JNCI Cancer Spectrum, 2025, 9(2), pkaf021
https://doi.org/10.1093/jncics/pkaf021
Advance Access Publication Date: February 8, 2025
Article

The use of large language models to enhance cancer clinical trial educational materials

Mingye Gao, MS[1], Aman Varshney, MS[2], Shan Chen, MS[3,4], Vikram Goddla [3,4], Jack Gallifant, MBBS[3,4], Patrick Doyle, MS[4], Claire Novack, BS[4], Maeve Dillon-Martin, BS[4], Teresia Perkins, BS[4], Xinrong Correia, BS[5], Erik Duhaime, PhD[5], Howard Isenstein, MA[6], Elad Sharon, MD[7], Lisa Soleymani Lehmann, MD, PhD[8], David Kozono, MD, PhD[4], Brian Anthony, PhD[1], Dmitriy Dligach, PhD[9], Danielle S. Bitterman, MD*[,3,4]

[1]Massachusetts Institute of Technology, Cambridge, MA 02139, United States
[2]Technical University of Munich, Munich 80333, Germany
[3]Artificial Intelligence in Medicine Program, Mass General Brigham, Harvard Medical School, Boston, MA 02115, United States
[4]Department of Radiation Oncology, Brigham and Women's Hospital/Dana-Farber Cancer Institute, Boston, MA 02115, United States
[5]Centaur Labs, Boston, MA 02116, United States
[6]Digidence, Bethesda, MD 20814, United States
[7]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, United States
[8]Department of Medicine, Mass General Brigham, Harvard Medical School, Boston, MA, United States
[9]Loyola University Chicago, Chicago, IL, United States

*Corresponding author: Danielle S. Bitterman, MD, Department of Radiation Oncology, Dana-Farber Cancer Institute/Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115, United States (dbitterman@bwh.harvard.edu).

Author Contributions: M. Gao and A. Varshney are co-first authors of this work.

Abstract

Background: Adequate patient awareness and understanding of cancer clinical trials is essential for trial recruitment, informed decision making, and protocol adherence. Although large language models (LLMs) have shown promise for patient education, their role in enhancing patient awareness of clinical trials remains unexplored. This study explored the performance and risks of LLMs in generating trial-specific educational content for potential participants.

Methods: Generative Pretrained Transformer 4 (GPT4) was prompted to generate short clinical trial summaries and multiple-choice question-answer pairs from informed consent forms from ClinicalTrials.gov. Zero-shot learning was used for summaries, using a direct summarization, sequential extraction, and summarization approach. One-shot learning was used for question-answer pairs development. We evaluated performance through patient surveys of summary effectiveness and crowdsourced annotation of question-answer pair accuracy, using held-out cancer trial informed consent forms not used in prompt development.

Results: For summaries, both prompting approaches achieved comparable results for readability and core content. Patients found summaries to be understandable and to improve clinical trial comprehension and interest in learning more about trials. The generated multiple-choice questions achieved high accuracy and agreement with crowdsourced annotators. For both summaries and multiple-choice questions, GPT4 was most likely to include inaccurate information when prompted to provide information that was not adequately described in the informed consent forms.

Conclusions: LLMs such as GPT4 show promise in generating patient-friendly educational content for clinical trials with minimal trial-specific engineering. The findings serve as a proof of concept for the role of LLMs in improving patient education and engagement in clinical trials, as well as the need for ongoing human oversight.

Background

Clinical trials are the gold standard for investigating management strategies that can potentially improve cancer patient outcomes. The experimental nature of clinical trials necessitates clear information and effective communication. However, patients have expressed the need for better resources to learn about their trial options.[1-4] Currently, the primary resources by which cancer patients and providers learn about clinical trial options are clinical trial registries, such as ClinicalTrials.gov.[5,6] ClinicalTrials.gov is a large, public database of clinical trials, but it primarily uses highly technical language that is geared toward a clinician and investigator audience, meaning they are often inaccessible to most patients. Therefore, there is still a lack of widely accessible materials to inform and educate potential participants about specific clinical trial options. This challenge is an important barrier to trial recruitment, informed consent, protocol adherence, and successful and timely accrual. At a more fundamental level, ensuring adequate information about clinical trials is imperative to ensure valid informed consent and widening access to treatment options within a clinical trial across sociodemographic backgrounds.

Large language models (LLMs) present a new opportunity to enhance trial processes via improved patient awareness and engagement. Prior studies have investigated LLMs to facilitate cancer patient education and communication, demonstrating promise but also risks arising from falsifications and fabrications.[7-11]

Received: December 5, 2024. **Revised:** January 21, 2025. **Accepted:** January 30, 2025

However, most work on using LLMs to improve clinical trial processes has focused on clinical trial matching,[3,4,12,13] and little research has investigated these models for enhancing informational and educational resources. The ability of LLMs to simplify and summarize texts, in particular, is an exciting avenue for improving education and awareness. For example, LLMs could be used to simplify and clarify the often complex and jargon-heavy information presented in clinical trial documents, including informed consent forms.[14,15] In addition, LLMs could support the development of new methods to support clinical trial education. For example, established methods to measure the quality of clinical trial understanding after informed consent, such as the validated Quality of Informed Consent questionnaire,[16] do not assess trial-specific details. LLMs could facilitate the development of trial-specific questionnaires, providing opportunities for patients to self-assess their understanding and a pathway for patient-specific education to address knowledge gaps.

In this study, we explored the potential and risks of using LLMs to generate informational and educational resources via secondary use of clinical trial materials. LLMs were used to generate short, plain-language summaries of a clinical trial and generate multiple-choice question-answer pairs from informed consent forms. Our findings provide proof of concept for the potential of LLMs to enhance informational and education resources for cancer clinical trials, which could improve patient engagement and support clinical trial processes.

Methods

Figure 1 illustrates our overall research approach of prompting Generative Pretrained Transformer 4 (GPT4) to generate new educational and informational resources for clinical trials: plain-language short summaries (to inform patients about a trial) and multiple-choice questions (to assess understanding of a trial).

Datasets

Informed consent forms for this study were collected using the ClinicalTrials.gov API;[17] PyMuPDF,[18] a publicly available Python library, was used to retrieve text from the PDF files.

For the summary generation task, 11 clinical trial informed consent forms were randomly selected from ClinicalTrials.gov for prompt engineering and initial evaluation by the research team (Table S1). The statistics of this set are shown in Figure S1. A larger set of informed consent forms were used for the large-scale questionnaire development. We selected 91 interventional cancer clinical trials registered between January 1, 2021, and April 15, 2024. This time frame was chosen to capture the most up-to-date practices in informed consent while still providing a substantial pool of studies for analysis. The distribution of informed consent form pages and number of tokens are shown in Figure S2.

Summary generation

We explored 2 approaches to generating the trial summary from consent forms: direct summarization from text and sequential extraction and summarization. In the direct summarization approach, the raw informed consent form text was provided to GPT4-0125 preview for summary generation with prompt instructions to include key information about the clinical trial, which were informed by the basic elements of informed consent in subpart A of the Revised Common Rule[19] (Figure S3). In the sequential summarization approach, the LLM was first prompted to extract the sections of text describing basic elements of informed consent, guided by subpart A of the Revised Common Rule from the raw informed consent form text (Table S2 and Figure S4, A). These extracted text sections were then used to generate a summary (Figure S4, B) with "temperature" and "top_p" set to 0. For both approaches, we set the sampling parameters temperature and top_p to 0 to ensure the generated summaries are consistent and deterministic, where temperature controls output randomness (0 being deterministic) and top_p limits token selection diversity (0 selecting only the most likely token).

Summary evaluation

To evaluate the quality of our 2 prompting approaches, 4 clinicians from the research team (JG, LL, DK, DB) evaluated the 11 clinical trial summaries generated using the 2 prompting methods. Each summary was evaluated for the binary presence or absence of key trial elements, readability, inaccuracies, biases,

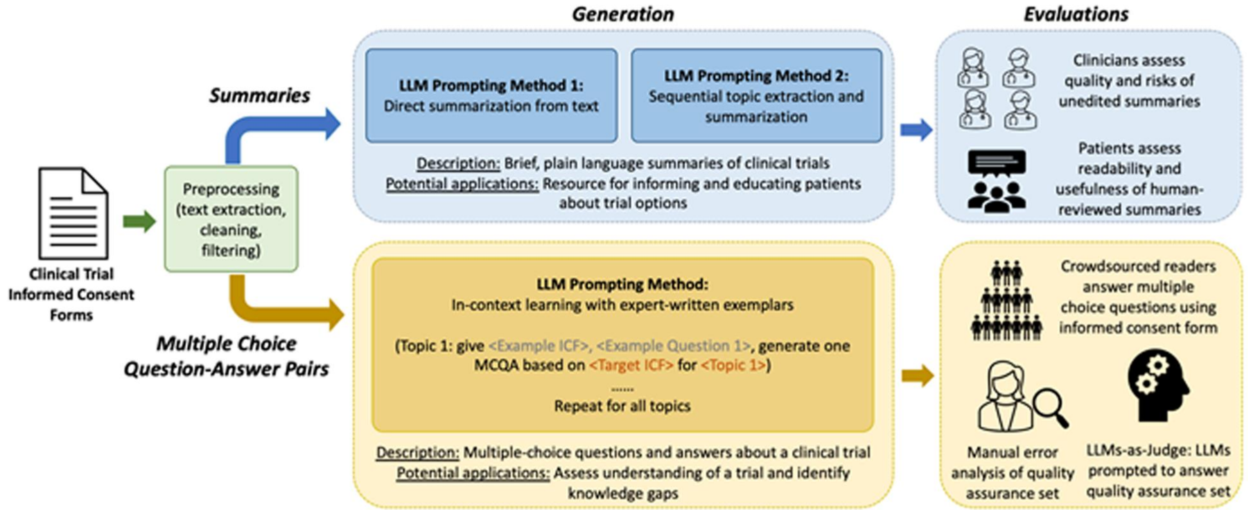
Generation
Evaluations
Summaries
LLM Prompting Method 1:
Direct summarization from text
LLM Prompting Method 2:
Sequential topic extraction and summarization
Description: Brief, plain language summaries of clinical trials
Potential applications: Resource for informing and educating patients about trial options
Clinicians assess quality and risks of unedited summaries
Patients assess readability and usefulness of human-reviewed summaries
Clinical Trial Informed Consent Forms
Preprocessing (text extraction, cleaning, filtering)
Multiple Choice Question-Answer Pairs
LLM Prompting Method:
In-context learning with expert-written exemplars
(Topic 1: give <Example ICF>, <Example Question 1>, generate one MCQA based on <Target ICF> for <Topic 1>)
......
Repeat for all topics
Description: Multiple-choice questions and answers about a clinical trial
Potential applications: Assess understanding of a trial and identify knowledge gaps
Crowdsourced readers answer multiple choice questions using informed consent form
Manual error analysis of quality assurance set
LLMs-as-Judge: LLMs prompted to answer quality assurance set

Figure 1. Illustration of the overall study design, including the approach to generating and evaluating summaries and multiple-choice question-answer pairs from clinical trial informed consent forms. Abbreviation: ICF = informed consent form; LLM = large language model; MCQA = multiple-choice question-answer.

and hallucinations. Overall summary quality was assessed using a 5-point Likert scale. The full survey, including the LLM-generated summaries for each trial, is available at www.github.com/BittermanLab/clinicaltrial-engagement-education.

Next, to explore patient perspectives of the generated summaries in a real-world setting, we invited patients undergoing informed consent for the Brigham and Womens's Hospital Radiation Oncology All-Department Biorepository to Accelerate New Discoveries (BROADBAND) research study (hereafter referred to as BROADBAND), a prospective secondary use protocol in the Department of Radiation Oncology at Brigham and Women's Hospital/Dana-Farber Cancer Institute, to participate in a survey evaluating 5 LLM-generated cancer clinical trial summaries generated using the sequential summarization approach, including a summary of BROADBAND as well as 4 other trials representing an observational trial, phase I trial, phase I-II trial, and phase III trial (Table S3). To avoid any risks of misinformation, all summaries underwent manual review and editing by an oncologist. In addition to minimizing any patient risk, this mimics the real-world application of such a summary, where a member of the clinical trial research team would review the GPT4-generated content before it reaches the patient. Participants completed the survey after the BROADBAND informed consent discussion, which provided the opportunity to assess the impact of the summary on the understanding of a clinical trial for informed consent. The survey tool was developed in RedCap, and participants could complete the survey on paper, a laptop in the clinic, or via an emailed link. Informed consent for the survey was waived as this study was deemed to be exempt human subjects research by the Mass General Brigham institutional review board (MGB protocol # 2024P000949).

Multiple-choice question-answer pair generation

GPT-4-1106 preview was used for multiple-choice question-answer generation. To ensure the quality of the generated multiple-choice question-answers, we adopted the in-context learning method: when generating a multiple-choice question-answer for each topic based on a target informed consent form, we fed an expert-created question-answer pair and its corresponding informed consent form text, along with the target informed consent form text, into GPT4. Fifteen multiple-choice question-answers focused on a subset of the basic elements of informed consent were manually written by a board-certified oncologist (DB) based on an exemplar informed consent form for a noncancer clinical trial[20] (see Table 1). Using these pairs as in-context examples, we engineered a multiturn prompt to generate multiple-choice question-answers (Figure S5). For each generation query, temperature and top_p were set to 0 to ensure the consistency of the outputs, and max_tokens was set to 3000. After filtering invalid generations (eg, responses returned by GPT4 that were not a multiple-choice question-answer), a total of 1335 multiple-choice question-answers were generated for 91 informed consent forms.

Multiple choice question-answer pair evaluation

Annotations of the GPT4-generated multiple-choice question-answers were obtained using DiagnosUs, a medical annotation crowdsourcing platform (Centaur Labs, Boston, MA, USA), which uses continuous performance monitoring and incentivization to optimize quality. Crowdsourced readers were provided the informed consent form for each set of multiple-choice question-answers and instructed to answer the GPT4-generated question stems based on information present in the informed consent form. In total, 504 crowdsourced readers participated and included individuals with a range of clinical and nonclinical backgrounds; details on the breakdown of self-reported clinical training are reported in Table S4. The number of human readers answering each GPT4-generated multiple-choice question-answer stem is defined as qualified reads. Evaluations are reported using metrics defined in Table 2.

The large number of generated multiple-choice question-answers precluded complete manual error analysis. Therefore, we defined a quality assurance set of 78 multiple-choice question-answers with difficulty of at least 0.6 and agreement of no more than 0.5 for manual error analysis (ie, multiple-choice question-answers with a majority of readers disagreement with the GPT4-assigned answer and with substantial disagreement between readers). Our goal with this quality assurance set was to gain a deeper understanding of why readers tended to disagree with the GPT4-assigned answer to understand failure modes. In addition, we used a multi-agent framework where 1 LLM verifies the GPT4-generated multiple-choice question-answers for the quality assurance test set. GPT-4o, Cohere R+, Gemini Pro 1, and Claude 3 Sonnet were prompted to answer the multiple-choice question-answers (Figure S6); temperature, top_p, and max_tokens were set to 0, 0, and 300, respectively, for the 4 LLMs.

Results

Clinician summary evaluation

For the 11 summaries evaluated by clinicians, both prompting approaches achieved comparable results for readability and topic content (Figure 2). There was slightly less evidence of inaccuracies, biases, and hallucinations using the sequential prompting approach. Quality was rated as acceptable or better in the majority of responses using both prompting approaches (Figure 3), although results varied substantially across trials and evaluators. We found that inaccuracies and hallucinations tended to occur for topics that were not described in the given informed consent form. Summaries generated using the sequential prompting method were preferred in 38.6% (17 of 44) responses, and summaries generated using the direct prompting method were preferred in 61.4% (27 of 44) responses.

Patient summary evaluation

A total of 13 patients completed the survey of the 5 cancer clinical trial summaries; Table 3 summarizes their characteristics, and Table 4 reports the key results. Complete results for each trial are in Table S5. Of note, a substantial portion of participants did not complete the survey for the non-BROADBAND trial summaries. Nevertheless, the majority of respondents found the summaries easy to understand and agreed or strongly agreed that the summaries provided sufficient information to decide about contacting the research team. Of the 13 participants, 11 (85%) reported that the summary improved their understanding of the BROADBAND trial after having completed the informed consent process for the trial.

Multiple-choice question-answer pair evaluation

Multiple-choice question-answers had an average of 5.21 (1.99) qualified reads. The majority answer agreed with the GPT4 answer in 1307 of 1335 (97.91%) of multiple-choice question-answers, with an average agreement of 86.87% and an average difficulty of 15.52%. Of note, the median difficulty and agreement were 0.0 and 1.0, respectively, demonstrating that all readers' answers matched the GPT4-assigned answer in more than half of

Table 1. Oncologist-written multi-choice question-answers used as in-context examples for automated generation with GPT4

Topic	Questions	Multi-choice options	Answer
A statement that the study involves research	True or false: This study involved research.	A) True; B) False	A
An explanation of the purposes of the research	What program is being evaluated in this research study?	A) A new blood pressure medication; B) The STOP program; C) Referral to a clinic neurologist only; D) Referral to a primary care physician only	B
The expected duration of the subject's participation	How long will each participant be involved in this study?	A) 6 months; B) 3 months; C) 12 months; D) 36 months	A
A description of the procedures to be followed	If you are randomly assigned to the STOP-Stroke group, what will occur at 1 week, 1 month, 3 months, and 5 months after enrollment?	A) Your blood pressure will be checked; B) You will be asked to complete a questionnaire; C) You will have a video visit with a nurse practitioner or stroke physician, social worker, and pharmacist and will receive educational text messages every other week; D) You will have a visit with your neurologist and primary care provider	C
The approximate number of subjects involved in the study	About how many patients will be enrolled in the study?	A) 25; B) 50; C) 75; D) 100	D
Identification of any procedures that are experimental	What is the experimental intervention in this study?	A) Follow-up with primary care to monitor your blood pressure every 2 weeks; B) 24-hour blood pressure monitoring each month; C) STOP program: video visits with a stroke prevention team and educational text messages; D) Educational text messages every other week alone	C
A description of any reasonably foreseeable risks or discomforts to the subject	The following is a possible risk of this study.	A) Breach of confidentiality; B) Worsening blood pressure; C) Increased risk of stroke; D) None of the above	A
A description of any benefits to the subject or to others that may reasonably be expected from the research	Select the benefits of participating in the study (you may select more than one).	A) You will receive educational materials about blood pressure control after stroke; B) Your blood pressure will be better controlled; C) Your stroke risk will be lower; D) You will receive 24-hour blood pressure monitoring	A, D
A disclosure of appropriate alternative procedures or courses of treatment, if any, that might be advantageous to the subject	What is the alternative option to participating in the study?	A) You will be offered a visit in the stroke clinic and will follow up with your primary doctor; B) You will have regular video visits with a stroke specialist; C) There is no alternative to participating; D) You will receive regular education about reducing your stroke risk	A
A statement describing the extent, if any, to which confidentiality of records identifying the subject will be maintained	If you participate in this study, who may review your personal health information? You may select more than 1 option.	A) Representatives of the UTHealth Sciences Center at Houston; B) The research sponsor; C) No one; D) Researcher from other hospitals	A, B
For research involving more than minimal risk, an explanation as to whether any compensation and an explanation as to whether any medical treatments are available, if injury occurs, and if so, what they consist of or where further information may be obtained	What treatment will be available for you if you are injured as a result of participating in the study?	A) No treatment; B) All needed facilities, emergency treatment, and professional services will be made freely available; C) All needed facilities, emergency treatment, and professional services will be made available but not free of charge; D) None of the above	C
Research, rights, or injury: An explanation of whom to contact for answers to pertinent questions about the research and research subjects' rights and whom to contact in the event of a research-related injury to the subject	Who can you contact with any questions, concerns, or input about the study?	A) The principal investigator and members of the study team; B) No one; C) Administrators at the UTHealth Sciences Center at Houston; D) Your primary physician	A
A statement that participation is voluntary, refusal to participate will involve no penalty or loss of benefits to which the subject is otherwise entitled, and the subject may discontinue participation at any time without penalty or loss of benefits, to which the subject is otherwise entitled	If you enroll in the study, when can you choose to stop participating?	A) You cannot stop participating once you enroll; B) Only within 1 week of enrolling; C) At any time during the study; D) Before you complete the first questionnaire	C
			B

(continued)

Table 1. (continued)

Topic	Questions	Multi-choice options	Answer
Any additional costs to the subject that may result from participation in the research	Who will be charged for medications, studies, or procedures recommended to you during this study?	A) The study sponsor; B) You or your insurance, because they will be considered standard of care; C) UTHealth, because they are running the study; C) None of the above	
The consequences of a subject's decision to withdraw from the research and procedures for orderly termination of participation by the subject	What will occur if you decide to stop being a part of the study?	A) You will not be able to continue receiving stroke care at UTHealth; B) There will be no change to the services available to you from UTHealth; C) You will not be eligible for clinical trials in the future; D) You will need to change your primary doctor	B

Abbreviations: GPT4 = Generative Pretrained Transformer 4; STOP = STROKE Telemedicine Outpatient Program; UTHealth = University of Texas Health.

Table 2. Definitions of metrics and associated terms used in multiple-choice question-answer evaluations

Term	Definition
Qualified reads	Number of crowdsourced readers answering each GPT4-generated multiple-choice question-answer stem
Difficulty	Qualified reads disagreeing with the GPT4 answer divided by total qualified reads
Agreement	Qualified reads with the majority answer (ie, the answer chosen by the majority of qualified reads) divided by total qualified reads
Accuracy	Percentage of multiple-choice question-answers where the majority answer matches the GPT4 answer

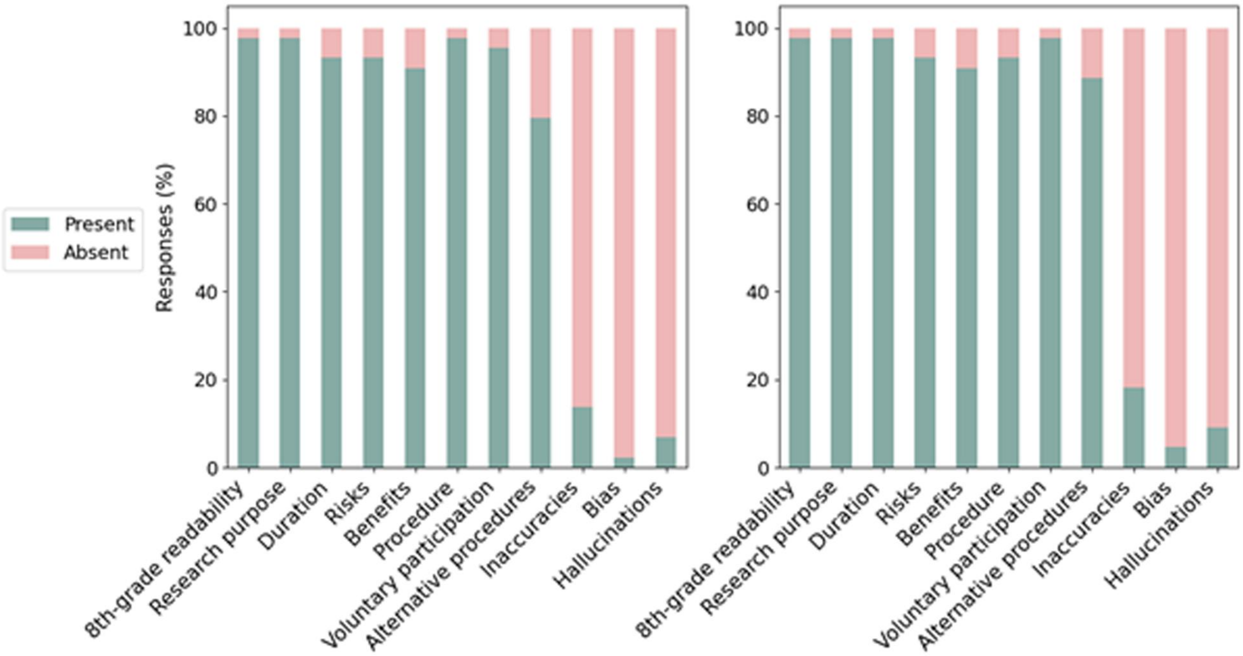

Figure 2. Clinician evaluation of 11 clinical trial summaries generated using the sequential prompting approach (**left**) and the direct prompting approach (**right**). There were 44 responses per item in the clinician evaluation summary.

all multiple-choice question-answers. Table S6 and Figure S7 provide detailed statistics of qualified reads, agreement, and difficulty. When broken down by multiple-choice question-answer topic, average agreement was lowest and difficulty highest for multiple-choice question-answers about contact information, expected duration, and alternative procedures (Table S7 and Figure S8).

We identified 78 of 1335 (5%) multiple-choice question-answers with difficulty of at least 0.6 and agreement of no more than 0.5 for our quality assurance set for manual error analysis. Manual review identified 5 error modes, which are summarized in Table 5. Of these, the majority of errors were human error, followed by errors in GPT4-generated multiple-choice question-answers, errors due to missing information in the informed consent forms, and errors arising from ambiguous language.

For most multiple-choice question-answers in the quality assurance set, the 4 LLMs disagreed with the multiple-choice question-answers when there were incorrect GPT4-assigned

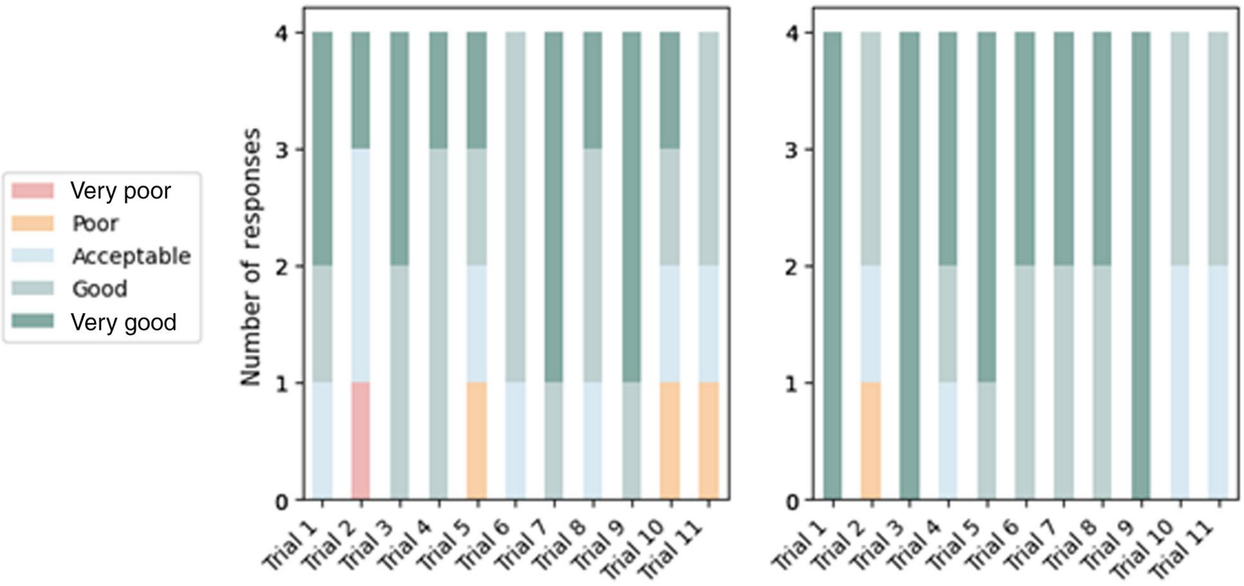

Figure 3. Clinician evaluation of overall quality of the 11 clinical trial summaries generated using the sequential prompting approach (**left**) and the direct prompting approach (**right**).

Table 3. Characteristics of clinical trial summary patient survey participants

Category	No. survey participants (n = 13)
Gender	
Male	10
Female	2
Missing	1
Age, y	
31-40	1
41-50	0
51-60	3
61-70	6
71-80	3
Race	
Asian	1
White	12
Ethnicity	
Hispanic	0
Non-Hispanic	13
Prior clinical trial experience	
Yes	4
No	9

labels, missing information in the informed consent forms, and ambiguous definitions. This suggests that testing the generated multiple-choice question-answers using other LLMs may be a promising avenue to assist humans in proofreading the quality of LLM-generated multiple-choice question-answers. More details on the LLM and human reader results on the quality assurance set are shown in Figures S9-S12.

Discussion

This study explored the potential of LLMs to automatically draft patient-friendly summaries and multiple-choice question-answers from informed consent forms. Experimental results demonstrated that LLMs can effectively draft readable and accurate summaries of informed consent forms with straightforward prompting techniques, with patient surveys suggesting potential roles in improving trial awareness and consent quality. Additionally, LLMs were able to generate high-quality multiple-choice question-answers based on the content of target informed consent forms. These results provide proof of concept for leveraging LLMs to accelerate the development of new, diverse, and scalable educational resources for clinical trial patient education while also highlighting key error modes that require ongoing human oversight and vigilance

Prior studies have demonstrated the potential of leveraging LLMs for clinical education, including the generation of patient-facing disease-specific information[21,22] and summarization and simplification of existing clinical documents.[23-32] In similar work, White et al.[29] used GPT3.5 to generate summaries of multiple clinical trials using the brief descriptions in ClinicalTrials.gov study records, although these were intended for researcher audiences. In addition, studies have shown the ability of LLMs to simplify language in consent forms for standard medical procedures.[33-36] Our findings build on this prior literature to demonstrate 2 new, promising applications of LLMs to support the unique informational and educational needs of patients learning about clinical trials.

There is a need for better patient education about clinical trial options.[37-39] Inadequate educational resources about clinical trials limit awareness of and engagement in trials and may contribute to the high rate of cancer trials that fail to accrue.[40,41] New diversified resources to educate patients about clinical trials could increase enrollment rates, improve patient understanding, and potentially broaden clinical trial access and diversity.[42-48] In fact, patients have expressed a need for more awareness about clinical research. During recruitment, patients and caregivers report a lack of familiarity with trial options and are more likely to have positive attitudes about participation if they learn about trials.[3,4,7] Past studies have explored novel approaches to improve education, primarily at the informed consent stage,[49-55] but scalability has previously been limited by the time and engineering expertise needed to develop trial-specific resources. Our findings show the potential of LLMs to lower the barrier to generating a diversity of educational resources from documents that

Table 4. Key results of patient surveys evaluating 5 cancer clinical trial summaries[a]

Survey item	Strongly disagree, No. (%)	Disagree, No. (%)	Neither agree nor disagree, No. (%)	Agree, No. (%)	Strongly agree, No. (%)	Missing, No. (%)
All 5 trial summaries						
This summary is easy to understand.	0 (0)	3 (5)	3 (5)	18 (28)	26 (40)	16 (25)
If I were researching trials, this summary provides enough information for me to decide if I would want to contact the research team to learn more about the trial.	0 (0)	2 (3)	10 (15)	21 (32)	21 (32)	11 (17)
BROADBAND summary only						
I believe that reading this summary improved my understanding of BROADBAND.	0 (0)	0 (0)	2 (15)	6 (46)	5 (38)	1 (8)

[a] Complete results of the survey, presented for each trial individually, are in Table S5.

Table 5. Error modes identified on qualitative error analysis of the quality assurance set

Error mode	Description	No. of MCQAs
Human error	The MCQA is correct, and information needed to correctly answer it is present in the ICF. Of these, 24 MCQAs had an answer that was not difficult to identify in the ICF, and 3 MCQAs included a large number of details (eg, >20 procedures, >1 page of potential risks), which made it particularly challenging to arrive at the correct answer.	27
Missing information in ICF	The topic corresponding to the MCQA is not included or explicitly demonstrated in ICF.	18
Error in GPT4-generated MCQA	The GPT4-generated MCQA included more than 1 correct answer when only 1 correct answer was assigned (n = 10) or the answer generated in the GPT4-generated MCQA was incorrect (n = 7).	17
Ambiguous definition	The definition of term(s) in MCQA or ICF is not clear enough to point to a correct label (eg, ICF mentions "patients should inform their physician if they get injured during the study"; what does "physician" mean? Does it mean the principal investigator of the study or their primary care doctor?).	13
Not in English	ICF is not in English, and readers cannot understand the MCQA	3

Abbreviations: ICF = informed consent forms; MCQA = multiple-choice question-answer.

are already created as a part of standard clinical trial conduct. Our prompting methods do not require significant engineering expertise to implement and are agnostic to the specifics of a given clinical trial.

While our findings demonstrate that GPT4 can, in general, follow prompt instructions to convert informed consent forms into new educational resources, they also highlight error modes necessitating ongoing human oversight and methods refinement. Though rare, the summaries and multiple-choice question-answers included inaccuracies and hallucinations, a known challenge of working with LLMs. These most often occurred when the informed consent form did not include adequate content requested in the prompt. This limitation may arise out of LLMs' alignment tuning, which leads them to prioritize helpfulness (ie, following users' instructions in the prompt) over factual accuracy—a key error mode to monitor for such LLM applications. Our sequential prompting method for summaries, which first extracted relevant informed consent form text and then summarized over the extracted text, appeared to mitigate but not completely alleviate this error mode. Further prompt refinement may improve these errors; however, human oversight is still needed to ensure accurate, comprehensive, and safe information when using LLMs for summarization, which may limit scalability.[56] At the same time, we note that even manually written patient-facing recruitment and consent materials standardly undergo additional human review by institutional review boards.

Implementing such LLM applications for clinical trial processes, and in health care more broadly, is currently limited by a lack of effective means for large-scale evaluation and ongoing monitoring of model performance. As above, while promising, even state-of-the-art LLMs such as GPT4 require a human in the loop to identify and resolve errors before they reach patients.[57] Nevertheless, our methods may lower the barrier to develop educational content because clinical trial staff may find it easier to review and revise LLM-generated drafts than to develop the content from scratch.

This study has several limitations. First, we evaluate a relatively limited number of trials and have a small number of human evaluators for the summaries, which limit generalizability. Additionally, the patients who agreed to participate in summary evaluations may not be reflective of the broader cancer population, including a lack of diversity. That said, ours is one of the very few studies that have assessed patients' perceptions of LLM outputs[58,59] and, to our knowledge, the only study evaluating patient perceptions of LLM content that relates to their own health care (ie, their understanding of the BROADBAND study). Further, the number of human evaluators in our study falls within the range of other studies evaluating LLMs for patient education.[60] Nevertheless, given these limitations, our results should be considered as early proof of concept, and larger-scale studies demonstrating safety, acceptability, and effectiveness are needed. In addition, we may not have used the optimal prompting approaches, and it is possible error rates could be reduced with additional prompt engineering. However, our goal was to understand the performance, behavior, and risks of widely available LLMs without significant additional engineering efforts, serving as a baseline for future technical innovation and advances. Similarly, including additional clinical trial materials, such as information from the ClinicalTrials.gov study records and trial protocols, may be a promising avenue to reduce the observed risk of inaccuracies and hallucinations by providing more substantive content on topics that may not be adequately described in an informed consent form. Finally, patients have diverse informational and educational needs. We explored LLMs to generate 2 types of educational resources, however, we did not explore personalizing the resources to individual preferences.

Our findings demonstrate a promising role of LLMs in patient-centered clinical trial education and informational resources. Future research should focus on optimizing output through advanced prompting techniques and automated oversight, understanding how differences in informed consent form and trial complexity impacts LLM output, investigating personalized approaches, and rigorously validating the safety and comparative effectiveness compared with existing approaches for clinical trial education.

Acknowledgments

The funders did not play a role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; and the decision to submit the manuscript for publication.

The authors would like to acknowledge the BROADBAND Research Project at the Brigham & Women's Hospital Department of Radiation Oncology for providing regulatory and personnel support for this project. The BROADBAND Project was in part made possible by the generous donations of Stewart Clifford, Fredric Levin, and their families.

Author contributions

Mingye Gao, MS (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing—original draft, Writing—review & editing), Aman Varshney, MS (Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing—original draft), Shan Chen, MS (Conceptualization, Data curation, Formal analysis, Investigation, Methodology), Vikram Goddla (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing—review & editing), Jack Gallifant, MBBS (Formal analysis, Investigation, Methodology, Supervision, Writing—review & editing), Patrick Doyle, MS (Investigation, Methodology, Project administration, Writing—review & editing), Claire Novack, BS (Data curation, Investigation, Methodology, Project administration, Writing—review & editing), Maeve Dillon-Martin, BS (Data curation, Investigation, Methodology, Writing—review & editing), Teresia Perkins, BS (Data curation, Investigation, Methodology, Project administration, Writing—review & editing), Xinrong Correia, BS (Data curation, Formal analysis, Investigation, Methodology, Resources, Validation, Writing—review & editing), Erik Duhaime, PhD (Formal analysis, Investigation, Methodology, Resources, Supervision, Validation, Writing—review & editing), Howard Isenstein, MA (Conceptualization, Methodology, Writing—review & editing), Elad Sharon, MD (Conceptualization, Formal analysis, Investigation, Supervision, Writing—review & editing), Lisa Soleymani Lehmann, MD, PhD (Conceptualization, Formal analysis, Investigation, Writing—review & editing), David Kozono, MD, PhD (Conceptualization, Formal analysis, Investigation, Methodology, Resources, Supervision, Writing—review & editing), Brian Anthony, PhD (Funding acquisition, Supervision, Writing—review & editing), Dmitriy Dligach, PhD (Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Writing—original draft), and Danielle S. Bitterman MD (Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Writing—original draft, Writing—review & editing).

Supplementary material

Supplementary material is available at *JNCI Cancer Spectrum* online.

Funding

The authors acknowledge financial support from the Google PhD Fellowship (SC), the Woods Foundation (DB, SC, JG), the NIH (NIH R01CA294033 [SC, JG, ES, DK, DB], NIH U54CA274516-01A1 [SC, DB], NIH R01LM012973 [DD]), and the American Cancer Society and American Society for Radiation Oncology, ASTRO-CSDG-24-1244514-01-CTPS Grant DOI: https://doi.org/10.53354/ACS.ASTRO-CSDG-24-1244514-01-CTPS.pc.gr.222210 (DB). This work was also conducted with support from UM1TR004408 award through Harvard Catalyst — the Harvard Clinical and Translational Science Center (National Center for Advancing Translational Sciences, National Institutes of Health) and financial contributions from Harvard University and its affiliated academic health-care centers. The content is solely the responsibility of the authors and does not necessarily represent the official views of Harvard Catalyst, Harvard University, and its affiliated academic health-care centers, or the National Institutes of Health. The authors thank Google Cloud research fund for Claude API inference costs.

Conflicts of interest

D.K.: Advisory and consulting, unrelated to this work: Genentech/Roche.

D.S.B.: Editorial, unrelated to this work: associate editor of Radiation Oncology, HemOnc.org (no financial compensation); advisory and consulting, unrelated to this work: MercurialAI (Dr. Danielle Bitterman has a financial interest in Mercurial AI, which is developing an AI platform to assist oncologists with reviewing patient records and treatment recommendations, and an AI chatbot to provide guidance and reassurance to cancer patients. Dr. Bitterman's interests were reviewed and are managed by Brigham and Women's Hospital and Mass General Brigham in accordance with their conflict of interest policies.).

H.I.: Principal, unrelated to this work; Digidence.

L.S.L.: Employee of Verily; advisor unrelated to this work: MediSensor Technologies, BellSant.

X.C.: Employee of Centaur Labs.

E.D.: Employee of Centaur Labs.

Data availability

All data and code are publicly available at www.github.com/BittermanLab/clinicaltrial-engagement-education.

References

1. Kumar G, Chaudhary P, Quinn A, Su D. Barriers for cancer clinical trial enrollment: A qualitative study of the perspectives of healthcare providers. *Contemp Clin Trials Commun*. 2022;28: 100939. https://doi.org/10.1016/j.conctc.2022.100939
2. BECOME Initiative Final Report. MBCA (Melanoma and Brain Cancer Alliance). Retrieved from https://www.mbcalliance.org/wp-content/uploads/BECOME-Final-Report-FULL.pdf
3. *Pretesting NIH Clinical Trial Awareness Messages: A Focus Study with Patients, Caregivers, and the General Public*. National Institutes of Health; 2011.
4. The need for awareness of clinical research. National Institutes of Health (NIH). Accessed May 30, 2015. https://www.nih.gov/health-information/nih-clinical-research-trials-you/need-awareness-clinical-research
5. N. L. of Medicine. ClinicalTrials.gov. 2023. Accessed November 23, 2023. https://clinicaltrials.gov/
6. Nass SJ, Moses HL, Mendelsohn J; Physician and Patient Participation in Cancer Clinical Trials. *A National Cancer Clinical Trials System for the 21st Century: Reinvigorating the NCI Cooperative Group Program*. National Academies Press (US); 2010.
7. Chen S, Kann BH, Foote MB, et al. Use of artificial intelligence Chatbots for cancer treatment information. *JAMA Oncol*. 2023;9: 1459-1462. https://doi.org/10.1001/jamaoncol.2023.2954
8. Chen S, Guevara M, Moningi S, et al. The effect of using a large language model to respond to patient messages. *Lancet Digital Health*. 2024;6:e379-e381.
9. Holstead RG. Utility of large language models to produce a patient-friendly summary from oncology consultations. *J Clin Oncol Oncol Pract*. 2024;20:1157-1159.
10. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*, 2023;29:721-732.
11. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *J Transl Med*, 2023;21:269.
12. Joshi V, Kulkarni AA. Public awareness of clinical trials: a qualitative pilot study in Pune. *Perspect Clin Res*. 2012;3:125-132.
13. Jin Q, Wang Z, Floudas CS, et al. Matching patients to clinical trials with large language models. *Nat Commun*. 2024;15:9074. https://doi.org/10.1038/s41467-024-53081-z
14. Hadden KB, Prince LY, Moore TD, James LP, Holland JR, Trudeau CR. Improving readability of informed consents for research at an academic medical institution. *J Clin Transl Sci*. 2017;1: 361-365. https://doi.org/10.1017/cts.2017.312
15. O'Sullivan L, Sukumar P, Crowley R, McAuliffe E, Doran P. Readability and understandability of clinical research patient information leaflets and consent forms in Ireland and the UK: a retrospective quantitative analysis. *BMJ Open*. 2020;10:e037994. https://doi.org/10.1136/bmjopen-2020-037994.
16. Joffe S, Cook EF, Cleary PD, Clark JW, Weeks JC. Quality of informed consent: A new measure of understanding among research subjects. *J Natl Cancer Inst*. 2001;93:139-147. https://doi.org/10.1093/jnci/93.2.139.
17. N. L. of Medicine. ClinicalTrials.gov API. 2023. Accessed November 23, 2023. https://clinicaltrials.gov/data-api/api
18. Artifex. PyMuPDF 1.24.9 documentation–pymupdf.readthedocs.io. 2023. Accessed November 23, 2023. https://pymupdf.readthedocs.io/en/latest/
19. Lecompte LL, Young SJ. Revised common rule changes to the consent process and consent form. *Ochsner J*. 2020;20:62-75.
20. ClinicalTrials.gov Identifier: NCT03923790. (n.d.). A Study to Evaluate the Safety and Efficacy of [Study Intervention/Drug Name, if applicable]. Retrieved from https://clinicaltrials.gov/study/NCT03923790?id=NCT03923790&rank=1
21. Rahimli Ocakoglu S, Coskun B. The emerging role of AI in patient education: a comparative analysis of LLM accuracy for pelvic organ prolapse. *Med Princ Pract*. 2024;33:330-337. https://doi.org/10.1159/000538538
22. Lambert R, Choo Z-Y, Gradwohl K, et al. Assessing the application of large language models in generating dermatologic patient education materials according to reading level: qualitative study. *JMIR Dermatol*. 2024;7:e55898.
23. Van Veen D, Van Uden C, Blankemeier L, et al.]Clinical text summarization: adapting large language models can outperform human experts. *Res Sq*. 2023;rs.3.rs-3483777. https://doi.org/10.21203/rs.3.rs-3483777/v1. Update in: *Nat Med*. 2024;30:1134-1142. https://doi.org/10.1038/s41591-024-02855-5.
24. Tariq A, Urooj A, Trivedi S, et al. Patient centric summarization of radiology findings using large language models. *medRxiv*, 2024: 2024.02. 01.24302145.
25. Lyu M, Peng C, Li X, et al. Automatic summarization of doctor-patient encounter dialogues using large language model through prompt tuning. arXiv, arXiv:2403.13089, 2024, preprint: not peer reviewed.
26. Mathur Y, Rangreji S, Kapoor R, et al. Summqa at mediqa-chat 2023: In-context learning with gpt-4 for medical summarization. In: *Proceedings of the 5th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2023:490–502.
27. Ghosh A, Tomar M, Tiwari A, Saha S, Salve J, Sinha S. From sights to insights: towards summarization of multimodal clinical documents. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics; 2024:13117-13129.
28. Liu Y, Ju S, Wang J. Exploring the potential of ChatGPT in medical dialogue summarization: A study on consistency with human preferences. *BMC Med Inform Decis Making*. 2024;24:75.
29. White R, Peng T, Sripitak P, et al. CliniDigest: A case study in large language model based large-scale summarization of clinical trial descriptions. In: *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*. 2023:396-402.
30. Phatak A, Savage DW, Ohle R, Smith J, Mago V. Medical text simplification using reinforcement learning (TESLEA): deep learning–based text simplification approach. *JMIR Med Inform*. 2022; 10:e38095.
31. JeblickK, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: An exploratory case study on simplified radiology reports. arXiv [cs.CL] 2022. Accessed January 3, 2024. http://arxiv.org/abs/2212.14882
32. BioLaySumm 2023 Shared Task: Lay Summarisation of Biomedical Research Articles. The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. Toronto, Canada: Association for Computational Linguistics2023. p. 468-477.
33. 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:e2336997. https://doi.org/10.1001/jamanetworkopen.2023.36997
34. Ali R, Connolly ID, Tang OY, et al. Bridging the literacy gap for surgical consents: An AI-human expert collaborative approach. *NPJ Digit Med*. 2024;7:63. https://doi.org/10.1038/s41746-024-01039-2

35. Vaira LA, Lechien JR, Maniaci A, et al. Evaluating AI-generated informed consent documents in oral surgery: a comparative study of ChatGPT-4, Bard Gemini advanced, and human-written consents. *J Craniomaxillofac Surg*. 2025;53:18-23. https://doi.org/10.1016/j.jcms.2024.10.002
36. Mirza FN, Tang OY, Connolly ID, et al. Using ChatGPT to facilitate truly informed medical consent. *Nejm AI*. 2024;1:AIcs2300145.
37. Hereu P, Pérez E, Fuentes I, Vidal X, Suñé P, Arnau JM. Consent in clinical trials: what do patients know? *Contemp Clin Trials*. 2010;31:443-446.
38. Juan-Salvadores P, Michel Gómez MS, Jiménez Díaz VA, Martínez Reglero C, Iñiguez Romo A. Patients' knowledge about their involvement in clinical trials. A non-randomized controlled trial. *Front Med (Lausanne)*. 2022;9:993086.
39. Pietrzykowski T, Smilowska K. The reality of informed consent: Empirical studies on patient comprehension-systematic review. *Trials*. 2021;22:57.
40. Peterson JS, Plana D, Bitterman DS, Johnson SB, Aerts HJWL, Kann BH. Growth in eligibility criteria content and failure to accrue among National Cancer Institute (NCI)-affiliated clinical trials. *Cancer Med*. 2023;12:4715-4724.
41. Unger JM, Vaidya R, Hershman DL, Minasian LM, Fleury ME. Systematic review and meta-analysis of the magnitude of structural, clinical, and physician and patient barriers to cancer clinical trial participation. *J Natl Cancer Inst*. 2019;111:245-255.
42. Rimel BJ. Clinical trial accrual: obstacles and opportunities. *Front Oncol*. 2016;6:103.
43. Malmqvist E, Juth N, Lynöe N, Helgesson G. Early stopping of clinical trials: Charting the ethical terrain. *Kennedy Inst Ethics J*; 2011;21:51-78.
44. Schwartz AL, Alsan M, Morris AA, Halpern SD. Why diverse clinical trial participation matters. *N Engl J Med*. 2023;388:1252-1254.
45. Clark LT, Watkins L, Piña IL, et al. Increasing diversity in clinical trials: overcoming critical barriers. *Curr Probl Cardiol* 2019;44:148-172.
46. National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Population Health and Public Health Practice, Committee on Community-Based Solutions to Promote Health Equity in the United States. *Communities in Action: Pathways to Health Equity*. National Academies Press; 2017.
47. Farb A, Viviano CJ, Tarver ME. Diversity in clinical trial enrollment and reporting-where we are and the road ahead. *JAMA Cardiol*. 2023;8:803-805.
48. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. *JAMA*. 2004;291:2720-2726.
49. Mazzochi AT, Dennis M, Chun HYY. Electronic informed consent: Effects on enrolment, practical and economic benefits, challenges, and drawbacks—a systematic review of studies within randomized controlled trials. *Trials*. 2023;24:127. https://pubmed.ncbi.nlm.nih.gov/36810093/
50. Golembiewski EH, Mainous AGI, Rahmanian KP, et al. An electronic tool to support patient-centered broad consent: a multi-arm randomized clinical trial in family medicine. *Ann Fam Med* 2021;19:16-23.
51. De Sutter E, Borry P, Geerts D, Huys I. Personalized and long-term electronic informed consent in clinical research: stakeholder views. *BMC Med Ethics* 2021;22:108. https://pubmed.ncbi.nlm.nih.gov/34332572/
52. Lajonchere C, Naeim A, Dry S, et al. An integrated, scalable, electronic video consent process to power precision health research: large, population-based, cohort implementation and scalability study. *J Med Internet Res* 2021;23:e31121.
53. Synnot A, Ryan R, Prictor M, Fetherstonhaugh D, Parker B. Audio-visual presentation of information for informed consent for participation in clinical trials. *Cochrane Libr*. 2014;2014:CD003717. https://pubmed.ncbi.nlm.nih.gov/24809816/
54. Kim YJ, DeLisa JA, Chung YC, et al. Recruitment in a research study via chatbot versus telephone outreach: a randomized trial at a minority-serving institution. *J Am Med Inform Assoc*, 2021;29:149-154.
55. Savage SK, LoTempio J, Smith ED, et al. Using a chat-based informed consent tool in large-scale genomic research. *J Am Med Inform Assoc* 2024;31:472-478.
56. Goodman KE, Yi PH, Morgan DJ. AI-generated clinical summaries require more than accuracy. *JAMA*. 2024;331:637-638. https://doi.org/10.1001/jama.2024.0555
57. Bitterman DS, Aerts HJWL, Mak RH. Approaching autonomy in medical artificial intelligence. *Lancet Digit Health*. 2020;2:e447-e449. https://doi.org/10.1016/S2589-7500(20)30187-4.
58. Kim J, Chen ML, Rezaei SJ, et al. Perspectives on artificial intelligence–generated responses to patient messages. *JAMA Netw Open*. 2024;7:e2438535. https://doi.org/10.1001/jamanetworkopen.2024.38535
59. Mannhardt N, Bondi-Kelly E, Lam B, et al. Impact of large language model assistance on patients reading clinical notes: a mixed-methods study. arXiv, arXiv:2401.09637, 2024, preprint: not peer reviewed.
60. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit Med*. 2024;7:258. https://doi.org/10.1038/s41746-024-01258-7

JNCI Cancer Spectrum, 2025, 9, 1–10
https://doi.org/10.1093/jncics/pkaf021
Article