Study Protocol

ChatGPT-generated versus expert-written answers to frequently asked questions about diabetes: an e-survey among all employees of a Danish diabetes center

Adam Hulman[1,2]*, Ole Lindgård Dollerup[1,2], Jesper Friis Mortensen[1,2], Kasper Norman[1], Henrik Støvring[1,3], Troels Krarup Hansen[1,2]

[1]Steno Diabetes Center Aarhus, Aarhus, Denmark; [2]Aarhus University, Aarhus, Denmark; [3]University of Southern Denmark, Odense, Denmark

*corresponding author (adahul@rm.dk)

Background

Towards the end of 2022, artificial intelligence (AI) chatbots were featured regularly in mainstream media demonstrating impressive abilities. ChatGPT (OpenAI, San Francisco, US) has played an important role in this by providing a user-friendly web-interface [chat.openai.com] that was launched on 30th of November. Up to now, more than a million people have used ChatGPT, a state-of-the-art conversational language model.

Several recent studies have shown that large language models encode clinical knowledge [Liévin et al. arXiv:2207.08143, Singhal et al. arXiv:2212.13138]. These studies used benchmark datasets including questions and answers from the medical domain and tested whether language models can answer the questions correctly. Some online surveys have been conducted on whether people can distinguish AI-generated images from real ones, but these have been mostly published on data science blogs, news websites, and were not conducted, evaluated and published with scientific rigor.

Our research interest is how a chatbot performs in answering frequently asked questions by patients in a diabetes clinic, more specifically whether AI-generated answers can be distinguished from answers written by human experts.

Aim & hypothesis

The aim of the study was to investigate ChatGPT's knowledge in the diabetes domain, more specifically in response to potential questions asked by patients about the disease, medication, diet and physical activity.

We hypothesized that participants (employees at a regional diabetes center), who have at least some and up to expert knowledge about diabetes, will not be able to distinguish between answers written by humans and generated by AI in response to diabetes-related questions. Our secondary hypothesis is that people with contact to patients as caregivers and those who previously tried ChatGPT might be better at identifying answers generated by AI.

Design

We followed the CHERRIES checklist [Eysenbach G, JMIR 2004;6:e34] when developing the protocol of the study.

Population

The target population was people who have basic knowledge about diabetes or exposure to the topic at the workplace. Therefore, all employees of Steno Diabetes Center Aarhus (full or part-time), including healthcare professionals (medical doctors, nurses, dieticians, etc.), researchers (also those affiliated to

Steno Diabetes Center Aarhus, but with another institute as primary affiliation), and administrative staff, formed the potential participants for the study. We identified 267 unique e-mail addresses where the invitations were sent.

Institutional Review Board approval

The study was registered in the database of research projects in the Central Denmark Region. Further ethical approval was not necessary in Denmark as the study only includes survey-based data collection.

Informed consent

Participants were informed on the opening page of the survey that they are participating in a research study and by submitting their answers, they agree with contributing data to the study.

Data protection

The questions did not include any sensitive information. Information on age was collected in broad categories (<30, 30-39, 40-49, 50<), so that participants cannot be identified when combining data with the other variables on participant characteristics (sex, patient contact as caregiver, ChatGPT use).

The survey was developed and distributed in the SurveyXact system (Rambøll, Copenhagen, Denmark) that adheres to GDPR regulations. After data collection, the data was exported and securely stored in the MidtX system (Alfresco Inc), which is specifically designed to store sensitive personal data in the Central Denmark Region. AH had access to the dataset.

Survey development & testing

We aimed to define a set of questions, with two answers each: (1) from a reliable source, written by a human expert, (2) generated by ChatGPT. The questions aimed to represent relevant diabetes-specific questions from patients, therefore the main inspirational resource was the 'Frequently Asked Questions' (FAQ) page of the Danish Diabetes Association. We did not consider questions that were only meaningful in the Danish context (e.g. asking for information on reimbursement; answers including links, phone numbers). Also, many of the questions were related to diet. To cover more diabetes-related topics, we created further questions, mostly based on the website of the Knowledge Center for Diabetes.

If a question was identified in among the FAQ, the corresponding answers were included as the human answers to the questions. Otherwise, we identified paragraphs that answered well-defined questions, that we formulated ourselves, and kept the text of the paragraph as close to the original as possible.

All questions are listed in the appendix including the source. Answers were shortened if necessary to keep their lengths between 45 and 65 words. These thresholds were chosen for practical reasons i.e. not to make the questionnaire too time consuming (<10mins), but to have enough complexity to make it at all feasible to identify AI-generated answers. The questions and answers were developed by OD & JFM (both with active clinical work and contact to patients), with comments and final edits by AH.

The AI-generated answers were generated using ChatGPT by OpenAI (version released on 09/01/23). Each answer was generated in a new chat window to avoid leakage of context between questions and answers. To give context to the questions, we used few-shot learning [Brown et al, NIPS 2020] when creating the prompts i.e. we described the context in the prompt and gave examples of the question-answer pairs before asking the questions (see below the actual prompt used). The method is a good way to guide the system about the format (length, sentences instead of lists), language (scientific, lay, etc.) of the answers. If

an answer exceeded this word limit, we evaluated whether any sentences (preferably the first or last) can be deleted while still answering the same question. If ChatGPT suggests to 'contact your doctor' or similar, that part will be removed, unless the human-written answer also has a similar component. Also, if ChatGPT refused to answer a question, we tried to modify the context information in the prompt. These alterations will be reported. In case we cannot get an answer, the question will be replaced with one of those used in the prompt as examples, and this information will be reported in the article. If the answer includes serious incorrectness, that will be removed to avoid the spread of misinformation, as the participants will not receive the results right after filling out the survey. This will be reported. Grammatical mistakes will be corrected and reported. The answers were generated after the list of questions and human answers were finalized.

The survey was pilot tested with three of the study collaborators (HS, KN, TKH) who did not participate in the development of the questions. Based on this, considering the time used for filling out the questionnaire, and the precision and power simulations described below, we decided to present ten questions.

We developed 13 questions including the following topics (number of questions): diabetes pathophysiology (2), insulin (2), complications (3), diet (3), physical activity (3). Ten were used in the survey, and three for inclusion in the prompt (more info about few-shot learning in the next paragraph). These three were randomly chosen from the topics that included three questions.

Distribution of the survey

Our study included a closed survey. Only employees of Steno Diabetes Center Aarhus (including part-time) were invited to participate. As the language of the questionnaire was Danish, good having Danish language skills was a requirement to participate. This information was stated when contacting employees. Participants were invited by e-mail including person-specific links to the survey. To advertise the survey and increase the participation rate, the CEO of Steno Diabetes Center Aarhus (TKH) sent out the e-mail inviting participants to fill out the survey.

Survey administration

The responses were collected using a web-based questionnaire created in SurveyXact. The opening page included the context and purpose of the survey, and practical information. Participation in the survey was voluntary. There were no incentives offered to participants. Data is planned to be collected for five working days from 23/01/23 to 27/01/23.

The order of the questions was randomized (except for the participants' basic information which were always the first three questions: job/role, sex, age), but all participants got the questions in this same order. The order of the answers was randomized at the individual level. The survey did not include adaptive questioning.

Each AI-related question and the corresponding two answers were presented together on a separate page. Participants had to answer the question to move to the next question, without an option to go back to a previous question and review or change the answer. There was no possibility to review the answers at the end. To complete the survey answers, all questions had to be answered.

Each participant received an individual specific link that allowed them to open and fill out the survey only once. This information was included in the e-mail used to distribute the link.

Response rates

We anticipate a 50% response rate (~135 responders).

Statistical methods

The study is designed as a non-inferiority trial where the aim is to show that survey participants are not able to pick the AI generated answer more often than 50%. In an exploratory simulation study, we investigated the magnitude of an inferiority margin that could be rejected with 80% or 90% power for given study designs defined by number of survey participants (n=100, 150, 200), number of questions to assess by each participant (k=10, 15) and variation in individual ability to identify the AI generated answer. The variation was modelled as individual probabilities of identifying AI generated answers following a normal distribution with mean of 50% and an SD of 1% or 5%, i.e. 95% prediction intervals for individual probabilities either being (48%; 52%) or (40%; 60%). To allow for the dependence of the multiple answers from the same participant, we used logistic regression with robust variance estimation with participant as cluster to analyze each generated dataset.

We found that with 100 participants and 10 questions, a non-inferiority margin of 54.6% would provide 80% power and 55.4% would provide 90% (Table 1). Margins were closer to 50% for all other investigated scenarios and were with 90% power below 56% for all scenarios with 100 participants. Margins were virtually identical when varying the SD governing between person variation. A margin of 55% with 90% power has the interpretation that the study can rule out larger deviations than 5% from a fair coin flip, 90% of the times such a study is conducted. Based on these simulations, we chose a 55% (0.55 probability) non-inferiority margin, which we also consider a reasonable margin from a scientific perspective. In the analysis of the study, we will use the same analytic approach as in the simulations (logistic regression with robust variance estimation).

Table 1 *Non-inferiority margin estimation in different simulation scenarios (n=number of participants; k=number of questions; SD=between individual variation in identifying the correct answer). Results correspond to 80% | 90% power.*

	k=10		k=15	
n	SD=0.01	SD=0.05	SD=0.01	SD=0.05
100	0.544 | 0.551	0.546 | 0.554	0.536 | 0.541	0.538 | 0.545
150	0.536 | 0.542	0.537 | 0.544	0.529 | 0.534	0.531 | 0.537
200	0.531 | 0.536	0.533 | 0.538	0.526 | 0.530	0.528 | 0.532

In a secondary analysis, we will investigate the effect of age, sex, contact with patients (expected ~50% of participants) and previous use of ChatGPT (expected ~10% of participants) in univariable analyses. We expect that patient contact and previous use of ChatGPT will be associated with a higher probability of identifying the AI-generated answers correctly, while age and sex will not. Our simulations show that we will be able to show a difference with 90% power and 0.05 alpha if it is at least 9% (>0.59 vs 0.50) for patient contact and at least 15% (>0.65 vs. 0.50) for previous ChatGPT use (Table 2). We will also compare group-wise estimates to the non-inferiority margin, although we are most likely underpowered to test this aspect. These analyses might give relevant insights, especially if ChatGPT will be found inferior. E.g. the association between previous ChatGPT use and a higher probability of identifying the correct answer might suggest that the style of the language, and not necessarily the actual content is a driver behind an inferior result.

Table 2 *Power simulation for different levels of difference in probabilities of identifying the correct answer. n_1 & n_2 represent the number of participants in non-exposed & exposed groups, respectively. We assumed n=150 participants, k=10 questions and a probability of 0.50 in the non-exposed group.*

	Probability of identifying the correct answer in the exposed group (n_2)						
n_1-n_2	0.54	0.56	0.58	0.60	0.62	0.64	0.66
75-75	31.9%	60.9%	85.2%	96.4%	99.7%	100%	100%
105-45	26.6%	52.9%	76.7%	92.5%	98.4%	99.8%	99.9%
135-15	16.8%	30.0%	45.1%	62.4%	77.9%	88.3%	95.3%

To explore heterogeneity between participants' ability to identify AI generated answers we will use a logistic regression model with a random effect for participant. In another exploratory analysis, we will estimate the probability of participants picking the AI generated answers correctly for each question.

Appendix

ChatGPT Prompt

Du arbejder som rådgiver i en diabetes klinik. Besvar patienternes spørgsmål om diabetes og hvordan diabetes påvirker deres liv. Lav svar mellem 45 og 65 ord.

Spørgsmål: Har det noget med min type 2 diabetes at gøre, at jeg er træt?

Svar: Man kan godt bliver træt, hvis blodsukkeret hele tiden ligger højt. Er det tilfældet, skal din behandling måske ændres. Halvdelen af alle dem, der har type 2 diabetes, skal med tiden have insulin. Det kan modvirke din træthed, hvis trætheden skyldes for højt blodsukker. Din egen læge kan måle dit langtidsblodsukker og se, om det er tilfældet.

Spørgsmål: Hvilken type træning er bedst, når jeg har type 2 diabetes?

Svar: Det anbefales at udføre en kombination af konditionstræning og styrketræning, da begge dele virker positivt på insulinfølsomheden og den generelle sundhed. Konditionstræning belaster dit kredsløb og hjerte, hvilket forbedrer din udholdenhed og kondition. Styrketræning belaster dit nervesystem og muskler, hvilket forbedrer din muskelstyrke. Spørg din læge eller træningsfysiolog til råds for individuelle retningslinjer.

Spørgsmål: Skal jeg holde mig fra fuldkorn, når jeg har diabetisk gastroparese?

Svar: Ja. Gastroparese er en delvis lammelse af mavesækken og forsinker tømningen af maveindhold til tarmen. Kostfibre og hele kerner, som vi får i større mængder fra fuldkornsbrød, bliver ikke nedbrudt i maven, og vil derfor give en yderligere forsinket mavetømning, der kan give ubehag som kvalme og opkast.

Spørgsmål: [insert question here]

Svar:

Questions (with source for human answer)

1. Hvor meget frugt må jeg spise, når jeg har diabetes? [diabetes.dk]
2. Skal jeg justere min insulinbehandling, når jeg er syg med feber? [diabetes.dk]
3. Hvordan opbevarer jeg insulin på en lang rejse? [diabetes.dk]
4. Skal jeg være bekymret for mine fødder, når jeg har diabetes? [diabetes.dk]
5. Hvorfor er mine blodsukre høje? [diabetes.dk]
6. Hvordan påvirker motion blodsukkeret når man har type 1 diabetes? [videncenterfordiabetes.dk]
7. Kan light-sodavand få mit blodsukker til at stige og påvirke min diabetes? [diabetes.dk]
8. Kan diabetes påvirke sexlivet? [diabetes.dk]
9. Hvordan påvirker forskellige former for træning typisk blodsukkeret hos personer med type 1 diabetes? [consensus statement on physical activity & type 1 diabetes]
10. Hvad er graviditetsdiabetes? [diabetes.dk]