

Original Paper

Patient-Centered Summarization Framework for AI Clinical Summarization: Mixed Methods Study

Maria Lizarazo Jimenez^{1*}, MD; Ana Gabriela Claros^{1*}, MD; Kieran Green¹, PhD; David Toro-Tobon^{1,2}, MD; Felipe Larios¹, MD; Sheena Asthana³, PhD; Camila Wenczenowicz¹, MD; Kerly Guevara Maldonado¹, MD; Luis Vilatuna-Andrango¹, MD; Ana Cristina Proano¹, MD; Satya Sai Sri Bandi¹, MBBS; Shubhangi Bagewadi¹, MBBS; Megan E Branda^{1,4}, MS; Misk Al Zahidy¹, MS; Saturnino Luz⁵, PhD; Mirella Lapata⁵, PhD; Juan P Brito^{1,2}, MD; Oscar J Ponce-Ponte^{1,6}, MD

¹Department of Medicine, Division of Endocrinology, Diabetes, Metabolism and Nutrition, Knowledge and Evaluation Research Unit, Care and AI Laboratory, Mayo Clinic, Rochester, MN, United States

²Division of Endocrinology, Diabetes, Metabolism, and Nutrition, Mayo Clinic, Rochester, MN, United States

³Centre for Health Technology, University of Plymouth, Plymouth, United Kingdom

⁴Division of Clinical Trials and Biostatistics, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, United States

⁵Usher Institute, University of Edinburgh, Edinburgh, Scotland, United Kingdom

⁶Derriford Hospital, University Hospitals Plymouth NHS Trust, Plymouth, England, United Kingdom

*these authors contributed equally

Corresponding Author:

Oscar J Ponce-Ponte, MD

Department of Medicine, Division of Endocrinology, Diabetes, Metabolism and Nutrition, Knowledge and Evaluation Research Unit, Care and AI Laboratory

Mayo Clinic

100 2nd St SW

Rochester, MN 55905

United States

Phone: 1 (507) 284-2511

Email: ponceoscarj@gmail.com

Abstract

Background: Large language models (LLMs) are increasingly demonstrating the potential to reach human-level performance in generating clinical summaries from patient-clinician conversations. LLMs are usually evaluated against clinical summaries that focus mainly on patients' biology and not on their biography (eg, preferences, values, wishes, and concerns). To achieve patient-centered care, artificial intelligence clinical summarization must incorporate patient-centered domains, implemented through patient-centered summaries (PCSs).

Objective: This study aimed to develop a framework to generate PCS that capture patients' values, preferences, and wishes while ensuring clinical utility for clinicians, and assess if current open-source LLMs can achieve human-level performance in generating PCS.

Methods: We developed a 4-step mixed methods process to define and evaluate PCS. First, 2 patient and public involvement and engagement groups were convened in the United Kingdom (10 patients and 8 clinicians), who participated in semistructured interviews exploring what personal and contextual information should be included in clinical summaries and how it should be structured for clinical use. Second, findings were translated into an annotation guideline, which was used by 8 clinician annotators to generate gold standard PCS from 88 transcribed patient-clinician consultations about the management of atrial fibrillation. Third, 16 consultations were used to iteratively develop and refine a prompt aligned with the annotation guideline. Finally, 5 LLMs (Llama-3.2-3B [Meta AI], Llama-3.1-8B [Meta AI], Mistral-8B [Mistral AI], Gemma-3-4B [Google DeepMind], and Qwen3-8B [Alibaba]) generated summaries from 72 consultations using zero-shot and few-shot prompting, which were evaluated against gold standard PCS using ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation–Longest Common Subsequence) and BERTScore (Bidirectional Encoder Representations from Transformers Score) and assessed for correctness, completeness, conciseness, patient-centeredness, and fluency.

Results: Patients emphasized that summaries should include (1) lifestyle routines and daily functioning as indicators of independence or disruption; (2) the presence and role of social support systems, especially during crises; (3) recent life events

or stressors, such as trauma, loss, or caregiving demands; and (4) care preferences, values, and communication styles that provide meaning or reflect autonomy. Clinicians sought summaries that included a concise functional baseline, psychosocial context, and emotional cues, preferably in a structured, clinically digestible format. In the 72 consultations (mean age 70, SD 11 y; 32/72, 44.4% female), the best zero-shot performance was observed with Mistral-8B (ROUGE-L 0.189) and Llama-3.1-8B (BERTScore 0.673). The best few-shot prompting was found with 3 examples using Llama-3.1-8B (ROUGE-L 0.206 and BERTScore 0.683).

Conclusions: The open-source LLMs we evaluated did not achieve human-level performance in generating PCSs. Without task-specific fine-tuning, current open-source LLMs cannot reach human-level performance in this task. Our framework serves as an innovative guideline for developing gold standard PCS for artificial intelligence clinical tasks.

J Med Internet Res 2026;28:e87061; doi: [10.2196/87061](https://doi.org/10.2196/87061)

Keywords: patient-centered; large language models; LLMs; qualitative research; mixed methods; artificial intelligence

Introduction

Clinical summarization is an essential component of modern health care, as it allows clinicians to efficiently aggregate, organize, and synthesize complex patient data into concise, actionable insights that support clinical reasoning, decision-making, and care coordination [1]. High-quality summaries help reduce information overload, promote safer care transitions, and improve communication among providers and with patients [2-6]. Yet, preparing these summaries is time-consuming and contributes to documentation burden, which has motivated the development of artificial intelligence (AI) systems to automate the task [7-9]. These systems use natural language processing to generate concise summaries from electronic health records (EHRs) and patient-clinician encounters [10,11]. Recent work in clinical dialogue summarization has implemented multistage pipelines and fine-tuned large language models (LLMs) to improve the accuracy and coherence of patient-clinician summaries, reaching almost human-level performance [11-13].

However, current AI models are trained on clinical notes that have increasingly evolved to capture biological complexity as well as provide enough documentation to justify billing and reimbursement [14-16]. This dual purpose has shaped current summaries and the datasets used to guide current AI scribes, making them biologically centered and prioritizing pathophysiological details over patient preferences, their context, and what truly matters to them [13]. As a result, these models often produce clinical summaries that focus on medical issues and the clinician's perspective, without fully incorporating patients' values, preferences, and concerns [17, 18].

Different forms of clinical summarization have demonstrated how patient-centered information can be integrated into routine care while having an impact on care outcomes [19-21]. For instance, including both patients and clinicians' perspectives on after-visit summaries provides actionable instructions, medication explanations, and tailored self-management guidance that support patient understanding and engagement [19]. In rheumatology, written consultation summaries cocreated with patients support continuity of care and improve health literacy [20]. Similarly, tools such as surveys or questionnaires designed to elicit patient "values" have been transformed into narrative summary reports

to communicate patients' values to health care professionals and to facilitate values-based discussions and shared decision-making (SDM) for chronic conditions, improving alignment between care plans and patient preferences and reducing decisional conflict [21]. Despite these examples, there is currently no structured framework that defines how to systematically generate and evaluate patient-centered summaries (PCSs) that capture patients' values and preferences while preserving clinical utility [22,23]. Additionally, the ability of current open-source LLMs to generate PCS has not yet been evaluated [13].

We conceptualize patient-centered summarization as a distinct subtype of AI clinical summarization, particularly relevant for patients with multimorbidity whose care decisions depend on goals, preferences, and lived context rather than condition-specific biomedical information alone [24]. Unlike traditional clinical summarization, which prioritizes biomedical extraction for billing and documentation [15,16,25], PCS emphasizes preferences, values, wishes, and concerns to support longitudinal decision-making and SDM [21,23]. This subtype is characterized by the need to capture both biomedical and patient-centered information, guided by structured patient-centered domains and requiring evaluation beyond conventional summarization metrics.

To ensure that PCS capture what matters to patients while remaining clinically useful, their definition must be context-dependent and informed by both patient and clinician perspectives through iterative feedback, enabling the development of adaptable, evolving gold standards for local benchmarking or model refinement [26,27]. Because this process can be time- and resource-intensive, we propose a patient and public involvement and engagement (PPIE)-driven approach and demonstrate the feasibility of this methodology [27-29]. Specifically, we (1) introduce a novel mixed methods framework for developing a context-aware, patient-centered, gold standard for clinical summarization, and (2) benchmark current open-source LLMs against gold standard PCS to establish a baseline for future improvement in this area.

Methods

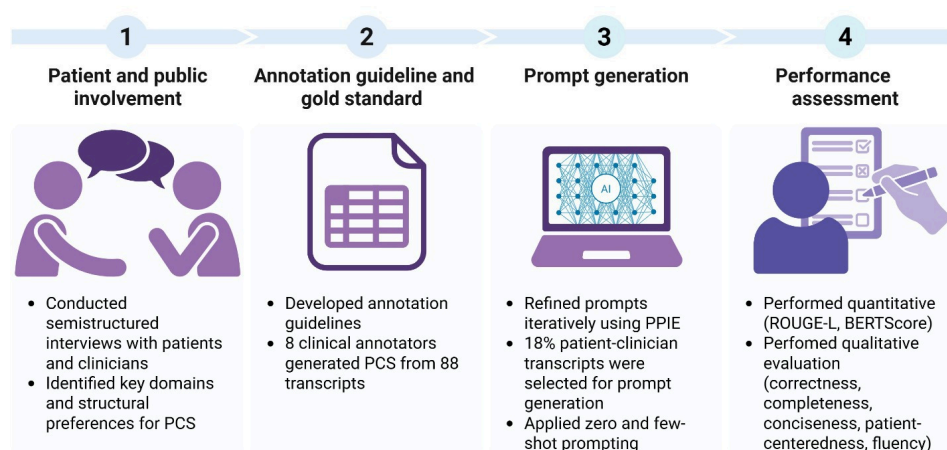
Overview

The framework to define, generate, and evaluate PCS followed a 4-step mixed methods design. First, we captured patients' and clinicians' perspectives into PCS by leveraging the PPIE methodology; 2 PPIE groups were convened to define the core components of a PCS through semistructured interviews. Second, these findings were translated into an annotation guideline, which 8 clinician annotators used to create gold standard PCS for 88 transcribed patient-clinician consultations. Third, a prompt was iteratively developed using 16 consultations and then used to generate PCS from

5 general-purpose, open-source LLMs (Llama-3.2-3B [Meta AI], Llama-3.1-8B [Meta AI], Mistral-8B, Gemma-3-4B [Google DeepMind], and Qwen3-8B [Alibaba Cloud]) with zero-shot and few-shot prompting techniques. Notably, none of these models were pretrained on clinical summaries from EHRs. Finally, the AI-generated summaries from the remaining 72 consultations were evaluated against the gold standard PCS using a quantitative analysis (ROUGE-L [Recall-Oriented Understudy for Gisting Evaluation–Longest Common Subsequence] and BERTScore [Bidirectional Encoder Representations from Transformers Score]) and a qualitative human assessment across 5 domains, including correctness and patient-centeredness. [Figure 1](#) provides an overview of the study workflow.

Figure 1. Overview of the 4-step study workflow for developing and evaluating patient-centered summaries (PCSs). The process included (1) patient and public involvement and engagement to define PCS domains, (2) development of annotation guidelines and creation of 88 gold standard PCSs, (3) prompt refinement using 16 consultations, and (4) benchmarking 5 open-source large language models on 72 held-out consultations using ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation–Longest Common Subsequence), BERTScore (Bidirectional Encoder Representations from Transformers Score), and qualitative human evaluation. PCS: patient-centered summary; PPIE: patient and public involvement and engagement.

Study methods



First Step

Patient and Public Involvement

Two PPIE groups were recruited from Devon, a southwest coastal region in the United Kingdom. Eligible participants included adult patients (aged 18 y or older) with capacity to provide informed consent and engage in PPIE discussions, as well as clinicians who regularly conducted face-to-face clinical appointments. Patients were excluded if they lacked decision-making capacity or could not meaningfully participate. Clinicians were excluded if they did not routinely provide direct patient care, meaning those with continuous contact with patients in the outpatient clinic.

Semistructured interviews for the PPIEs were led by KG, who developed and iteratively improved and simplified the interview questions with input from experts in patient-centeredness. These questions explored the importance of capturing (1) personal and lifestyle information, (2) support systems and beliefs, (3) health goals and challenges, (4) emotional and mental health, (5) sources of meaning, and for clinicians, an additional question about how they preferred to receive

and prioritize the information discussed from an appointment. These domains were derived and synthesized from integrative models of patient-centered care and from expert-validated frameworks that prioritize dimensions, such as the unique person, patient involvement, provision of information, patient-clinician communication, and empowerment [30,31]. These questions were implemented so that patients discussed what personal and contextual information should be captured in clinical summaries, while clinicians reflected on how such information could be structured for clinical use.

Data Analysis

Interview notes were handwritten during or shortly after sessions and thematically coded using the 5 patient-centered domains, along with an additional category for structural preferences. Data were stratified by participant age groups (young [18–30 y], middle-aged [31–60 y], and older [61–91 y]) and role (patient or clinicians) and analyzed through iterative thematic analysis to identify common and divergent priorities. Relevant quotes were categorized and introduced into each subheading. The number of interviews was

determined by the relevance and depth of contributions rather than by strict saturation thresholds, consistent with PPIE research guidelines [28,29]. Because the objective was to inform the development of the patient-centered summarization framework rather than exhaustively identify themes, interviews continued until additional data were unlikely to meaningfully alter the structure or domains of the framework, consistent with the concept of information power [32]. Data collection occurred within 1 month and depended on the availability of general practice sites able to support participation.

Second Step

Annotation Guideline and Gold Standard PCS

The results from the first step were used to develop the annotation guideline, which incorporated input from both the clinicians' PPIE to guide the desired format and the patients' PPIE to provide illustrative quotes and categories within each key area, as illustrated in [Multimedia Appendix 1](#).

A team of 8 clinician annotators applied the guideline after a calibration phase to ensure shared understanding. In this phase, pairs worked on the same transcript, identifying background topics, issues, the plan, and patient-centered elements, and generating summaries until their outputs showed similar structure and complete inclusion of PCS elements. Disagreements were resolved through discussion.

Annotators generated 88 PCS from nonscripted patient-clinician consultations obtained from a SDM clinical trial in atrial fibrillation [33]. A total of 922 encounters were originally recorded. Scripted encounters from the intervention arm were excluded because clinicians followed the decision aid step by step, limiting exploration of other aspects of patients' lives. Encounters were also excluded when essential components of the consultation structure, such as a clear discussion and agreement on a management plan, were missing. After applying these criteria, the team purposively selected 88 eligible encounters to ensure diversity in patient and clinician characteristics as well as in communication styles. Original clinical notes were not used as a comparator, as they are clinician-oriented and focus primarily on biomedical aspects. In contrast, transcriptions captured the full dialogue, including patients' perspectives and contextual details, providing a comprehensive and ethically accessible source for developing PCS.

Data Analysis

We quantified interannotator agreement using percent agreement and Gwet AC1, consistent with previous clinical annotation studies, which adjust for chance agreement while being less sensitive to label imbalance and the κ paradox [34,35]. For each consultation, 2 annotators independently generated summaries. Interrater agreement was calculated by coding the presence of information items and patient-centered elements using binary ratings (1=present, 0=absent). Percent agreement and Gwet AC1 were calculated across all coded elements. Final mean percent agreement was 92.6% with an

overall mean Gwet AC1 of 0.91 (SD 0.12), and 92% and 0.88 for patient-centered elements, indicating excellent reliability.

Third Step

Prompt Generation

Out of 88 transcribed patient-clinician conversations, we selected 16 (18%) for prompt generation. The prompt [Multimedia Appendix 2](#) was iteratively improved based on the annotation guideline, with instructions for generating each section of the clinical summary refined through comparison between AI-generated outputs and gold standard PCS.

Open-source models were selected because the study data included sensitive real-world clinical conversations, and ensuring full deidentification of personally identifiable information would have introduced additional privacy risks and administrative burdens. The use of closed-source or externally hosted LLMs is constrained by institutional data governance policies, as these systems require transmission of data to third-party servers. Although some institutions permit their use under specific agreements, such arrangements limit reproducibility and broader applicability across settings. Accordingly, proprietary models and certain recently released systems (eg, DeepSeek) were excluded under our institution's strict information security and vendor risk management policies. To ensure compliance with data privacy requirements and maximize generalizability in real-world health care settings, all selected models supported local, offline inference, allowing full control over data handling and ensuring data security.

Data Analysis

Performance during the prompt refinement phase was assessed using a dual approach. First, we performed a qualitative alignment check against the gold standard summaries. Second, quantitative metrics, specifically ROUGE-L for lexical similarity and BERTScore for semantic alignment, were calculated to provide objective feedback for iterative prompt adjustments

Fourth Step

Performance Assessment

The 5 selected open-source LLMs generated summaries from 72 transcribed patient-clinician consultations using zero-shot and few-shot (1-3 examples) prompting. For all models, hyperparameters were standardized, with temperature fixed at 1 and the maximum number of new tokens set to 2048, as the gold standard PCS did not exceed this length. The study evaluated pretrained open-source models without additional training or fine-tuning. During inference, we used a batch size of 1 and greedy search as the decoding strategy and used a single NVIDIA L4 GPU with 24 GB of memory.

Data Analysis

Quantitative performance was evaluated against the gold standard PCS using ROUGE-L for lexical similarity and BERTScore for semantic alignment. ROUGE-L was calculated using the standard F_1 formulation, which measures

the longest common subsequence between the model output and the reference summary; scores range from 0 to 1, with higher values indicating greater overlap (eg, a ROUGE-L of 0.2 reflects relatively limited lexical similarity) [36]. BERTScore, also normalized between 0 and 1, was computed using contextual embeddings to capture semantic similarity; higher scores reflect closer alignment in meaning between system and reference summaries [37].

In the descriptive analysis, we first quantified the amount of patient-centered information present in the source transcripts. Each consultation was manually labeled according to our patient-centered framework, and the number of patient-centered statements per encounter was recorded. Across the 72 consultations, this number ranged from 0 to 16.

A qualitative assessment was performed using the framework proposed by Van Veen et al [38], which we adapted to evaluate summaries across 5 domains—correctness, completeness, conciseness, patient-centeredness, and fluency. To assess patient-centeredness, we identified which elements of the PCS framework were present in each summary using the labeled transcripts and quantified the amount of patient-centered content included. Furthermore, 5 pilot summaries were then rated and discussed to calibrate the approach, and afterward 2 reviewers (MLJ and AGC) independently rated all summaries. This assessment was conducted in a blinded strategy, as evaluators did not know whether the summaries had been generated by the model or by human experts. For this purpose, summaries were anonymized, randomized, and labeled only as summary A and summary B. Each domain was scored on a scale from –5 to +5, where negative values favored the gold standard PCS, positive values favored the AI-generated summaries, and 0 indicated no difference; directionality was assigned after unblinding to identify which summary corresponded to each source. Scoring was guided by predefined rules; for correctness, we reviewed whether the information in each summary was factually accurate according to the transcript. After unblinding, instances where the model introduced information that was not present in the source, such as inventing medications or clinical details without supporting context, were classified as hallucination and rated –5 if more than 2 instances were present or –3 if 1 instance was identified. For patient-centeredness, we compared summaries with tagged transcripts and assigned 0 if both contained a similar amount of patient-centered content, –3 or +3 if 1 included less than

half of the tagged content, and –5 or +5 if 1 included more than half, depending on which summary achieved this.

To assess interrater reliability between the 2 evaluators (MLJ and AGC), we conducted a pilot evaluation using 3 videos, each rated across 5 parameters on an ordinal scale ranging from –5 to +5. Interrater agreement was quantified using the quadratic-weighted Cohen κ coefficient, which accounts for the ordinal nature of the data by assigning smaller penalties to minor rating differences. The resulting κ value was 0.84, indicating excellent agreement [39].

Ethical Considerations

The study was conducted in accordance with the International Conference on Harmonization Good Clinical Practice guidelines. The study protocol was approved by the Mayo Clinic Institutional Review Board (IRB#25-007747) and included approval for the secondary analysis of data derived from the atrial fibrillation clinical trial. Written informed consent was obtained from all participants involved in both the PPIE groups and the original clinical trial.

Regarding participant compensation, patients did not receive direct financial remuneration for their participation in the PPIE groups or the original clinical trial. However, no additional travel, logistical, or incidental expenses were incurred outside of their routine clinical care. No participant suffered financial detriment. Clinicians were compensated for their time according to a fixed hourly institutional rate to offset the direct loss of billable clinical hours required for participation.

Results

Patient-Centered Summarization Framework

Through the PPIE participants, including 8 clinicians and 10 patients representing diverse age groups (Table 1), we identified essential themes to guide patient-centered summarization according to patients (Table 2 and Multimedia Appendix 3). Their insights informed the development of a structured framework with the following five domains: (1) lifestyle and daily routines, (2) support systems and beliefs, (3) health goals and challenges, (4) emotional and mental health, and (5) sources of meaning.

Table 1. Demographics of PPIE^a participants. The table summarizes the composition of the PPIE groups, including clinicians and patients, stratified by age group, sex distribution, and percentage of the total sample.

Group	Age range (y)	Sex, n (% of group)	Total, n (% of all)
General practitioners	— ^b	—	7 (38.9)
District nurses	—	—	1 (5.5)
Young patients	18-30	• Male: 2 (66.7) • Female: 1 (33.3)	3 (16.7)
Middle-aged patients	31-60	• Male: 1 (33.3) • Female: 2 (66.7)	3 (16.7)

Group	Age range (y)	Sex, n (% of group)	Total, n (% of all)
Older patients	61-91	- Male: 3 (75) - Female: 1 (25)	4 (22)
Total patients	—	- Male: 6 (60) - Female: 4 (40)	10 (55.5)

^aPPIE: patient and public involvement and engagement.

^bNot available.

Table 2. Themes identified using the patient-centered framework. Themes identified from the patient and public involvement and engagement interviews, differentiated by older, middle-aged, and younger patients. The table summarizes perspectives across 5 domains: lifestyle and daily routines, support systems and access, events and life stressors, care preferences, and sources of meaning or value.

Category	Older patients (61-91 y)	Middle-aged patients (31-60 y)	Younger patients (18-30 y)
Lifestyle and daily routines	Active in managing health (eg, monitoring blood pressure and volunteering) [Patient A, C]	Shared limited lifestyle details (eg, physical job and retired builder) [Patients H, I]	Limited lifestyle information; concerns about chronic pain [Patient E]
Support systems and access	Reliance on friends for transport; more care needed from loved ones [Patient C]	Family support during crises; challenges with relatives' health [Patients H, I]	Emotional support from family and friends; support sometimes is hard because close friends are often busy [Patient D]
Events and life stressors	Bereavement and chronic illness impact	Bereavement and family illness noted but not shared often	Extrinsic factors like trauma, relationships, and work stress
Care preferences	Preference for continuity and face-to-face care; avoid technology for engagement [Patient J]	Limited specific preferences	Specific preferences based on condition (eg, medication avoidance) [Patients D]
Sources of meaning or value	Community contributions and self-management	Less emphasis on sources of meaning unless stressed	Importance of work and family as motivators

Lifestyle and Daily Routines

Older patients emphasized the importance of documenting daily routines as indicators of independence and health engagement. For example, they shared efforts to manage blood pressure, engage in volunteering, or maintain physical activity despite mobility challenges. In contrast, middle-aged and younger patients provided fewer details unless health issues clearly affected their routines. When symptoms disrupt daily functioning, such as chronic pain or fatigue, younger patients consider lifestyle information more relevant to record.

Support Systems and Beliefs

Across age groups, patients valued recording the presence and quality of social support, especially during health crises. Older patients described logistical challenges, such as transportation and reliance on friends, while younger and middle-aged patients referred to emotional support from family as relevant, although sometimes inadequate. These narratives underscored how both practical and emotional support systems influenced access to care and coping strategies.

Health Goals and Challenges

Patients of all ages described life stressors affecting their physical and emotional well-being. Younger participants highlighted the impact of trauma, relationship conflict, caregiving roles, and system-level stressors (eg, child protection fears). Older patients emphasized bereavement or caregiving burdens as key stressors. Some participants expressed a desire for clinicians to recognize how these

challenges affected motivation, resilience, and engagement with care.

Emotional and Mental Health

Preferences for emotional support and mental health disclosure varied. Some younger patients shared significant psychological distress, including suicidality or depression linked to life events. Others wanted clinicians to recognize emotional triggers but preferred sensitive handling. Older patients voiced frustration when distress was dismissed, especially when their usual coping mechanisms failed. The findings revealed that acknowledging emotional states is crucial to understanding health behavior and care needs.

Sources of Meaning and Personal Values

Older patients identify meaning through community engagement, independence, and contribution, with some explicitly linking these to care preferences, such as avoiding residential care or choosing not to be resuscitated. Middle-aged and younger participants are less often named sources of meaning, unless those values (eg, work and parenting) were under threat. When identified, these sources of meaning shaped their outlook, motivation, and desire for personalized care planning.

Clinicians also expressed strong support for integrating person-centered information into clinical summaries, emphasizing the importance of understanding patients' functional baselines, psychosocial context, and emotional well-being. Many described their role not only as medical providers but also as informal therapists or coaches, often encountering complex cases rooted in bereavement, loneliness, trauma, or caregiving stress. Several clinicians advocated for routinely documenting elements such as

support networks, mental health history, and care preferences (eg, avoiding invasive treatments), as well as patients' ideas, concerns, and expectations.

At the same time, clinicians noted practical considerations for summarization and implementation. Most preferred brief, structured summaries using bullet points, incorporating psychosocial data when relevant. Suggestions for system integration included embedding person-centered content into existing EHR tabs or flagging key insights with visual icons. Some clinicians proposed advanced uses of AI to visualize longitudinal patterns of psychosocial and health-related events or to generate personalized follow-up letters that were humane and accessible. However, a few raised philosophical concerns about whether AI could meaningfully capture the relational and dynamic aspects of healing, particularly the transformative role of the patient-clinician interaction itself.

Baseline Characteristics

A total of 72 transcribed patient-clinician consultations from the atrial fibrillation trial were analyzed, with general characteristics described in Table 3. The mean patient age was 70 (SD 11) years, and 44.4% (32/72) were female. Most consultations took place at Mayo Clinic Rochester (25/72, 34.7%). While all encounters were related to atrial fibrillation, they occurred in different clinical settings, most commonly cardiology appointments (40/72, 55.6%), followed by thrombophilia clinic (9/72, 12.5%). The most prevalent comorbidity was a history of hypertension, reported in 87.5% (63/72) of participants.

Table 3. Demographic and clinical characteristics of patients participating in the atrial fibrillation conversations. Characteristics are reported for the 72 patient-clinician encounters included in the analysis, including demographic variables, clinical context, appointment characteristics, and care settings.

General characteristics	Total (N=72)
Age (y), mean (SD)	70 (11)
Sex, n (%)	
Female	32 (44.4)
Male	40 (55.6)
BMI (kg/m ²), mean (SD)	33.0 (7.71)
Location, n (%)	
Mayo Clinic, Rochester, Minnesota	25 (34.7)
Park Nicollet, Minnesota	19 (26.4)
Hennepin County Medical Center	12 (16.7)
Alabama	4 (5.6)
Mississippi	12 (16.7)
Appointment length (min), mean (SD)	33.8 (17.5)
Appointment type, n (%)	
Emergency department	1 (1.4)
Primary care or family medicine	4 (5.6)
Inpatient	4 (5.6)
Cardiology	40 (55.6)
Thrombophilia clinic	9 (12.5)
Parkside ambulatory	2 (2.8)
Brooklyn Center Ambulatory center	7 (9.7)
Anticoagulation Clinic	3 (4.2)
Other	2 (2.8)
History of	
Hypertension, n (%)	63 (87.5)
Congestive heart failure, n (%)	13 (18.1)
Stroke, n (%)	12 (16.7)
Stroke type, n (%)	
TIA ^a	3 (25)
Ischemic stroke	9 (75)
Vascular disease, n (%)	21 (29.2)
Diabetes, n (%)	30 (41.7)
Renal disease, n (%)	10 (13.9)

General characteristics	Total (N=72)
Liver disease, n (%)	4 (5.6)
MI ^b , n (%)	5 (6.9)
PAD ^c , n (%)	2 (2.8)

^aTIA: transient ischemic attack.
^bMI: myocardial infarction.
^cPAD: peripheral artery disease.

Quantitative Evaluation

In the quantitative evaluation, the best performance in the zero-shot setting was achieved by Mistral-8B on ROUGE-L (0.189) and Llama-3.1-8B on BERTScore (0.673). As few-shot attempts were introduced, performance improved across all models. With 1 example, Llama-3.1-8B outperformed all others, achieving a ROUGE-L of 0.201 and a BERTScore of 0.680. The best-performing model overall was Llama-3.1-8B with 3-shot prompting, which achieved a ROUGE-L of 0.206 and a BERTScore of 0.683.

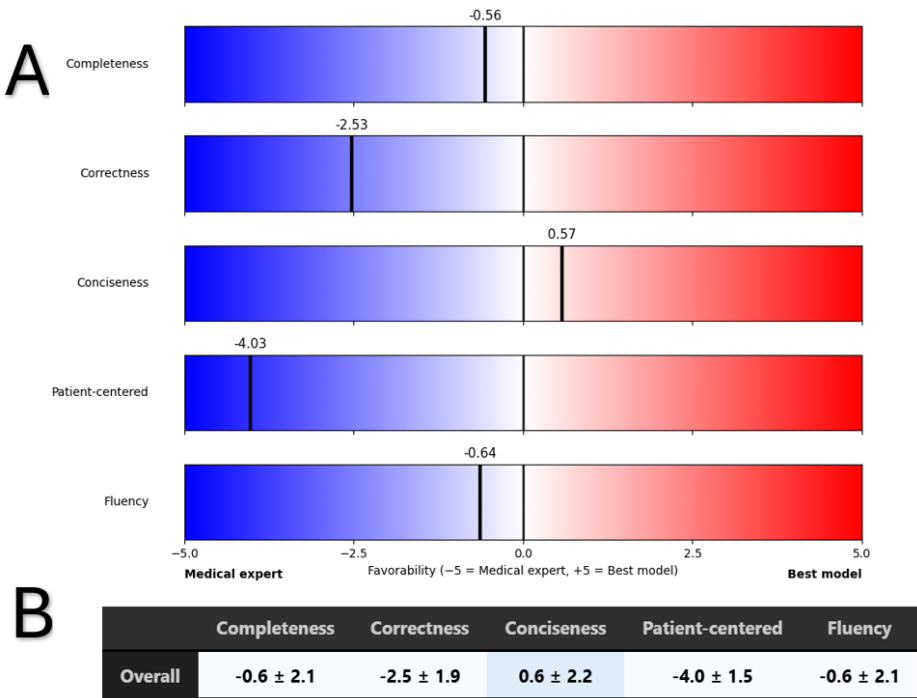
Qualitative Evaluation

Only 1 consultation had no patient-centered content, while among the remaining 71, 10 included a single domain and 61 incorporated multiple domains. Patient-centered topics spanned lifestyle and daily routines, support systems, life stressors, care preferences, and sources of meaning.

The qualitative human evaluation compared the AI-generated summaries from the best-performing model against

the gold standard PCS (Figure 2). The 2 approaches performed similarly in terms of completeness (mean -0.6, SD 2.1) and fluency (mean -0.6, SD 2.1). However, the gold standard PCS were rated significantly higher for correctness (mean -2.5, SD 1.9), as the model occasionally hallucinated procedures and medications not mentioned in the consultations. Conciseness slightly favored the model (mean 0.6, SD 2.2). The most substantial difference was in patient-centeredness, where gold standard PCS vastly outperformed the model (mean -4.0, SD 1.5), reflecting their superior ability to capture the nuanced details patients shared about their daily routines, support networks, and personal values. On average, transcripts contained (mean 3.6, SD 2.5) patient-centered domains, whereas AI-generated summaries captured only (mean 0.8, SD 0.9), and nearly half (35/72, 49%) contained none. Illustrative qualitative comparisons between gold standard and AI-generated PCS are provided in Multimedia Appendix 4.

Figure 2. Qualitative comparison between artificial intelligence-generated summaries and gold standard patient-centered summaries across consultations (n=72). (A) Diverging bar plots show mean domain scores (range -5 to +5) for completeness, correctness, conciseness, patient-centeredness, and fluency; negative values favor the gold standard summaries and positive values favor the artificial intelligence-generated summaries. (B) Table summarizing mean±SD scores for each domain. Gold standard summaries showed higher correctness and patient-centeredness, while conciseness slightly favored the model.



Discussion

Principal Findings and Relevance

In this study, we developed a PCS framework informed by PPIE and evaluated the performance of open-source LLMs in capturing patient-centered elements in clinical summaries. Through PPIE, patients identified 5 key domains as essential to their care—lifestyle and daily routines, support systems, health goals and challenges, emotional and mental health, and sources of meaning and personal values. Clinicians similarly supported the inclusion of patient-centered information, highlighting the importance of understanding patients' functional baselines and psychosocial context while emphasizing the value of brief, structured, and clinically accessible summaries. When applied to real-world atrial fibrillation consultations, patient-centered information was better captured in the gold standard PCS and was inconsistently present in AI-generated summaries, with a mean patient-centeredness rating of -4.0 (SD 1.5) on the -5 to $+5$ evaluation scale. Although commonly used summarization metrics, such as ROUGE-L and BERTScore, were consistently low across all AI-generated PCS, qualitative evaluation of the best AI-generated PCS with few-shot prompting were comparable with the gold standard PCS in conciseness, completeness, and fluency but were inferior in correctness (ie, hallucinations) and patient-centered information.

Standard clinical AI summarization research primarily targets clinician-oriented documentation, such as visit or discharge notes, emphasizing biomedical information extraction and documentation efficiency [13,18]. For example, recent studies generate structured clinical notes from patient-clinician conversations using LLMs and prioritize diagnoses, medications, and care plans for clinician use, while earlier approaches focused on compressing clinical records for professional documentation [40–42]. In contrast, our PCS framework explicitly incorporates patient-identified domains, positioning PCS as a methodologically distinct subtype of AI clinical summarization, rather than a conventional documentation-focused task.

Previous research across multiple health care settings demonstrates the value of patient-centered written summaries for supporting patient understanding and engagement. In outpatient care, patient-centered after-visit summaries have been associated with improved patient awareness and medication-related behaviors, although they frequently exhibit suboptimal design features, including high reading grade levels and poor usability [19,43]. Similarly, the introduction of patient-centered information in cancer consultation summaries has been linked to improved patient-centered communication and perceived quality of care, yet these summaries remain inconsistently implemented in practice [44,45]. In this context, creating a rapid, contextualized approach to create a framework to generate PCS might be impactful for patient care outcomes.

The domains identified in our framework align with previous work aiming to incorporate holistic patient

perspectives into care documentation and planning [46,47]. For instance, in our PCS framework, older participants emphasized independence and daily functioning, reflecting well-documented age-specific stress perception and priorities. As highlighted by the American Psychological Association, declines in functional abilities increase reliance on environmental supports to maintain autonomy and quality of life, reinforcing older adults' focus on daily routines and health-related function [48]. In contrast, younger participants in our study discussed emotional burdens and system-level stressors, including work, finances, and relationships, patterns consistent with evidence that younger adults report higher total daily stress and are more affected by interpersonal and occupational demands [49,50]. These age-specific differences between older adults (eg, independence) and younger adults (eg, occupational stress) imply that a “one-size-fits-all” AI prompting is insufficient. Future models should be able to adapt to each patient's values and preferences to generate PCS.

In our study, clinicians expressed strong support for systematically capturing person-centered data. They emphasized that understanding patients' functional baselines, emotional challenges, and preferences are central to care, according to the biopsychosocial model widely encouraged in primary care literature [23]. At the same time, clinicians highlighted practical constraints, advocating for summaries that are brief, have structure, and are clinically accessible. They suggested integrating person-centered content into discrete sections of the EHR (eg, icons or lifestyle tabs). These preferences align with previous studies where clinicians have expressed the need for EHR systems to facilitate intuitive, less burdensome documentation interfaces and to allow for both structured and narrative content, supporting both efficient data entry and the nuanced, person-centered aspects of care [51].

In our evaluation, gold standard PCS were notably stronger in patient-centeredness and correctness compared with AI-generated summaries, underscoring key limitations of current LLMs in clinical summarization. Despite modest gains with few-shot prompting, the models showed limited lexical similarity (ROUGE-L) and semantic alignment (BERTScore) with gold standard PCS. These findings contrast with previous work demonstrating strong LLM performance in biomedical clinical summarization and suggest that current models remain poorly equipped to capture personal, emotional, and value-based information [38]. This limitation likely reflects the predominance of biomedical and pathophysiological content in model training data [52]. The lack of dedicated datasets, annotation standards, and evaluation tools for PCS further contributes to this problem, making it difficult for models to capture the social, emotional, and functional information that matters most to patients and clinicians.

Hallucinations and factual inaccuracies remain a known limitation of LLMs and may introduce clinically unsafe information, such as incorrect medications, diagnoses, or management plans [53]. In this study, factual correctness was systematically evaluated by comparing the information

from the generated summaries either by AI or humans to the information from the transcribed conversation. The addition of information not present on the transcriptions (ie, hallucinations) was more prevalent in the AI-generated PCS. Therefore, these summaries should not be used without clinician oversight.

Limitations and Future Directions

Several limitations must be considered. First, the PCS framework was developed using PPIE conducted in the UK population and evaluated in patient-clinician consultations focused on a single clinical condition from the United States; this pragmatic decision was driven by the availability of a high-quality, IRB-approved dataset. While the core principles of patient-centeredness are universal, the specific expression and prioritization of patient concerns can be shaped by cultural and health system contexts. For instance, patients within the United Kingdom's National Health Service may emphasize different priorities than navigating the US insurance system, where financial considerations can be a more prominent part of the illness experience [54]. As such, while our framework provides a robust foundational model, adaptation and validation across diverse health care systems are needed.

Second, because interviews were designed to inform framework development rather than achieve thematic saturation, the sample size and recruitment timeframe may have limited the breadth of perspectives captured. A larger or more diverse PPIE sample could identify additional priorities that might further refine the framework.

Third, while this study evaluated 5 widely used open-source LLMs to ensure transparency and reproducibility, the selection of models was also shaped by institutional data constraints. Because the study involved sensitive real-world patient-clinician conversations, the use of closed-source or externally hosted models requiring data transmission to third-party servers was not permitted. In clinical practice, strict data governance often precludes the transmission of sensitive patient-clinician audio transcripts to third-party cloud providers [55]. By benchmarking open-source models that can be deployed on-premise, we establish a baseline for

secure, HIPAA (Health Insurance Portability and Accountability Act)-compliant AI summarization, a prerequisite for real-world adoption.

Fourth, commonly used quantitative metrics such as ROUGE and BERTScore [36,37] have known limitations, as they emphasize lexical and semantic overlap and do not fully capture patient-centeredness. This limitation motivated the use of domain-specific categories and structured qualitative human evaluation by using our patient-centered framework, which shows that AI-generated PCS can be still useful regardless of their automated metrics scores.

Finally, in this initial evaluation, we did not use retrieval-augmented generation, as our primary focus was on the PPIE and qualitative components. Future AI systems should be able to leverage recorded clinical conversations to derive each patient's values, preferences, and wishes, and incorporate this information into summarization AI models to truly generate individualized PCS [56]. The deployment of such systems should be evaluated in pragmatic clinical trials to assess their impact on care outcomes and should be guided and overseen by health care professionals to avoid potential risks such as hallucinations. In future studies, we will compare fine-tuning and retrieval-augmented generation approaches and explore the development of our own model using a larger dataset.

Conclusion

This study shows that producing truly patient-centered clinical summaries remains a major challenge for current open-source LLMs. Using a mixed methods framework grounded in patient and clinician input, we identified the key domains of a PCS. The tested models failed to capture these dimensions, particularly factual correctness and nuanced patient values, with ROUGE-L and BERTScore results well below human benchmarks. Our work provides a validated framework, a reference standard for future research, and a pathway toward patient-centered AI summarization. Future work should focus on developing AI summarization models that can include individualized patient-centered elements in clinically meaningful summaries and evaluate its impact on clinical and care outcomes in clinical trials.

Acknowledgments

We would like to acknowledge the patients and general practitioners who participated in the Patient and Public Involvement and Engagement groups in the United Kingdom, whose insights were essential to developing the patient-centered framework.

Funding

This work was supported by the National Institute for Health and Care Research (NIHR; Award ID: NIHR207380) and the Torbay Medical Research Fund (Grant 145).

Data Availability

The data generated in this study include identifiable patient-clinician encounter recordings and linked electronic health record data. Due to privacy and ethical restrictions, these data cannot be made publicly available. Deidentified data may be made available upon reasonable request to the corresponding author and with appropriate institutional approvals.

Authors' Contributions

MLJ, AGC, KG, JPB, and OJPP conceived the study idea. MLJ and AGC led the protocol drafting, system design, and manuscript preparation. KG, SA, SL, and ML led the patient and public involvement groups in the United Kingdom. FL, CW,

KGM, LVA, CPV, SSSB, and SB contributed to the development of the annotation guideline and creation of the gold-standard PCS. DTT, OJPP, and MAZ worked on prompting design and data analysis. MEB performed the statistical analysis. JPB and OJPP provided overall supervision, clinical oversight, and critical revision of the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

DTT is a consultant for Immunovant but reports no conflicts of interest related to this publication. All other authors have no conflicts to declare.

Multimedia Appendix 1

Annotation guideline.

[\[DOCX File \(Microsoft Word File\), 20 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Prompt.

[\[DOCX File \(Microsoft Word File\), 15 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

Illustrative examples of how patients in different age groups described their experiences across thematic categories (Lifestyle and daily routines, support systems and access, events and life stressors, care preferences, and sources of meaning or value), corresponding to the summary presented in Table 2.

[\[DOCX File \(Microsoft Word File\), 21 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

Comparison of gold standard patient-centered summaries (PCS) and model-generated patient-centered summaries. Representative excerpts (“chunks”) are presented side by side to illustrate qualitative differences in clinical content, structure, and inclusion of protocol-defined patient-centered elements. Two illustrative cases (Example #1 and Example #2) are shown. To protect patient privacy, full summaries are not displayed. Selected excerpts are organized by summary sections (Background, Issues, Plan and PCS domains). Domain-specific qualitative evaluation scores (completeness, correctness, conciseness, fluency, and patient-centeredness) are reported after each example.

[\[DOCX File \(Microsoft Word File\), 34 KB-Multimedia Appendix 4\]](#)

References

1. Feblowitz JC, Wright A, Singh H, Samal L, Sittig DF. Summarization of clinical information: a conceptual model. *J Biomed Inform.* Aug 2011;44(4):688-699. [doi: [10.1016/j.jbi.2011.03.008](https://doi.org/10.1016/j.jbi.2011.03.008)] [Medline: [21440086](https://pubmed.ncbi.nlm.nih.gov/21440086/)]
2. Palanisamy B, Chakrabarti A, Singh A, Hassija V, Chalapathi GSS, Singh A. From information overload to lucidity: a survey on leveraging GPTs for systematic summarization of medical and biomedical artifacts. *IEEE Access.* 2025;13:7902-7922. [doi: [10.1109/ACCESS.2024.3521596](https://doi.org/10.1109/ACCESS.2024.3521596)]
3. Smith CJ, Beachy M, Vasey A, Langenhan T, Shiffermiller J. Evaluation of an enhanced discharge summary template: building a better handoff document. *J Clin Outcomes Manag.* 2017;24(12):542-549. URL: <https://cdn.mdedge.com/files/s3fs-public/Document/November-2017/JCOM02412542.PDF> [Accessed 2026-06-09]
4. Hoyer EH, Odonkor CA, Bhatia SN, Leung C, Deutschendorf A, Brotman DJ. Association between days to complete inpatient discharge summaries with all-payer hospital readmissions in Maryland. *J Hosp Med.* Jun 2016;11(6):393-400. [doi: [10.1002/jhm.2556](https://doi.org/10.1002/jhm.2556)] [Medline: [26913814](https://pubmed.ncbi.nlm.nih.gov/26913814/)]
5. Crucefix AL, Fleming APL, Lebus CS, Slowther AM, Fritz Z. Sharing a written medical summary with patients on the post-admission ward round: a qualitative study of clinician and patient experience. *J Eval Clin Pract.* Dec 2021;27(6):1235-1242. [doi: [10.1111/jep.13574](https://doi.org/10.1111/jep.13574)] [Medline: [33960593](https://pubmed.ncbi.nlm.nih.gov/33960593/)]
6. Wang M, Wang M, Yu F, Yang Y, Walker J, Mostafa J. A systematic review of automatic text summarization for biomedical literature and EHRs. *J Am Med Inform Assoc.* Sep 18, 2021;28(10):2287-2297. [doi: [10.1093/jamia/ocab143](https://doi.org/10.1093/jamia/ocab143)] [Medline: [34338801](https://pubmed.ncbi.nlm.nih.gov/34338801/)]
7. Shah SJ, Devon-Sand A, Ma SP, et al. Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden. *J Am Med Inform Assoc.* Feb 1, 2025;32(2):375-380. [doi: [10.1093/jamia/ocae295](https://doi.org/10.1093/jamia/ocae295)] [Medline: [39657021](https://pubmed.ncbi.nlm.nih.gov/39657021/)]
8. Gorenshtein A, Perek S, Vaisbuch Y, Shelly S. AI-generated neurology consultation summaries improve efficiency and reduce documentation burden in the emergency department. *Sci Rep.* Nov 6, 2025;15(1):38868. [doi: [10.1038/s41598-025-22769-7](https://doi.org/10.1038/s41598-025-22769-7)] [Medline: [41198773](https://pubmed.ncbi.nlm.nih.gov/41198773/)]

9. Kripalani S, LeFevre F, Phillips CO, Williams MV, Basaviah P, Baker DW. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. *JAMA*. Feb 28, 2007;297(8):831-841. [doi: [10.1001/jama.297.8.831](https://doi.org/10.1001/jama.297.8.831)] [Medline: [17327525](https://pubmed.ncbi.nlm.nih.gov/17327525/)]
10. Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. *J Am Med Inform Assoc*. Sep 2015;22(5):938-947. [doi: [10.1093/jamia/ocv032](https://doi.org/10.1093/jamia/ocv032)] [Medline: [25882031](https://pubmed.ncbi.nlm.nih.gov/25882031/)]
11. Fraile Navarro D, Coiera E, Hambly TW, et al. Expert evaluation of large language models for clinical dialogue summarization. *Sci Rep*. Jan 7, 2025;15(1):1195. [doi: [10.1038/s41598-024-84850-x](https://doi.org/10.1038/s41598-024-84850-x)] [Medline: [39774141](https://pubmed.ncbi.nlm.nih.gov/39774141/)]
12. Savage T, P Ma S, Boukil A, et al. Fine-tuning methods for large language models in clinical medicine by supervised fine-tuning and direct preference optimization: comparative evaluation. *J Med Internet Res*. Sep 23, 2025;27:e76048. [doi: [10.2196/76048](https://doi.org/10.2196/76048)] [Medline: [40986888](https://pubmed.ncbi.nlm.nih.gov/40986888/)]
13. Bednarczyk L, Reichenpfader D, Gaudet-Blavignac C, et al. Scientific evidence for clinical text summarization using large language models: scoping review. *J Med Internet Res*. May 15, 2025;27:e68998. [doi: [10.2196/68998](https://doi.org/10.2196/68998)] [Medline: [40371947](https://pubmed.ncbi.nlm.nih.gov/40371947/)]
14. Arko Iv L, Hudelson C, Kumar J, et al. Documenting care with AI: a comparative analysis of commercial scribe tools. *Stud Health Technol Inform*. Aug 7, 2025;329(337-41):337-341. [doi: [10.3233/SHTI250857](https://doi.org/10.3233/SHTI250857)] [Medline: [40775875](https://pubmed.ncbi.nlm.nih.gov/40775875/)]
15. Lavoie-Gagne O, Woo JJ, Williams RJ 3rd, Nwachukwu BU, Kunze KN, Ramkumar PN. Artificial intelligence as a tool to mitigate administrative burden, optimize billing, reduce insurance- and credentialing-related expenses, and improve quality assurance within health care systems. *Arthroscopy*. Aug 2025;41(8):3270-3275. [doi: [10.1016/j.arthro.2025.02.038](https://doi.org/10.1016/j.arthro.2025.02.038)] [Medline: [40120727](https://pubmed.ncbi.nlm.nih.gov/40120727/)]
16. Keszthelyi D, Gaudet-Blavignac C, Bjelogrić M, Lovis C. Patient information summarization in clinical settings: scoping review. *JMIR Med Inform*. Nov 28, 2023;11:e44639. [doi: [10.2196/44639](https://doi.org/10.2196/44639)] [Medline: [38015588](https://pubmed.ncbi.nlm.nih.gov/38015588/)]
17. Camaradou JCL, Hogg HDJ. Commentary: patient perspectives on artificial intelligence; what have we learned and how should we move forward? *Adv Ther*. Jun 2023;40(6):2563-2572. [doi: [10.1007/s12325-023-02511-3](https://doi.org/10.1007/s12325-023-02511-3)] [Medline: [37043172](https://pubmed.ncbi.nlm.nih.gov/37043172/)]
18. Ganzinger M, Kunz N, Fuchs P, et al. Automated generation of discharge summaries: leveraging large language models with clinical data. *Sci Rep*. May 12, 2025;15(1):16466. [doi: [10.1038/s41598-025-01618-7](https://doi.org/10.1038/s41598-025-01618-7)] [Medline: [40355506](https://pubmed.ncbi.nlm.nih.gov/40355506/)]
19. Federman AD, Sanchez-Munoz A, Jandorf L, Salmon C, Wolf MS, Kannry J. Patient and clinician perspectives on the outpatient after-visit summary: a qualitative study to inform improvements in visit summary design. *J Am Med Inform Assoc*. Apr 1, 2017;24(e1):e61-e68. [doi: [10.1093/jamia/ocw106](https://doi.org/10.1093/jamia/ocw106)] [Medline: [27497793](https://pubmed.ncbi.nlm.nih.gov/27497793/)]
20. Tvedten OG, Croker A, Lawrence M, Williams K, Perry N, Croker J. Personalised written consultation summaries for patients: an “up-close, in-depth, inside-out” exploration of a rheumatologist’s patient-centred strategy. *Patient Educ Couns*. Jul 2022;105(7):2362-2370. [doi: [10.1016/j.pec.2021.11.026](https://doi.org/10.1016/j.pec.2021.11.026)] [Medline: [34920911](https://pubmed.ncbi.nlm.nih.gov/34920911/)]
21. Cole AC, Liu CC, Bissram J, et al. Presenting summary reports to patients with chronic health conditions to support values-based discussions and decision-making: a literature review. *Patient Educ Couns*. Oct 2025;139:109256. [doi: [10.1016/j.pec.2025.109256](https://doi.org/10.1016/j.pec.2025.109256)] [Medline: [40645062](https://pubmed.ncbi.nlm.nih.gov/40645062/)]
22. Heckemann B, Chaaya M, Jakobsson Ung E, Olsson DS, Jakobsson S. Finding the person in electronic health records. A mixed-methods analysis of person-centered content and language. *Health Commun*. Apr 2022;37(4):418-424. [doi: [10.1080/10410236.2020.1846275](https://doi.org/10.1080/10410236.2020.1846275)] [Medline: [33213210](https://pubmed.ncbi.nlm.nih.gov/33213210/)]
23. Butler JM, Gibson B, Patterson OV, et al. Clinician documentation of patient centered care in the electronic health record. *BMC Med Inform Decis Mak*. Mar 12, 2022;22(1):65. [doi: [10.1186/s12911-022-01794-w](https://doi.org/10.1186/s12911-022-01794-w)] [Medline: [35279157](https://pubmed.ncbi.nlm.nih.gov/35279157/)]
24. Schuttner L, Theis M, Wong E, et al. Patient-centered prioritization of health care processes for multimorbidity. *JAMA Netw Open*. Dec 1, 2025;8(12):e2549693. [doi: [10.1001/jamanetworkopen.2025.49693](https://doi.org/10.1001/jamanetworkopen.2025.49693)] [Medline: [41396612](https://pubmed.ncbi.nlm.nih.gov/41396612/)]
25. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. Apr 2023;616(7956):259-265. [doi: [10.1038/s41586-023-05881-4](https://doi.org/10.1038/s41586-023-05881-4)] [Medline: [37045921](https://pubmed.ncbi.nlm.nih.gov/37045921/)]
26. Sittig DF, Boxwala A, Wright A, et al. A lifecycle framework illustrates eight stages necessary for realizing the benefits of patient-centered clinical decision support. *J Am Med Inform Assoc*. Aug 18, 2023;30(9):1583-1589. [doi: [10.1093/jamia/ocad122](https://doi.org/10.1093/jamia/ocad122)] [Medline: [37414544](https://pubmed.ncbi.nlm.nih.gov/37414544/)]
27. Maurer M, Mangrum R, Hilliard-Boone T, et al. Understanding the influence and impact of stakeholder engagement in patient-centered outcomes research: a qualitative study. *J Gen Intern Med*. Apr 2022;37(Suppl 1):6-13. [doi: [10.1007/s11606-021-07104-w](https://doi.org/10.1007/s11606-021-07104-w)] [Medline: [35349017](https://pubmed.ncbi.nlm.nih.gov/35349017/)]
28. Staniszewska S, Brett J, Simera I, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. *Res Involv Engagem*. 2017;3(13):13. [doi: [10.1186/s40900-017-0062-2](https://doi.org/10.1186/s40900-017-0062-2)] [Medline: [29062538](https://pubmed.ncbi.nlm.nih.gov/29062538/)]
29. NHS Health Research Authority. Public involvement guidance for researchers. URL: <https://www.hra.nhs.uk/planning-and-improving-research/best-practice/public-involvement/guidance-for-researchers> [Accessed 2026-02-05]

30. Scholl I, Zill JM, Härter M, Dirmaier J. An integrative model of patient-centeredness - a systematic review and concept analysis. *PLoS ONE*. 2014;9(9):e107828. [doi: [10.1371/journal.pone.0107828](https://doi.org/10.1371/journal.pone.0107828)] [Medline: [25229640](https://pubmed.ncbi.nlm.nih.gov/25229640/)]
31. Zill JM, Scholl I, Härter M, Dirmaier J. Which dimensions of patient-centeredness matter? - Results of a web-based expert Delphi survey. *PLoS ONE*. 2015;10(11):e0141978. [doi: [10.1371/journal.pone.0141978](https://doi.org/10.1371/journal.pone.0141978)] [Medline: [26539990](https://pubmed.ncbi.nlm.nih.gov/26539990/)]
32. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. *Qual Health Res*. Nov 2016;26(13):1753-1760. [doi: [10.1177/1049732315617444](https://doi.org/10.1177/1049732315617444)] [Medline: [26613970](https://pubmed.ncbi.nlm.nih.gov/26613970/)]
33. Noseworthy PA, Branda ME, Kunneman M, et al. Effect of shared decision-making for stroke prevention on treatment adherence and safety outcomes in patients with atrial fibrillation: a randomized clinical trial. *J Am Heart Assoc*. Jan 18, 2022;11(2):e023048. [doi: [10.1161/JAHA.121.023048](https://doi.org/10.1161/JAHA.121.023048)] [Medline: [35023356](https://pubmed.ncbi.nlm.nih.gov/35023356/)]
34. Chung P, Swaminathan A, Goodell AJ, et al. Verifying facts in patient care documents generated by large language models using electronic health records. *NEJM AI*. Jan 1, 2026;3(1):AIdbp2500418. [doi: [10.1056/AIdbp2500418](https://doi.org/10.1056/AIdbp2500418)]
35. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol*. 1990;43(6):543-549. [doi: [10.1016/0895-4356\(90\)90158-1](https://doi.org/10.1016/0895-4356(90)90158-1)] [Medline: [2348207](https://pubmed.ncbi.nlm.nih.gov/2348207/)]
36. Lin CY. ROUGE: a package for automatic evaluation of summaries. *ACL Anthology*; 2004. URL: <https://aclanthology.org/W04-1013.pdf> [Accessed 2026-06-09]
37. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. *arXiv*. Preprint posted online on Apr 21, 2019. [doi: [10.48550/arXiv.1904.09675](https://doi.org/10.48550/arXiv.1904.09675)]
38. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med*. Apr 2024;30(4):1134-1142. [doi: [10.1038/s41591-024-02855-5](https://doi.org/10.1038/s41591-024-02855-5)] [Medline: [38413730](https://pubmed.ncbi.nlm.nih.gov/38413730/)]
39. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. Mar 1977;33(1):159-174. [doi: [10.2307/2529310](https://doi.org/10.2307/2529310)] [Medline: [843571](https://pubmed.ncbi.nlm.nih.gov/843571/)]
40. Chen Q, Hu Y, Peng X, et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nat Commun*. 2025;16(1):3280. [doi: [10.1038/s41467-025-56989-2](https://doi.org/10.1038/s41467-025-56989-2)]
41. Abdellaoui C, Redjda A, Seroussi B. Generative-AI-based approaches for information extraction from clinical notes: a scoping review. *Stud Health Technol Inform*. Jun 26, 2025;328:193-197. [doi: [10.3233/SHTI250700](https://doi.org/10.3233/SHTI250700)] [Medline: [40588908](https://pubmed.ncbi.nlm.nih.gov/40588908/)]
42. Williams CYK, Subramanian CR, Ali SS, et al. Physician- and large language model-generated hospital discharge summaries. *JAMA Intern Med*. Jul 1, 2025;185(7):818-825. [doi: [10.1001/jamainternmed.2025.0821](https://doi.org/10.1001/jamainternmed.2025.0821)] [Medline: [40323616](https://pubmed.ncbi.nlm.nih.gov/40323616/)]
43. Salmon C, O'Connor R, Singh S, et al. Characteristics of outpatient clinical summaries in the United States. *Int J Med Inform*. Oct 2016;94:75-80. [doi: [10.1016/j.ijmedinf.2016.06.005](https://doi.org/10.1016/j.ijmedinf.2016.06.005)] [Medline: [27573314](https://pubmed.ncbi.nlm.nih.gov/27573314/)]
44. Swoboda CM, Fareed N, Walker DM, Huerta TR. The effect of cancer treatment summaries on patient-centered communication and quality of care for cancer survivors: a pooled cross-sectional HINTS analysis. *Patient Educ Couns*. Feb 2020;103(2):301-308. [doi: [10.1016/j.pec.2019.08.036](https://doi.org/10.1016/j.pec.2019.08.036)] [Medline: [31477514](https://pubmed.ncbi.nlm.nih.gov/31477514/)]
45. Blanch-Hartigan D, Chawla N, Beckjord EI, et al. Cancer survivors' receipt of treatment summaries and implications for patient-centered communication and quality of care. *Patient Educ Couns*. Oct 2015;98(10):1274-1279. [doi: [10.1016/j.pec.2015.06.005](https://doi.org/10.1016/j.pec.2015.06.005)] [Medline: [26146237](https://pubmed.ncbi.nlm.nih.gov/26146237/)]
46. Vale MD, Perkins DW. Discuss and remember: clinician strategies for integrating social determinants of health in patient records and care. *Soc Sci Med*. Dec 2022;315:115548. [doi: [10.1016/j.socscimed.2022.115548](https://doi.org/10.1016/j.socscimed.2022.115548)] [Medline: [36403352](https://pubmed.ncbi.nlm.nih.gov/36403352/)]
47. Watson BN, Estenson L, Eden AR, et al. Person-centered care planning for people living with or at risk for multiple chronic conditions. *JAMA Netw Open*. Oct 1, 2024;7(10):e2439851. [doi: [10.1001/jamanetworkopen.2024.39851](https://doi.org/10.1001/jamanetworkopen.2024.39851)] [Medline: [39418021](https://pubmed.ncbi.nlm.nih.gov/39418021/)]
48. American Psychological Association. Guidelines for psychological practice with older adults. *Am Psychol*. Jan 2014;69(1):34-65. [doi: [10.1037/a0035063](https://doi.org/10.1037/a0035063)] [Medline: [24446841](https://pubmed.ncbi.nlm.nih.gov/24446841/)]
49. Neupert SD, Bellingtier JA. Daily stressor forecasts and anticipatory coping: age differences in dynamic, domain-specific processes. *J Gerontol B Psychol Sci Soc Sci*. Jan 1, 2019;74(1):17-28. [doi: [10.1093/geronb/gby043](https://doi.org/10.1093/geronb/gby043)] [Medline: [29912449](https://pubmed.ncbi.nlm.nih.gov/29912449/)]
50. Stawski RS, Sliwinski MJ, Almeida DM, Smyth JM. Reported exposure and emotional reactivity to daily stressors: the roles of adult age and global perceived stress. *Psychol Aging*. Mar 2008;23(1):52-61. [doi: [10.1037/0882-7974.23.1.52](https://doi.org/10.1037/0882-7974.23.1.52)] [Medline: [18361654](https://pubmed.ncbi.nlm.nih.gov/18361654/)]
51. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc*. 2011;18(2):181-186. [doi: [10.1136/jamia.2010.007237](https://doi.org/10.1136/jamia.2010.007237)] [Medline: [21233086](https://pubmed.ncbi.nlm.nih.gov/21233086/)]

52. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digit Med*. Jul 29, 2023;6(1):135. [doi: [10.1038/s41746-023-00879-8](https://doi.org/10.1038/s41746-023-00879-8)] [Medline: [37516790](https://pubmed.ncbi.nlm.nih.gov/37516790/)]
53. Jung KH. Large language models in medicine: clinical applications, technical challenges, and ethical considerations. *Healthc Inform Res*. Apr 2025;31(2):114-124. [doi: [10.4258/hir.2025.31.2.114](https://doi.org/10.4258/hir.2025.31.2.114)] [Medline: [40384063](https://pubmed.ncbi.nlm.nih.gov/40384063/)]
54. Desai M, Rachet B, Coleman MP, McKee M. Two countries divided by a common language: health systems in the UK and USA. *J R Soc Med*. Jul 2010;103(7):283-287. [doi: [10.1258/jrsm.2010.100126](https://doi.org/10.1258/jrsm.2010.100126)] [Medline: [20595532](https://pubmed.ncbi.nlm.nih.gov/20595532/)]
55. Okyere Boadu R, Wireko Adu V, Okyere Boadu KA, et al. Examine frameworks policies and strategies for effective information governance in healthcare organizations. *PLoS ONE*. 2025;20(7):e0327496. [doi: [10.1371/journal.pone.0327496](https://doi.org/10.1371/journal.pone.0327496)] [Medline: [40644477](https://pubmed.ncbi.nlm.nih.gov/40644477/)]
56. Al Zahidy M, Guevara Maldonado K, Vilatuna Andrango L, et al. Longitudinal and multimodal recording system to capture real-world patient-clinician conversations for AI and encounter research: protocol for an observational study. *JMIR Res Protoc*. 2025;15:e84688-e84688. [doi: [10.2196/84688](https://doi.org/10.2196/84688)]

Abbreviations

AI: artificial intelligence

BERTScore: Bidirectional Encoder Representations from Transformers Score

EHR: electronic health record

IRB: Institutional Review Board

LLM: large language model

PCS: patient-centered summary

PPIE: patient and public involvement and engagement

ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation–Longest Common Subsequence

SDM: shared decision-making

Edited by Alicia Stone; peer-reviewed by Chengzhi Zhang, Tohida Rehman; submitted 04.Nov.2025; final revised version received 07.Apr.2026; accepted 08.Apr.2026; published 03.Sep.2026

Please cite as:

Lizarazo Jimenez M, Claros AG, Green K, Toro-Tobon D, Larios F, Asthana S, Wenczenowicz C, Guevara Maldonado K, Vilatuna-Andrango L, Proano AC, Bandi SSS, Bagewadi S, Branda ME, Al Zahidy M, Luz S, Lapata M, Brito JP, Ponce-Ponte OJ

Patient-Centered Summarization Framework for AI Clinical Summarization: Mixed Methods Study

J Med Internet Res 2026;28:e87061

URL: <https://www.jmir.org/2026/1/e87061>

doi: [10.2196/87061](https://doi.org/10.2196/87061)

© Maria Lizarazo Jimenez, Ana Gabriela Claros, Kieran Green, David Toro-Tobon, Felipe Larios, Sheena Asthana, Camila Wenczenowicz, Kerly Guevara Maldonado, Luis Vilatuna-Andrango, Ana Cristina Proano, Satya Sai Sri Bandi, Shubhangi Bagewadi, Megan E Branda, Misk Al Zahidy, Saturnino Luz, Mirella Lapata, Juan P Brito, Oscar J Ponce-Ponte. Originally published in the *Journal of Medical Internet Research* (<https://www.jmir.org>), 03.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the *Journal of Medical Internet Research* (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.