Journal of Patient-Reported Outcomes

COMMENTARY

Open Access

A new generation of patient-reported outcome measures with large language models

Jan Henrik Terheyden[1*], Maren Pielka[2,3], Tobias Schneider[2,3], Frank G. Holz[1] and Rafet Sifa[2,3]

Abstract

Background Patient-reported outcome measures (PROMs) are cornerstones of patient-centered clinical medicine and reflect patients' abilities, difficulties, perceptions and behaviors. The highly structured questionnaire format of PROMs currently limits their real-world validity and acceptability to patients, which becomes increasingly relevant with the high clinical interest in PROM data. In this short commentary, we aim to demonstrate the potential use of large language models (LLMs) in the context of PROM data collection and interpretation.

Main body The popularization of LLMs enables the development of a new generation of PROMs generated and administered through digital technology that interact with patients and score their responses in real time based on artificial intelligence. LLM-PROMs will need to be developed with multi-stakeholder input and careful validation against established PROMs. LLM-PROMs could complement traditional PROMs particularly in real-world clinical applications.

Conclusion LLM-PROMs could allow quantifying patient-relevant dimensions based on less structured contents and foster the use of patient-reported data in digital, clinical applications of PROMs.

Keywords Patient-reported outcome measures, Large Language models, Generative artificial intelligence, Digital medicine

Background

Patient-reported outcome measures (PROMs) are central to patient-centered clinical medicine and assess health domains such as health-related quality of life, symptoms and health behaviors, allowing practitioners to tailor treatment approaches to patient needs [1–3]. PROMs are increasingly used as trial endpoints, in quality assessment of healthcare programs and during routine care, as they are ideal candidates for obtaining health information outside clinical settings, e.g. during remote monitoring of chronic conditions [1–3]. However, the high degree of standardization necessary for the development of reliable PROM instruments implies that patients are asked to complete structured, inflexible questionnaires [4, 5].

*Correspondence:
Jan Henrik Terheyden
jan.terheyden@ukbonn.de
[1]University Hospital Bonn, Department of Ophthalmology, Venusberg-Campus 1, 53127 Bonn, Germany
[2]Bonn-Aachen International Center for Information Technology, University of Bonn, Friedrich- Hirzebruch-Allee 6, 53115 Bonn, Germany
[3]Media Engineering Department, Fraunhofer IAIS, Schloss Birlinghoven 1, 53757 Sankt Augustin, Germany

The advent of large language models (LLMs) is one of the breakthroughs in current artificial intelligence (AI) technology and can help to transform healthcare at a large scale [6, 7]. Popularized with the software ChatGPT (OpenAI, San Francisco, CA) [8], a plurality of public and private LLMs have been proposed for use in healthcare settings, ranging from diagnostic and therapeutic approaches to research and medical training purposes [9]. Despite this, individualizing and measuring quantitative, patient-reported outcome domains through LLMs has not yet gained attention. Thus, we aim to demonstrate the potential use of large language models (LLMs) in the context of PROM data collection and interpretation in this short commentary. The newly introduced term LLM-PROM describes a hybrid system generating individualized, open-ended PROM items and numerically interpreting patient responses.

Main text

Since standardized patient questionnaires were first used to measure health outcomes in the 1960s, the understanding of PROM design, validation, application, and interpretation has made significant progress [1]. PROM contents are developed by clinicians, researchers and policymakers with qualitative input from patients, medical experts and the medical literature [1, 10]. Based on this, PROMs can be applied to a broad range of people, including the general population and those with specific medical conditions (generic and condition-specific measures) [1]. The highly structured format of PROMs including items and response options is based on these qualitative development steps. This structured format is widely accepted in the pharmacoregulation context in which PROMs were first developed and are commonly used. One major advantage of this structured format is to ensure that the items and subscales are interpreted in the same way across different populations so that treatment and time changes can be assessed at an inter-individual level. For clinical trials, ensuring comparable responses and broad item coverage remains essential for assessing the patient relevance of new treatments. Requirements of PROM use in routine healthcare differ from this and administering a complex questionnaire to patients' needs to be justified by feasibility and added benefit. PROMs that implement a higher degree of personalization in sub-populations (e.g. in the context of individually relevant activities of daily living, comprehensibility as per individual language level) promise to generate an increasing impact in the context of clinical care. However, PROMs are prone to missing data [11], which can severely impact the efficacy and effectiveness of outcome measurement in health care programs.

Table 1 Conceptualization of large language model - patient-reported outcome measures

PROM component	Equivalent in LLM-PROM
Item	Algorithmically generated, open-ended question content valid in the context of a given medical condition that is directed to a patient
Response (per response scale)	Patient's reply to a given LLM-PROM item validated for the given use scenario

LLM, large language model; PROM, patient-reported outcome measure

Significant efforts are put into developing short forms of PROMs that may be faster to administer during routine care but can also lose precision compared to the full-length instrument. In the context of latent trait models, item banking has been introduced to make the assessments better targeted. Computer adaptive testing (CAT) deconstructs and individualizes PROMs into combinations of single items [1]. While this approach may reduce the complexity of an assessment, it may not necessarily capture all content domains relevant to an individual patient. As opposed to PROMs, patient-generated outcome measures (PGOMs) summarize individualized questionnaire tools that are developed with an individual person with a high effort to create one measure per individual patient, targeting their respective needs [12–14]. PGOMs have been suggested as a complement to PROMs or to support treatment decisions but not as a replacement of existing instruments for real-world healthcare [13]. Overall, the foundational principles underlying the definition of a PROM have remained largely unchanged from their original conceptualization, which could limit their application in real-world care settings.

Natural language processing (NLP) describes making natural languages (as opposed to software code) readable and computable by machines [15]. NLP is the foundation for the development of LLMs, i.e. AI models that were trained to generate and interpret human language [16]. While the prognostic value of clinician-reports can be limited, patient-reports provide a rich information source about various domains regarding symptoms, quality of life and health behaviors [17]. The quantification of patients' perspectives currently requires rigid instruments that are highly structured. The resulting assessment could potentially be limited by individual respondents' motivations and backgrounds (e.g. educational and cultural backgrounds not covered in validation studies) [18]. One of the main capabilities of AI is the interpretation of unstructured data. Using LLMs and NLP to generate and interpret personalized patient interactions could lead to a new generation of PROMs.

LLM-PROMs are LLM-based psychometric measurement instruments capturing patient-reported outcome data. LLM-PROMs combine two core functionalities: Item generation and patient-report interpretation, which are both conducted by algorithms (Table 1).

Therefore, LLM-PROMs combine individualized open-ended items (e.g. text messages, audio material read to

the patient) with the interpretation of responses by a machine learning algorithm trained to derive quantitative metrics out of its responses (Fig. 1). The suggested framework of LLM-PROMs is shown in Fig. 1 and consists of three layers. The foundation of LLM-PROMs are text contents, which could be derived e.g. from existing PROMs or qualitative datasets of patient or expert interviews, or focus group discussions. The second layer consists of the LLM which is supplied with a predefined input (prompt) that captures essential aspects of the intended contents of the PROM. Examples of contexts captured are in line with existing PROMs and may include health-related quality of life, symptom burden or health behaviors. The result of this becomes the interaction between a patient and a digital bot (e.g. chat bot, voice bot) with a focus on the concept of the LLM-PROM (e.g. health-related quality of life). In contrast to conventional PROMs, the linguistic context of LLM-PROM items can be dynamically adapted to the respondent during the process, e.g. with regards to preferences, language use or cultural background. LLM-PROMs share similarities with existing adaptive PROMs (e.g. CAT-based systems) but are based on open-ended items and can introduce items not covered by an existing item bank. Based on the between a patient and an LLM-PROM, a machine learning algorithm ("interpreter") infers quantitative metrics capturing the very aspects for which structured questionnaire responses are a current requirement.

Like PROMs, LLM-PROMs must be rigorously tested for objectivity, reliability, validity, and responsiveness before clinical use. In this context, existing PROMs will likely need to remain the gold standard for establishing convergent validity and LLM-PROMs will need to prove predictive validity and responsiveness prior to clinical use. Furthermore, the administration burden of LLM-PROMs will need to be evaluated precisely before they can be implemented in routine health care delivery. While AI approaches have been used to score data from existing PROMs [19], the developments in LLMs seen today create the opportunity to introduce the dimension of personalized medicine into PROM assessment which could be useful for guiding treatment decisions, capturing symptoms and care needs as well as adherence and safety monitoring. Thus, it could become possible to obtain a new category of quantitative metrics from written and spoken natural language. This might not only reduce the administration burden of existing PROMs but contribute to the overall goal of a more patient-centered healthcare through LLM-PROMs.

The development of LLM-PROMs requires scientifically robust methods to ensure the content validity of the instruments. The importance of this is highlighted by the fact that the phrasing of items used in LLM-PROMs is not driven by work with patients directly but originates

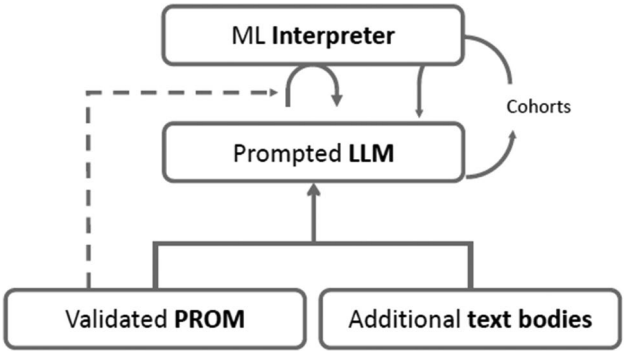

Fig. 1 Development framework of a large-language model patient-reported outcome measure (LLM-PROM). LLM, large language model; ML, machine learning; PROM, patient-reported outcome measure

from prompts to the large language model. However, the opportunity to include large bodies of text into the back-end of the LLM-PROM framework (Fig. 1) may hypothetically even increase the patient-relevance of LLM-PROMs over traditional PROMs. Biases of LLM-PROMs will be an important topic of research. For instance, the sentiment ("mood") in patient responses may be affected by priming effects. Furthermore, biases towards sociodemographic groups (e.g., based on sex, age, ethnicity) [20] will be important topics to consider during the development of LLM-PROMs. We therefore propose using heterogeneous training sets for AI models. They should particularly reflect the populations in which the LLM-PROMs are intended to be used (e.g., in terms of sex, age, ethnicity). Specific types of biases that require to be addressed include minority bias, missing data bias and informed mistrust [21, 22]. Furthermore, we suggest to develop a safety filter network to correct for these biases. Such a filter mechanism could be a separate application that detects inappropriate LLM-PROM items before these are displayed to the respondent. Currently, the sizes of data sets required to effectively capture PROM variables with machine learning algorithms are largely unclear and it may be troublesome to train machine learning interpreters to cover rare conditions since large databases will be required. Further research will be needed to stratify the types of cohorts needed to rigorously validate an LLM-PROM and ensure its real-life validity for healthcare applications.

Current digital applications of PROMs mostly reflect their pen and paper equivalents, while the term digital transition could imply using digitization to rethink existing processes from the start. The integration of LLM-PROMs into digital chat bots could allow novel, personalized and patient-centered models of telemedicine by facilitating the collection of PROM data in existing care pathways. This may not only hold true for remote monitoring applications in chronic conditions and postoperative care but further impact digital treatment

regimens in behavioral health. While applications of LLM-PROMs in clinical and research contexts could be implemented in the near future, their use in clinical trials may remain challenging as no structured regulatory pathways for use of AI in the context of PROMs exist yet. For the integration of LLM-PROMs into clinical care pathways, their interaction with electronic health records will need to be demonstrated, posing potential implementation challenges.

A large language model is prompted based on an existing, validated patient-reported outcome measure (source PROM) and additional text bodies. A machine learning algorithm is trained to estimate the LLM-PROM score based on individuals' free-text responses based on a user interaction and source PROM scores.

LLM-PROMs could hypothetically alter the adoption of personalized medicine during outcome assessment since individual contributors to e.g. health-related quality of life or health behaviors could be targeted to the individual patient. This might be beneficial from a content validity perspective, pending empirical validation. Furthermore, LLM-PROMs could potentially improve the comprehensibility of PROM assessment given the inter-individual differences in language levels and health literacy. Lastly, the use of digital technology with LLM-PROMs promises to hypothetically reduce the loss of PROM data during routine healthcare supervision by addressing patient needs more individually than traditional, highly structured PROMs. Overall, LLM-PROMs remain in the early stages of development, and further scientific evaluation is required to determine their validity and effectiveness.

In the early stages of a new field, a conceptualization framework of LLM-PROMs will be a key requirement. Similarly to PROM development guidelines, scientific standards should direct the strategical development of LLM-PROMs for the healthcare, psychometrics and AI communities. First and foremost, knowledge from existing PROMs needs to be used to develop application-specific vocabularies and safety control mechanisms since the use of LLM-PROMs implies that AI algorithms communicate directly with patients. Similarly to the use of AI in medical imaging, using LLMs to measure patient-reported data will generate a diagnostic black box less available for external evaluation and auditing than conventional and model-based approaches. This issue is under debate for most use cases of AI in healthcare and no final conclusions can be drawn at this point, since broader societal involvement is needed. Integrating patient organizations at the core of these developments will be a key factor ensuring that LLM-PROMs actually foster patient-centered care.

Conclusions

Generative AI and LLMs hold the potential for the development of a novel type of PROMs. Ensuring content validity, appropriateness and psychometric robustness of these instruments will be key enablers of their success. Healthcare providers, researchers, patients and organizations will need to align on a conceptual framework for the development of LLM-PROMs at an early stage.

Abbreviations

AI	Artificial intelligence
LLM	Large language model
NLP	Natural language processing
PROMs	Patient-reported outcome measures

Acknowledgements

None.

Author contributions

JHT drafted the commentary manuscript and MP, TS, FGH and RS critically revised it. All authors contributed significantly to the work.

Funding

None.

Data availability

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

JHT: Heidelberg Engineering, Optos, Zeiss, CenterVue, Novartis, Okko. MP: None. TS: None. FGH: Allergan, Annexon, Alzheon, Apellis, Astellas, Bayer, Boehringer-Ingelheim, Bioeq/Formycon, CenterVue, Roche/Genentech, 4D Molecular Therapeuticcs, Geuder, Grayburg,, Heidelberg Engineering, IvericBio/Astellas, Janssen, LinBiosciences, NightStarX, Novartis, Optos, Oxurion, Pixium Vision, Stealth BioTherapeutics, Carl Zeiss Meditec, Grade Reading Center. RS: None.

Received: 19 April 2024 / Accepted: 13 March 2025
Published online: 24 March 2025

References

1. Churruca K, Pomare C, Ellis LA et al (2021) Patient-reported outcome measures (PROMs): A review of generic and condition-specific measures and a discussion of trends and issues. Health Expect 24:1015–1024. https://doi.org/10.1111/hex.13254
2. Snyder CF, Aaronson NK (2009) Use of patient-reported outcomes in clinical practice. Lancet 374:369–370. https://doi.org/10.1016/S0140-6736(09)61400-8
3. Penedo FJ, Oswald LB, Kronenfeld JP et al (2020) The increasing value of eHealth in the delivery of patient-centred cancer care. Lancet Oncol 21:e240–e251. https://doi.org/10.1016/S1470-2045(20)30021-8
4. Atkinson TM, Schwartz CE, Goldstein L et al (2019) Perceptions of response burden associated with completion of Patient-Reported outcome assessments in oncology. Value Health 22:225–230. https://doi.org/10.1016/j.jval.2018.07.875
5. Huberts AS, Koppert LB, Benschop JAM et al (2024) Facilitators and barriers in the implementation and adoption of Patient-Reported outcomes

measurements in daily practice. Value Health 27:1235–1242. https://doi.org/10.1016/j.jval.2024.05.020
6. Ghassemi M, Oakden-Rayner L, Beam AL (2021) The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 3:e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9
7. Briganti G, Le Moine O (2020) Artificial intelligence in medicine: today and tomorrow. Front Med (Lausanne) 7:27. https://doi.org/10.3389/fmed.2020.00027
8. OpenAI, Achiam J, Adler S et al (2023) GPT-4 Technical Report. arXiv
9. Thirunavukarasu AJ, Ting DSJ, Elangovan K et al (2023) Large Language models in medicine. Nat Med 29:1930–1940. https://doi.org/10.1038/s41591-023-02448-8
10. Prinsen CAC, Mokkink LB, Bouter LM et al (2018) COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 27:1147–1157. https://doi.org/10.1007/S11136-018-1798-3
11. Gomes M, Gutacker N, Bojke C et al (2015) Addressing missing data in Patient-Reported outcome measures (PROMS): implications for the use of PROMS for comparing provider performance. Health Econ 25:515–528. https://doi.org/10.1002/hec.3173
12. Lin Y, Yu Y, Zeng J et al (2020) Comparing the reliability and validity of the SF-36 and SF-12 in measuring quality of life among adolescents in China: a large sample cross-sectional study. Health Qual Life Outcomes 18:360. https://doi.org/10.1186/s12955-020-01605-8
13. Patel KK, Veenstra DL, Patrick DL (2003) A review of selected patient-generated outcome measures and their application in clinical trials. Value Health 6:595–603. https://doi.org/10.1046/j.1524-4733.2003.65236.x
14. Ruta DA, Garratt AM, Leng M et al (1994) A new approach to the measurement of quality of life. The Patient-Generated index. Med Care 32:1109–1126. https://doi.org/10.1097/00005650-199411000-00004
15. Nadkarni PM, Ohno-Machado L, Chapman WW (2011) Natural Language processing: an introduction. J Am Med Inf Assoc 18:544–551. https://doi.org/10.1136/amiajnl-2011-000464
16. Naveed H, Khan AU, Qiu S et al (2023) A Comprehensive Overview of Large Language Models
17. Bottomley A, Coens C, King M et al (2009) Is patient self-reporting more accurate than clinician reporting of symptoms for predicting survival in patients with cancer? Meta-analysis of 30 closed EORTC randomized controlled trials. JCO 27:9597. https://doi.org/10.1200/jco.2009.27.15_suppl.9597
18. Porter I, Gonçalves-Bradley D, Ricci-Cabello I et al (2016) Framework and guidance for implementing patient-reported outcomes in clinical practice: evidence, challenges and opportunities. J Comp Eff Res 5:507–519. https://doi.org/10.2217/cer-2015-0014
19. Cruz Rivera S, Liu X, Hughes SE et al (2023) Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies. Lancet Digit Health 5:e168–e173. https://doi.org/10.1016/S2589-7500(22)00252-7
20. Omiye JA, Lester JC, Spichak S et al (2023) Large Language models propagate race-based medicine. NPJ Digit Med 6:195. https://doi.org/10.1038/s41746-023-00939-z
21. Ueda D, Kakinuma T, Fujita S et al (2024) Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol 42:3–15. https://doi.org/10.1007/s11604-023-01474-3
22. Carey S, Pang A, de Kamps M (2024) Fairness in AI for healthcare. Future Healthc J 11:100177. https://doi.org/10.1016/j.fhj.2024.100177

Publisher's note