Simulating Diverse Patient Populations Using Patient Vignettes and Large Language Models

Kerstin Denecke and Daniel Reichenpfader
Institute for Patient-centered Digital Health
Bern University of Applied Sciences, Biel, Switzerland
{daniel.reichenpfader, kerstin.denecke}@bfh.ch

Abstract

Ensuring equitable access to digital therapeutics (DTx) is essential to avoid healthcare inequalities in an era of increasing digitization. This requires DTx to be tested with users from diverse populations, which is often not realistic due to time and resource constraints. In this paper, we propose the use of large language models (LLMs) to simulate diverse patients. Specifically, we manually create a patient vignette that characterizes a specific population group. Variations of this vignette are used for role-prompting a commercial LLM, GPT-4, instructing the LLM to take on the role described in the patient vignette and act accordingly. We investigate if the LLM stays in its given role. To do this, we simulate a medical anamnesis interview with the role-prompted LLM and analyze its responses for compliance, coherence, correctness, containment, and clarification. Our results show that GPT-4 generates compliant, coherent and clinically valid responses, including information that is not explicitly stated in the provided patient vignette.

Keywords: Large Language Models, Inclusive Design, Accessibility, Patient Vignettes, Simulation

1. Introduction

Digital therapeutics (DTx) promise to transform patient care and outcomes (Dang et al., 2020). As these digital interventions become more widespread, it is important to ensure that their design is inclusive and accessible to diverse user groups (Rivera-Romero et al., 2022). The principle of inclusivity not only enhances the usability of DTx across different demographic groups, but also underpins the effectiveness and equity of DTx. However, testing a DTx with a broad spectrum of patients is not only time consuming, but also requires significant financial resources, limiting the scope and frequency of these essential evaluations. Furthermore, the recruitment process is inherently susceptible to selection bias, skewing the sample and potentially missing critical user needs and preferences which undermines the goal of inclusive design. Beyond, the participation of vulnerable groups often requires adaptations in the testing procedure (Peute et al., 2022).

Given these limitations, there is a need for innovative methods that can simulate a wide range of patient populations. Specifically, this article aims to explore the potential of Large Language Models (LLMs) as a tool for simulating various user groups based on patient vignettes. If LLMs are a reliable method to simulate patient populations, they could contribute to the development of more inclusive and effective DTx relying on verbal communication, such as chatbots or conversational agents.

A vignette is a short, carefully written description of a person or situation (Schoenberg and Ravdal, 2000). They are a useful tool for health education, evaluating health professionals, conducting health research (Evans et al., 2015), and evaluating symptom checkers (Ben-Shabat et al., 2022). Benoit already investigated the ability of LLMs to generate and rewrite vignettes (Benoit, 2023). In contrast to their work, we are not interested in developing a text vignette using an LLM, but in using an LLM to simulate the patient characterized by a vignette. Campillos-Llanos et al. already created a system that simulates patients, but it is based on terminology-rich resources instead of LLMs (Campillos-Llanos et al., 2020). We assume that LLMs might have the potential to simplify the development of such a system, having recently demonstrated human-level performance on various tasks, e.g. for medical question answering (Singhal et al., 2023) or for provision of medical information (Cocci et al., 2024). LLMs can be instructed to follow a certain role (Kong et al., 2023) such as the role of a teacher, physician etc. This approach to role-prompting will be used in this paper. Specifically, we will consider the following research questions:

- Which aspects are needed to accurately simulate a patient?
- Do LLMs stay in the role defined by a vignette and answer accordingly? Do LLMs instructed to follow a role provide meaningful information that is not explicitly contained in the patient vignette?

This paper reports on the methodology and validation results based on a single patient vignette.

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 20–25

2. Methods

In order to answer the above-mentioned research questions, we follow a 5-step process, see Figure 1. First, relevant aspects needed for creating patient vignettes are collected based on a selective literature review. These aspects comprise, e.g., demographic data, past medical history and current symptoms or medical problems. Second, an example for each aspect of a patient vignette is drafted.

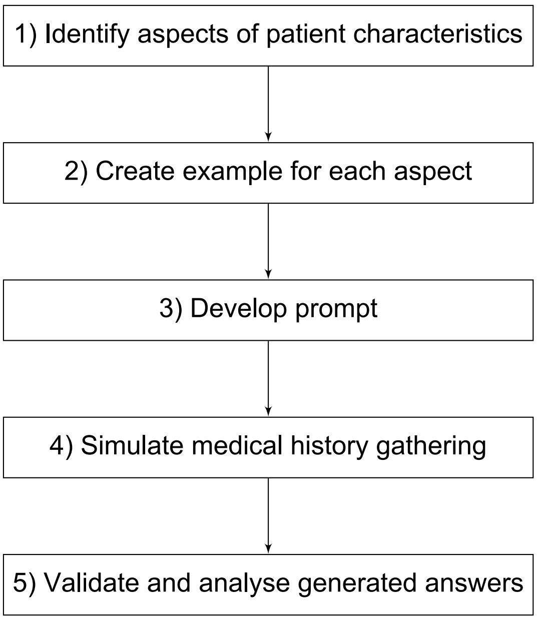
1) Identify aspects of patient characteristics
2) Create example for each aspect
3) Develop prompt
4) Simulate medical history gathering
5) Validate and analyse generated answers

Figure 1: Methodology

Next, a prompt template is developed that instructs the LLM used for validation to impersonate the personality described within the patient vignette. The prompt development process is based on the work of Wang et al. (Wang et al., 2024) and the recommendations of OpenAI (OpenAI, 2024).

For the validation process, GPT-4 by OpenAI is used due to its accessibility and unprecedented performance. Reproducibility is ensured by implementing a Python application that executes the validation. Using the patient vignette and the prompt template from the previous steps, we will simulate a medical anamnesis interview between a physician and a patient: For this simulation, the patient is impersonated by the LLM while the medical history questions are manually entered into the system. For each interaction, the previous interactions are appended to a conversation history and included in the model input. For conducting the medical interview, we define a script based on the recommendations to conduct a medical history interview suggested by Füeßl et Middeke (Füeßl and Middeke, 2022).

We carry out two ablation studies, removing specific parts of the patient vignette (e.g. secondary information), in order to investigate whether the LLM is able to infer information which is not explicitly stated in the vignette. After completion of the interview, each turn of conversation is assessed according to the following assessment categories. The assessment is performed by both authors. Additionally, a qualitative analysis of generated responses is carried out.

- Compliance: The model output complies with the defined patient vignette.
- Coherence: The model output coheres with previous outputs.
- Correctness: The model output is clinically meaningful and realistic.
- Containment: The model output is explicitly contained in the patient vignette.
- Clarification: The model output contains a question asking for clarification.

3. Results

We make all results as well as the source code publicly available as a Git repository via Zenodo (doi:10.5281/zenodo.10889465). The total costs for prompt template development and the validation of three variants amounted to USD 3,58.

3.1. Patient Vignette Development

Based on six sources, we identified 16 dimensions to be included in a patient vignette, see Table 1. We distinguish two categories of information: Primary information is directly asked by the health professional. Secondary information is usually not asked directly, but might have a major impact on communication: For example, Clack et al. investigate personality differences between clinicians and patients and their implications on the patient-clinician relationship. Their findings indicate that different types of personality can cause miscommunication during the consultation process (Clack et al., 2004). Redelmeier et al. review the OCEAN taxonomy, an evidence-based model to understand personalities, and state that spontaneous impressions formed by clinicians could induce incorrect clinical judgements (Redelmeier et al., 2021). Pérez-Stalbe and El-Toukhy identify factors associated with poor patient-clinician communication (Pérez-Stable and El-Toukhy, 2018). Bartz et al. review the role of factors related to sex and gender in healthcare (Bartz et al., 2020). Chipidza et al. give recommendations on how to evaluate and treat angry patients (Chipidza et al., 2016).

Dimension	Source
Primary information	
Current symptoms Past medical history Current medication Triggering factors Psychosocial aspects Family anamnesis Occupational anamnesis	(Füeßl and Middeke, 2022)
Secondary information	
Personality traits	(Redelmeier et al., 2021; Clack et al., 2004)
Communication style	(Clack et al., 2004)
Health literacy	(Pérez-Stable and El-Toukhy, 2018)
Race, geographic location and country of origin	(Pérez-Stable and El-Toukhy, 2018)
Sex and gender	(Bartz et al., 2020)
Emotion	(Chipidza et al., 2016)
Language proficiency	(Pérez-Stable and El-Toukhy, 2018)
Digital literacy	(Pérez-Stable and El-Toukhy, 2018)
Socioeconomic status	(Pérez-Stable and El-Toukhy, 2018)

Table 1: Dimensions of information contained in a patient vignette

Below, we show excerpts from the developed patient vignette for the dimensions *current symptoms*, *past medical history*, *health literacy* and *emotional state*. For the complete vignette, we refer to the Git repository.

- Current symptoms: *You have a headache and a fever*
- Past medical history: *You have a history of migraines*
- Health literacy: *You are very knowledgeable about your condition*
- Emotional state: *You are feeling anxious and depressed*

The following task prompt was developed iteratively and additionally self-improved by asking GPT-4 for optimisation: *Imagine that you are in the shoes of a patient during a medical consultation. You are about to engage in a detailed conversation with a healthcare provider who is taking your medical history, also known as an anamnesis. Below, you will find specific information about your health, lifestyle, and medical background. Use this information to respond accurately and thoughtfully to the healthcare provider's inquiries. Remember, your role is to embody the patient's experience, drawing from the details provided. Your responses should reflect the depth and nuances of the concerns, experiences, and medical history of a real patient.* ***<Dimensions are inserted here>****. As the consultation wraps up, remember to stay true to the character and information you have been given. If the healthcare provider asks for details not explicitly mentioned, use your imagination to provide realistic and considerate answers that align with the character's background and current health scenario. Should any question seem unclear or unfamiliar based on your role as the patient, don't hesitate to ask for further clarification, just as a real patient might seek to understand their healthcare provider's inquiries fully.*

3.2. Anamnesis Simulation and Validation of Role-prompted LLM

The anamnesis simulation consisted of eleven questions posed to the LLM in total, see below:

1. Tell me more about your symptoms.
2. Can you give me more details regarding your headache?
3. Tell me more regarding its localization and spread.
4. Tell me more about its quality.
5. Tell me more about its severity.
6. Are you currently taking any medication?
7. Have you noticed any factors that trigger your symptoms?
8. Do you currently face difficult situations in your life?
9. Are there any diseases that run in your family?
10. What is your occupation?
11. Are you taking the pill?

In total, three variations of the patient vignette were investigated: As baseline, the complete patient vignette was used. For the first ablation study, we only kept primary information according to Table 1 and removed all secondary information. For the second ablation study, only current symptoms were retained as primary information and all secondary information was retained. The results of all three variants are summarised in Table 2. Across all three simulation variants, GPT-4 generated answers that complied with each vignette,

Variant	Compliance	Coherence	Correctness	Containment	Clarification	Average word count of model answers
Baseline	100 %	100 %	100 %	64 % (n=7/11)	0 %	53
Ablation 1	100 %	100 %	100 %	45 % (n=5/11)	0 %	41
Ablation 2	100 %	100 %	100 %	9 % (n=1/11)	0 %	78

Table 2: Validation results: For each variant, the same eleven questions were posed to the role-prompted model.

that cohered with previous answers given and that were clinically meaningful and realistic. It becomes apparent that the model makes up large proportions of its output, realistically adding information to the provided vignette. This effect is strongest in ablation 2; ten of the eleven generated responses contained information that was not included in the patient vignette. The model did not ask for clarification. Interestingly, the model used a scale from one to ten to answer the question about the intensity of symptoms. We can also see that the model tends to negate specific questions, e.g. regarding the use of oral contraception, instead of making up an answer. For example, in ablation 1 the model gave the following answer: *No, I am not currently taking any form of contraceptive pill. Other than the Tylenol for my headaches, I'm not on any other medication.* However, in case the model adds information to the provided vignette, it shows coherence when doing so: For example, the model mentioned the use of ibuprofen twice during the anamnesis simulation (ablation 2). On average, GPT-4 generates the longest answers with only minimal primary information (ablation 2) and the shortest answers when omitting secondary information (ablation 1).

4. Discussion and Outlook

In this paper, we show a first approach to simulating various patient populations based on manually drafted patient vignettes. We identified 16 dimensions to be included in a patient vignette. GPT-4 generates compliant, coherent and clinically valid responses and succeeds in adding additional information not contained in the patient vignette.

The role-prompted LLM comprehensively described the headache that was mentioned as symptom in the vignette. While the vignette only contained the term ”headache” and, in case the medical history was included, the term ”migraine”, the description of pain was very detailed, even including a rating of the pain on a scale. In this sense, we can conclude that the LLM acted well in its defined role. However, it remains open to study whether these extensions and elaborations of the symptoms are biased or follow certain stereotypes. Furthermore, it is still unknown whether more complicated vignettes reflecting complex clinical cases can still be accurately simulated.

Furthermore, it is interesting that the generated answers are longer when less primary information is provided in the vignette. Thus, when the LLM lacks a clear guidance, it fills the gaps as requested in our prompt, but with a higher risk of losing its role and adding information that does not fit accordingly. In none of the three variants, the LLM asked for clarification, although the prompt suggested this. A reason might be that the questions for medical history taking were pretty simple. However, other researchers have already found that LLMs are unable to ask for clarification and, therefore, to play a proactive role (Deng et al., 2023).

This paper reports work in progress and thus has some limitations: We conducted this study with only one patient vignette that was created by a medical informatician without clinical validation. Similarly, the assessment of generated answers was carried out by both authors who have a background in medical informatics, but no medical training. The literature considered for identifying the aspects considered in the vignette was collected in a selective non-systematic literature research and did not use a consensus-based approach. In future work, when developing more vignettes, we will follow the recommendations for vignette content provided by Evans et al. (Evans et al., 2015). Instead of inventing patient histories, synthetic patient data could be used (Guillaudeux et al., 2023). Furthermore, our approach is based on GPT-4, a commercial LLM. Future research might focus on investigating whether similar results can be achieved with open source LLMs such as BioMistral, a set of LLMs based on Mistral being further pre-trained on texts from PubMed Central (Labrak et al., 2024).

We highlight additional open research topics: LLMs might deny impersonating specific patient vignettes due to the practice of model alignment, where undesired or harmful behaviour is reduced during the training process. Also, mimicking certain personality features might be impossible (e.g. becoming aggressive). In this way, the approach will fail to properly simulate a patient. Furthermore, the patient vignette used for the three variants was rather short. It must be noted that the length of the vignette as well as the simulated conversation are directly proportional to model cost. This is because the costs of the commercial model are calculated on the basis of the input and output length. For each request, the entire conversation history is attached as model input, accumulating over time.

We envision two use cases for the application of the methodology tested in this paper, including educational purposes and evaluation of DTx. Simulations are used to train health professionals to act appropriately in critical situations or, generally, in patient interactions. A frequently chosen approach is to hire actors who simulate patients. With our approach, patients could be simulated by a role-prompted LLM, augmented by text-to-speech generation. The interaction could take place between the LLM and the health professional in training. For such a use case, it is less important that all the information provided is correct in a clinical sense (patients might also be inconsistent in their statements). It is more important that the main characteristics of the role are maintained, i.e. the health literacy level or cognitive abilities. Our evaluation corresponds to the general principles of simulation-based learning (Herold-Majumdar et al., 2023). In these settings, the simulated interaction takes place and is analysed afterwards. It still has to be assessed whether our approach is effective for such educational purposes.

Another potential application area is using the role-prompting-based simulation to evaluate DTx that are centred on communication. For example, conversational agents could be tested with such simulated patients. This would allow challenging the DTx with a diversity of user inputs, in different language capabilities, health literacy levels, etc. For this scenario, it still has to be clarified how role-prompted LLMs react to ambiguous or unclear input. To support this, we plan to develop a patient vignette generator where the different characteristics can be selected from a predefined list and the clinical validity of the generated patient vignette can be ensured. This vignette can then directly be used for role-prompting in an LLM. We conclude that there is potential in using LLMs together with patient vignettes to simulate interactions with patients. A more in-depth analysis is required to systematically identify potentials and limitations.

5. Bibliographical References

Deborah Bartz, Tanuja Chitnis, Ursula B. Kaiser, Janet W. Rich-Edwards, Kathryn M. Rexrode, Page B. Pennell, Jill M. Goldstein, Mary Angela O'Neal, Meryl LeBoff, Maya Behn, Ellen W. Seely, Hadine Joffe, and JoAnn E. Manson. 2020. Clinical Advances in Sex- and Gender-Informed Medicine to Improve the Health of All: A Review. *JAMA Internal Medicine*, 180(4):574–583.

Niv Ben-Shabat, Gal Sharvit, Ben Meimis, Daniel Ben Joya, Ariel Sloma, David Kiderman, Aviv Shabat, Avishai M Tsur, Abdulla Watad, and Howard Amital. 2022. Assessing data gathering of chatbot based symptom checkers-a clinical vignettes study. *International Journal of Medical Informatics*, 168:104897.

James RA Benoit. 2023. ChatGPT for Clinical Vignette Generation, Revision, and Evaluation. *MedRxiv*, pages 2023–02.

Leonardo Campillos-Llanos, Catherine Thomas, Éric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. 2020. Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. *Natural Language Engineering*, 26(2):183–220.

Fallon Chipidza, Rachel S. Wallwork, Traci N. Adams, and Theodore A. Stern. 2016. Evaluation and Treatment of the Angry Patient. *The Primary Care Companion for CNS Disorders*, 18(3):10.4088/PCC.16f01951.

Gillian B Clack, Judy Allen, Derek Cooper, and John O Head. 2004. Personality differences between doctors and their patients: Implications for the teaching of communication skills. *Medical Education*, 38(2):177–186.

Andrea Cocci, Marta Pezzoli, Mattia Lo Re, Giorgio Ivan Russo, Maria Giovanna Asmundo, Mikkel Fode, Giovanni Cacciamani, Sebastiano Cimino, Andrea Minervini, and Emil Durukan. 2024. Quality of information and appropriateness of ChatGPT outputs for urology patients. *Prostate Cancer and Prostatic Diseases*, 27(1):103–108.

Amit Dang, Dimple Arora, and Pawan Rane. 2020. Role of digital therapeutics and the changing future of healthcare. *Journal of Family Medicine and Primary Care*, 9(5):2207–2213.

Kerstin Denecke, Richard May, and Octavio Rivera-Romero. 2024. Transformer models in healthcare: A survey and thematic analysis of potentials, shortcomings and risks. *Journal of Medical Systems*, 48(1):23.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621, Singapore. Association for Computational Linguistics.

Spencer C Evans, Michael C Roberts, Jared W Keeley, Jennifer B Blossom, Christina M Amaro, Andrea M Garcia, Cathleen Odar Stough, Kimberly S Canter, Rebeca Robles, and Geoffrey M Reed. 2015. Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application in icd-11 field studies. *International journal of clinical and health psychology*, 15(2):160–170.

Hermann Füeßl and Martin Middeke. 2022. Bestandteile der Anamnese. In *Duale Reihe Anamnese Und Klinische Untersuchung*, 7. edition. Thieme, Stuttgart.

Marjory Gordon. 2022. *Pflegeassessment Notes*, 2. edition. Hogrefe, Bern.

Morgan Guillaudeux, Olivia Rousseau, Julien Petot, Zineb Bennis, Charles-Axel Dein, Thomas Goronflot, Nicolas Vince, Sophie Limou, Matilde Karakachoff, Matthieu Wargny, et al. 2023. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *NPJ Digital Medicine*, 6(1):37.

Astrid Dorothea Herold-Majumdar, Selina Baumann, Kathrin Hofman, Julia Kämmer, Debora Küllsen, and Valentina Müller. 2023. *Klinisches Simulationslernen in Der Pflege: Die Skills-Lab-Methode*, 1. edition. Hogrefe, Bern.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains.

Siddika S Mulchan, Megan Miller, Christopher B Theriault, William T Zempsky, and Adam Hirsh. 2022. A systematic approach to developing virtual patient vignettes for pediatric health equity research. *Health Equity*, 6(1):862–872.

OpenAI. 2024. Prompt Engineering. https://platform.openai.com.

Eliseo J. Pérez-Stable and Sherine El-Toukhy. 2018. Communicating with diverse patients: How patient and clinician factors affect disparities. *Patient Education and Counseling*, 101(12):2186–2194.

Linda W Peute, Gaby-Anne Wildenbos, Thomas Engelsma, Blake J Lesselroth, Valentina Lichtner, Helen Monkman, David Neal, Lex Van Velsen, Monique W Jaspers, and Romaric Marcilly. 2022. Overcoming challenges to inclusive user-based testing of health information technology with vulnerable older adults: Recommendations from a human factors engineering expert inquiry. *Yearbook of medical informatics*, 31(01):074–081.

Donald A. Redelmeier, Umberin Najeeb, and Edward E. Etchells. 2021. Understanding Patient Personality in Medical Care: Five-Factor Model. *Journal of General Internal Medicine*, 36(7):2111–2114.

Octavio Rivera-Romero, Elia Gabarron, Talya Miron-Shatz, Carolyn Petersen, and Kerstin Denecke. 2022. Social media, digital health literacy, and digital ethics in the light of health equity. *Yearbook of Medical Informatics*, 31(01):082–087.

Nancy E Schoenberg and Hege Ravdal. 2000. Using vignettes in awareness and attitudinal research. *International journal of social research methodology*, 3(1):63–74.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. 2024. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine*, 7(1):1–9.