Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Conversational AI Tool for ICF-Informed Physiotherapy: A Worked Example of a Formative Mixed-Method Evaluation from Patient and Physiotherapist Perspectives

Marco Griseri

marco.griseri@supsi.ch

University of Applied Sciences and Arts of Southern Switzerland

Andrea Franchini

University of Applied Sciences and Arts of Southern Switzerland - University of Italian Switzerland

Tiziano Negrisolo

Fisioterapia Kinesis

Andrea Pina

Fisioterapia Kinesis

Alessandro Schneebeli

University of Applied Sciences and Arts of Southern Switzerland

Fabio Rinaldi

University of Applied Sciences and Arts of Southern Switzerland - University of Italian Switzerland

Marco Barbero

University of Applied Sciences and Arts of Southern Switzerland

---

Research Article

Keywords: artificial intelligence, physiotherapy, International Classification of Functioning, Disability and Health, usability, user experience, technology acceptance

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11470004/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. Some authors (GM, FA, RF and BM) were involved in the design and development of the prototype evaluated in this study. This role was inherent to the

study aim of conducting a formative evaluation of the system. The authors declare that they have no financial competing interests.

A Conversational AI Tool for ICF-Informed Physiotherapy: A Worked Example of a Formative Mixed-Method Evaluation from Patient and Physiotherapist Perspectives

, Marco Griseri[1*], Andrea Franchini[2], Tiziano Negrisolo[3], Andrea Pina[3], Alessandro Schneebeli[1], Fabio Rinaldi[2], Marco Barbero[1]

[1*] Rehabilitation Research Laboratory, Department of Business Economics, Health and Social Care, University of Applied Sciences and Arts of Southern Switzerland, Via Violino 11, Manno, 6928, Ticino, Switzerland .

[2] Dalle Molle Institute for Artificial Technology, University of Applied Sciences and Arts of Southern Switzerland-University of Southern Switzerland, Via La Santa 1, Lugano, 6962, Ticino, Switzerland .

[3] Fisioterapia Kinesis, Via Vincenzo Vela 9, Mendrisio, 6850, Ticino, Switzerland .

*Corresponding author(s). E-mail(s): marco.griseri@supsi.ch;

Abstract

Background: Implementing the biopsychosocial perspective and the International Classification of Functioning, Disability and Health in routine physiotherapy remains challenging because comprehensive assessment requires time, clinical expertise, and integration within existing workflows. We developed a conversational prototype to address these barriers. This study evaluated its user experience, usability, and acceptability from patient and physiotherapist perspectives and provides a worked example of a multi-actor formative mixed-methods evaluation.

Methods: A single-centre formative mixed-methods study was conducted in a Swiss physiotherapy clinic. Thirty patients completed a conversational interview before a simulated initial consultation; ten physiotherapists each reviewed three AI-generated reports. Patient outcomes included interview completion, time,

study-specific ISO 9241-11 usability indices, User Experience Questionnaire scores, acceptability, and open-ended feedback. Physiotherapists completed the System Usability Scale, a context-adapted Technology Acceptance Model questionnaire, and open-ended questions. Quantitative findings were summarised descriptively. Inductive thematic analysis captured concerns; deductive framework analysis explored findings prioritised using established benchmarks where applicable and study-specific criteria otherwise. Joint displays integrated findings into refinement proposals.

Results: ISO 9241-11 usability indices and objective measures were reported as moderate. Patients reported favourable perspicuity and novelty, while efficiency and stimulation were below average and attractiveness and dependability were rated as bad against User Experience Questionnaire benchmarks. Twenty-two patients (73.3%) supported future implementation. Physiotherapists reported moderate usability (mean System Usability Scale score 66.0, SD 13.3), perceived ease of use of 3.50/5, perceived usefulness of 2.83/5, and behavioural intention to use of 3.25/5. Qualitative findings indicated that the interview could support patient reflection, disclosure, and consultation preparation. However, repetitive or insufficiently tailored questions, conversational reliability problems, and incomplete or insufficiently prioritised reports reduced usability and perceived clinical value. Integration identified discrepancies between successful completion and interaction burden, and between basic learnability and uncertain clinical usefulness.

Conclusions: The prototype showed potential as an acceptable and usable support tool for ICF-informed physiotherapy but requires further refinement before routine implementation. Proposals include adaptive questioning, improved conversational reliability, concise and clinically prioritised reports, traceability to patient responses, and explicit physiotherapist verification of AI-generated content. The multi-actor mixed-methods approach enabled the identification of actionable refinements that would not have been otherwise apparent.

Keywords: artificial intelligence, physiotherapy, International Classification of Functioning, Disability and Health, usability, user experience, technology acceptance

Background

The biopsychosocial (BPS) model, originally proposed by Engel, conceptualizes health and illness as arising from the dynamic interaction of biological, psychological, and social factors [1]. In rehabilitation, it is widely regarded as an important foundation for person-centered assessment and intervention, supporting more comprehensive, individualized, and context-sensitive care [2]. The World Health Organization's International Classification of Functioning, Disability and Health (ICF) operationalizes this perspective by providing an internationally recognized framework and common language for describing functioning, disability, and contextual factors [3].

When translated into adapted clinical tools and routines, the ICF has been reported to support interdisciplinary communication, structure clinical reasoning and goal setting, and facilitate information exchange across rehabilitation services and return-to-work stakeholders [4–6]. However, its routine implementation remains inconsistent and is often constrained by the complexity of the classification, time and documentation demands, limited knowledge and training, and poor integration within established clinical workflows [7–10]. These implementation challenges highlight the need for a tool that can help clinicians translate the ICF from a conceptual framework into a feasible component of routine practice.

We therefore developed a large language model (LLM)-based conversational prototype to support the collection of patient-reported BPS information before the first physiotherapy consultation. The prototype conducts a semi-structured interview with the patient, informed by validated ICF Core Sets for musculoskeletal conditions [11], and generates an ICF-based structured summary for the physiotherapist's review. By shifting the initial collection and structuring of patient-reported BPS information from the physiotherapist to a conversational interface, the prototype aims to facilitate ICF implementation while addressing barriers related to assessment complexity, time demands, and integration into routine clinical workflows. Co-designed with patients and physiotherapists, the prototype helps patients articulate BPS concerns before consultation while leaving clinical judgement and decisions to physiotherapists, consistent with WHO principles of autonomy, stakeholder engagement, and accountability in AI for health [12].

Digital and AI-enabled healthcare formative evaluation should examine the experience of direct interaction with technology as well as its perceived fit with the tasks, roles, and context in which it is intended to be used [13, 14]. In this study, user experience (UX) referred to patients' broader appraisal of the conversational interaction, including pragmatic and hedonic qualities [15]. Usability referred to the extent to which specified users could achieve specified goals effectively, efficiently, and with satisfaction in a specified context of use [16]. Acceptability referred to stakeholders' perceptions that the intervention was agreeable, satisfactory, or appropriate for potential use [17]. These related but non-interchangeable outcomes were particularly relevant for this prototype because patients directly interacted with the conversational interview, whereas physiotherapists reviewed its AI-generated output and considered its clinical usefulness and fit with the consultation workflow.

Formative usability evaluation aims to identify use-related problems and inform refinement and does not provide population-level performance estimates with precision. In digital health, quantitative measures can indicate which aspects of the system warrant attention, whereas qualitative accounts can explain user difficulties, contextual conditions, and concerns not captured by predefined measures. Their explicit integration can generate more actionable conclusions than parallel reporting alone [18, 19]. Accordingly, this study had two aims: first, to assess the UX, usability, and acceptability of an LLM-based conversational prototype from patient and physiotherapist perspectives within a simulated physiotherapy pathway; and second, to provide a worked

methodological example of a multi-actor formative evaluation that integrates subjective, task-performance, and qualitative evidence to identify transparent proposals for prototype refinement.

Methods

Study design and setting

This single-centre formative mixed-methods study evaluated the usability, UX, and acceptability of an LLM-based conversational prototype within a pre-consultation workflow in outpatient physiotherapy. The study also aimed to provide a worked methodological example of a multi-actor evaluation, illustrating how subjective ratings, objective interview completion and duration measures, and qualitative feedback can inform proposals for prototype refinement. The study was conducted at Fisioterapia Kinesis, an outpatient physiotherapy clinic in Mendrisio, Canton Ticino, Switzerland, between May and July 2026.

This study was conducted simulating a pre-consultation workflow. The consultations were conducted solely for study purposes and were not part of participants' routine rehabilitation care. The simulation was intended to preserve key features of the intended workflow; therefore, it did not assess sustained use, organizational adoption, or effects on routine care delivery.

Prototype development

Co-design and needs assessment

The prototype was developed through a co-design process. Co-design actively involves end-users and relevant stakeholders to inform digital health intervention development and identify potential adoption barriers [20].

Between May and August 2025, patients and physiotherapists contributed to a preliminary needs assessment through group-specific questionnaires and a mixed stakeholder focus group. A total of 71 patients and 66 physiotherapists participated in the co-design. A multidisciplinary development team, including professional physiotherapy association delegates (Physioswiss- Associazione Cantonale Ticino), rehabilitation researchers, and natural language processing experts, reviewed these findings. Quantitative results from this preliminary consultation can be found in Additional file 1, Table A1 and A2.

Participants' views on barriers encountered during initial physiotherapy consultations, priorities for support, perceived benefits, and concerns regarding an AI-based solution informed the prototype's design requirements and intended role. These findings were translated into design requirements concerning the prototype's clinical scope, its intended role in the pre-consultation pathway, and safeguards to preserve professional clinical responsibility.

Clinical framework and intended use

The prototype was designed to support the collection and structuring of patient-reported BPS information before an initial outpatient physiotherapy consultation. The patient interaction consisted of an adaptive semi-structured interview. Following a small set of introductory questions [21], the interview was oriented towards a specific musculoskeletal ICF Core Set when applicable [22–29]; otherwise, it was guided by the ICF Core Set for Rehabilitation [30].

Subsequent questions explored the ICF categories relevant to the selected framework. Before the consultation, physiotherapists received a structured report comprising a narrative account of the interview, a list of the ICF categories explored with their supporting rationale, and a BPS analysis of the patient's context.

The prototype was intended to support pre-consultation information gathering and clinical reflection without providing diagnoses or treatment recommendations. Responsibility for clinical interpretation and decision-making remained with the physiotherapist.

System architecture and model allocation

The prototype comprised four components: a front end providing the patient-facing interface, a back end coordinating application logic, a database for data storage, and an AI system conducting the interview. The front end provided a conversational interface supporting text and voice input. Additional features included support for multiple conversations and feedback on system-generated messages through positive or negative ratings and optional comments. The AI system managed the interview by processing incoming patient messages and determining subsequent actions. Specifically, it selected questions, determined whether to continue or conclude the interview, and initiated report generation and delivery upon completion. Four AI models supported these tasks: gpt-5.2, gpt-5-mini, gpt-4.1-mini, and whisper-1. At the time of prototype development, gpt-5.2 was considered the most capable model for the principal interview tasks [31]. Its use was therefore restricted to two critical operations: jointly predicting the interview state and generating the next response, and producing the final report. The lower-cost, lower-latency gpt-5-mini model summarised ongoing conversations when they approached a predefined size limit, thereby limiting the amount of conversational context processed in subsequent steps. The gpt-4.1-mini model translated patient messages into English and system responses back into the patient's language, establishing English as the internal working language of the AI component. Although Italian is not generally considered a low-resource language, self-translation was adopted on the basis of reported performance benefits for some non-English tasks [32]. The whisper-1 model transcribed patient speech, enabling live dictation through the user interface. Overall, this allocation of tasks across models was intended to balance task performance, computational latency, and financial cost.

The turn loop

Each interview followed an iterative cycle: the patient responded to a system-generated question, the AI system evaluated the response and selected the next action, and the application executed that action. At the start of each conversation, a predefined welcome message informed patients of chatbot-related risks and explained how to begin the interview. After the first patient message, the AI system followed a workflow guided by the following objectives: 1. Establish the patient's reason for seeking assistance 2. Determine whether the condition was acute or post-acute 3. Identify the patient's principal disability. 4. Select the most relevant or pressing core domain of the ICF to investigate 5. Explore the relevant subdomains within the selected core domain 6. Investigate additional relevant core domains, as appropriate 7. Conclude the interview when no further relevant exploration was required and generate the report.

Development of the interview strategy

The implemented interview used a semi-structured approach, replacing an earlier, fully structured method. The earlier method traversed the ICF hierarchy, accessed through a graph database, from general to specific concepts. During the conversation, deterministic, depth-first queries to the graph database supplied the questions to be asked by the AI system. In the initial implementation, this fully structured, ontology-driven approach produced interviews that were judged to be tedious, repetitive, and excessively time-consuming. Its rigid traversal strategy limited the prioritization of patients' most pressing concerns because each conceptual branch had to be explored before the system could proceed to another, even when some topics were only marginally relevant. These limitations may have reflected implementation choices rather than inherent shortcomings of ontology-driven interviewing. Further development could therefore investigate whether a more deterministic approach can retain sufficient flexibility to accommodate unexpected conversational developments. The subsequent semi-structured approach was based on the hypothesis that the ICF concepts and hierarchy could serve to elicit guidance throughout the interview without imposing a fixed sequence of questions. At each turn, the AI system updated an internal interview state using the conversation history. This state included the predicted condition; its acute or post-acute status; the core ICF domains already investigated; the domain currently under investigation; the current step index; relevant ICF codes and labels; and whether the interview had concluded, together with the reason for its conclusion. Interview-state tracking and constrained generation were used to guide the system's selection of subsequent actions in accordance with the interview schema. For example, when selecting a subdomain, the system was provided only with options belonging to the current parent domain. Similarly, the available core domains were restricted according to whether the condition was classified as acute or post-acute.

Interview completion and reporting

When the AI system determined that the interview was complete, it notified the patient but allowed the conversation to continue if the patient wished to provide additional information. Each time the interview was deemed complete, the system generated a

report structured according to the biopsychosocial model. Report generation incorporated the full conversation and the recorded sequence of interview states. The interview start and end timestamps and the reasons for termination were also recorded; the timestamps were used to estimate interview duration. The report was then sent to the study lead as an encrypted PDF.

Participants and recruitment

Patients receiving treatment at, or scheduled for, an initial outpatient physiotherapy consultation at Fisioterapia Kinesis were recruited between May and July 2026. Physiotherapists employed at the clinic were recruited in May 2026. An employed physiotherapist (PA) approached patients, whereas the clinic manager (NT) informed physiotherapists. Both groups received study information and access to electronic informed consent.

Eligible patients were adults known to the clinic for a musculoskeletal condition who could provide consent, access the internet, and independently complete the web-based interview in Italian or English. Eligible physiotherapists were licensed, could read Italian or English, and had a digital device for reviewing AI-generated summaries.

Previous exposure to the prototype during co-design excluded both groups. Patients were also excluded for severe cognitive impairment or visual, hearing, or motor impairment preventing use of the tool. Physiotherapists were excluded if they planned to be absent during data collection. Participants were enrolled upon electronic consent, and those who completed all study procedures were included in the analysis.

Consent and ethical approval

All participants received written study information and had the opportunity to obtain further verbal clarification from the project leader before enrolment. Electronic informed consent was obtained from all patients and physiotherapists through the REDCap platform [33], hosted by SUPSI, before the start of any study procedure.

The study was conducted in accordance with the Declaration of Helsinki and the Swiss Human Research Act and Human Research Ordinance. Ethical approval was obtained from the competent ethics committee (Swissethics - Comitato Etico Cantonale Ticino) before study initiation (2025-02227; Rif CE TI 5003).

Study procedures and data collection

Pre-consultation

After enrolment, patients received individual credentials and a link to access the prototype on a SUPSI-hosted platform. They completed the pre-consultation interaction and subsequently completed questionnaires on sociodemographic characteristics, digital literacy [34], and their experience of using the prototype, including the User Experience Questionnaire (UEQ) [15].

Physiotherapists completed a baseline questionnaire on sociodemographic characteristics, digital literacy, and attitudes towards AI. Once the patient interaction was completed, a simulated initial physiotherapy consultation was scheduled with a physiotherapist not involved in the patient's ongoing or previous rehabilitation care. The structured report generated by the prototype was made available to the assigned physiotherapist through password-protected access before the consultation. All questionnaires were administered electronically through individual REDCap links.

Simulated physiotherapy consultation

Before meeting the patient, the physiotherapist reviewed the prototype's report. The simulated consultation was conducted autonomously according to usual clinical practice; clinical interpretation and decision-making remained the physiotherapist's responsibility.

Immediately after each consultation, physiotherapists completed open-ended questions about their experience of reviewing and using the AI-generated report. These responses were collected to identify context-specific facilitators, barriers, and proposals for prototype refinement. Patients completed a single-item question on overall acceptability.

Each physiotherapist conducted three simulated consultations with different patients and, after the third consultation, completed the System Usability Scale (SUS) [35] and a context-adapted Technology Acceptance Model (TAM) questionnaire [36]. All questionnaires were administered electronically through individual REDCap links.

The study procedures and timing of outcome assessments for patients and physiotherapists are summarised in Figure 1.

Outcome measures

Usability, UX, and acceptability were assessed using complementary subjective, objective, behavioural, and qualitative measures, as these related but distinct constructs cannot be adequately characterised by a single instrument or metric [37–39]. The composition of all outcomes, outcome measures and indices is provided in Additional file 1, Table A3.

Patient-reported measures

Patient-facing usability was assessed according to ISO 9241-11 [16] using study-specific Likert-scale items to construct indices of effectiveness, efficiency, and satisfaction.

Interview completion rate, defined as the percentage of patients able to fully complete the interview, and interview completion time were reported as additional measures of effectiveness and efficiency, respectively [40]. Interview completion time was calculated as the sum of the time differences between two consecutive messages. Time intervals exceeding 10 minutes were excluded from this sum.

UX was assessed with the 26-item User Experience Questionnaire (UEQ), yielding scores for attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty

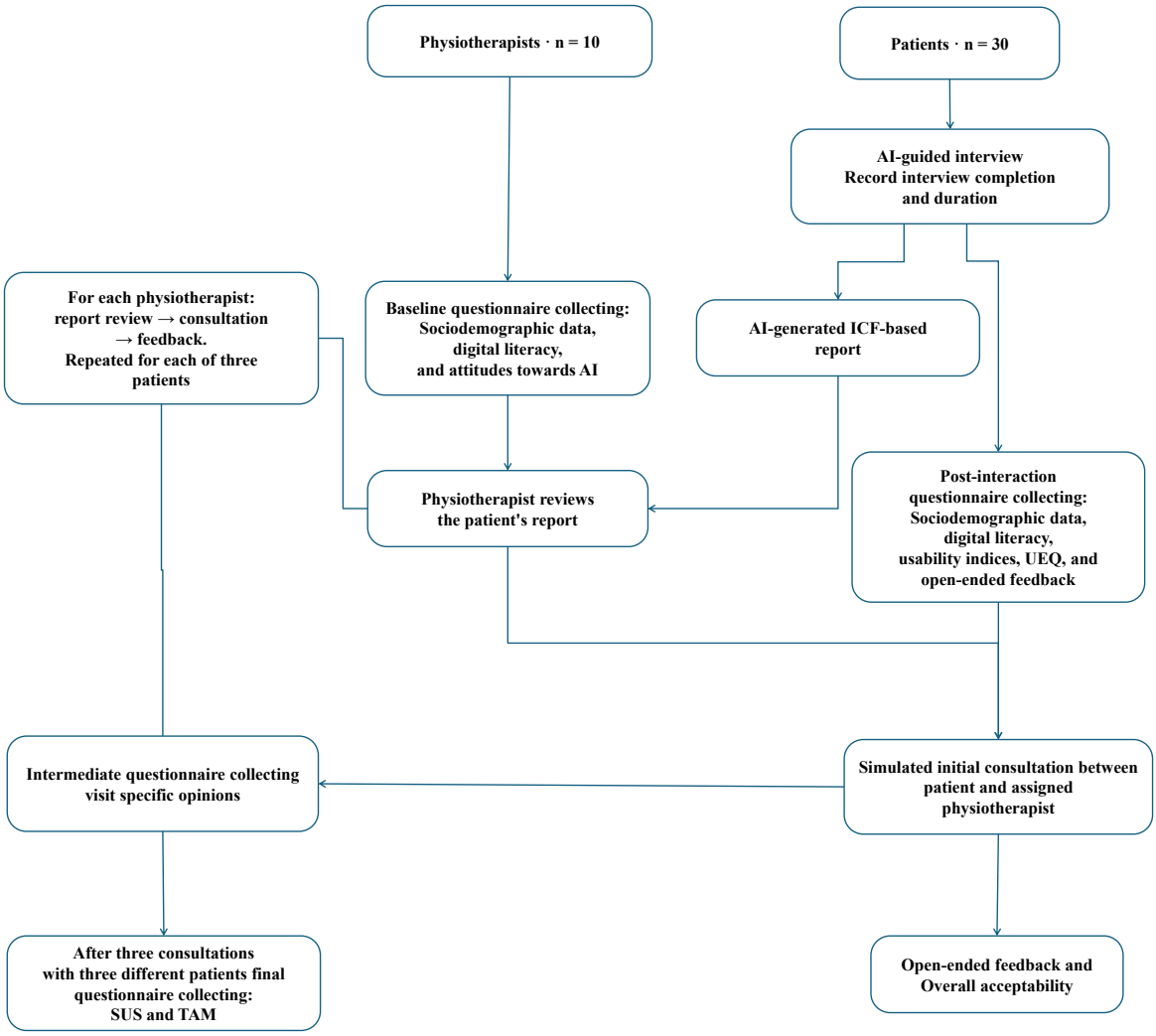


Figure 1: Study procedures and timing of outcome assessments for patients and physiotherapists. UEQ, User Experience Questionnaire; SUS, System Usability Scale; TAM, Technology Acceptance Model; ICF, International Classification of Functioning, Disability and Health.

[15, 41]. After the consultation, patients completed a single-item acceptability measure. Open-ended questions, informed by ISO 9241-11, were included to contextualize quantitative findings [42].

Physiotherapist-reported measures

Physiotherapist technology acceptance was assessed after the third simulated consultation using a context-adapted Technology Acceptance Model (TAM) questionnaire measuring perceived usefulness (PU), perceived ease of use (PEOU), and behavioural intention to use (BI) [36, 43]. The SUS was administered concurrently as a standardised global measure of perceived usability [35, 44]. Open-ended questions, informed by TAM constructs, were included to contextualize quantitative findings [42].

Data analysis

Quantitative analysis

Analyses were conducted separately for patients and physiotherapists in R [45]. Given the formative prototype-evaluation design and sample size, analyses were primarily descriptive and exploratory, focusing on score distributions and variability.

Data were checked for valid ranges and missingness, and negatively worded items were reverse scored where applicable. Composite scores were calculated as the means of available component items without imputation. Distributional characteristics were examined using skewness and kurtosis, histograms, boxplots, Q–Q plots, and Shapiro–Wilk tests. Potential outliers were assessed visually using scatterplots.

Patient outcomes, including completion metrics, ISO 9241-11 usability indices, UEQ scores, and acceptability, were summarised descriptively. For physiotherapists, PU, PEOU, BI, and SUS were summarised descriptively. For patient ISO 9241-11 indices and physiotherapist TAM indices, constructs with a mean score of 3 or lower were flagged as potential targets for refinement. This was a study-specific prioritization rule. UEQ and SUS were interpreted using established scoring and benchmark procedures [35, 46–49].

Qualitative analysis and mixed-methods integration

Open-ended responses were analysed in QualCoder [50] using complementary inductive and deductive approaches. Two researchers (GM and SA) independently coded the full dataset, resolved differences through discussion, and documented analytic decisions in an audit trail [51].

Inductive thematic analysis was conducted separately for patients and physiotherapists without reference to the quantitative results. Codes were grouped and refined into themes describing perceived benefits, concerns, and conditions for appropriate use [52].

In parallel, a deductive framework analysis [51] mapped responses to quantitative findings proposed for prototype refinement: patient UEQ scales below benchmark, physiotherapist SUS dimensions below benchmark, and TAM constructs with mean scores of 3/5 or lower.

To support the study's purpose as a worked methodological example, joint displays documented the links between selected quantitative findings, qualitative evidence, integrated interpretations, and proposed refinements [19, 53]. The integration approach

allowed overlap among refinement implications derived from different meta-inferences, recognising that the constructs assessed were conceptually distinct but interrelated. A single proposed refinement could therefore address issues associated with more than one construct scoring below its respective benchmark. The proposals were intended to inform subsequent prototype development; their effectiveness was not evaluated in this study.

Sample size

A convenience sample was recruited for this formative prototype evaluation. The target of 30 patients was selected to provide descriptive quantitative usability indicators and a sufficiently broad set of user accounts to identify recurrent usability barriers and opportunities for improvement. In formative usability studies, modest samples can identify recurrent problems [54].

Ten physiotherapists each completed three simulated consultations with different patients, providing repeated exposure to the prototype and a range of qualitative accounts of how the AI-generated report could be used in simulated initial consultations.

Results

Participant characteristics and data completeness

Thirty patients and ten physiotherapists were included in the quantitative analyses. Participant characteristics are summarised in Table 1.

Table 1: *Participant characteristics. Values are reported as mean (SD), median (IQR), or n (%), as appropriate. Educational level was scored on a 7-point scale among patients (†) and a 4-point scale among physiotherapists (‡) . Digital literacy and attitude towards AI (§) were scored on 5-point scales. IQR: interquartile range; AI: artificial intelligence.*

Characteristic	Patients (n = 30)	Physiotherapists (n = 10)
Age, years, mean (SD)	41.4 (15.0)	38.8 (8.1)
Gender identity, n (%)		
Women	14 (46.7)	2 (20.0)
Men	11 (36.7)	8 (80.0)
Missing	5 (16.7)	0 (0.0)
Educational level, mean (SD)	4.3 (1.34)†	2.3 (0.86)‡
Digital literacy, median (IQR)	4.2 (0.86)§	3.8 (0.50)§
Attitude towards AI, median (IQR)	—	4.0 (0.00)§

UEQ subscale scores were available for 28 to 30 patients. One value was missing for efficiency and stimulation (3.3% each), and two values were missing for dependability and novelty (6.7% each). All physiotherapists completed the SUS and TAM questionnaires.

Patients usability, user experience, and acceptability

Within the ISO 9241-11 framework, satisfaction was the highest-rated dimension, with a mean score of 3.83 out of 5 (SD 0.55). Effectiveness had a mean score of 3.51 out of 5 (SD 0.77), and efficiency had a mean score of 3.55 out of 5 (SD 0.54).

Twenty-six of 30 patients completed the prototype interview (86.7%). Interview completion time had a median of 41 minutes (IQR 33.75; range 15–158 minutes). Figure 2 presents the full patient usability results.

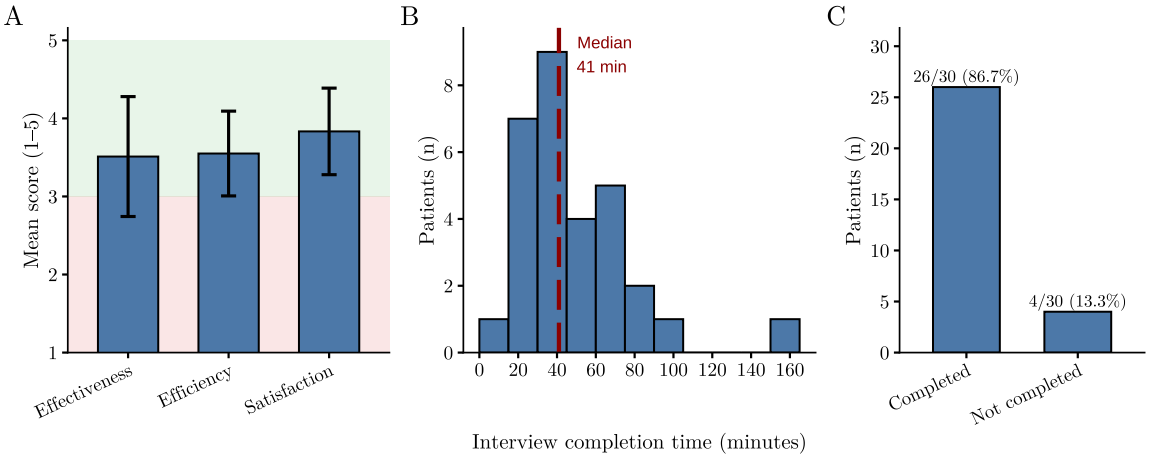


Figure 2: Patient usability outcomes. (A) Mean ISO 9241-11 effectiveness, efficiency, and satisfaction scores; error bars indicate standard deviations. The red and green background bands denote scores at or below, and above, the scale midpoint of 3, respectively. (B) Distribution of AI-guided interview completion times among patients; the dashed vertical line indicates the median completion time. (C) AI-guided interview completion status among all patients. Higher ISO scores indicate more favourable usability.

UX was most favourable for perspicuity (mean 1.32, 95% CI 0.97 to 1.67) and novelty (mean 1.06, 95% CI 0.61 to 1.51), both classified as above average according to the UEQ benchmark. Efficiency (mean 0.65, 95% CI 0.21 to 1.08) and stimulation (mean 0.51, 95% CI 0.01 to 1.01) were classified as below average. Attractiveness (mean 0.50, 95% CI 0.20 to 0.80) and dependability (mean 0.74, 95% CI 0.41 to 1.07) were classified as bad. Complete UEQ results and their integration with qualitative findings are reported in Table 2.

Patients' final acceptability judgement was generally favourable: 22 of 30 participants (73.3%) reported that they would support implementation of the prototype, three (10.0%) would not support implementation, and five (16.7%) were uncertain.

Physiotherapists usability and technology acceptance

The mean SUS score was 66.0 (SD 13.29), close to but below the conventional reference score of 68. Exploratory item-level comparisons against reference values associated with an overall SUS score of 68 showed relatively lower ratings for intention to use the system frequently (mean 3.00, SD 0.82), perceived ease of use (mean 3.30, SD 0.95), functional integration (mean 3.10, SD 0.74), and perceived consistency (mean 3.40, SD 0.84).

Simplicity (mean 3.50, SD 1.08) and independence (mean 4.10, SD 0.99) were close to their respective reference values. Rapid learnability (mean 3.80, SD 0.79), comfort (mean 3.90, SD 0.99), confidence in use (mean 3.80, SD 0.63), and familiarity (mean 4.50, SD 0.53) met or exceeded their respective reference values.

The mean TAM score was 2.83 out of 5 (SD 0.82) for perceived usefulness (PU), 3.50 out of 5 (SD 0.57) for perceived ease of use (PEOU), and 3.25 out of 5 (SD 1.11) for behavioural intention to use (BI). Among the five physiotherapists with a mean BI score of 3 or lower, four indicated that they would reconsider adopting the prototype if it were modified in response to their feedback.

Physiotherapist-reported usability and technology-acceptance findings are reported and integrated with qualitative findings in Table 3.

Qualitative results

The qualitative corpus comprised 179 open-ended responses from 40 participants (30 patients and 10 physiotherapists), corresponding to 179 inductively coded and 171 deductively coded segments. Deductive coding classified 119 segments as positive, 14 as negative but not informative, 65 as revealing a problem, and 39 as proposing a solution.

Patients inductive findings

The inductive analysis identified two potential benefits of the prototype and two conditions for its appropriate implementation. First, patients described the interaction as providing space to reflect on their condition and articulate concerns at their own pace. This theme included self-reflection and awareness, new perspectives on the condition, feeling heard, and having more time or less pressure to disclose concerns. Completing the interaction outside the time constraints of a face-to-face consultation was described as helping patients connect different aspects of their symptoms, organise their thoughts, and identify issues to discuss with their physiotherapist.

> "I had a sort of 'bird's-eye view' of my problem, which helped me reflect a lot and look at it from a different perspective."

Second, patients described the prototype as potentially useful for preparing for physiotherapy consultations. They considered that it could collect preliminary information, provide physiotherapists with an initial overview, and support subsequent patient-physiotherapist communication.

> "It can help the physiotherapist form an idea of the problem through data collection before seeing the patient."

Patients also identified conditions that constrained these potential benefits. Participants reported that the current question set did not always address relevant contextual information, including work, everyday activities, social circumstances, and characteristics of the presenting problem. They suggested more adaptive and clinically specific questions and, where appropriate, the possibility of using images to describe symptoms more concretely.

> "It did not ask about my social context, my profession, or what I do in everyday life."

Finally, patients positioned the prototype as a complementary and case-dependent tool. Its appropriateness was perceived to vary according to the individual patient and clinical presentation. Although the prototype could support reflection and preparation for the first consultation, participants emphasised that it should not replace physical examination, clinical judgement, human interaction, or the therapeutic relationship.

> "For the continuation of the visit, I think a physiotherapist who can 'get hands-on' with the problem is essential."

Physiotherapists inductive findings

The inductive analysis identified three themes concerning the conditions under which the AI-generated report could provide clinically meaningful support. First, physiotherapists described the clinical value of the report as conditional on the completeness, relevance, accuracy, and organisation of its content. Some reports were considered sufficiently detailed to provide an initial overview of the patient. Others omitted clinically important information, focused on issues of limited relevance to rehabilitation, repeated content, or contained interpretations inconsistent with the physiotherapist's assessment. The report was therefore viewed as potentially useful but as requiring clinical prioritisation and verification.

> "Unfortunately, the information was rather incomplete: there was no reference to medication use, which could help me understand how the patient tried to manage the pain independently. The patient's goals, occupation, and sporting and recreational activities were not explored either."

Second, physiotherapists highlighted the situated validity of patient-generated information. They reported that the completeness and interpretability of the output depended on whether patients understood the task, recognised which aspects of their history were relevant, and could describe their condition accurately. They also noted that information collected before the appointment might no longer reflect the patient's current status in acute or subacute presentations.

> "If patients were not instructed, they struggled to provide a complete account of their history and deliberately omitted some problems or illnesses because they did not consider them part of the study."

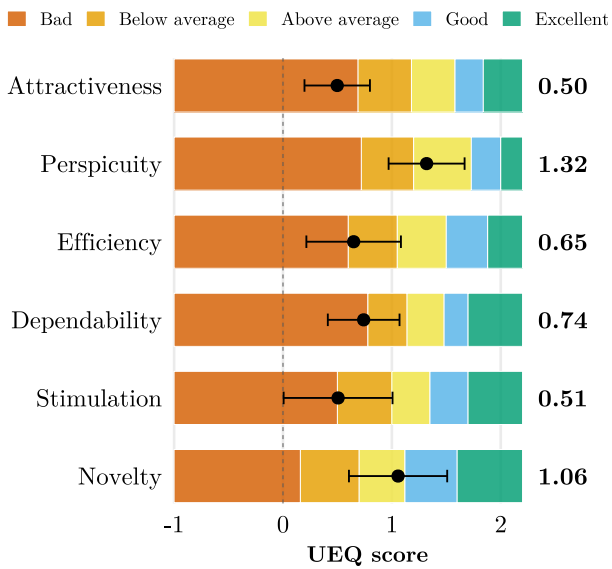
Third, physiotherapists described the prototype as a potential supplementary channel for information gathering and consultation preparation. It could provide an initial overview, help patients organise their account, reduce some of the time required for preliminary data collection, and facilitate disclosure among patients who were less talkative or found it difficult to describe their problems during a conventional consultation.

> "I found it useful for a patient who was not very talkative; it gave me a lot of information that I would have struggled to obtain."

Quantitative-qualitative integration

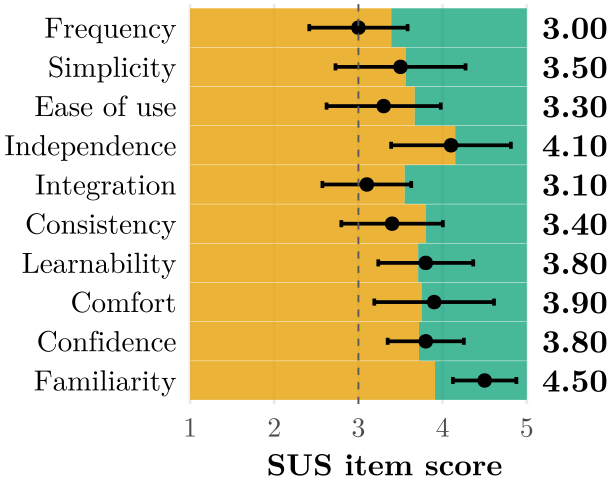
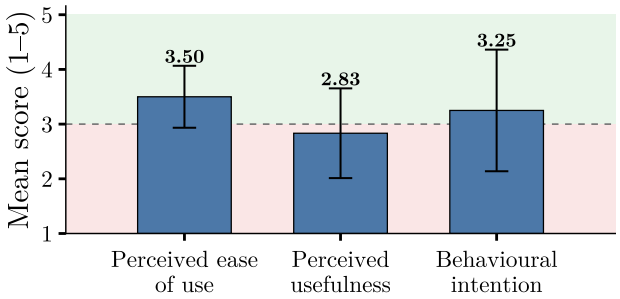
The joint displays in Table 2 and Table 3 integrate the quantitative constructs prioritised for refinement with qualitative accounts deductively linked to the corresponding constructs. Qualitative findings consistent with less favourable quantitative evaluations were considered alongside positive or contrasting accounts to identify convergences, divergences, and implications for prototype refinement.

Table 2: Joint display integrating patient-reported UEQ findings with qualitative findings

Quantitative findings	Constructs prioritised for refinement	Convergent qualitative findings	Divergent qualitative findings	Meta-inference
 **UEQ results.** Attractiveness: $M = 0.50$ (95% CI 0.20–0.80; bad); Efficiency: $M = 0.65$ (0.21–1.08; below average); Dependability: $M = 0.74$ (0.41–1.07; bad); Stimulation: $M = 0.51$ (0.01–1.01; below average).	**Attractiveness**	Participants described the prototype as experimental, immature, cumbersome, or still requiring substantial improvement. Some also perceived the interaction as impersonal or distant from human communication.	Other participants described the prototype as useful, interesting, original, or better than expected, indicating that an unfavourable overall evaluation did not preclude perceived potential value.	The low Attractiveness score reflects an interaction that showed promise but did not yet feel sufficiently mature or patient-centred. Refinement should improve conversational quality, presentation, and responsiveness while preserving the originality and potential usefulness recognised by some participants.
	Efficiency	Participants reported unexpectedly long sessions, repetitive questions, unnecessary interaction steps, and, in some cases, the need to repeat the conversation. Some considered the interaction more time-consuming than speaking directly with a physiotherapist.	Other participants valued having sufficient time to formulate responses without pressure, reflect on their condition, and disclose information that might otherwise be forgotten or omitted.	Efficiency should not be pursued solely by shortening the interaction. Adaptive branching and removal of repetitive or non-relevant questions should reduce unnecessary burden while preserving the unhurried space that supported reflection and disclosure for some patients.
	Dependability	Participants reported voice-transcription failures, conversational stalls, misunderstandings, repeated interactions, unclear interface controls, and instances in which their responses appeared to have been interpreted inaccurately or incompletely.	Some participants nevertheless experienced the interaction as responsive and understandable and considered the questions relevant or similar to those used in physiotherapy assessment.	Dependability requires both technical reliability and transparent handling of patient responses. Refinement should improve voice input, prevent or recover from conversational stalls, clarify interface controls, and allow patients to verify that their responses have been recorded and interpreted as intended.
	Stimulation	Participants described repetitive, monotonous, insufficiently natural, or emotionally "cold" interactions. Long or repetitive exchanges sometimes reduced engagement and willingness to continue.	Other participants reported curiosity, interest, engagement, and new perspectives on their condition, suggesting that reflective prompts could provide intrinsic value beyond information collection.	The dialogue should become more varied, responsive, and personally relevant while retaining prompts that encourage patients to reflect on their condition. Reducing repetition and impersonal exchanges may improve engagement without removing the reflective elements valued by some participants.

Note. UEQ = User Experience Questionnaire; CI = confidence interval. Points indicate mean UEQ scale scores and error bars indicate 95% confidence intervals. Coloured bands show scale-specific UEQ benchmark categories: bad, below average, above average, good, and excellent. Constructs listed in the second column were below the UEQ benchmark. Convergent findings are qualitative accounts consistent with the less favourable quantitative evaluation; divergent findings are positive or contrasting accounts associated with the same construct.

Table 3: Joint display integrating physiotherapist-reported usability and technology-acceptance findings with qualitative findings

Quantitative findings	Constructs prioritised for refinement	Convergent qualitative findings	Divergent qualitative findings	Meta-inference
 **SUS Item-level findings.** Frequency $M = 3.00$ (SD 0.82); Simplicity $M = 3.50$ (SD 1.08); Ease of use $M = 3.30$ (SD 0.95); Independence $M = 4.10$ (SD 0.99); Integration $M = 3.10$ (SD 0.74); Consistency $M = 3.40$ (SD 0.84). Ease of use and independence constructs were not included in the joint display because no relevant qualitative accounts were identified.	**Frequency**	Use was perceived as selective and case-dependent and not appropriate for all patients or routine application. Its value depended on the clinical presentation and information already available.	The tool could be useful for first consultations, complex presentations, limited prior information, or patients who disclosed less during face-to-face interaction.	Define targeted pre-consultation use cases, particularly where preliminary information may meaningfully support the first visit.
	Simplicity	Reports were described as long, repetitive, or difficult to synthesise. Excessive detail and limited prioritisation increased the effort needed to identify clinically relevant information.	Some reports were considered sufficiently complete and well organised to provide a useful initial overview of the patient.	Provide a concise, clinically prioritised summary with optional access to detailed information, reducing cognitive burden without losing relevant content.
	Integration	Workflow fit was limited when information was incomplete, overly detailed, insufficiently prioritised, or outdated by the time of consultation. Physiotherapists also considered clinical verification necessary.	In selected cases, the report supported first-visit preparation, offered an initial patient overview, and reduced preliminary information-gathering time.	Position the prototype as a timely and verifiable pre-consultation information source, with concise outputs that fit existing clinical routines.
	Consistency	Reports sometimes omitted important information, focused on less relevant aspects, or included interpretations that were inaccurate or inconsistent with the clinician's assessment.	In other cases, the anamnesis was considered complete and sufficiently accurate to support a useful clinical overview.	Distinguish patient-reported information from AI-generated interpretation, flag uncertainty and missing information, and retain clinician verification as essential.
 **Technology Acceptance Model findings.** Perceived usefulness $M = 2.83$ (SD 0.82).	**Perceived usefulness**	Incomplete, poorly targeted, or inaccurate information limited clinical value and could reduce confidence in the report. Omissions concerned patient goals, work and daily activities, medication use, coping strategies, pain characteristics, and psychological or emotional factors; some reports also included interpretations that clinicians considered insufficiently relevant or inconsistent with their assessment.	In selected consultations, reports supported preparation for the first visit by providing an initial patient overview and highlighting issues to explore further. They were considered particularly useful when prior medical information was limited or when patients were less talkative or found it difficult to organise and communicate their account during a face-to-face consultation.	Improve completeness, clinical relevance, and actionability through case-sensitive questioning and clinically prioritised summaries. The report should distinguish patient-reported information from AI-generated interpretations, make missing or uncertain information visible, and support clinician verification when pre-consultation information is likely to add value.

Note. SUS = System Usability Scale; TAM = Technology Acceptance Model; SD = standard deviation. Higher scores indicate more favourable evaluations. Convergent findings are qualitative accounts consistent with the less favourable quantitative evaluation; divergent findings are positive or contrasting accounts associated with the same construct.

Discussion

From the patient perspective, the findings revealed a tension between the reflective value of the pre-consultation interview and the burden introduced by its current interaction design. Although most patients completed the interview and perspicuity and novelty were the most favourably evaluated UX dimensions, the 41-minute median completion time, four incomplete interviews, and below-benchmark ratings for efficiency, stimulation, attractiveness, and dependability indicate that completion did not necessarily represent the absence of burden. Qualitative accounts helped clarify this apparent discrepancy: some patients valued having sufficient time and less interpersonal pressure to organise their thoughts, reflect on their condition, and identify concerns before meeting the physiotherapist, whereas repetitive or insufficiently tailored questions, conversational stalls, transcription failures, and unclear controls made parts of the interaction feel unnecessarily complex. The design priority should therefore not be to minimise completion time indiscriminately, but to reduce interaction experienced as redundant or unproductive while preserving prompts that facilitate reflection and disclosure. This interpretation is consistent with evidence that personalisation may improve satisfaction, engagement, and dialogue quality in healthcare conversational agents [55]. Patients' generally favourable views about implementation should consequently be understood as conditional: they saw the prototype as a potentially useful preparation and data collection tool, but not as a substitute for physical examination, clinical judgement, or the therapeutic relationship.

From the physiotherapists' perspective, the main limitation was the uncertain clinical value. The moderate SUS score and comparatively favourable PEOU were accompanied by lower PU, suggesting that physiotherapists could understand and operate the prototype without necessarily being convinced that it would improve their work. Qualitative findings explained this distinction: reports were potentially valuable when they provided an initial overview, supported preparation, or captured information from patients who disclosed less readily in person, but their usefulness diminished when content was incomplete, repetitive, insufficiently prioritised, or inconsistent with the physiotherapist's assessment. Under these conditions, the prototype risked transferring clinical work, as physiotherapists still needed to identify relevant information, investigate omissions, and verify the generated interpretation. This emphasis on clinical value is consistent with health-technology adoption research, in which PU is a particularly consistent determinant of clinicians' intention to use a system, while design limitations, technical concerns, and additional time demands can constrain adoption even when a technology is relatively easy to use [56–58]. The findings therefore support neither universal deployment nor rejection of the prototype, but a more selective and conditional role in situations where preliminary information gathering could add value, such as initial consultations, limited prior clinical information, complex presentations, or difficulties with face-to-face disclosure. Refinement should consequently focus on making the report more concise, clinically prioritised, complete, and relevant to the consultation.

The prototype's potential contribution to ICF implementation lies in redistributing its preparatory work across the patient, the AI system, and the physiotherapist. In the

present study, the conversational interview created opportunities for patients to reflect on BPS aspects of their condition; however, the omissions, unsupported interpretations, and variable clinical relevance identified by physiotherapists show that using an ICF-informed structure does not by itself ensure a clinically meaningful assessment. Because the prototype mediates between the patient's account and the physiotherapist's interpretation, the report should distinguish patient-reported information from AI-generated organisation or interpretation, preserve traceability to the source responses, and make missing or uncertain information visible. These safeguards are particularly important because evaluations of LLM-generated clinical summaries identify factual accuracy, omissions, hallucinations, relevance, and groundedness as distinct properties [59].

The methodological contribution of this study lies in providing a worked example of a multi-actor formative evaluation of an AI-enabled clinical workflow. The evaluation examined the patient-AI-physiotherapist pathway by assigning complementary outcomes and methods to the actors' different roles: patients directly experienced the conversational interview, whereas physiotherapists interpreted its output and considered its potential clinical use. Combining behavioural indicators, standardised and study-specific measures, and open-ended accounts enabled different aspects of usability and UX to be examined, consistent with the view that no single method adequately captures the interactions among users, tasks, technologies, and clinical contexts [39]. Importantly, quantitative and qualitative findings were not merely reported in parallel: predefined criteria identified candidate priorities for refinement, qualitative analysis clarified their mechanisms, contextual conditions, and exceptions, and joint displays were used to examine convergence and divergence and formulate actionable meta-inferences. Such explicit integration is a defining strength of mixed-methods usability evaluation because it can generate insights beyond those obtainable from either data source alone [18, 19]. This process revealed findings that a single overall score could have obscured. The workflow simulation also preserved key elements of the intended pathway while allowing the evaluation to remain feasible at an early development stage, although it cannot substitute for testing within routine care. The resulting approach may therefore be useful for formative studies of healthcare technologies in which the person providing information and the professional interpreting the generated output are different users.

Several limitations should be considered when interpreting these findings. First, the small convenience samples were recruited from a single outpatient clinic and included participants with generally favourable digital literacy and, among physiotherapists, favourable baseline attitudes towards AI. The findings should therefore be interpreted as formative and descriptive and may not transfer to users with lower digital confidence, different clinical routines, or different organisational and linguistic contexts. Moreover, excluding people with severe sensory, motor, or cognitive limitations restricted the study's ability to examine accessibility and equity.

Second, embedding the prototype in simulated consultations did not reproduce the time pressures, competing demands, documentation requirements, or consequences of routine care; the study therefore did not examine sustained implementation. Furthermore,

open-ended survey responses enabled all participants to provide feedback but offered limited opportunities for probing and clarification. Finally, physiotherapists identified omissions, inconsistencies, and unsupported interpretations in the generated reports, but report quality was not independently evaluated against interview transcripts using predefined criteria. These observations indicate areas of potential clinical concern; they do not provide estimates of the frequency or severity of errors. Future evaluations should assess factual consistency, omissions, unsupported inferences, clinical relevance, groundedness, and ICF mapping through structured expert review, as recommended in emerging methods for evaluating LLM-generated clinical summaries [59, 60].

Conclusions

This formative mixed-methods evaluation found that the prototype was a promising but still immature tool. The prototype supported reflection and preliminary information gathering, yet its interaction and AI-generated reports require refinement before routine clinical use. Integrating patient and physiotherapist perspectives with quantitative and qualitative evidence identified important discrepancies between the experience of using the interview and the clinical value of its output. This multi-actor mixed-methods approach generated concrete proposals for the next development cycle.

List of abbreviations

AI: artificial intelligence BPS: biopsychosocial BI: behavioural intention ICF: International Classification of Functioning, Disability and Health PEOU: perceived ease of use PU: perceived usefulness SUS: System Usability Scale TAM: Technology Acceptance Model UEQ: User Experience Questionnaire

Declarations

Ethics approval and consent to participate

The study was conducted in accordance with the Declaration of Helsinki and the Swiss Human Research Act and Human Research Ordinance. Ethical approval was obtained from the competent ethics committee (Swissethics - Comitato Etico Cantonale Ticino) before study initiation (2025-02227, Rif CE TI 5003). Electronic informed consent was obtained from all participants before any study procedure.

Clinical trial number

Clinical trial number: not applicable.

Consent for publication

All participants provided informed consent for the use of their anonymized study data for research publication.

Availability of data and materials

The datasets generated and analysed during the current study are not publicly available because they contain potentially identifiable qualitative data, but anonymized data may be available from the corresponding author on reasonable request and subject to applicable ethical and institutional requirements.

Competing interests

Some authors (GM, FA, RF and BM) were involved in the design and development of the prototype evaluated in this study. This role was inherent to the study aim of conducting a formative evaluation of the system. The authors declare that they have no financial competing interests.

Funding

This work was supported by the competitive national funding Innosuisse with the Innovation Booster Co-Designing Human Services program and by an internal SUPSI voucher for competitive grant proposal preparation.

Authors' contributions

GM, FA, RF and BM designed the co-design phase and developed the prototype. GM and BM designed the study. GM, NT and PA collected the data. GM analyzed the quantitative data. GM and SA analyzed the qualitative data and made the mixed-methods integration. GM and BM wrote the manuscript. All authors contributed to the interpretation of the findings, critically revised the manuscript, and approved the final version.

Acknowledgements

The authors thank the patients and physiotherapists who participated in this study and Fisioterapia Kinesis for supporting study recruitment and the simulated consultation procedures. The authors also thank the professional association Physioswiss - Associazione Cantonale Ticino for having supported the co-design phase.

References

[1] Engel GL. The need for a new medical model: a challenge for biomedicine. Science. 1977;196(4286):129–36. https://doi.org/10.1126/science.847460.

[2] Wade DT, Halligan PW. The biopsychosocial model of illness: a model whose time has come. Clin Rehabil. 2017;31(8):995–1004. https://doi.org/10.1177/0269215517709890.

[3] World Health Organization. International Classification of Functioning, Disability and Health (ICF). Geneva, Switzerland: World Health Organization; 2001.

Available from: https://iris.who.int/server/api/core/bitstreams/fbb08b1d-ebe0-44a9-8108-c0d4476ebc5d/content.

[4] Biz MCP, Chun RYS. Operationalization of the International Classification of Functioning, Disability and Health, ICF, in a Specialized Rehabilitation Center. Codas. 2020;32(2):e20190046. https://doi.org/10.1590/2317-1782/20192019046.

[5] Johansen T, Kvaal AM, Konraethsdottir AD. Developing and Implementing ICF-Based Tools for Occupational Rehabilitation Supporting the Communication and Return to Work Process Between Sickness Absentees, Clinical Team and Jobcentre Contacts. Front Rehabil Sci. 2022;3:830067. https://doi.org/10.3389/fresc.2022.830067.

[6] Tempest S, Jefferson R. Engaging with clinicians to implement and evaluate the ICF in neurorehabilitation practice. NeuroRehabilitation. 2015;36(1):11–5. https://doi.org/10.3233/NRE-141185.

[7] Barradell S, Scholten I. How, and to what end, is the WHO-ICF framework represented in physiotherapy? Insights from a qualitative research synthesis. Physiother Theory Pract. 2025;41(4):792–809. https://doi.org/10.1080/09593985.2024.2370345.

[8] Maini M, Nocentini U, Prevedini A, Giardini A, Muscolo E. An Italian experience in the ICF implementation in rehabilitation: preliminary theoretical and practical considerations. Disabil Rehabil. 2008;30(15):1146–52. https://doi.org/10.1080/09638280701478397.

[9] Paltamaa J, van Lingen E, Haumer C, Kidritsch A, Aerts I, Mutanen L. Specific ICF training is needed in clinical practice: ICF framework education is not enough. Front Rehabil Sci. 2024;5:1351564. https://doi.org/10.3389/fresc.2024.1351564.

[10] Wiegand NM, Belting J, Fekete C, Gutenbrunner C, Reinhardt JD. All talk, no action?: the global diffusion and clinical implementation of the international classification of functioning, disability, and health. Am J Phys Med Rehabil. 2012;91(7):550–60. https://doi.org/10.1097/PHM.0b013e31825597e5.

[11] Bickenbach J, Cieza A, Selb M, Stucki G, editors. ICF Core Sets: Manual for Clinical Practice. 2nd ed. Göttingen, Germany: Hogrefe Publishing; 2021.

[12] World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva, Switzerland: World Health Organization; 2021. Available from: https://iris.who.int/handle/10665/341996.

[13] Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. J Med Internet Res. 2017;19(11):e367. https:

//doi.org/10.2196/jmir.8775.

[14] Harte R, Glynn L, Rodriguez-Molinero A, Baker PM, Scharf T, Quinlan LR, et al. A Human-Centered Design Methodology to Enhance the Usability, Human Factors, and User Experience of Connected Health Systems: A Three-Phase Methodology. JMIR Hum Factors. 2017;4(1):e8. https://doi.org/10.2196/humanfactors.5443.

[15] Laugwitz B, Held T, Schrepp M. Construction and Evaluation of a User Experience Questionnaire. In: Holzinger A, editor. HCI and Usability for Education and Work. vol. 5298 of Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2008. p. 63–76.

[16] International Organization for Standardization. ISO 9241-11:2018. Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. Geneva, Switzerland: International Organization for Standardization; 2018. Available from: https://www.iso.org/standard/63500.html.

[17] Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health. 2011;38(2):65–76. https://doi.org/10.1007/s10488-010-0319-7.

[18] Alwashmi MF, Hawboldt J, Davis E, Fetters MD. The Iterative Convergent Design for Mobile Health Usability Testing: Mixed Methods Approach. JMIR Mhealth Uhealth. 2019;7(4):e11656. https://doi.org/10.2196/11656.

[19] Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res. 2013;48(6 Pt 2):2134–56. https://doi.org/10.1111/1475-6773.12117.

[20] Kilfoy A, Hsu TC, Stockton-Powdrell C, Whelan P, Chu CH, Jibb L. An umbrella review on how digital health intervention co-design is conducted and described. NPJ Digit Med. 2024;7(1):374. https://doi.org/10.1038/s41746-024-01385-1.

[21] Chester EC, Robinson NC, Roberts LC. Opening clinical encounters in an adult musculoskeletal setting. Man Ther. 2014;19(4):306–10. https://doi.org/10.1016/j.math.2014.03.011.

[22] Boonen A, Braun J, van der Horst Bruinsma IE, Huang F, Maksymowych W, Kostanjsek N, et al. ASAS/WHO ICF Core Sets for ankylosing spondylitis (AS): how to classify the impact of AS on functioning and health. Ann Rheum Dis. 2010;69(1):102–7. https://doi.org/10.1136/ard.2008.104117.

[23] Cieza A, Schwarzkopf S, Sigl T, Stucki G, Melvin J, Stoll T, et al. ICF Core Sets for osteoporosis. J Rehabil Med. 2004;36(Suppl 44):81–6. https://doi.org/10.1080/16501960410016028.

[24] Cieza A, Stucki G, Weigl M, Disler P, Jackel W, van der Linden S, et al. ICF Core Sets for low back pain. J Rehabil Med. 2004;36(Suppl 44):69–74. https://doi.org/10.1080/16501960410016037.

[25] Cieza A, Stucki G, Weigl M, Kullmann L, Stoll T, Kamen L, et al. ICF Core Sets for chronic widespread pain. J Rehabil Med. 2004;36(Suppl 44):63–8. https://doi.org/10.1080/16501960410016046.

[26] Dreinhofer K, Stucki G, Ewert T, Huber E, Ebenbichler G, Gutenbrunner C, et al. ICF Core Sets for osteoarthritis. J Rehabil Med. 2004;36(Suppl 44):75–80. https://doi.org/10.1080/16501960410015498.

[27] Grill E, Ewert T, Chatterji S, Kostanjsek N, Stucki G. ICF Core Sets development for the acute hospital and early post-acute rehabilitation facilities. Disabil Rehabil. 2005;27(7-8):361–6. https://doi.org/10.1080/09638280400013974.

[28] Grill E, Zochling J, Stucki G, Mittrach R, Scheuringer M, Liman W, et al. International Classification of Functioning, Disability and Health (ICF) Core Set for patients with acute arthritis. Clin Exp Rheumatol. 2007;25(2):252–8.

[29] Hernandez-Lazaro H, Mingo-Gomez MT, Ceballos-Laita L, Medrano-de-la Fuente R, Jimenez-Del Barrio S. Validation of the international classification of functioning, disability, and health (ICF) core sets for musculoskeletal conditions in a primary health care setting from physiotherapists' perspective using the Delphi method. Disabil Rehabil. 2023;45(15):2458–2468. https://doi.org/10.1080/09638288.2022.2096128.

[30] Prodinger B, Cieza A, Oberhauser C, Bickenbach J, Ustun TB, Chatterji S, et al. Toward the International Classification of Functioning, Disability and Health (ICF) Rehabilitation Set: A Minimal Generic Set of Domains for Rehabilitation as a Health Strategy. Arch Phys Med Rehabil. 2016;97(6):875–84. https://doi.org/10.1016/j.apmr.2015.12.030.

[31] OpenAI. Introducing GPT-5.2 [Internet]; 2025. Available from: https://openai.com/index/introducing-gpt-5-2/.

[32] Etxaniz J, Azkune G, Soroa A, Lopez de Lacalle O, Artetxe M. Do Multilingual Language Models Think Better in English? [Preprint]; 2023. Available from: https://arxiv.org/abs/2308.01223. arXiv:2308.01223.

[33] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42(2):377–81. https://doi.org/10.1016/j.jbi.2008.08.010.

[34] Jacob C, Sezgin E, Sanchez-Vazquez A, Ivory C. Sociotechnical Factors Affecting Patients' Adoption of Mobile Health Tools: Systematic Literature Review and

Narrative Synthesis. JMIR Mhealth Uhealth. 2022;10(5):e36284. https://doi.org/10.2196/36284.

[35] Brooke J. SUS: A 'Quick and Dirty' Usability Scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. Usability Evaluation in Industry. London: Taylor & Francis; 1996. p. 189–194.

[36] Cheah WH, Mat Jusoh N, Aung MMT, Ab Ghani A, Mohd Amin Rebuan H. Mobile Technology in Medicine: Development and Validation of an Adapted System Usability Scale (SUS) Questionnaire and Modified Technology Acceptance Model (TAM) to Evaluate User Experience and Acceptability of a Mobile Application in MRI Safety Screening. Indian J Radiol Imaging. 2023;33(1):36–45. https://doi.org/10.1055/s-0042-1758198.

[37] Perski O, Short CE. Acceptability of digital health interventions: embracing the complexity. Transl Behav Med. 2021;11(7):1473–1480. https://doi.org/10.1093/tbm/ibab048.

[38] Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res. 2017;17(1):88. https://doi.org/10.1186/s12913-017-2031-8.

[39] Yen PY, Bakken S. Review of health information technology usability study methodologies. J Am Med Inform Assoc. 2012;19(3):413–22. https://doi.org/10.1136/amiajnl-2010-000020.

[40] Georgsson M, Staggers N. Quantifying usability: an evaluation of a diabetes mHealth system on effectiveness, efficiency, and satisfaction metrics with associated user characteristics. J Am Med Inform Assoc. 2016;23(1):5–11. https://doi.org/10.1093/jamia/ocv099.

[41] Schrepp M. User Experience Questionnaire Handbook [Handbook]; 2015. Available from: https://doi.org/10.13140/RG.2.1.2815.0245.

[42] Boland MR, Rusanov A, So Y, Lopez-Jimenez C, Busacca L, Steinman RC, et al. From expert-derived user needs to user-perceived ease of use and usefulness: a two-phase mixed-methods evaluation framework. J Biomed Inform. 2014;52:141–50. https://doi.org/10.1016/j.jbi.2013.12.004.

[43] Davis FD. User acceptance of information systems : the technology acceptance model (TAM). Ann Arbor, MI: University of Michigan, School of Business Administration, Division of Research; 1987. 529. Available from: https://hdl.handle.net/2027.42/35547.

[44] Bangor A, Kortum PT, Miller JT. An Empirical Evaluation of the System Usability Scale. International Journal of Human–Computer Interaction. 2008;24(6):574–594. https://doi.org/10.1080/10447310802205776.

[45] R Core Team. R: A Language and Environment for Statistical Computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing; 2024. Available from: https://www.R-project.org/.

[46] Hyzy M, Bond R, Mulvenna M, Bai L, Dix A, Leigh S, et al. System Usability Scale Benchmarking for Digital Health Apps: Meta-analysis. JMIR Mhealth Uhealth. 2022;10(8):e37290. https://doi.org/10.2196/37290.

[47] Lewis JR, Sauro J. Item Benchmarks for the System Usability Scale. Journal of Usability Studies. 2018;13(3):158–167.

[48] Sauro J, Lewis JR. Quantifying the User Experience: Practical Statistics for User Research. 2nd ed. Morgan Kaufmann; 2016. Available from: https://books.google.it/books?id=USPfCQAAQBAJ.

[49] Schrepp M, Hinderks A, Thomaschewski J. Construction of a Benchmark for the User Experience Questionnaire (UEQ). International Journal of Interactive Multimedia and Artificial Intelligence. 2017;4(4):40–44. https://doi.org/10.9781/ijimai.2017.445.

[50] Curtain C, Dröge K, Missaghieh-Poncet J, Salomón L. QualCoder (version 2.8.2) [Computer software]; 2025. Available from: https://qualcoder.org/doc/en/#citation.

[51] Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol. 2013;13:117. https://doi.org/10.1186/1471-2288-13-117.

[52] Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology. 2006;3(2):77–101. https://doi.org/10.1191/1478088706qp063oa.

[53] Haynes-Brown TK, Fetters MD. Using Joint Display as an Analytic Process: An Illustration Using Bar Graphs Joint Displays From a Mixed Methods Study of How Beliefs Shape Secondary School Teachers' Use of Technology. International Journal of Qualitative Methods. 2021;20:1609406921993286. https://doi.org/10.1177/1609406921993286.

[54] Sauro J, Lewis JR. What sample sizes do we need? Part 2: formative studies. In: Quantifying the User Experience: Practical Statistics for User Research. 2nd ed. Morgan Kaufmann; 2016. p. 143–183.

[55] Kocaballi AB, Berkovsky S, Quiroz JC, Laranjo L, Tong HL, Rezazadegan D, et al. The Personalization of Conversational Agents in Health Care: Systematic Review. J Med Internet Res. 2019;21(11):e15360. https://doi.org/10.2196/15360.

[56] Gagnon MP, Desmartis M, Labrecque M, Car J, Pagliari C, Pluye P, et al. Systematic review of factors influencing the adoption of information and communication technologies by healthcare professionals. J Med Syst. 2012;36(1):241–77. https://doi.org/10.1007/s10916-010-9473-4.

[57] Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform. 2010;43(1):159–72. https://doi.org/10.1016/j.jbi.2009.07.002.

[58] Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A Systematic Review of the Technology Acceptance Model in Health Informatics. Appl Clin Inform. 2018;9(3):604–634. https://doi.org/10.1055/s-0038-1668091.

[59] Croxford E, Gao Y, Pellegrino N, Wong K, Wills G, First E, et al. Current and future state of evaluation of large language models for medical summarization tasks. Npj Health Syst. 2025;2. https://doi.org/10.1038/s44401-024-00011-2.

[60] Asgari E, Montana-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med. 2025;8(1):274. https://doi.org/10.1038/s41746-025-01670-7.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx