times more errors than human transcription. There was also a marked difference in the number of errors between the two AI modes we assessed. Transcription accuracy is critical since it affects data validity in a study. Comparable studies used text comparison tools to assess a close match/ match/ no match (Louw, 2021), did not tabulate errors (Eftekari, 2024), showed wide variability in WER ranging from 3.3% to 28.1% (Kuhn et al, 2023) or “cannot draw any strong conclusions” (Rao, 2023, p.4). Nonetheless, errors in AI transcription persist across all studies and are variable. Therefore, the selection of an AI tool for transcription should go beyond a time/cost assessment and prioritize transcription accuracy. This may involve researchers piloting several AI tools to assess transcription accuracy prior to use.

Our results on the accuracy of transcription align with other studies that underscore the importance of reviewing and editing AI generated transcripts to improve data validity (Bokhove & Downey, 2018; Eftekari, 2024; Louw, 2021). Some errors in AI transcription can only be corrected manually (e.g., homophones, transliteration of acronyms, proper nouns or sounds, cleaned colloquial grammar), until AI technology improves. This may involve listening to audio recordings and making corrections. While this is also advised for human generated transcripts, AI transcription has significantly more errors. AI transcription can therefore provide an effective *first draft* transcript (step 1) that needs to be carefully reviewed and edited manually (step 2). Thus, AI tools become part of a two-step *transcription process*, rather than a method to generate a final transcript ready for analysis. This two-step process enables researchers to benefit from the efficiency of AI generated transcription and maintain data accuracy. Our study demonstrates why the review of AI transcription is warranted – by tabulating and categorizing all transcription errors found in AI tools. However, with the seemingly high accuracy of AI transcription, the concern remains that AI generated transcription is seen as producing data ready for analysis, and the critical step of review and editing is omitted (Kuhn et al, 2023; Louw, 2021). This would significantly compromise data validity, reducing the value of qualitative research. Instead, AI transcription could be viewed as part of the process of generating data rather than the final product (Lapadat, 2000).

Some AI transcription errors found in our study could be reduced, which would improve the accuracy of the first draft and decrease time needed for manual correction. Based on our results we suggest some considerations for the distinct types of transcription errors we identified.

Grammar Errors

Our results show that AI generated transcription is not completely verbatim (e.g., colloquial language and grammar is corrected, false starts are cleaned, pauses and laughter are omitted), therefore considerable editing is still needed to reflect verbatim speech. It's important for researchers to acknowledge this outcome and assess whether it impacts the analytic approach to be used in a study. Some analytic approaches (e.g., grounded theory; conversation analysis) rely on transcription of verbatim transcripts with colloquial speech, while for other approaches (e.g., a thematic analysis to generate survey items) the colloquial nature of speech is less important than the content. When a verbatim colloquial transcription is needed, AI generated transcripts will take longer to correct since grammar errors comprise 26-28% of all errors. However, when a verbatim colloquial transcript is not critical, the time needed to edit AI transcription will be significantly reduced, making AI transcription an efficient option. Therefore, the efficiency of AI transcription relates to the type of data needed for analysis. We also found a greater number of acronyms, Arabic

word and proper noun errors in AI modes, reflecting a potential limitation in AI performance with speaker accents.

Speaker Errors

Our results show the causes of speaker errors to be incorrect speaker designation, crosstalk, and speaker interjections. To minimize speaker errors, interviewers could be trained to refrain from verbal interjections/acknowledgments. To minimize crosstalk during an interview and for study participants to be provided with similar guidance prior to an interview. Other strategies may improve the accuracy of speaker designation in AI transcription (see Table 5). Managing speaker errors would have a significant impact on the accuracy of AI generated transcription. For example, preventing speaker errors in MAXQDA transcription would reduce the number of errors to be close to human transcription (2,235 errors in MAXQDA without speaker errors, versus 2,018 human). These results show how targeted efforts to reduce specific types of errors in AI transcription can significantly improve transcription accuracy, thereby eroding the main disadvantage of AI transcription – the high number of errors and additional time needed for error correction.

Word Errors

While some word errors seem less problematic than others, all word errors matter because they compromise the accuracy of a verbatim transcript. In our study, many passages of text contained several minor word errors that taken together influence the clarity of the passage and therefore data validity. For example, the AI generated passage below has several minor errors (e.g., incorrect words/phrase/word conjugation and added words) that cumulatively lead to a significant lack of clarity, creating meaningless data, compared with the corrected passage below.

Transcribed by AI:

So it's a multiple we make NCD medication available is a better quality of care, and we can provide we are not looking at complex medication. So, I don't think it's essential drug list around the travelers to make it available for everyone. (Otter.ai).

Corrected Transcription:

So, the more we make NCD medication available the better quality of care we can provide. We are not looking at complex medications, we are looking at the essential drug list, chronic drug list to make it available for everyone.

Therefore, all errors can potentially compromise data validity, and none can be evaluated in isolation. The context in which an error occurs is central to understanding its impact, which means it would be inappropriate to assume that one category of errors is inherently less consequential than another. This underscores the importance of manually reviewing and editing AI-generated transcription.

Based on our study results, we developed recommendations to maximize the accuracy and efficiency of AI transcription in qualitative research (Table 5). We recommend planning for AI transcription at all stages of a qualitative study from study design, data collection, transcription and reporting with the guidance below. As AI transcription tool advance, our results highlight several features that could improve accuracy. These include allowing users to enter study-specific terms (e.g., places, organizations, acronyms), select spelling and grammar conventions (US vs UK English, word contractions) and indicating the study's global region to improve place name

recognition and speaker identification. AI tools should also flag inaudible sections for targeted manual review.

Table 5

Recommendations for Using AI Transcription Tools in Qualitative Research

Research Stage	Recommendation for using AI Transcription
Study Design	Know AI transcription is not fully verbatim (e.g., pauses, laughter, speech stutters, false starts are not transcribed, and grammar is corrected). Assess whether verbatim transcription is needed for the analytic approach to be used. Pilot different AI transcription tools, to assess transcription efficiency vs. accuracy, the nature of errors and correction time needed. Ensure AI transcription cost and error correction time is included in study timeline and budget. Review how AI transcription services manage data security to ensure ethical compliance with research. This may involve additional consent from study participants to use AI transcription¹.
Data Collection	Ensure recording have limited background noise to improve audio quality (e.g., pilot recording device) Train interviewers to generate clear dialogue: - speak first on every recording to ensure consistent speaker designations across transcripts (e.g., Speaker 1 is always the interviewer) - speak slowly and clearly to avoid inaudible speech and ask study participant to repeat unclear, softly spoken or mumbled speech to be audible on recording. - replace verbal acknowledgements (e.g., Ok, I understand) with visual acknowledgment (e.g., head nodding, eye contact) to reduce speaker designation errors. - reduce speaker interjections causing crosstalk errors (e.g., wait until the interviewee finishes speaking before talking again) begin audio recording after introductions and avoid using participants name on the audio recording to maintain participant anonymity when uploaded to AI platform. Provide guidance for interviewees (before recording begins) to: - allow the interviewer to speak first on the recording to enable consistent speaker designation by AI across transcripts. - speak slowly and clearly to avoid inaudible speech. - avoid verbal acknowledgement interjections (e.g., yea, OK) - allow one person to speak at a time to reduce cross talk (e.g., wait until the interviewer has finished speaking before responding)

Table 5

Continued

Research Stage	Recommendation for using AI Transcription
AI Transcription & Editing	Use a two-step process to generate then edit AI-generated transcripts. When Generating AI Transcriptions: Check available settings on AI tools (e.g., selecting US vs British English, entering proper nouns, etc.) Check AI platform has ethical data storage for recordings and transcripts (e.g., data security measures) before uploading files¹. Ensure audio recording filenames do not contain identifiers (e.g., participant's name) before uploading to AI platform. When Editing AI-Generated Transcription: Review each AI generated transcript to ensure data validity, specifically to: - cross check all transcripts with the recording - correct all transcription errors. - check speakers are correctly identified. - mark sections of the recording that are unclear as “inaudible.” - check for homophone errors. - check place names and foreign words for accuracy. - edit grammar correction by AI to be verbatim. - be attentive to hallucination text generated by AI. - be attentive to transliteration of acronyms. Provide a list of place names, organization names, acronyms, and common topic words related to the study, to use when checking AI-generated transcript to improve accuracy.
Report Writing	Report that AI was used for transcription in publications. Identify specific AI tool used (e.g., name, version, date) Report how AI generated transcripts were checked for accuracy.

Note. ^a Samuel and Wassenaar (2024)

Study Limitations

Our results are based on the cost of AI transcription and institutional pay rates at the time of the study; however, these costs may change or be variable per institution. Regardless of any minor cost variation in future, our results provide important guidance. Our study assessed two AI transcription tools, however, results for other AI tools may differ. Also, AI technology is fast-changing and future improvements may benefit the accuracy of verbatim transcription. Correction-time calculations represent an average for each mode of transcription. Finally, our data comprised

English language transcripts by Arabic speakers, which may have influenced AI transcription performance in our study and is an area that warrants further research.

Conclusion

Our study provides a detailed assessment of the accuracy and efficiency of three modes of transcription, to provide empirically based guidance for researchers on selecting transcription options. Our results show that AI tools can offer an efficient option for generating the first draft of verbatim transcripts for qualitative research; however, manual review and editing remains essential. We provide guidance on how to reduce common transcription errors found in AI transcription to generate a more effective first draft. We also recommend being transparent on whether AI transcription was used in a qualitative study and for researchers to conduct specific tasks at different stages of the study to improve the accuracy of AI generated transcription.

Author Contributions

MH: study conception and design; data acquisition, analysis, and interpretation; drafting the manuscript.

BK: data acquisition, analysis, and interpretation; critical review of the manuscript.

EY: data acquisition, analysis, and interpretation; critical review of the manuscript.

NA: data acquisition; critical review of manuscript.

GF: study conception and design; data acquisition, analysis, and interpretation; critical review of the manuscript.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

This study received ethical approval from the Institutional Review Boards of Emory University in the USA (approval no. STUDY00006226) on 07/13/2023 and Modern University for Business and Science, Lebanon (approval no. MU-20230612-42) on 06/12/2023.

Informed Consent Form

Participants who contributed data for the main study upon which this analysis is based received, reviewed, and signed an informed consent document. This form is available by email request to the corresponding author (Hennink, mhennin@emory.edu).

Data Availability Statement

The dataset analyzed for the current study is not publicly available due to its use in further analysis for separate manuscripts. Data may be available later upon reasonable request.

Acknowledgements

We would like to express our sincere gratitude to Tyler Goodlet (Emory University) and Nathalie Fakhreddine (Modern University for Business and Science) who contributed to generating data used in this study, in particular human transcriptions of audio recordings.

Conflicts of Interest

The author(s) declare no conflicts of interest with respect to the research, authorship, and/or publication of this article.

AI Statement

We have not used any AI tool or technology in the preparation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Battaglia, F. (2024). From “listen and repeat” to “listen and revise”: How to transcribe interviews offline quickly and for free using voice recognition software. *International Journal of Qualitative Methods*, 23, 1–15. <https://doi.org/10.1177/16094069241247473>
- Bokhove, C., & Downey, C. (2018). Automated generation of ‘good enough’ transcripts as a first step to transcription of audio-recorded data. *Methodological Innovations*, 11(2). <https://doi.org/10.1177/2059799118790743>
- Bucholtz, M. (2000). The politics of transcription. *Journal of Pragmatics*, 32, 1439–1465. [https://doi.org/10.1016/S0378-2166\(99\)00094-6](https://doi.org/10.1016/S0378-2166(99)00094-6)
- Da Silva, J. (2021). Producing ‘good enough’ automated transcripts securely: Extending Bokhove and Downey (2018) to address security concerns. *Methodological Innovations*, 14(1), 1–11. <https://doi.org/10.1177/2059799120987766>
- Denham, M. A., & Onwuegbuzie, A. J. (2013). Beyond words: Using nonverbal communication data in research to enhance thick description and interpretation. *International Journal of Qualitative Methods*, 12, 670–696. <https://doi.org/10.1177/160940691301200137>
- Eftekhari, H. (2024). Transcribing in the digital age: Qualitative research practice utilizing intelligent speech recognition technology. *European Journal of Cardiovascular Nursing*, 23(5), 553–560. <https://doi.org/10.1093/eurjcn/zvae013>
- Green, J., & Thorogood, N. (2018). *Qualitative methods for health research*. Sage.
- Hammersley, M. (2010). Reproducing or constructing? Some questions about transcription in social research. *Qualitative Research*, 10, 553–569. <https://doi.org/10.1177/1468794110375230>

- Hennink, M., & Weber, M. B. (2013). Quality issues of court reporters and transcriptionists for qualitative research. *Qualitative Health Research*, 23(5), 700–710. <https://doi.org/10.1177/1049732313481502>
- Hennink, M., Hutter, I., & Bailey, A. (2020). *Qualitative research methods* (2nd ed.). Sage.
- Jaffe, A. (2007). Variability in transcription and the complexities of representation, authority, and voice. *Discourse Studies*, 9(6), 831–836. <http://www.jstor.org/stable/24049463>
- Johnson, B. E. (2011). The speed and accuracy of voice recognition software-assisted transcription versus the listen-and-type method: A research note. *Qualitative Research*, 11(1), 91–97. <https://doi.org/10.1177/1468794110385966>
- Johnson, J., Adkins, D., & Chauvin, S. (2020). A review of the quality indicators of rigor in qualitative research. *American Journal of Pharmaceutical Education*, 84(1), Article 7120. <https://doi.org/10.5688/ajpe7120>
- King, N., & Horrocks, C. (2010). *Interviews in qualitative research*. Sage.
- Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2023). Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing*, 16(4), Article 25. <https://doi.org/10.1145/3636513>
- Kvale, S., & Brinkmann, S. (2009). *Interviews: Learning the craft of qualitative research interviewing* (2nd ed.). Sage.
- Lapadat, J. C. (2000). Problematizing transcription: Purpose, paradigm and quality. *International Journal of Social Research Methodology*, 3(3), 203–219. <https://doi.org/10.1080/13645570050083698>
- Louw, S. (2021). Automated transcription software in qualitative research. In Proceedings of the International Conference: DRAL4 2021 (pp. 1–12). School of Liberal Arts, King Mongkut's University of Technology Thonburi. https://sola.pr.kmutt.ac.th/dral2021/index.php/proceedings-2/?utm_source=chatgpt.com
- Marshall, D., & Naff, D. (2024). The ethics of using artificial intelligence in qualitative research. *Journal of Empirical Research on Human Research Ethics*, 19(3), 92–102. <https://doi.org/10.1177/15562646241262659>
- Matheson, J. L. (2007). The voice transcription technique: Use of voice recognition software to transcribe digital interview data in qualitative research. *Qualitative Report*, 12(4), 547–560. <https://doi.org/10.46743/2160-3715/2007.1611>
- McMullin, C. (2023). Transcription and qualitative methods: Implications for third sector research. *Voluntas*, 34(1), 140–53. <https://doi.org/10.1007/s11266-021-00400-3>
- Meanwell, E., Walsh, A., Berg, P., & Tuna, E. (2024). Automated transcription service: Enabling transcription at scale for social science research. In Practice and experience in advanced research computing 2024: Human powered computing (Article 9, pp. 1–7). Association for Computing Machinery. <https://doi.org/10.1145/3626203.3670545>
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage.
- Nascimento, L. D. S., & Steinbruch, F. K. (2019). “The interviews were transcribed”, but how? Reflections on management research. *RAUSP Management Journal*, 54(4), 413–429. <https://doi.org/10.1108/RAUSP-05-2019-0092>
- Ochs, E. (1979). Transcription as theory. In E. Ochs & B. B. Schiefflin (Eds.), *Developmental pragmatics* (pp. 43–72). Academic Press. <https://doi.org/10.1017/S0047404500009465>

- Oliver, D. G., Serovich, J. M., & Mason, T. L. (2005). Constraints and opportunities with interview transcription: Towards reflection in qualitative research. *Social Forces*, 84, 1273–1289. <https://doi.org/10.1353/sof.2006.0023>
- Park, J., & Zeanah, A. E. (2005). An evaluation of voice recognition software for use in interview-based research: A research note. *Qualitative Research*, 5(2), 245–251. <https://doi.org/10.1177/1468794105050837>
- Point, S., & Baruch, Y. (2023). (Re)thinking transcription strategies: Current challenges and future research directions. *Scandinavian Journal of Management*, 39(2), Article 101272. <https://doi.org/10.1016/j.scaman.2023.101272>
- Poland, B. (1995). Transcription quality as an aspect of rigor in qualitative research. *Qualitative Inquiry*, 1(3), 290–310. <http://dx.doi.org/10.1177/107780049500100302>
- Rao, A. (2023). Transcribing educational videos using Whisper: A preliminary study on using AI for transcribing educational videos [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2307.03200>
- Samuel, G., & Wassenaar, D. (2024). Joint editorial: Informed consent and AI transcription of qualitative data. *Journal of Empirical Research on Human Research Ethics*, 20(1–2), 3–5. <https://doi.org/10.1177/15562646241296712>
- Sponholz, J., Weilinghoff, A., & Schopf, J. (2025). Halving transcription time: A fast, user-friendly, and GDPR-compliant workflow to create AI-assisted transcripts for content analysis [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2503.13031>
- Stolcke, A., & Droppo, J. (2017). Comparing human and machine errors in conversational speech transcription [Preprint]. *comarXi*. <https://doi.org/10.48550/arXiv.1708.08615>
- Tilley, S., & Powick, K. (2002). Distanced data – Transcribing other people’s research tapes. *Canadian Journal of Education*, 27(2), 291–310. <https://cje-rce.ca/index.php/cje-rce/article/view/2835>
- Tolle, H., Castro, M. del M., Wachinger, J., Putri, A. Z., Kempf, D., Denking, C. M., & McMahon, S. A. (2024). From voice to ink (Vink): Development and assessment of an automated, free-of-charge transcription tool. *BMC Research Notes*, 17, Article 95. <https://doi.org/10.1186/s13104-024-06749-0>
- Tracy, S. (2010). Qualitative quality: Eight big-tent criteria for excellent qualitative research. *Qualitative Inquiry*, 16(10), 837–851. <https://doi.org/10.1177/1077800410383121>
- Van Maanen, J. (1979). The fact of fiction in organizational ethnography. *Administrative Science Quarterly*, 24, 539–550. <https://doi.org/10.2307/2392360>
- Wald, M. (2006). Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. *Interactive Technology and Smart Education*, 3, 131–141. <https://doi.org/10.1108/17415650680000058>

Notes on Contributors

Monique Hennink, PhD. Dr. Hennink is a Professor in the Hubert Department of Global Health at Emory University. She is internationally recognized for her expertise in qualitative research, with over 30 years of experience conducting qualitative studies on public health issues. Dr. Hennink has authored leading textbooks on qualitative methods and developed the QUALWORKS training program for public health professionals. In addition to methodological research, her studies focus on social and cultural influences on health behaviors, with emphasis on sexual and reproductive health, sanitation, and chronic disease management.

Bahaa Kazzi. Bahaa Kazzi is a third-year medical student at the Emory University School of Medicine. His research in public and global health focuses on preventable and infectious diseases. His work also tackles the financial and psychological burden of cancer care in the United States. Bahaa believes that a multi-cultural perspective has the potential to significantly improve medical education and community development projects, especially in the Middle East and Central/South America.

Erin Yoon. Erin Yoon studied at the Emory College of Arts and Sciences and graduated in May 2025 with a Bachelor of Arts in Human Health. Throughout her undergraduate studies, she was involved in research and volunteered at community health clinics. She hopes to pursue a career in medicine and public health.

Nael Alami. Vice President at the Modern University for Business and Science (MUBS), Lebanon, is a molecular neuroscientist working at the intersection of neuroscience, global health, education technology, and Learning Sciences. He is the founder of MUBS's Virtual Exchange Program, which expands educational access and cross-cultural understanding for underserved learners. His scholarship includes research investigating mechanisms involved in motor neuron degeneration, global health, and the use of innovative educational technologies to promote communication and empathy.

Ghada Farhat, MPH, PhD. Dr. Farhat is an Associate Professor in the Hubert Department of Global Health at Emory University. She is an epidemiologist whose work focuses on the prevention and control of noncommunicable diseases (NCDs), with a particular emphasis on refugee and displaced populations. Her programmatic work centers on strengthening health systems through workforce training and capacity strengthening initiatives. She currently serves as Vice Chair for Academic Affairs in the Hubert Department of Global Health.

ORCID

Monique Hennink, <https://orcid.org/0000-0002-5853-5791>

Erin Yoon, <https://orcid.org/0009-0009-9407-7300>

Nael Alami, <https://orcid.org/0000-0002-1802-8250>

Ghada Farhat, <https://orcid.org/0009-0005-0152-5209>