The role of patient-reported outcome measures in trials of artificial intelligence health technologies: analysis of ClinicalTrials.gov records (1997 – 2022)

Pearce, Finlay J ; Cruz Rivera, Samantha; Liu, Xiaoxuan; Manna, Elaine ; Denniston, Alastair; Calvert, Melanie

DOI:
10.1016/S2589-7500(22)00249-7

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Pearce, FJ, Cruz Rivera, S, Liu, X, Manna, E, Denniston, A & Calvert, M 2023, 'The role of patient-reported outcome measures in trials of artificial intelligence health technologies: analysis of ClinicalTrials.gov records (1997 – 2022): a systematic evaluation', *The Lancet Digital Health*, vol. 5, no. 3, pp. e160–e167. https://doi.org/10.1016/S2589-7500(22)00249-7

The role of patient-reported outcome measures in trials of artificial intelligence health technologies: a systematic evaluation of ClinicalTrials.gov records (1997–2022)

Finlay J Pearce, Samantha Cruz Rivera, Xiaoxuan Liu, Elaine Manna, Alastair K Denniston, Melanie J Calvert

The extent to which patient-reported outcome measures (PROMs) are used in clinical trials for artificial intelligence (AI) technologies is unknown. In this systematic evaluation, we aim to establish how PROMs are being used to assess AI health technologies. We searched ClinicalTrials.gov for interventional trials registered from inception to Sept 20, 2022, and included trials that tested an AI health technology. We excluded observational studies, patient registries, and expanded access reports. We extracted data regarding the form, function, and intended use population of the AI health technology, in addition to the PROMs used and whether PROMs were incorporated as an input or output in the AI model. The search identified 2958 trials, of which 627 were included in the analysis. 152 (24%) of the included trials used one or more PROM, visual analogue scale, patient-reported experience measure, or usability measure as a trial endpoint. The type of AI health technologies used by these trials included AI-enabled smart devices, clinical decision support systems, and chatbots. The number of clinical trials of AI health technologies registered on ClinicalTrials.gov and the proportion of trials that used PROMs increased from registry inception to 2022. The most common clinical areas AI health technologies were designed for were digestive system health for non-PROM trials and musculoskeletal health (followed by mental and behavioural health) for PROM trials, with PROMs commonly used in clinical areas for which assessment of health-related quality of life and symptom burden is particularly important. Additionally, AI-enabled smart devices were the most common applications tested in trials that used at least one PROM. 24 trials tested AI models that captured PROM data as an input for the AI model. PROM use in clinical trials of AI health technologies falls behind PROM use in all clinical trials. Trial records having inadequate detail regarding the PROMs used or the type of AI health technology tested was a limitation of this systematic evaluation and might have contributed to inaccuracies in the data synthesised. Overall, the use of PROMs in the function and assessment of AI health technologies is not only possible, but is a powerful way of showing that, even in the most technologically advanced health-care systems, patients' perspectives remain central.

Introduction

Research into artificial intelligence (AI) technologies has growing international interest, with health care at its forefront.[1,2] AI, which is used as a term to encompass a range of subdivisions including machine learning and natural language processing, performs tasks associated with human intelligence. AI can support health-care professionals in delivering better and faster patient-centred care, and reduce strain on health-care services.[1] AI health technologies include computer-aided diagnosis and clinical decision support systems that aid health-care providers, but also AI health technologies used by patients, such as AI-enabled smartphone apps. These apps can empower patients to increasingly take their health care into their own hands and include smartphone-supported home medical testing or medication adherence.[3,4]

Patient-reported outcomes (PROs) refer to the report of the status of a patient's health condition, such as physical, psychological, and wellbeing. PROs can be used to assess the impact of disease and treatment on symptom burden and health-related quality of life from the patient's perspective. PROs are assessed in trials with questionnaires referred to as PRO measures (PROMs).[5] PROMs can be classified as disease-specific or generic. Disease-specific PROMs can be tailored to specific health conditions, populations, or functions, whereas generic PROMs capture general aspects of health-related quality of life irrespective of the health condition.[6,7] Examples of disease-specific PROMs include the European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire (QLQ-C30)[8] and the Paediatric Asthma Quality of Life Questionnaire (PAQLQ),[9] whereas generic PROMs include the EuroQoL five-dimension (EQ-5D) questionnaire[10] and the 36-Item Short Form Health Survey (SF-36).[11]

PROM data collected in clinical trials can inform shared decision making, development of clinical guidelines, and facilitate patient communication at an individual level.[12,13] Additionally, PROM data can be used to inform product labelling claims and inform regulators on the efficacy and tolerability of treatments.[5,14] For example, PROMs have been used in oncology clinical trials to show benefits to patients' symptoms and quality of life, supporting pharmaceutical labelling and approval.[15] A review of clinical trials registered on ClinicalTrials.gov between 2007 and 2013 found that 27% of clinical trials used PROMs.[16] Up-to-date research on the overall use of PROMs in clinical trials registered on ClinicalTrials.gov is not available; however, PROM uptake is likely to have increased because of a greater awareness of the benefits of PROMs and regulatory guidance.[5,17] The inclusion of PROMs in clinical trials of AI health technologies offers the incorporation of

Lancet Digit Health 2023; 5: e160–67

Medical School (F J Pearce BMedSc), **Centre for Patient Reported Outcomes Research, Institute of Applied Health Research** (S Cruz Rivera PhD, E Manna BSc, Prof A K Denniston PhD, Prof M J Calvert PhD), **Birmingham Health Partners Centre for Regulatory Science and Innovation** (S Cruz Rivera, Prof A K Denniston, Prof M J Calvert), **Data-Enabled Medical Technologies and Devices Hub** (S Cruz Rivera, Prof A K Denniston, Prof M J Calvert), **Academic Unit of Ophthalmology, Institute of Inflammation and Ageing** (X Liu MBChB PhD, Prof A K Denniston), **and National Institute for Health and Care Research Applied Research Collaboration West Midlands** (Prof M J Calvert), **University of Birmingham, Birmingham, UK; National Institute for Health and Care Research Birmingham Biomedical Research Centre** (Prof M J Calvert), **and National Institute for Health and Care Research Surgical Reconstruction and Microbiology Centre** (Prof M J Calvert), **University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK** (X Liu, Prof A K Denniston); **Health Data Research UK, London, UK** (Prof A K Denniston, Prof M J Calvert); **National Institute for Health and Care Research Biomedical Research Centre for Ophthalmology, Moorfields Hospital London NHS Foundation Trust and Institute of Ophthalmology, University College London, London, UK** (Prof A K Denniston, Prof M J Calvert); **National Institute for Health and Care Research Birmingham–Oxford Blood and Transplant Research Unit in Precision Transplant and Cellular Therapeutics, Birmingham, UK** (Prof M J Calvert)

Correspondence to:
Dr Samantha Cruz Rivera, Centre for Patient Reported Outcomes Research, Institute of Applied Health Research, University of Birmingham, Birmingham B15 2TT, UK
s.rivera@bham.ac.uk

See Online for appendix 1

patients' perspectives as an important metric through which these technologies can be assessed. Previous qualitative research has suggested that a barrier to implementing these AI devices in regular clinical practice is poor patient acceptance and understanding.[18] In addition to the use of PROMs as a trial outcome to assess the effectiveness of an AI technology as a health intervention, AI models can use PRO data in their regular functions outside of clinical trials.[19] For example, PRO data can act as an input for an AI model, either through large PRO datasets or a specific PRO data input for analysis. On the other hand, PRO data can be used as an output. For instance, AI can be developed to predict a particular PROM score. Currently, evidence relating to the use of PROMs in clinical trials of AI health technologies is scarce. Therefore, the aim of this Review is to explore the current use of PROMs in clinical trials of AI health technologies, identify the percentage of trials of AI health technologies that use PROMs, and identify how PROMs are being used (ie, as an input or output), considering the different PROMs being used and the types of AI technologies being developed.

Methods

Search strategy and selection criteria

We systematically searched the ClinicialTrials.gov registry in accordance with previous methodologies.[16,20] A reviewer (FJP) searched for all interventional trials that tested an AI health technology, from registry inception to Sept 20, 2022. Although the term AI might be used variably, for the purposes of our systematic evaluation we searched for trials using the following terms because they are more commonly adopted for high-capacity machine learning techniques that have emerged in the past decade: "Artificial Intelligence" OR "AI" OR "Machine Learning" OR "Deep Learning" OR "Natural Language Processing" OR "Neural Network". The search strategy was limited to the English language because the database only includes registries in English. We excluded observational studies, patient registries, and expanded access reports. We applied no restrictions on trial population, diseases and conditions studied, or the intended use case for the AI health technology being trialled. We removed exact duplicates but we did not exclude separate trials testing the same AI device with different methods. We did not include the term patient-reported outcome and synonyms in the search strategy to avoid missing any relevant trials, as a consequence of the inconsistent use of the terms.[21] We limited the search to interventional studies only and excluded trials not assessing an AI-enabled health technology. Originally, we planned to include only clinical trials of AI health technologies that used one or more PROMs as an outcome measure. However, we expanded the eligibility criteria to include all clinical trials of AI health technologies to allow for comparison between trials of AI health technologies that used PROMs and those that did not. We included eligible trials regardless of their trial status (ie, completed, recruiting, not yet recruiting, withdrawn, suspended, or terminated). We did not include trials reporting proxy-reported outcomes, unless they also reported a validated PROM. We excluded trials if it was not possible to extract enough details from the ClinicalTrials.gov trial records to satisfy the data extraction form fields (eg, specific AI health technology being trialled). Detailed information of the 627 included trials of AI health technologies is given in appendix 1.

Data screening

We downloaded the retrieved studies as a comma-separated-values file and converted it into an Excel workbook. Two reviewers (FJP and SCR) independently did the screening of all studies after a pilot aimed at improving the process. The pilot comprised 10% of the included studies. Discrepancies were resolved through discussion, with the involvement of a third reviewer (MJC) when needed.

Data extraction and analysis

One reviewer (FJP) extracted data using an agreed data extraction table. A second reviewer (SCR) independently screened all the data and the level of agreement on inclusion of AI studies was 100%. Any queries or discrepancies related to PROMs (eg, PROM was disease-specific or generic) were solved through a meeting with a third reviewer (MJC).

We extracted the following data for all trials of AI health technologies, independent of whether they included PROMs: clinical area or speciality addressed by the AI health technology; intended use population (ie, patient or health-care provider); and intended use of the AI health technology along the patient pathway (ie, for diagnosis, treatment, monitoring, or supportive care). Additional classifications were prevention, screening, and other (where the use-case did not fit into any other category); type of AI algorithm used; and whether PROMs were used as an input or output for the AI model.

Additionally, we extracted the software or hardware the AI was applied to (eg, a smartphone app or medical device) and a summary of the function and purpose of the AI health technology. For PROs used as trial outcomes, we also recorded whether the corresponding PROM was used as a primary or secondary trial endpoint. We identified and cross-referenced potential PROMs with the Patient-Reported Outcome and Quality of Life Instruments Database (PROQOLID).[22] If we did not find a PROM on PROQOLID, we searched its validation study on PubMed. However, if we did not identify a validation study, we excluded the clinical trial from the PROM analysis. We included visual analogue scales (VASs), patient-reported experience measures (PREMs), and usability measures when they were completed by trial participants. We excluded tools not classified as PROMs, such as clinician-completed tools, composite measures,

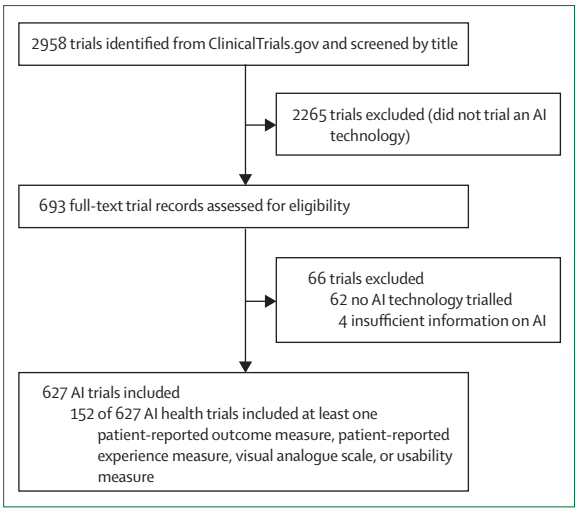

Figure 1: **Study selection**
AI=artificial intelligence.

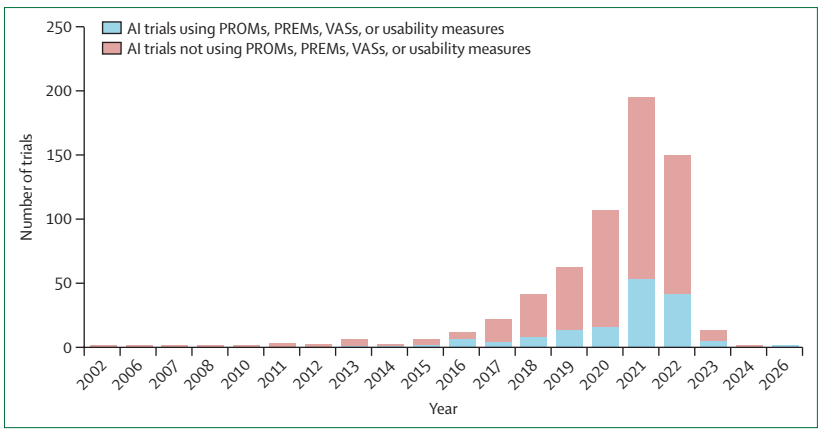

Figure 2: **Yearly use of PROMs, PREMs, VASs, and usability measures in trials of AI health technologies**
For the years 2003–05, and 2009, we did not identify trials of AI technologies registered on ClinicalTrials.gov that used PROMs. For the years 2023, 2024, and 2026, the indicated trials are the trials registered on ClinicalTrials.gov that are due to commence recruitment in those years. AI=artificial intelligence. PREM=patient-reported experience measure. PROM=patient-reported outcome measure. VAS=visual analogue scale.

and health-care use measures. We categorised each PROM as a disease-specific or generic measure.

Results

The ClinicalTrials.gov search yielded 2958 trial records. We screened titles for AI (figure 1), resulting in the inclusion of 693 trials for full-text assessment. We excluded 62 trials because the trial intervention did not assess an AI health technology and we excluded four trials because they did not provide detailed information to satisfy the required data extraction fields. We included the remaining 627 trials in the analysis, of which nine have been published,[23–31] and the remaining ones are either ongoing trials or their results have not been published (appendix 2 pp 2–15). Of these 627 trials, 152 (24%) included at least one specific PROM, PREM, VAS, or usability measure as a trial outcome.

Use of trials of AI health technologies per year

The number of trials that used at least one PROM, PREM, VAS, or usability measure as a trial endpoint increased from four trials in 2017, to 53 trials in 2021 (figure 2). The total number of registered clinical trials of AI health technologies increased in the same period, from 8 trials to 142 trials. The lower number of trials in 2022 than in 2021 is due to the search strategy being limited to September, 2022.

Trials of AI health technologies per country

211 (34%) of the 627 trials that used AI health technologies were conducted in the USA, followed by 85 (14%) in China, 39 (6%) in France, 30 (5%) in Spain, 29 (5%) in Canada, and 26 (4%) in the UK (appendix 2 pp 16–17). 108 trials did not provide information on location and were classified according to the location of the trial sponsor, responsible party, or participant eligibility criteria. We could not establish the trial location of one trial.

The distribution by country of clinical trials of AI health technologies that used PROMs differs from the distribution of all clinical trials of AI health technologies. Among the six countries with the highest numbers of clinical trials of AI health technologies, the highest rate of PROM inclusion was in the UK (9 [35%] of their trials), followed by Canada (10 [34%]), the USA (61 [29%]), Spain (8 [27%]), France (9 [23%]), and China (4 [5%]).

Trials of AI health technologies per clinical area

We classified the 627 clinical trials of AI health technologies according to the clinical specialities and areas the AI health technology was designed for. The three most common clinical areas were digestive system health (97 [15%]), oncology (81 [13%]), and mental and behavioural health (70 [11%]). The inclusion of PROMs as a trial endpoint differs on the basis of the clinical area of the AI health technology trial. The clinical area with the highest rate of PROM inclusion in AI health technology trials was musculoskeletal health, such as osteoarthritis (16 [57%] of 28 trials in this area); followed by mental and behavioural health (38 [54%] of 70); endocrine, nutritional, and metabolic health, such as diabetes (22 [46%] of 48); and neurological system health (18 [41%] of 44; figure 3).

In the 152 trials that included PROMs, PREMs, VASs, or usability measures, we identified 219 unique PROMs (appendix 2 pp 26–31). Of these measures, 149 (68%) were classified as generic PROMs and 70 (32%) as disease-specific. Additionally, we identified 26 unique PREMs and usability measures (appendix 2 p 32) and 15 unique VASs. We excluded 30 measures because we did not find them on PROQOLID and did not identify a corresponding validation study on PubMed. Of the 152 trials that included PROMs, PREMs, VASs, or

See Online for appendix 2

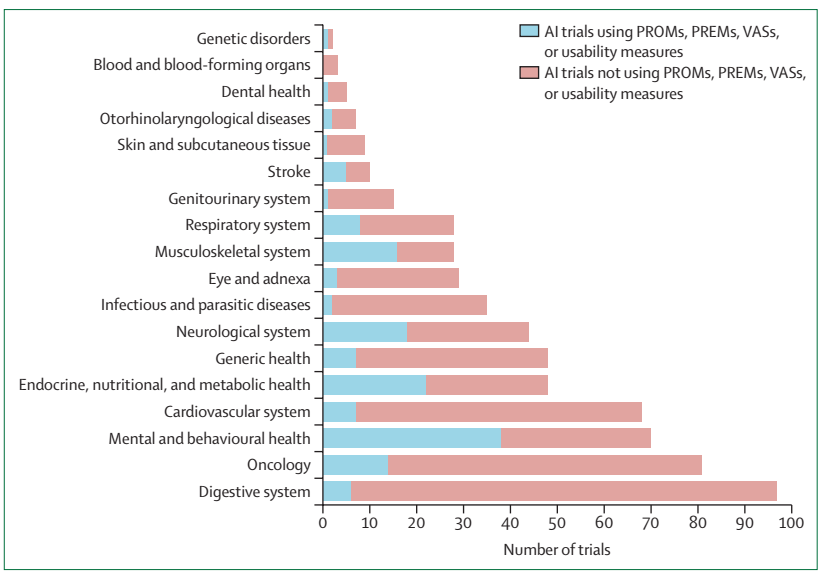

Figure 3: **Use of PROMs, PREMs, VASs, and usability measures in trials of AI health technologies per clinical area**
AI=artificial intelligence. PREM=patient-reported experience measure. PROM=patient-reported outcome measure. VAS=visual analogue scale.

usability measures, 35 (23%) used one or more PROM to assess a primary endpoint, 75 (49%) to assess a secondary endpoint, and 42 (28%) to assess both primary and secondary endpoints. The PROMs mostly used to assess trial outcomes were EQ-5D (20 [13%] of 152 trials); the Patient Health Questionnaire, 9-item version[32] (PHQ-9; 19 [13%] trials); the General Anxiety Disorder-7 (GAD-7) questionnaire[33] (18 [12%] trials); the Hospital Anxiety and Depression Scale[34] (HADS; 7 [5%] trials); and SF-36 (7 [5%] trials). PHQ-9 is a generic tool used to screen for symptoms of depression and GAD-7 is a generic tool for symptoms of anxiety; their frequent use might partly be due to the large proportion of clinical trials of AI health technologies in mental and behavioural health (table 1).

Of the 627 AI health technologies trialled, 24 (4%) trials tested AI health technologies that captured PRO data to be processed by the AI system as an input (eg, a clinical decision support system that recommends the best treatment option for a patient, on the basis of their PROM data). Two (0·3%) trials tested AI health technologies that, in addition to using PRO data as an input for the AI system, also provide predictions of PRO status as an output (table 1).

The role of AI and the intended use population in trials of AI health technologies

We extracted data regarding the characteristics of the AI health technology being trialled. We categorised each trial according to the overall role of the AI health technology and its intended use population (ie, the group that interacts with, or operates, the AI interface: patients or health-care providers, or both). At least one PROM, PREM, VAS, or usability measure was included in 90 (49%) of the 183 trials of AI health technologies designed for disease treatment; 32 (50%) of the 64 trials of AI health technologies used for monitoring or supportive care; 12 (6%) of the 217 trials of AI health technologies used for diagnosis; 95 (54%) of the 176 clinical trials of AI health technologies that tested devices or technologies designed for patient use; and 31 (8%) of the 395 trials of AI health technologies used by health-care providers (table 2).

For the 152 trials of AI health technologies that incorporated one or more PROM, PREM, VAS, or usability measure as a trial endpoint, we extracted additional data regarding the specific form and function of the AI health technology: smart-device applications were the most common AI health technology, followed by clinical decision support systems and chatbots (appendix 2 pp 18–24). 44 (29%) trials that used at least one PROM, PREM, VAS, or usability measure exclusively tested AI-enabled smart-device applications. Of these, 33 (75%) were intended for patient use. These smart-device applications were most often designed as treatment interventions in mental and behavioural health (15 of 44 trials).

23 (15%) of 152 trials tested clinical decision support systems designed to support patients or health-care providers, or both, in making health-care decisions on the basis of different parameters such as patient physiology or characteristics. 13 (57%) of these trials were designed for use by health-care providers, and 11 (48%) were involved in informing decisions regarding disease treatment. The most common clinical areas for clinical decision support systems were mental and behavioural health (6 [26%] of 23 trials), musculoskeletal health (5 [22%]), and endocrine, nutritional, and metabolic health (5 [22%]). Chatbots appeared in 14 (9%) of 152 trials and were primarily used in mental and behavioural health.

183 trials reported testing a technology that used only machine learning (appendix 2 p 25). Of these, 48 (26%) used at least one PROM, PREM, VAS, or usability measure. In 315 trials, the investigated technology or algorithm was simply described as artificial intelligence.

Discussion

This Review summarises for the first time the role of PROMs in trials of AI health technologies with data from a large international trials registry. The data show that the number of trials of AI health technologies incorporating PROMs is growing rapidly, specifically in countries such as Canada, France, Spain, the UK, and the USA. The main trials of AI health technologies including PROMs focused on disease treatment and monitoring or supportive care.

The use of PROMs in the assessment of AI health technologies as a trial endpoint (7% of trials of AI health technologies) falls behind the rate of PROM use (27%)

	Number of trials
Most commonly used PROMs	
EuroQoL five-dimension five-level questionnaire[27]	20/152 (13%)
Patient Health Questionnaire, 9-item version[28]	19/152 (13%)
General Anxiety Disorder-7[29]	18/152 (12%)
Hospital Anxiety and Depression Scale[30]	7/152 (5%)
36-Item Short Form Health Survey[31]	7/152 (5%)
Other	
Visual analogue scales	15/152 (10%)
Patient-reported experience measures and usability measures	39/152 (26%)
Endpoint positioning of PROMs	
Primary	35/152 (23%)
Secondary	75/152 (49%)
Both primary and secondary	42/152 (28%)
Classification of PROMs	
Generic	149/219 (68%)*
Disease-specific	70/219 (32%)*
PROMs as input or output	
Input	24/627 (4%)
Input and output	2/627 (0·3%)
Output	0

PROM=patient-reported outcome measure. *For generic and disease-specific PROMs, data are number of measures and not trials.

Table 1: **PROMs characteristics in trials of artificial intelligence health technologies that used at least one PROM, visual analogue scale, patient-reported experience measure, or usability measure**

	Trials (n=627)	Trials that used at least one PROM, PREM, VAS, or usability measure
Role of artificial intelligence		
Diagnosis	217	12 (6%)
Treatment	183	90 (49%)
Monitoring or supportive care	64	32 (50%)
Prevention	66	11 (17%)
Screening	76	5 (7%)
Other	21	2 (10%)
Intended use population		
Patients	176	95 (54%)
Health-care provider	395	31 (8%)
Patients and health-care provider	56	26 (46%)

PREM=patient-reported experience measure. PROM=patient-reported outcome measure. VAS=visual analogue scale.

Table 2: **Artificial intelligence trials according to the role of artificial intelligence and the intended use population, with the corresponding number of trials including at least one PROM, PREM, VAS, or usability measure**

across all clinical trials registered on ClinicalTrials.gov between 2007 and 2013;[16] the overall inclusion rate of PROMs in clinical trials of AI health technologies was 25% between December, 2014, and September, 2022.

PROMs were most commonly used for AI technologies involved in mental health and long-term conditions, such as diabetes and osteoarthritis, for which assessment of health-related quality of life and symptom burden is particularly important.

The substantial use of PRO data collection among these clinical areas highlights the increased recognition of PRO data in clinical decisions to inform diagnosis and monitor improvement of symptoms. Many trials in endocrine, nutritional, and metabolic health were intended for patients with long-term conditions, such as diabetes, and aimed to enable self-management and lifestyle changes to improve quality of life and reduce symptom burden. PROM use also varies by geographical location. Notably, uptake of PROs is particularly low in trials of AI health technologies in China, which is consistent with a review that called for further expansion of PRO use in China to avoid missing important data.[35]

PROM use was the lowest in trials of AI-assisted diagnostic devices; however, the importance of PROMs varies according to context. PROMs should have a clear rationale for assessment, whether included as an input, output, or trial endpoint. Similarly to trials of other interventions, patient involvement in the co-design of trials of AI health technologies can help ensure that studies are designed to meet patients' needs. Notably, AI health technologies are sometimes met by a lack of acceptance or understanding from patients.[18] In this systematic evaluation, we identified very few trials that used a patient-completed questionnaire that was specifically designed to assess the usability or acceptability of the AI health technology being trialled. These types of questionnaires could be used to identify and address patient concerns on the accessibility of AI health technologies intended for patient use.

Transparency in the reporting of trial protocols, including the description of the intervention and the demographics of patients included, is important for any trial. In particular, transparency is crucial in trials of AI health technologies in which the performance of two apparently similar technologies might vary widely, and the performance might generalise poorly between different population groups. Several trials in this Review did not report specific key information on the functioning of the AI health technology being tested, the type of algorithm being used, and details regarding the interface through which it is interacted with, thus presenting concerns on its reproducibility. These features are important to understand AI health technologies and to assess the validity of the trial, and would be reported if registrants complied with the reporting guidelines for trial protocols involving PROs (SPIRIT-PRO) and AI health technologies (SPIRIT-AI).[36,37] However, these fields are not currently mandated on ClinicalTrials.gov.

Measures to ensure inclusive recruitment leading to a suitably diverse study population are an important part of ethical trial design and are recommended by the US Food and Drug Administration,[38] but do not feature on

ClinicalTrials.gov. For instance, the reporting of a participant's ethnicity as a mandatory field on ClinicalTrials.gov is absent. In the context of AI, this requirement is particularly important because of issues of poor generalisability when AI health technologies are trained and tested on patient groups that are not representative of the intended use population, usually because of a lack of diversity. PRO data collection also needs to be inclusive and should consider the needs of underserved groups, such as minority ethnic or socioeconomically disadvantaged groups and those with particular health statuses, because they might experience specific barriers or prejudices that restrict their inclusion in research.[39]

Capturing PRO data to assess AI health technologies in clinical trials allows people who commission these technologies to have a wider view of their impact, including their effect on patients' quality of life and health-related quality of life. The use of PROM data as an input to an AI health technology helps ensure that some aspects of the patient's own perspective are part of the set of variables from which the algorithm makes its predictions. The use of PROs as an output from the AI health technology (ie, when the algorithm output involves the prediction of a PRO status such as improvement in a symptom score or quality of life measure) might help support interventions and treatment decisions that result in health outcomes that better align with patient priorities. Several considerations for the inclusion of PROs in trials of AI health technologies are given in the panel.

Limitations

Trial records having inadequate detail regarding the PROMs used or the type of AI health technology tested was an issue and might have contributed to inaccuracies in the data synthesised. Several trials did not provide adequate information regarding trial endpoint assessment via PROMs, including information on the specific PROM used, or on whether a PROM was patient-reported or proxy-reported for PROMs that could be used in both cases. In trials of advanced illness, some PROMs were probably completed by proxies.

Four trials reported the AI health technology that was being trialled; however, they did not provide enough information required to populate the data extraction form created. Because of the absence of reporting standards on ClinicalTrials.gov, trial registry entries often do not report enough information for assessment, which could have led to inaccuracies in the data extracted, particularly relating to the function of the AI health technology being tested. We did not assess the protocols of the trials included because many protocols are not available and results are yet to be reported for a number of studies.

Additionally, the primary focus of this Review was the emerging high-capacity machine learning technology (eg, deep learning and neural networks), which is typically indexed under the search terms included. We acknowledge an absence of consensus in the community on whether methods such as logistic regression should also fall under the definition of AI; however, this aspect was not the primary focus of the Review. Therefore, we might have missed trials assessing methods such as logistic regression that were not indexed under any of the search terms.

For AI-enabled health technologies that captured PROMs as an input for the algorithm or provided a prediction in PROM format as an output, almost none of the trials specified the particular measures captured or the PROM format being predicted by the AI model.

Conclusions

This systematic evaluation of the ClinicalTrials.gov trial registry provides new insights into the role PROMs have in both the assessment of AI health technologies in clinical trials and their use in AI models. The use of PROs provides a way to ensure that the patient's own perspective and priorities are represented in the function of these technologies, and in the process that assesses them. The use of PROs data in the function and assessment of AI health technologies is not only possible, but is a powerful way of showing that, even in the most technologically advanced health-care system, the patient perspective remains central. In accommodating individual patient perspectives and placing them at the centre of AI development, patient concerns and anxieties surrounding AI can potentially be addressed and support for AI implementation in health care will grow.

Panel: Considerations for the inclusion of patient-reported outcomes in clinical trials of artificial intelligence (AI) health technologies

- Clearly define the nature of the AI health technology[36]
- At the design stage, clearly state the intended use and the intended use population;[36] intended use is a technical term with substantial regulatory meaning, and the lack of clarity on this term at an early stage is a recurrent problem in design
- Carefully consider measures to ensure the inclusion of underserved groups
- Involve patients and members of the public in the co-design of AI trials and selection of patient-reported outcome measures (PROMs)[36,37]
- Carefully select and define PROMs on the basis of the target population[36]
- Clearly define whether a PROM will be completed by a patient alone or via a proxy (ie, carer or health-care provider)[36]
- Transparently report the demographics of the included participants
- Openly and transparently report trial protocols by use of international guidelines for clinical trials of AI health interventions (SPIRIT-AI) and patient-reported outcomes (SPIRIT-PRO)[36,37]
- Ensure trial protocols are made more readily available to avoid duplication of research and waste of research resources

Contributors

MJC and SCR conceptualised and supervised the study. MJC, SCR, and FJP contributed to the design of the study and had access to the data. FJP developed the search strategy, carried out the screening of the search results, did the data extraction, and drafted the manuscript. All authors critically revised and edited the manuscript for important intellectual content.

Declaration of interests

SCR receives funding from UK SPINE and European Regional Development Fund DEMAND Hub and has received support to attend the Organisation for European Cancer Institutes 2022 conference as an invited speaker. MJC is Director of the Birmingham Health Partners Centre for Regulatory Science and Innovation, Director of the Centre for Patient Reported Outcomes Research, and is a National Institute for Health and Care Research (NIHR) Senior Investigator. MJC receives funding from the NIHR, UK Research and Innovation (UKRI), NIHR Birmingham Biomedical Research Centre, the NIHR Surgical Reconstruction and Microbiology Research Centre, NIHR, Applied Research Collaboration (ARC) West Midlands, UK SPINE, Research England, European Regional Development Fund DEMAND Hub at the University of Birmingham and University Hospitals Birmingham NHS Foundation Trust, and the NIHR Birmingham–Oxford Blood and Transplant Research Unit in Precision Transplant and Cellular Therapeutics. MJC also receives funding from Health Data Research UK, Innovate UK (part of UKRI), Macmillan Cancer Support, UCB Pharma, Janssen, GSK, Gilead Sciences, European Commission, European Federation of Pharmaceutical Industries and Associations, and The Brain Tumor Charity. MJC has received personal fees from Aparito, CIS Oncology, Takeda Pharmaceuticals, Merck, Daiichi Sankyo, Glaukos, GSK, the Patient-Centered Outcomes Research Institute, Genentech, and Vertex Pharmaceuticals, outside of the submitted work. MJC has received lecture fees from the University of Maastricht, Maastricht, Netherlands. In addition, a family member owns shares in GSK. The views expressed in this Review are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. All other authors declare no competing interests.

Acknowledgments

The study was funded by the University of Birmingham. The funder had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

References

1 Zhang D, Maslej N, Brynjolfsson E, et al. The AI Index 2022 annual report. AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University. March, 2022. https://aiindex.stanford.edu/report/(accessed July 7, 2022).

2 He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019; **25:** 30–36.

3 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; **25:** 44–56.

4 Leddy J, Green JA, Yule C, Molecavage J, Coresh J, Chang AR. Improving proteinuria screening with mailed smartphone urinalysis testing in previously unscreened patients with hypertension: a randomized controlled trial. *BMC Nephrol* 2019; **20:** 132.

5 US Department of Health and Human Services, US Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Center for Devices and Radiological Health. Guidance for industry—patient-reported outcome measures: use in medical product development to support labeling claims. December, 2009. https://www.fda.gov/media/77832/download (accessed July 7, 2022).

6 Black N. Patient reported outcome measures could help transform healthcare. *BMJ* 2013; **346:** f167.

7 Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. *Med Care* 1989; **27** (suppl)**:** S217–32.

8 Kaasa S, Bjordal K, Aaronson N, et al. The EORTC core quality of life questionnaire (QLQ-C30): validity and reliability when analysed with patients treated with palliative radiotherapy. *Eur J Cancer* 1995; **31A:** 2260–63.

9 Juniper EF, Guyatt GH, Feeny DH, Ferrie PJ, Griffith LE, Townsend M. Measuring quality of life in children with asthma. *Qual Life Res* 1996; **5:** 35–46.

10 Feng YS, Kohlmann T, Janssen MF, Buchholz I. Psychometric properties of the EQ-5D-5L: a systematic review of the literature. *Qual Life Res* 2021; **30:** 647–73.

11 Brazier JE, Harper R, Jones NM, et al. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ* 1992; **305:** 160–64.

12 Till JE, Osoba D, Pater JL, Young JR. Research on health-related quality of life: dissemination into practical applications. *Qual Life Res* 1994; **3:** 279–83.

13 Mitchell K. How do patient-reported measures contribute to value in health care? Institute for Healthcare Improvement, Aug 7, 2014. http://www.ihi.org/communities/blogs/_layouts/15/ihi/community/blog/itemview.aspx?List=7d1126ec-8f63-4a3b-9926-c44ea3036813&ID=92 (accessed July 1, 2022).

14 Sanders C, Egger M, Donovan J, Tallon D, Frankel S. Reporting on quality of life in randomised controlled trials: bibliographic study. *BMJ* 1998; **317:** 1191–94.

15 Coon CD. The use of patient-reported outcomes in demonstrating safety and efficacy in oncology. *Clin Ther* 2016; **38:** 756–58.

16 Vodicka E, Kim K, Devine EB, Gnanasakthy A, Scoggins JF, Patrick DL. Inclusion of patient-reported outcome measures in registered clinical trials: evidence from ClinicalTrials.gov (2007–2013). *Contemp Clin Trials* 2015; **43:** 1–9.

17 US Food and Drug Administration. FDA patient-focused drug development guidance series for enhancing the incorporation of the patient's voice in medical product development and regulatory decision making. 2018. https://www.fda.gov/Drugs/DevelopmentApprovalProcess/ucm610279.htm (accessed June 30, 2022).

18 Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. *Lancet Digit Health* 2021; **3:** e599–611.

19 Cruz Rivera S, Liu X, Hughes SE, et al. Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies. *Lancet Digit Health* 2023; **5:** e168–73.

20 Kyte D, Retzer A, Ahmed K, et al. Systematic evaluation of patient-reported outcome protocol content and reporting in cancer trials. *J Natl Cancer Inst* 2019; **111:** 1170–78.

21 Maruszczyk K, Aiyegbusi OL, Cardoso VR, et al. Implementation of patient-reported outcome measures in real-world evidence studies: analysis of ClinicalTrials.gov records (1999–2021). *Contemp Clin Trials* 2022; **120:** 106882.

22 ePROVIDE. Patient reported outcome and quality of life instrument database. https://eprovide.mapi-trust.org/advanced-search (accessed Nov 10, 2021).

23 Zwolan TA, Presley R, Chenier L, Buck B. Investigation of an outcomes-driven, computer-assisted approach to CI fitting in newly implanted patients. *Ear Hear* 2021; **42:** 558–64.

24 Zhu T, Uduku C, Li K, Herrero P, Oliver N, Georgiou P. Enhancing self-management in type 1 diabetes with wearables and deep learning. *NPJ Digit Med* 2022; **5:** 78.

25 Piette JD, Newman S, Krein SL, et al. Patient-centered pain care using artificial intelligence and mobile health tools: a randomized comparative effectiveness trial. *JAMA Intern Med* 2022; **182:** 975–83.

26 Mohr DC, Tomasino KN, Lattie EG, et al. IntelliCare: an eclectic, skills-based app suite for the treatment of depression and anxiety. *J Med Internet Res* 2017; **19:** e10.

27 Mohr DC, Schueller SM, Tomasino KN, et al. Comparison of the effects of coaching and receipt of app recommendations on depression, anxiety, and engagement in the IntelliCare platform: factorial randomized controlled trial. *J Med Internet Res* 2019; **21:** e13609.

28 Livovsky DM, Veikherman D, Golany T, et al. Detection of elusive polyps using a large-scale artificial intelligence system (with videos). *Gastrointest Endosc* 2021; **94:** 1099–109.

29 Jaroszewski AC, Morris RR, Nock MK. Randomized controlled trial of an online machine learning-driven risk assessment and intervention platform for increasing the use of crisis services. *J Consult Clin Psychol* 2019; **87:** 370–79.

30 Eng DK, Khandwala NB, Long J, et al. Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: a prospective multicenter randomized controlled trial. *Radiology* 2021; **301:** 692–99.

31 Buegler M, Harms R, Balasa M, et al. Digital biomarker-based individualized prognosis for people at risk of dementia. *Alzheimers Dement (Amst)* 2020; **12:** e12073.

32 Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001; **16:** 606–13.

33 Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med* 2006; **166:** 1092–97.

34 Snaith RP. The Hospital Anxiety and Depression Scale. *Health Qual Life Outcomes* 2003; **1:** 29.

35 Zhou H, Yao M, Gu X, et al. Application of patient-reported outcome measurements in clinical trials in China. *JAMA Netw Open* 2022; **5:** e2211644.

36 Cruz Rivera S, Liu X, Chan A-W, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med* 2020; **26:** 1351–63.

37 Calvert M, Kyte D, Mercieca-Bebber R, et al. Guidelines for inclusion of patient-reported outcomes in clinical trial protocols: the SPIRIT-PRO extension. *JAMA* 2018; **319:** 483–94.

38 US Department of Health and Human Services, US Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. Enhancing the diversity of clinical trial populations—eligibility criteria, enrollment practices, and trial designs guidance for industry. November, 2020. https://www.fda.gov/media/127712/download (accessed Oct 25, 2022).

39 National Institute for Health and Care Research. Improving inclusion of under-served groups in clinical research: guidance from INCLUDE project. Aug 7, 2020. https://www.nihr.ac.uk/documents/improving-inclusion-of-under-served-groups-in-clinical-research-guidance-from-include-project/25435 (accessed Oct 25, 2022).