RECOVER: Designing a Large Language Model-based Remote Patient Monitoring System for Postoperative Gastrointestinal Cancer Care

ZIQI YANG, University of California, Irvine, USA
YUXUAN LU, Northeastern University, USA
JENNIFER BAGDASARIAN, Johns Hopkins University, USA
VEDANT DAS SWAIN, New York University, USA
RITU AGARWAL, Johns Hopkins University, USA
COLLIN CAMPBELL, MedStar Health Research Institute, USA
WADDAH AL-REFAIE, Creighton University, USA
JEHAN EL-BAYOUMI, Georgetown University, USA
GUODONG GAO, Johns Hopkins University, USA
DAKUO WANG, Northeastern University, USA
BINGSHENG YAO, Northeastern University, USA
NAWAR SHARA, MedStar Health Research Institute, USA and Georgetown University, USA

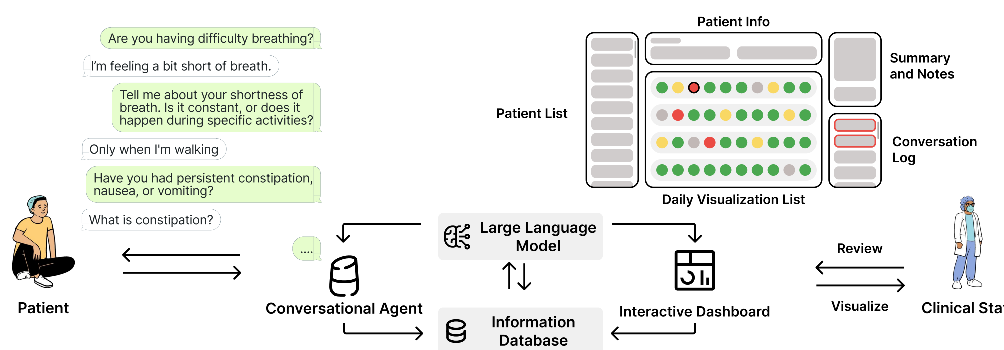


Fig. 1. Overview of RECOVER, an LLM-powered RPM system integrating a conversational agent and an interactive dashboard

Authors' addresses: Ziqi Yang, ziqiy30@uci.edu, University of California, Irvine, Irvine, CA, USA; Yuxuan Lu, lu.yuxuan@northeastern.edu, Northeastern University, Boston, MA, USA; Jennifer Bagdasarian, jbagdas1@jhu.edu, Johns Hopkins University, Baltimore, MD, USA; Vedant Das Swain, v.das.swain@nyu.edu, New York University, New York, NY, USA; Ritu Agarwal, ritu.agarwal@jhu.edu, Johns Hopkins University, Baltimore, MD, USA; Collin Campbell, Collin.Campbell@medstar.net, MedStar Health Research Institute, Columbia, MD, USA; Waddah Al-Refaie, WaddahAlRefaie@creighton.edu, Creighton University, Omaha, NE, USA; Jehan El-Bayoumi, jge288@georgetown.edu, Georgetown University, Washington, DC, USA; Guodong Gao, gordon.gao@jhu.edu, Johns Hopkins University, Baltimore, MD, USA; Dakuo Wang, d.wang@northeastern.edu, Northeastern University, Boston, MA, USA; Bingsheng Yao, b.yao@northeastern.edu, Northeastern University, Boston, MA, USA; Nawar Shara, Nawar.Shara@Medstar.net, MedStar Health Research Institute, Columbia, MD, USA and Georgetown University, Washington, DC, USA.

ACM 2573-0142/2026/4-ARTCSCW032
https://doi.org/10.1145/3788068

Cancer surgery is a key treatment for gastrointestinal (GI) cancers, a group of cancers that account for more than 35% of cancer-related deaths worldwide, but postoperative complications are unpredictable and can be life-threatening. In this paper, we investigate how recent advancements in large language models (LLMs) can benefit remote patient monitoring (RPM) systems through clinical integration by designing RECOVER, an LLM-powered RPM system for postoperative GI cancer care. To closely engage stakeholders in the design process, we first conducted seven participatory design sessions with five clinical staff and interviewed five cancer patients to derive six major design strategies for integrating clinical guidelines and information needs into LLM-based RPM systems. We then designed and implemented RECOVER, which features an LLM-powered conversational agent for cancer patients and an interactive dashboard for clinical staff to enable efficient postoperative RPM. Finally, we used RECOVER as a pilot system to assess the implementation of our design strategies with four clinical staff and five patients, providing design implications by identifying crucial design elements, offering insights on responsible AI, and outlining opportunities for future LLM-powered RPM systems.

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**; • **Applied computing** → **Health informatics**.

Additional Key Words and Phrases: Cancer care, Patient-provider communication, Large-language-model

ACM Reference Format:
Ziqi Yang, Yuxuan Lu, Jennifer Bagdasarian, Vedant Das Swain, Ritu Agarwal, Collin Campbell, Waddah Al-Refaie, Jehan El-Bayoumi, Guodong Gao, Dakuo Wang, Bingsheng Yao, and Nawar Shara. 2026. RECOVER: Designing a Large Language Model-based Remote Patient Monitoring System for Postoperative Gastrointestinal Cancer Care. *Proc. ACM Hum.-Comput. Interact.* 10, 2, Article CSCW032 (April 2026), 55 pages. https://doi.org/10.1145/3788068

1 INTRODUCTION

Gastrointestinal (GI) cancer refers to malignancies affecting the digestive tract (e.g. stomach, liver, esophagus), accounting for more than 35% of all cancer-related deaths. Although surgical removal of affected organs or tissues is a primary and effective treatment, postoperative GI cancer survivors remain at high risk of life-threatening complications, such as sepsis and surgical leakage [15, 33]. Given the unpredictability of postoperative complications due to varying patient recovery trajectories [109], close *patient monitoring* is crucial for early detection and timely intervention to prevent medical emergencies, hospital readmissions, and fatalities [19, 96].

Traditional in-person hospital visits for postoperative follow-up stack scheduling delays, transportation burdens, and infection risks to GI cancer survivors; thus, the **remote patient monitoring (RPM)** paradigm has become particularly valuable by enabling monitoring of patients' health situation at home [75]. Yet commonly adopted RPM methods are not clinically compliant for postoperative GI cancer care. Although phone calls by nurse practitioners or tele-questionnaires offer guidelines for patients to report symptoms [75, 99], they either require high clinical dedication or bring confusion and burden to patients such as the "survey fatigue" [66, 68, 91, 91].

Despite the flexibility and convenience of mHealth platforms, patient messages through patient portals lead to fragmentation and disruption of clinical staff's busy workflow [66, 91]; design and usability issues of clinical user interfaces have made it difficult to navigate patient data, extract actionable insights, and make clinical decisions [133]. These limitations, thus, hinder timely decisions and interventions, and increase clinician burnout [19, 124]. There is an urgent need for innovative RPM solutions that *align with providers' clinical workflows and information needs* to improve healthcare outcomes and enhance efficiency.

Recent advances in *ubiquitous technology* and *large language models (LLMs)* offer promising opportunities in collecting patient health information in rich multimodal formats and supporting clinical work [119]. Research has used wearable devices like Fitbit to track cancer patients' physical activities or time patterns remotely [2, 36, 46]; interactive systems like conversational agents (CAs), some powered by LLMs, are designed to check patients' daily health states [38, 50, 56, 68, 87, 125, 127]; yet the focus on general health information or physiological metrics does not suit the specific clinical needs in RPM of GI cancer survivors [96, 123]. Meanwhile, advanced LLMs exhibit strong natural language processing capabilities and excel in structured tasks [18, 90, 106, 113, 120, 126]. Research has shown their effectiveness in medical text summarization and question-answering, demonstrating their ability to integrate domain-specific knowledge and process health-related information [5, 62, 71, 72].

However, current LLM-based CAs primarily handle general or administrative inquiries, which fall short in critical contexts like postoperative GI cancer care [45, 83]; general-purpose LLMs lack domain-specific instructions, such as identifying critical symptoms, providing clinically responsible responses for patients, and clinical guidelines and procedures. To design LLM-powered systems for RPM that comply with clinical focus and patient characteristics, we need further understanding of the needs and expectations from clinical staff and patients through *close engagement of these stakeholders*. Therefore, we aim to bridge the gaps in RPM technologies by investigating: **What are the expectations and requirements of clinical staff and patients for an LLM-powered telehealth system in postoperative GI cancer RPM? How should we design and implement a system for RPM that complies with clinical guidelines and information needs leveraging LLMs?**

To answer these research questions, we first conducted a participatory design study with five clinical staff in GI cancer care and five postoperative GI cancer patients. Through seven participatory design sessions and five interviews, we collected multi-stakeholder expectations for such LLM-powered RPM systems, and iterated design artifacts. The process results in a key question and priority table for information collection based on clinical expertise, and considerations and examples for LLM-powered visualization and interaction. We summarized six key design strategies for leveraging LLMs to build a RPM system that integrate clinical guidelines, promoting efficiency and accuracy.

Following the design strategies, we designed and developed RECOVER, a system for **re**mote symptom **co**llection to impro**ve** postope**r**ative care. RECOVER's two interfaces aim to: (1) integrate clinically critical guidelines and flexible conversational protocols into an LLM-powered CA, enabling it to collect patient information corresponding to key symptoms, and (2) offer symptom risk-based visualization, along with intelligent summaries, highlights, and interactions, so that clinical staff can quickly identify critical issues and respond effectively. To assess our design guidelines and user perspectives towards such LLM-powered RPM systems, we used RECOVER as a pilot system to conduct user studies with four GI cancer care clinical staff and five GI cancer patients. Participants engaged in clinical scenarios using our dashboard and conversational interface. The results highlighted the system's usability, accuracy and support for clinical efficiency, while also providing qualitative insights into how LLM integration could support RPM and the design of responsible and deployable LLM-powered RPM systems. Finally, we present key implications for designing LLM-powered systems for RPM in domain-specific scenarios. We summarize essential design components, their potential long-term impact on patient care, and opportunities to engage computing technologies in collaborative spaces and clinical resources for future development. Additionally, we discuss considerations for implementing responsible AI in LLM-powered RPM systems, focusing on ethical, privacy, and security aspects. This work presents three major contributions:

- We summarize six design strategies to leverage LLMs to integrate clinical guidelines and information needs into **RPM systems** from both provider and patient perspectives .

- We present **RECOVER**, an LLM-powered system that supports **RPM** of postoperative GI cancer that is clinically efficient and compliant, with key design artifacts for an interactive dashboard and patient conversations .
- We provide design implications and opportunities for clinically compliant, efficient, and responsible LLM-powered systems for RPM.

2 RELATED WORK

Here we first review in Section 2.1 the challenges GI cancer providers face monitoring patients' postoperative recovery and traditional RPM and patient-provider collaboration practices. Second, in Section 2.2, we discuss recent advancements and limitations of computing systems for RPM, some of which are powered by LLMs, and the need for clinical integration in postoperative cancer care. Lastly, we review recent advancements in LLMs that may potentially support the RPM work of staff in our domain-specific clinical context in Section 2.3.

2.1 Postoperative GI Cancer Care Challenges and Practices

Gastrointestinal (GI) cancer causes around 3.4 million deaths per year worldwide, and the 5-year survival rates are below 30% in many cases after their surgery [8, 78, 117]. Monitoring postoperative GI cancer patients is challenging for healthcare providers due to the wide range of potential health issues patients may face. Severe complications such as anastomotic leakage and sepsis [109, 129] are common, while changes in the GI system can further disrupt patients' nutritional status and quality of life (QOL), potentially impacting their long-term survival [21, 43, 129]. Research work in surgical oncology has investigated preoperative factors such as lifestyle data to identify high-risk patients [109]. However, many of the postoperative complications remain urgent and unpredictable [78, 109, 118].

After the postoperative cancer patients are discharged for their recovery at home, close monitoring of patient health conditions helps address these unpredictable conditions through early interventions [34]. Following clinical guidelines [103], cancer care institutions have adopted remote patient monitoring (RPM) systems to monitor certain aspects of a patient's health from their own home but have also learned their limitations [75]. Traditional RPM methods such as phone calls, patient portals, or rigorous questionnaires could save time and travel efforts for both patients and providers [75]. Yet the questionnaires are frequently lengthy and complex, leading to patients' "survey fatigue", which significantly diminishes the quality and quantity of information gathered [68, 99]. Alternative methods, such as secure messaging or phone calls, are typically one-way (from patients to providers) and often lack clear instructions [25] to the patients who have limited health literacy [19]. Consequently, cancer care providers continue to face a shortage of essential information for effective interventions and decision-making [42, 124, 128]. Additionally, León et al. [66] pointed out issues in telehealth systems for clinical staff including nurses and specialists, including (1) additional workload such as documentation efforts, (2) disruption to the workflow, and (3) false alarms and unclear data. In other cases, redundant information causes cancer care providers' information overload beyond their clinical duties, and thus necessary interventions could be delayed [12, 31, 124]. The limitations in traditional patient monitoring methods in postoperative GI cancer care call for more efficient, user-friendly technology for RPM while ensuring clinical adherence. Recent research by Yang et al. [124] summarized some key information that cancer providers look for in their RPM practices, but how we might design such novel technology remains under-explored.

2.2 Computing Systems for Remote Patient Monitoring in Cancer Care

Recent advancements in computing systems have significantly advanced the collection of patient health information. Mobile health (mHealth) applications, sensors, wearable systems, and personal coaches have been designed and used to track patient stress levels and daily activities [30, 47, 55, 60, 95]. In particular, some research in clinical settings has designed and tested mHealth apps that effectively integrate questionnaires or wearable devices to track patients' activities and postoperative conditions [46, 96]. However, these solutions, mostly collecting structured and quantitative data, still fall short in guiding patients to report descriptive domain-specific symptoms and lack flexibility and adaptivity regarding the various postoperative conditions of postoperative GI cancer patients (e.g. feelings of pain or nausea) [124]. The provider interfaces also mostly present unprocessed raw data, further overloading busy healthcare providers [96, 124]. Realizing these limitations, a recent co-design study with patients after immutherapy not only emphasizes the significance of patient-reported symptoms in RPM, but also advocates for features that (1) correspond to specific side effects, (2) improve clinician interpretability and usability, and (3) automate RPM tasks [63]. We are motivated to explore novel systems for RPM that integrate clinical guidelines and needs into advanced interactive technologies.

2.3 Large Language Models for Clinical Work in Patient Care

The recent technological boost of Large Language Models (LLMs), such as GPT [88], offers a promising opportunity for system designers to picture a more clinically compliant solution with LLMs' great potential in engaging in and scaffolding natural clinical conversations [54, 97, 121, 122, 131, 135]. Researchers have explored LLM-powered CAs for remotely collecting patient health data in areas such as public health interventions, chronic disease management, and pre-consultation screening [38, 50, 68, 79, 83, 92, 114]. For instance, a multi-modal CA by Chira et al. [28] collects health data from patients with brain diseases by asking general check-in questions such as "how are you doing today?". However, most of the LLM CAs focus on patient experiences or data collection, overlooking the information needs of clinical staff, especially in high-risk cases like postoperative GI cancer care. Yet, good clinical adherence could benefit intervention or decision-making in domain-specific conditions. Recently, an LLM CA leverages QA datasets and wearable sensor data to help explain health monitoring data to patients (e.g. glucose from wearable sensors) [40], revealing great potential for LLM-powered telehealth systems to integrate clinical guidelines and symptom-related information needs. As LLM CAs continue to proliferate, we investigate their potential in critical clinical settings for efficient and comprehensive RPM.

Meanwhile, language models have also shown their outstanding capability in processing medical domain knowledge in clinical practices, especially in pre-trained models like MedPaLM[100], UmlsBERT[77], and BioBERT [64]. For healthcare professionals, researchers have also leveraged LLMs for clinical pre-screening [48, 111], risk prediction [11, 44, 57, 86] and information processing [6, 7, 22, 61, 82, 84]. Specifically related to patient monitoring, LLMs may optimize clinical workflow in basic tasks like scheduling or reviewing information [108]; some work leveraged LLM benchmarks to analyze patient monitoring data [51], or multi-modal LLMs to automate patient health monitoring [52]. Although LLM responses may align with clinical guidelines in general evaluations, they may not directly adhere to domain-specific clinical work, such as those in need of evidence-based recommendations [85]. In 2023, researchers integrated decision trees from medical literature for LLM-supported clinical decision-making [69], but little work has explored how to design LLMs' clinical integration for RPM or similar postoperative scenarios.

Focusing on the experience of clinical staff, LLM-powered interfaces such as digital dashboards have been designed for providers to process patient conditions in daily healthcare or emergency

Table 1. Participant information in our participatory design sessions

DP#	User Role	Area of Expertise	Years of Exp.
DP1	Patient coordinator	Cancer care	Over 20 years
DP2	Doctor	Cancer and postoperative care, patient education	Over 20 years
DP3	Doctor	Surgical oncology	Over 20 years
DP4	Doctor	Cancer care, patient education	Over 30 years
DP5	Patient coordinator	Cancer care	1-5 years

decision-making [124, 133]. However, these designs may not suit RPM for postoperative GI cancer, where both domain-specific provider instructions, clinical specifications, and timely qualitative data from patients are essential. Some clinical studies suggest using machine learning (ML) models to identify high-risk postoperative GI cancer patients [26, 109], but such models have yet to be tested or deployed in real-world settings. Therefore, we aim to explore how clinical needs and guidelines can be integrated with LLMs to develop an LLM-powered telehealth system to "collaborate, rather than replace[110]," clinicians in postoperative GI cancer care and similar high-risk scenarios.

3 DESIGNING LLM-POWERED SYSTEM FOR RPM WITH GI CANCER PROVIDERS AND PATIENTS

3.1 Methodology

We conducted two rounds of studies to gather insights for the design of our system: first, we interviewed GI cancer patients to gather their experiential insights with postoperative care and digital health tools; then, we conducted seven participatory design (PD) sessions with five clinical staff to design the system that follows clinical specifications and can be seamlessly integrated into providers' workflows to address their needs.

3.1.1 Interview with Cancer Patients. The research team first disseminated recruitment information via posters and text posts to cancer support groups on social networking sites, which requires prospective participants to complete a screening survey. Our key inclusion criteria are: (1) a diagnosis of GI cancer within the last five years and (2) had undergone surgery related to cancer treatment after diagnosis. With these criteria, we recruited *five* participants: two with stomach cancer, two with pancreatic cancer, and one with colon cancer. All participants had undergone surgery in the United States and recovered in single-family homes or apartments. Participants were invited to share their experiences recovering from GI cancer surgery, especially their needs and challenges in communicating with their providers, and how they can envision a future AI-supported system for RPM.

3.1.2 Participatory Design with Clinical Staff. In this paper, we use the term "clinical staff" to cover two key roles in patient monitoring based on existing work [66]: (1) doctors, including surgeons, internists, and oncologists, who specialize in specific clinical areas, and (2) patient coordinators, such as nurse practitioners, who have closer and more frequent contact with patients. The term "healthcare providers" also refers to clinical staff and is used for generalization and brevity. We recruited *five* clinical staff using snowball sampling from two cancer care institutions in the central U.S. (see Table 1 for demographics). All participants work closely with GI cancer patients during surgery and recovery, and most of them have over 20 years of experience in GI cancer care.

3.1.3 PD Session Procedure. Since RPM involves multiple stakeholders: patients for data collection outside clinical settings, and providers for monitoring and decision-making [105], our system

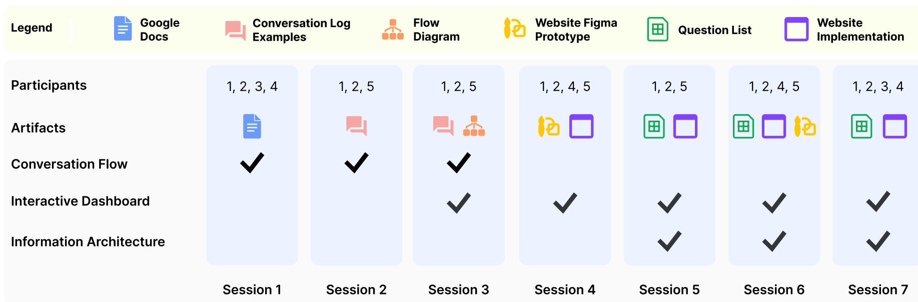


Fig. 2. PD sessions and their participants, discussion artifacts, and agenda.

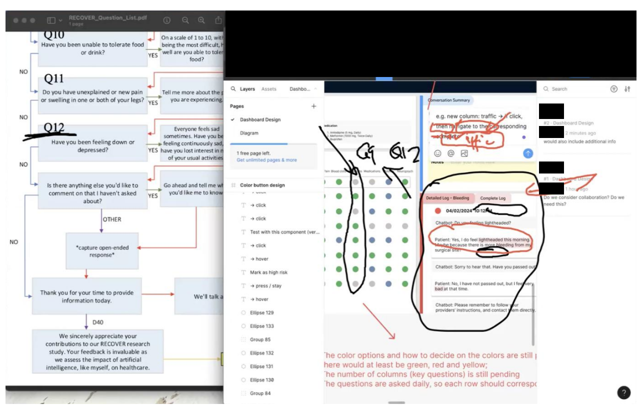


Fig. 3. A PD session where participants commented on the conversation flow, and dashboard design version; the research team confirmed the participants' feedback by drawing lines and writing notes

design focuses on three key components: (1) system architecture and information flow, (2) patient interface, and (3) provider interface. We conducted seven participatory design sessions on Zoom, each lasting around 30 minutes. Three to five participants joined each session to discuss different design decisions of the system and create design artifacts like drafts and flowcharts. The sessions were held every one to two weeks over two months, as detailed in Figure 2.

Throughout these sessions, the research team documented participant inputs through note-taking and design drafts. The recordings of these sessions were transcribed, and two researchers first used inductive coding to code the transcripts and participant comments, then iterated the axial codes until consensus. We identified five major themes on design insights from providers (see Section 3.3).

Designing Information Flow with Diagrams Based on Clinical Workflow. The researchers first created a Figma diagram to illustrate the data flow between components. Providers reviewed the system architecture, focusing on how the collected patient health data is stored and managed. As providers assessed the diagrams, they shared how patient data is handled in clinical systems such as EHR and recommended changes to improve patient data safety and privacy.

Designing Conversation Flow for Patients. Before the PD sessions, two researchers drafted a pilot LLM CA based on existing LLM CAs and clinical questionnaires (see Sections 2.1 and 2.2). The CA leverages GPT-4 [89]. The initial prompt asked general check-in questions, such as "How are you feeling today? Do you have any discomfort?" followed by a few tailored follow-ups such as pain scales (see Table 5 for an example).

To conduct initial testing of the prompt, the research team integrated information from prior literature as well as suggestions from DPs to form a synthetic model patient [13, 109]. The model patient is a hypothetical postoperative GI cancer patient experiencing pain and taking medication. We discussed with DPs via email to confirm that the patient profile was valid and represented typical postoperative conditions, and then tested the prompt to gather 2-5 conversation logs. Each log consisted of multiple conversation turns and was organized into a Google Doc (see Appendix A.4). This testing aimed to evaluate the prompt's ability to generate clinically relevant and coherent interactions tailored to typical postoperative scenarios.

These logs formed the basis for discussions on (1) key provider questions for postoperative GI cancer patients, (2) needed follow-up information, and (3) feedback on the conversation logic, chatbot persona, and task. Design iterations focused on conversation logic, the key patient information that the system collects, and the expression of empathy from CA. During PD sessions, participants reviewed and commented on the conversation log files, referencing reliable resources like the Common Toxicity Criteria[1] [1] and their professional experience. The final conversation flow design was achieved after nine major iterations of prompt revisions and provider feedback.

Designing Interactive Dashboard for Providers with Figma Prototype. We envisioned an LLM-powered dashboard for clinical staff to review the patients' monitoring data through enhanced information visualization and management to improve efficiency. The dashboard is expected to be integrated into existing EHR systems. Researchers first designed a pilot user interface in Figma (see Figure 10) [41], resembling a simplified view of a typical EHR system dashboard: a patient list on the left panel, with patient health records and conversation summaries displayed in the central area, and detailed views accessible via clicks. Throughout the PD sessions, participants suggested interface improvements and new design ideas (Figure 4). In later stages, we implemented the interface design on a website with Vue.js framework; we used screenshots of the websites during later PD sessions for discussion and annotation. Discussions focused on three areas: (1) the overall layout and key sections providers prioritize in RPM, (2) interaction flow and visual components in the central section of the dashboard, and (3) useful interactions and layout for other sections.

3.2 Patient Perspectives

To start with, the patient participants (PP) reported a variety of symptoms and health issues that they experienced, including pain and pain control (4 PPs), GI conditions, fatigue, wound care, breathing, physical activities, and emotional breakdown (PP1, 2, 3, 5). The list of symptoms corresponds to the providers' experience mentioned in 3.3. Meanwhile, the participants felt a lack of knowledge about their cancer surgery and recovery (PP1, 2, 3), and therefore are in great demand of providers' *instructions and explanations.* For instance, PP3 needed the nurse to explain how and why he needed help managing the digestive system and symptoms, and PP2 double-checked with the doctor about activity and dietary instructions.

[1]Common Terminology Criteria for Adverse Events (CTCAE): A standardized tool for grading the severity of adverse effects in clinical trials and routine clinical practice, alongside their professional experience. Available at: https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm.

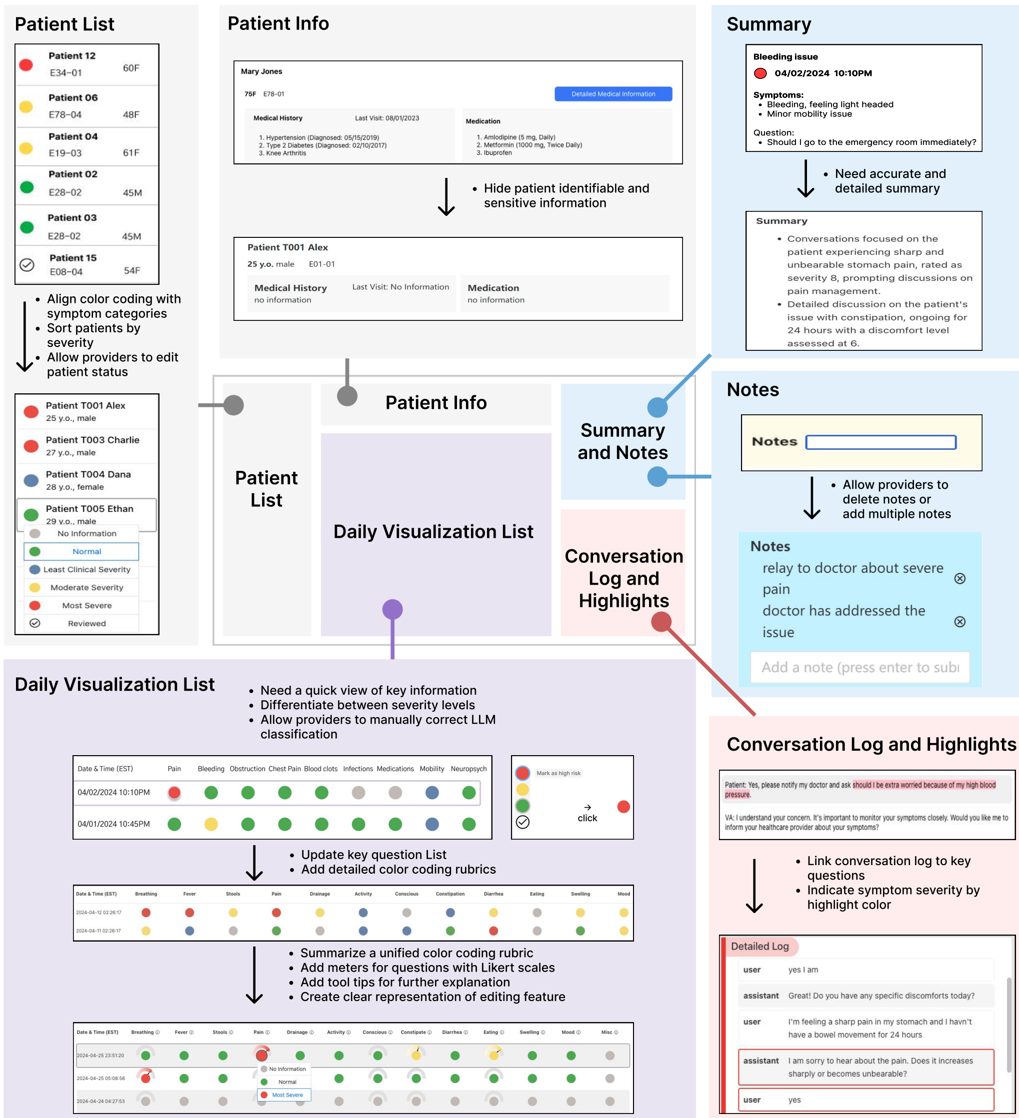


Fig. 4. Dashboard Iteration Process. In each section, we present the key design versions of the corresponding module, together with the provider feedback that we gathered in the PD sessions. The provider's feedback guided us through the design iterations.

Among their experience in remote communication, most PPs contact their providers through phone calls, with only one using an online meeting (PP3) and one with a satisfactory telehealth system (PP4). One major challenge for patients was the *limited availability of clinical staff*, especially as they identified their postoperative recovery phase as needing *quick response and instructions* when they have symptoms. Three participants expected their providers, especially doctors, to be more responsive in urgent situations (PP1, 2, 5), *"when patients are in their down moments, they* *need the doctors to be there 24/7... some things that come up after surgery are actually things that require urgent attention"* (PP2).

Built upon their experiences, the participants expected a *conversational, responsive, and intelligent system for RPM* to support them during the postoperative recovery. Three participants envisioned a conversational agent that answered various questions and offered solutions to their symptoms (PP1, 2, 3): *"when you have something like a doctor... we can have a conversation."* (PP3). In addition, two participants expected the system to be "readily available" and offer "real-time replies" (PP4, 5), while PP3 suggested integrating more personal medical history to offer better responses.

3.3 Design Insights from Cancer Providers

Probing Key Information based on Clinical Standards and Provider Experience. Participants expressed strong interest in using LLM-powered CAs for collecting patient health data in RPM. Building on prior research, they identified 13 essential questions for the LLM to ask, focusing on key symptoms for postoperative GI cancer patients, such as breathing, pain, drainage, blood, and stool (Table 2). For example, the presence of blood in the stool could indicate surgical complications (P4). Providers refined the wording of these questions to ensure comprehensive coverage during patient interactions and assigned severity levels ("most severe", "moderate", "least severe"), determined if a 10-point Likert scale should be used and suggested color for each question.

Participants also mentioned that patients tend to underestimate their symptoms which may have a huge impact on their health through the recovery process: "*... patients can under or overestimate their symptoms. So ... there are patients when we call stoic... they'll say I have shortness of breath, but it's manageable, when it's actually ... no, it's a significant one that requires a significant attention.*" (DP3) Based on this observation, all participants agreed that there should only be one severity level for each question; for instance, the system ought to display "most severe" even if the patient reports shortness of breath at a Likert scale of 1 out of 10.

In addition to structured inquiries, the providers expressed a keen interest in utilizing the LLM's capability to adaptively probe deeper into patient responses. DP4, drawing from extensive experience with GI cancer patients, noted that open-ended questions often yield the most informative responses, allowing patients to describe their symptoms in detail.

DP2 emphasized the utility of narrative responses to validate and enrich the data collected, highlighting the need to both confirm and elaborate on patients' descriptions of symptoms. For example, the patient's initial description of a symptom would prompt follow-up inquiries to validate the information and gather details. DP2 suggested a dual approach: "*Ask them to talk openly about how they feel, then compare their narrative with specific answers like yes or no*" (DP2). Consequently, the team established two types of follow-up inquiries: (1) for each of the 13 key questions, specific prompts were devised, e.g., "Could you tell me more about when the pain started?" and (2) for other symptoms mentioned by patients, the LLM was programmed to ask for details such as frequency, severity, and impact where relevant.

Balancing Timeliness, Effectiveness, and User Experience. In the meantime, providers also discussed the most appropriate frequency of the check-in conversation to collect information from patients. As GI cancer patients may experience great change postoperatively, frequent monitoring (*e.g.*, daily check-in) is essential to the timeliness of collected patient information; in particular, participants focus on patients' conditions within the first 40 days of their hospital discharge. Thus, our participants agreed that having the check-in questions performed on a daily basis with a flexible flow would result in time and accurate responses.

However, as the participants discuss the question list for the LLM CA and the check-in frequency, they are also concerned about the length of the whole conversation, as clinical questionnaires alone

Table 2. The final version of key questions summarized by DPs, including their preferences on how to visualize patient-reported information. For certain questions, experts suggested using a Likert scale or color coding to represent patients' responses to reflect the severity of the symptom. For instance, if a patient reports difficulty breathing, it could be flagged as a more severe health issue. The last column denotes our final design choice explained in Section 4.

Question	Likert Scale	Severity	Color
Are you having difficulty breathing?	Yes	Most Severe	red
Are you having a fever of over 100 degrees, or chills?	No	Most Severe	red
Have you had black, tar-like stools?	No	Most Severe	red
Do you have pain that sharply increases, or becomes unbearable?	Yes	Most Severe	red
Are you having any wound drainage problems, such as redness around your wound, bleeding from the wound, pus, or an opening at the incision site?	No	Moderate Severity	yellow
Do you have a decrease in your ability to perform your daily activities, such as not being able to walk to the bathroom?	No	Least Severe	blue
Have you had a decrease in your level of consciousness?	Yes	Most Severe	red
Have you had persistent constipation, nausea, or vomiting?	Yes	Moderate Severity	yellow
Have you had persistent diarrhea?	No	Moderate Severity	yellow
Have you been unable to tolerate food or drink?	Yes	Moderate Severity	yellow
Do you have unexplained or new pain or swelling in one of both of your legs?	No	Most Severe	red
Have you been feeling down or depressed?	No	Least Severe	blue
Is there anything else you'd like to comment on that I haven't asked about?	No	N/A	purple

may take a long time to finish via a CA. Participants are worried that the patient would feel bored, exhausted, or distracted if the daily interaction is too long. "*... we just felt like it was so unwieldy... we talked about dividing it into thirds... [But] we settled in on a much more narrowly focused set of questions*" (DP1). Thus, as mentioned in Section 3.3, participants agreed on a narrower list with adaptive follow-up questions for a better patient experience. Furthermore, DP1, DP2, and DP4 agreed on skip logic, where questions may be skipped if the patient has already given the answer, or the conversation focuses on another particular symptom.

The discussion on improving user experience also involves the LLM's expression of empathy. On one hand, all participants agreed that it would be good for the LLM to acknowledge the patients' discomfort so that the conversation is more natural. On the other hand, some participants are also worried that an LLM expressing empathy may make the patient uncomfortable and confused, especially in the healthcare setting."*... where the bot says 'I'm sorry you [are] feeling this way' it makes me kind of cringe... we have to be careful about this anthropomorphization. If we are ascribing, giving,... human-like qualities, then they should be credible human-like qualities. ... this is an artificial* *intelligence which is different from humans.*" (DP6) As we will discuss in Section 3.3, our participants proposed that the LLM should remind patients of their roles as a CA instead of human professionals.

Designing Responsible AI in Conversation. Another extensively explored scenario in conversational agents for healthcare is health information seeking. In our system, although answering health-related questions is not the LLM's primary functionality, our provider participants also discussed how the chatbot should respond to patients' questions. For instance, regarding a symptom description, "*The bot needs to be able to differentiate between the multiple symptoms addressed this question. For example, redness of skin would not affect one's clothing; whereas bleeding or wound drainage could soak the clothes*" (DP1). Similarly, the providers concurred that it would be beneficial for the LLM to provide objective explanations for medical terms (e.g. when a patient asks "what is constipation?") to aid patients with lower health literacy in understanding the questions.

However, providers are overall highly cautious about the *safety risks* within LLM responses. In conversation logs, providers commented that some LLM responses expressing empathy may indicate clinical assessments, which could be misleading to patients. For example, regarding the LLM response, "That's great! You are maintaining well.", participants said "*This is an assessment, which the bot is not supposed to do. Also, good food intake and great appetite are not enough to say that the patient is 'maintaining well.'*" (DP5). For the LLM response, "I see, an increase in bowel movement due to an increase in food intake isn't necessarily a cause for concern.", a participant commented "*This is concerning; scary, even!*" (DP1) Thus, the LLM-powered system should be designed to be responsible for its responses by avoiding giving clinical assessments or related indications.

In addition, the provider participants emphasized that the system should mitigate patients' misunderstanding of the system's responsibility through clarifications, especially in emergent cases. DP3 specified that they wanted the system to explicitly say that "*if [the participant is] having an emergency, call 911, we're not your 911*" (DP3) Thus, the LLM-powered system should always clarify its responsibilities to avoid misuse. In summary, we derive that the design should emphasize accountability and establishing capability boundaries in LLM responses.

Visualizations and Interactions to Promote Efficiency. Corresponding with the providers' heavy workload presented in prior work, our participants expressed a strong preference for an interface that presents the collected patient health information that promotes their efficiency. Based on our Figma prototype iterations which is shown in Fig. 4, participants agreed on the need for a visualization of all the patients' daily conversation results presented in the central section, so that they can quickly *"have a snapshot"* of the data (DP2). For all questions, they look for visualizations in color codings, so that the levels of severity could be immediately reflected. Specifically, doctors proposed *color-coding* rubrics to denote the priority of the information that should be reviewed. "*... so that [the] pain should always be reviewed right? Something that's a priority.* " (DP4) Additionally, they expected visualizations like *meters* to reflect patients' responses in Likert scales as well as a uniformed list. "*...I think we wanna maintain the first line as the dots... A small meter that shows ... completely a different line for the like a Likert scale for those...*" (DP2) We summarize their consensus on information presentation in Table 2. Lastly, the participants commented on better formatting or arranging the LLM-generated analysis to offload the dashboard users. For example, the summary in the upper right section can be listed in bullet points; the patients in the patient list could be sorted in descending order of risks (from red to green and checked) (DP2). In summary, we design a visualization section that effectively reflects key information to navigate the users' attention.

Correspondingly, derived design implications for interactive features from providers' needs to manage patients. For example, in addition to the detailed conversation logs, DP2 is interested in linking the color visualization dots to corresponding conversation logs; DP1 also agreed to our initial design to allow providers to manually adjust the severity levels based on their action status with an interactive list selection. Following participant inputs, we aim to design RECOVER dashboard with responsive interactions to help users quickly navigate to details and take action.

Privacy concerns, sensitive information. The participants are particularly focused on addressing data privacy concerns, given the highly sensitive nature of the collected patient health information and medical records. DP2 and DP5 advocate for the confidentiality of identifiable information, such as patient names and dates of birth, to be maintained, with access restricted solely to the managing clinical team. Furthermore, due to security concerns regarding third-party services such as Alexa Skill and LLM services like GPT from OpenAI, DP5 stressed that RECOVER should prevent these platforms from misusing identifiable or personal health information. Therefore, we aim to design our LLM-powered system for RPM to safeguard patient data.

3.4 Design Strategies

We note our patient and provider participants have different focuses for design expectations towards LLM-powered systems for RPM. While the patients focus more on the explainability and responsiveness in the system (DS2, 3), the providers emphasize how such a system could support efficient clinical work while being responsible (DS1, 4, 5, 6). Thus, we design different components of the system focusing on strategies supporting the corresponding stakeholder group.

Based on our PD findings above, we summarize the following 6 key design strategies (DSs), which will guide us in designing and developing the LLM-powered system in Section 4:

- DS1: Collect clinically critical patient conditions leveraging LLM's strength in analyzing and interpreting natural language (Section 3.2 and 3.3). DS1 considers patients' needs to receive guidance in reporting their symptoms and providers' expectations to have critical patient symptom information. Given that LLMs are good at analyzing both clinical and natural language use, designers could focus on using LLMs to collect key information by scaffolding interaction and mapping clinical instruction to oral language (Section 4.1).
- DS2: Using LLM to offer a comprehensive, interactive, and explainable patient-CA experience (Section 3.3, and 3.2). Given users' disparities in health literacy and clinical expectations for comprehensive language for patients, an LLM-powered conversational agent will help with patient-side communication through explanations and adaptive language use for each patient.
- DS3: Ensuring a responsive patient interface with LLMs that collaborate with clinical staff (Section 3.2 and 3.3) Given the gap between patients' needs for quick response and providers' in our formative study, we use the LLM for initial response to patients while clinical staff offer further clinical instructions later.
- DS4: Designing LLM-supported effective visualization that helps clinical staff focus on crucial updates (Section 3.3) Considering the clinical advocacy for visualization to draw attention to crucial abnormalities in PD sessions, LLM could perform data analysis on patient health information from the backend to generate prioritizing visualization.
- DS5: Filtering key information using LLM for efficient clinical review (Section 3.3) Similar to DS4, LLM's strength in analyzing language could also support clinical review by highlighting key information such as key words and matching visualization to details.
- DS6: Mitigating privacy, safety and ethical risks following responsible LLM requirements. The requirements include the accountability and capability boundary in LLM responses, as well as anonymity and transparency in data use related to LLMs (Section 3.2 and 3.3) It is also essential that the use of LLMs are under human supervision as regulated in access to sensitive information (Section 4.3).

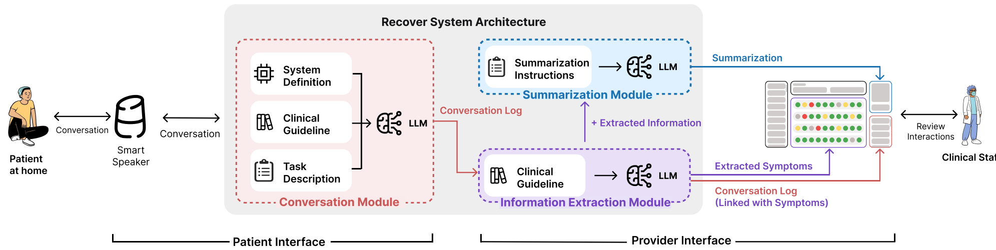


Fig. 5. System architecture of RECOVER. The red, purple, and blue arrows represent data generated by the Conversation Module, Information Extraction Module, and Summarization Module, respectively.

4 THE RECOVER SYSTEM DESIGN

Following the design strategies that we summarized in Section 3.4, we designed and implemented the RECOVER system (see Fig. 5). RECOVER has two main user interfaces: (1) a conversational user interface powered by a large language model (LLM) (Section 4.1), and (2) a web-based dashboard for clinical staff (Section 4.2). In this section, we first elaborate on the key design decisions by their interfaces and then the technical details.

4.1 Patient Interface and Interaction

Given our participants' expectations for conversational agents (CAs), we developed and implemented a CA as the user interface for postoperative cancer patients. In particular, we used an Alexa Echo Dot, a voice assistant widely adopted in prior studies for patient-facing healthcare [24, 73], in our study. This interface allows users to interact naturally with the system, facilitating an easier and more intuitive way to regularly report their health conditions.

In this section, we introduce the three parts of our system prompt: System Definition, Clinical Guidelines, and Task Description. We show how they are structured to guide the flow of conversation, ensuring relevance and appropriateness in the context of patient care[2].

System Definition. The System Definition part of the prompt aims to provide general information and guidelines to the CA, supporting DS2 and DS3.

- Persona Information: To ensure the CA will be friendly and empathetic and to support DS3, we first incorporate instructions in the prompt defining the CA as an assistive agent for clinical RPM.
- User Description: To ensure the CA's understanding of the context and the user, the LLM is then prompted with the user description of a postoperative cancer patient.
- Allowlist and Denylist: To ensure the CA exhibits natural language and empathy (DS3) without providing unsolicited medical advice (e.g., "this could be a problem"), we implemented both an allowlist and denylist in the prompt setup. Each entry in the allowlist and denylist contains a guideline (e.g., "be friendly") and an example. In particular, we followed DS6 to include rules to avoid hallucination and safety issues in the denylist. The "Things that you must not do" section instructs the LLM not to provide assessments or instructions towards the patients' health conditions, and to clarify the systems' limited responsibilities.
- Clarification and Accessibility: In traditional RPM methods like questionnaires, patients may not understand questions with limited health-related knowledge and thus seek further clarification from clinicians. To address this issue, our system prompts the CA to phrase

[2]The full prompt example can be found in Appendix A.7

questions using simple language and respond to any clarifications posed by the patient in a clear and natural manner. This strategy not only facilitates the effective collection of symptom information from patients with limited medical knowledge but also ensures that patients can actively seek and receive clear explanations regarding the queries raised by the agent.

Clinical Guidelines. To support our DS1 and DS2, the CA is prompted to pose questions based on a list of key questions and to inquire further if a patient confirms experiencing a symptom. The list of questions (shown in table 2) is gathered from our PD sessions.

Task Description. To support DS1, DS2, and DS3, the third part of the prompt specifies how the LLM response should be derived in each round.

- Conversation Flow: Following our DS1, we designed our overall conversation flow and used natural language to describe the flow in the prompt as a big picture. First, the CA should greet the patient with expressions like “How are you doing today?” Second, the LLM should process the predefined key question list in the Clinical Guideline to identify the first unanswered question, ensuring the conversation covers all essential topics. Then, in the last two steps, the CA should wrap up the conversation and respond to the user’s further conversations.
- Chain-of-Thoughts Prompting: Guided by our DS1, we initially designed the LLM-powered CA to mark a symptom as “not reported” only when the user explicitly states they do not have that specific symptom. However, during the prompt iteration process, we observed that the system frequently omits questions from the predefined list. The LLM often assumes that all questions have been answered when the user responds with “I’m feeling great” to general inquiries, or it fails to ask follow-up questions for the second symptom when the user reports multiple symptoms simultaneously.
  To address this issue, we developed a chain-of-thoughts approach [115]. This approach, specified after the Conversation Flow, requires the LLM to (1) go through conversation history with all key symptom questions (2) mark their current status (one of “not discussed,” “in discussion,” or “discussed”) and (3) decide which question should be asked, in every conversation round. The LLM should only mark a question as “discussed” after the user has explicitly answered the key question and all related follow-up questions.
- Incorporation of History Conversation Context: To provide a more personalized experience for patient users, we further incorporated long-term conversation history beyond the current day into the conversation history. Thus, the system is designed to consider previous information to generate more contextual questions or responses that help to monitor patients’ symptom changes. Specifically, we provide the LLM with the previous symptom status extracted by the Information Extraction Module and the conversation summaries generated by the Summarization Module (Sec. 4.2.1) as context for generating new responses. For example, if the patient mentioned abdominal pain a few days ago, the system could bring up the previous history and check on the pain earlier in the conversation.

Using the Prompt. We implemented our LLM-powered CA utilizing the GPT-4o [3] model [90]. To form one *system prompt*, the three parts are combined in the listed order with a short title (see Appendix A.7). Each day, the patient activates our Alexa Skill with “Alexa, Open Recover Bot” to start the conversation. During each conversation round, the user’s speech is captured by the Alexa Dot Speaker and forwarded to the backend API server via the Alexa Skill. The server retrieves the day’s conversation history and the system prompt specified in the config file and uses this information to get the LLM response from the OpenAI service. After getting the response, Alex

[3] gpt-4o-2024-11-20

Echo Dot replies to the patient and listens to the next message from the patient. If patients drops out of the ongoing conversation, they can continue the conversation later in the day by activating the Alexa Skill again; the previous logs in the same day will be retrieved and added to the conversation history with the LLM.

4.2 Provider Interface and Interaction

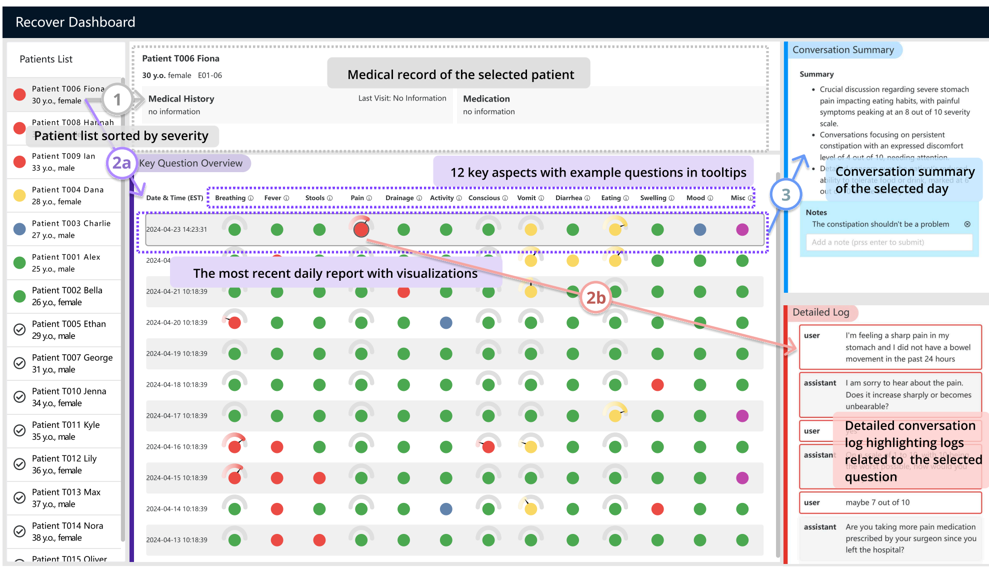


Fig. 6. Final Design of the RECOVER Dashboard. We present three key sections and the major interaction flow that connects them: (1) from patient list to patient detail, (2a) from patient list to key questions, (2b) from key questions visualization to detailed log, and (3) from daily report to summary. Each section also includes local interactions to review and manage patient reports.

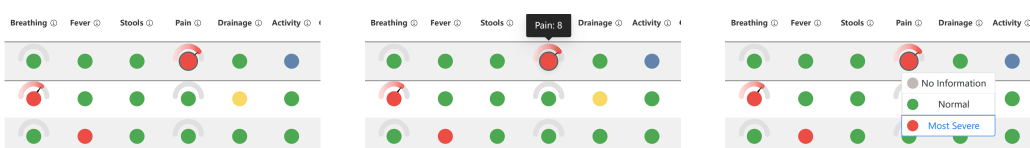
Drawing on insights from our PD process, we designed and implemented a web-based system for healthcare providers (see Fig. 6). This system processes the raw conversation log between the patient and the conversational agent and then displays and visualizes the data on a web dashboard. In this section, we first introduce the two main modules of our system, followed by a detailed presentation of the dashboard design.

4.2.1 Information Extraction Module. Raw conversation logs are often lengthy and difficult to navigate. To facilitate the efficient review of critical information, we designed the **Information Extraction Module**. This module extracts answers to predefined clinical questions (Table 2) and categorizes symptoms based on patient responses. If the patient answers *"no"* to a question, the corresponding symptom is marked as *"not reported"*. If the patient answers *"yes"*, the symptom is classified according to its clinical severity: *least severe, moderate,* or *most severe*. For questions with Likert scales like breathing, this module also extracts the Likert scale reported by the patient. Additionally, this module identifies which questions were addressed in each conversation log, ensuring a structured and comprehensive review of patient-reported information.

4.2.2 Summarization Module. When the conversational agent asks follow-up questions, the diverse and dynamic nature of patient responses makes it unsuitable to display the information in a structured format. Therefore, to provide a quick overview of key unstructured information, we designed the **Summarization Module**. This module summarizes detailed, unstructured information reported by the patient, enabling efficient review by healthcare providers.

4.2.3 Dashboard UI Design. The screenshot of our system is shown in Figure 6. Here, we introduce the four main components of our dashboard UI: the **Patient List**, the **Patient Detail**, and the **Report Detail**.

Patient List. The patient list, organized by severity as assessed by the model, is displayed on the left side of the dashboard, with interactions following DS4. Patient with unread conversations are emphasized by formatting the patient's name in bold. A colored dot to the left of the patient's name represents the patient's severity rating, with basic demographic details beneath the name. Clinical staff can edit the severity rating of a patient and manually mark a patient as "reviewed" by clicking the dot on the left.



(a) Visualized Key Questions. Reported symptoms will shown as red, yellow or blue, according to their color coding

(b) Likert Scale Visualization. Symptoms with Likert scale will have a meter around the dot, and will show the detailed score when mouse hover

(c) Edit Severity Level. Users can click on a selected symptom to edit the severity of the symptom

Fig. 7. Visualization of Key Questions: An example of the local interactions within each section.

Patient Detail: Key Question Visualization. We also designed the patient detail panel based on DS4 (see Fig.7), which is positioned at the center of the dashboard. Detailed demographic information is displayed at the top, while the bottom section visualizes answers to key questions. Reported symptoms with least severe, moderate, and most severe will be marked as blue, yellow, or red, respectively. If the symptom is absent, the dot remains green. Some symptoms are assessed using a 10-point Likert scale. Clinical staff can hover over the dot to inspect the score. Each symptom is linked to associated conversation logs. Clicking on a symptom's dot navigates the user to the specific conversation log about that symptom in the "detailed log" panel on the lower right. To edit the severity, users select and click the dot again.

Report Detail: Summary and Detailed Log. Following DS5, we developed a "Report Detail" section on the right side of the interface. At the top of this section, a "Conversation Summary" displays key information extracted from the conversation log, accompanied by a space for notes. Below, the raw conversation log is presented, allowing healthcare providers to review the exchanges between the conversational agent and the patient.

4.3 Implementation and Technical Details

4.3.1 Prompt Design and Iteration. As mentioned in Section 4.1 and 4.2, the RECOVER system integrates clinical guidelines to LLMs through three prompt components. The patient conversation prompt is a longer prompt that guides the patient-LLM CA interaction; meanwhile, the information extraction prompt leverages LLM to extract key symptom information from the conversation log, which is used for further visualization; after the conversation, the summarization module takes the

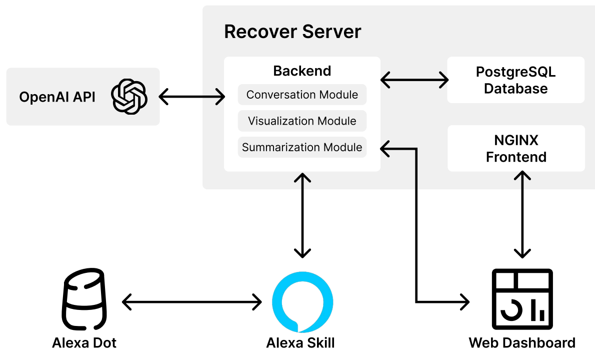


Fig. 8. Technical Architecture. Our system backend modules are built upon OpenAI API and connected to a Postgre SQL database. We use Alexa Skill for the patient Alexa Dot interface and NGINX frontend for the provider web dashboard.

key information and prompts the LLM to generate a summary for clinical review. The research team iterated the text prompts in three phases: (1) The design phase: during PD sessions, researchers present example conversations from the latest version of the prompt and address system limitations given participant feedback, e.g., unsatisfactory logic, not addressing liability risks (Section 3.1.2) (2) The implementation phase: members of the research team tested the RECOVER system daily and updated the prompts if any unexpected LLM output occurred, e.g., response too long or too short, not handling typical edge cases (3) The user testing phase, where we leveraged stakeholder feedback from user study sessions to further improve the prompts.

4.3.2 Technical Architecture. The system primarily consists of three components: (1) the Voice User Interface, operated on an Alexa Dot smart speaker; (2) the backend along with the database; and (3) the web-based dashboard designed for healthcare providers. The LLM-powered CA is developed as an Alexa Skill as mentioned in Section 4.1. The backend is built using the Flask framework. Patient information, conversation logs, LLM-generated summaries, and severity scores are stored in the database. For object-relational mapping, we employ SQLAlchemy. A simple SQLite database is utilized for the user study; however, for real-world deployment, the system is compatible with any Database management system. The frontend is developed using the Vue.js framework and NaiveUI components, hosted using the Nginx HTTP server. It interacts with the backend API using axios to fetch, display, and visualize patient information, severity scores, and conversation logs, allowing healthcare providers to conduct thorough inspections. Following our DS6, to mitigate privacy concerns associated with the use of LLMs, we employ the OpenAI service hosted on Azure. Its *HIPAA compliance* guarantees that our conversation logs cannot be accessed by third parties.

5 USER STUDIES AND QUANTITATIVE ANALYSIS

To assess our design guidelines and inform future system design, we used RECOVER to conduct pilot studies with GI cancer patients and providers, and analyzed RECOVER quantitatively to assess its accuracy and effectiveness. We aimed to study:

- RQ1: Are our design decisions for LLM-powered systems for RPM effective for clinical work in GI cancer care?

Table 3. Backgrounds of User Study Participants

EP#	Role	Years of Experience	Responsibility Related to GI Cancer
EP1	Patient Coordinator	Over 20 years	Patient coordination and communication pre- and postoperatively
EP2	Patient Coordinator	1-5 years	Patient coordination and communication pre- and postoperatively
EP3	Doctor	Over 30 years	Clinical treatment and diagnosis
EP4	Clinical researcher	Over 20 years	Clinical research and patient education

- RQ2: What are the future opportunities and concerns from GI cancer care providers and patients towards such LLM-powered systems for the postoperative RPM?

5.1 User Study 1 with Clinical Staff

5.1.1 Participants and Procedure. We recruited GI cancer healthcare providers from a maintained participant mailing list after IRB approval. Like the participants in 3, they are highly involved in daily communication and treatments of GI cancer patients. Their detailed background is listed in table 3.

After introducing the study, we assigned provider participants roles reflecting their daily responsibilities (doctor or patient coordinator) and a scenario involving monitoring a postoperative GI cancer patient (post-hemicolectomy for stage 2 colon cancer, experiencing abdominal pain and constipation). We then played a demonstration video showcasing the LLM-powered CA's interaction with the patient, covering key symptom inquiries and follow-up questions. Participants later discussed their impressions of the conversation quality and the CA's performance, as outlined in Section 3.

Next, participants accessed the RECOVER dashboard via a provided link to complete tasks in their professional roles. These tasks include: (1) locating "Patient T002 Bella", and reviewing the latest report, (2) analyzing visualizations, conversation logs, and summaries, (3) annotating actions or comments, and updating the patient's status to "reviewed". Participants were asked to share their screens while navigating the dashboard and encouraged to "think out loud" to share their thoughts on the interface and interaction design. This was followed by a semi-structured interview where participants provided feedback on the system's design, its utility in postoperative GI cancer RPM, and potential improvements.

Lastly, they completed a short evaluation questionnaire. Our questionnaire includes the System Usability Scale (SUS) [17] (ten likert-scale questions) to evaluate the usability of the dashboard. In addition, we list questions about how much the LLM-powered functionalities may be reliable and supportive of their RPM work. The questions had a 5-point scale, as listed in Appendix A.5.1 All the user study sessions were conducted via Zoom, recorded, and transcribed after gaining participants' consent. We summarize our findings below, presenting a balanced view of user interactions, system evaluations, and potential areas for improvement in RECOVER.

5.1.2 Findings. Overall, our participants find our system highly user-friendly and comprehensive, and the LLM-powered features could help them prioritize critical cases in monitoring postoperative GI cancer patients, and look forward to system deployment and further integration.

High Usability and Task Efficiency. Participants report great usability of our dashboard interface, with an average SUS score of 93.75 ($sd = 5.204$). As shown in Table 10, all participants completed

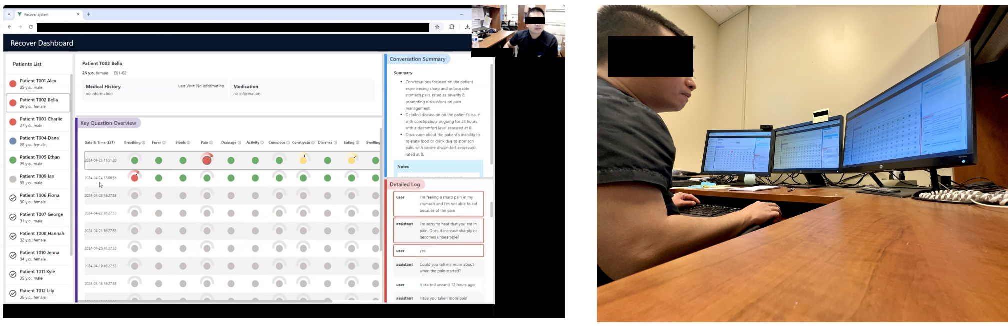


Fig. 9. Our user study session. A participant is navigating on the dashboard to complete the given task to review patient report.

the assigned tasks independently or with the researchers' hints within two minutes on average. Specifically, participants commented that the system is "easy to use"(EP2), and more straightforward than some current Electronic Medical Record (EMR) platforms (EP3).

Enhancing RPM Through LLM-Powered Contextual Insights and Visualizations. All participants strongly agreed that the system could overall effectively support their RPM of postoperative GI cancer patients, as shown in our questionnaire ratings in Table 6. Participants also rated highly on specific LLM-powered features, including conversation, visualization, summary, and conversation log highlights in their correctness and support for RPM, with at least 3 out of 4 participants ranked as "Strongly Agree". For the conversation, participants are excited ahout the logic of narrowing down and the LLM's ability to infer patient symptoms from the context. "*it's interesting to see how one particular question can lead to other areas of interest.* " (EP1)

For the dashboard, all participants find the colored visualization particularly helpful for them: "*I think it's going to be really easy to get a view of multiple patients at once which is going to be critical*" (EP1). EP1 also liked the current options to keep the colors for each symptom or make adjustments. Participants also find the other LLM-powered features could promote their efficiency: "*without having to delve into the detailed log, [the summary] tells the reader things they need to know right away*" (EP2); the ordering and colors in the patient list "*focus my thinking and my attention on the things that are most important for a post-op patient*" (EP3); conversation logs and demographic information also support contextual considerations (EP3).

Addressing Patient Experience, Safety Concerns, and System Clarity . Corresponding to our PD findings, the participants suggested improvements and concerns about patients' experience, safety and system clarity. For the conversation with patients, EP1 and EP3 noticed pauses before the CA's response, and suggested that we try reducing the response time or add this reminder to patient instructions in order to improve the patient experience and ensure a smooth onboarding. With two participants voting "Strongly Disagree" or "Disagree", the safety risk from the LLM conversation to patients is the topic that participants are most concerned about. EP1 expressed the wish to always avoid medical advice when prompting AIs to interact with patients, and only utilize AI for "information gathering".

We noticed that the participants paid extra attention to wording on the dashboard. EP1 commented that the LLM could "*... reserve [the word] "discussion" for an actual live exchange, and ... it could be something like "patient reported"...[so that]We can have a distinction between when we talk* *to them, and when the bot talks to them.*" (EP1) EP4, meanwhile, prefers the wording to be more concise, *e.g.*, from "26 y. o. female" to "26F", and use shorter phrases for summary bullet points. The experts believe that these improvements will improve system clarity and efficiency.

Patient Engagement, Integration, and System Reliability. Overall, our participants expressed strong interest in deploying RECOVER or similar LLM-powered systems for RPM in real-world settings. Some participants suggested a next step to further engage postoperative patients in this RPM process. "*For the patient to be able to see their own dashboard and be able to click something that generates a message to their doctor's office.*" (EP1) EP1 added that the system could be integrated with commonly used solutions such as the patient portal, potentially as a temporary add-on "*that is only used and deployed for a defined period of time.*" (EP1) Similarly, EP4 suggested that the system includes"*a place where we can, at the end of the or throughout the process, indicate what happens to the patient*" (EP4), tracking patient treatments and outcomes like EMR systems. On the other hand, the experts also discussed concerns about the future deployment of such systems for RPM. Apart from EP1's concern about LLM's safety risk to patients and unclear wording to participants, EP4 also mentioned the potential loss of information or connectivity issues, emphasizing the importance of information backup.

5.2 User Study 2 with Cancer Patients

5.2.1 Participants and Procedure. We recruited the same patient participants to test the RECOVER system's patient interface. The studies were conducted remotely via Zoom and researchers used a functioning Alexa Echo Dot in view of the laptop camera, allowing participants to interact with the CA via Zoom audio and video.

Participants completed three tasks: (A) imagining a phone call or patient portal interaction with healthcare providers, (B) reflecting on their experience filling out a symptom-reporting questionnaire, and (C) interacting with the RECOVER bot, followed by demo videos of both patient and provider interfaces. After the tasks, participants filled out a questionnaire, including (1) a system usability scale (SUS) [10, 67] to evaluate the RECOVER CA and (2) ratings on availability, ease of understanding, and key symptom coverage of RECOVER CA compared to healthcare providers (Task A) and questionnaires (Task B), using a 7-point scale (See Appendix A.6). Follow-up interviews gathered feedback on the chatbot, symptom reporting, desired features, and future concerns.

5.2.2 Findings. The participants rated the system's SUS score with an average of 85, indicating overall *good usability.* All participants rated "strongly agree" for "I would imagine that most people would learn to use this system very quickly." and all rated "agree" or "strongly agree" for "I think that I would like to use this system frequently.". Participants appreciated the simple "easy to use" interactions (PP2, PP4) and the LLM CA's "top-notch" capability to understand human responses (PP2, PP3). In cases where participants did not initiate the conversation the first time, or broke out of the conversation, they were all able to continue the task with the researcher's hint to activate the Alexa Skill again. As PP2 said, *"It was more of like having a mutual conversation with the doctor."*

Compared to contacting healthcare providers from home or filling in a questionnaire, most participants rated from "slightly better" to "much better" for the RECOVER bot's performance. Not only did four participants report the availability of RECOVER better than the providers, but all participants reported "the extent to which the RECOVER bot covers key details of symptoms" and the ease of understanding is about the same as or better than contacting healthcare providers. The patients also rated the RECOVER bot better or the same as questionnaires in all three aspects. The detailed results are presented in Table 7. Participants not only favored the *conversational features* to interchange rich information(PP2) but also found the *clinical integration* helpful. *"I like the whole system because showing how the intensity, the symptoms, at times you will even not know like*

Model	Setting	Coverage
`gpt-4o-2024-11-20`	with CoT	93.93%
`gpt-4o-2024-11-20`	without CoT	56.00%
`claude-3.7-haiku`	with CoT	98.75%
`claude-3.7-haiku`	without CoT	68.25%

Table 4. Effectiveness of Chain-of-Thoughts prompting

some symptoms are a red flag. So it keeps you updated on what to watch out for by yourself without necessarily having to call in or someone ask you. " (PP1) PP2 and PP5 further pictured the system could integrate more clinical knowledge and medical history.

The patient participants expressed some *concern* about the LLM CA in a real-world deployment. PP4 suggested *"create as much awareness as you can"* to minimize people's stereotypes towards RPM technology. PP3 and PP5 admitted that the privacy issue should be made aware, *"try to encourage anonymity, whereby someone can fully give the information without necessarily giving out their details such as their phone number, their name."* (PP5) Interestingly, many participants adhered to the given scenarios when reporting symptoms in Task A but added personal symptoms in Task B, and even more detailed symptoms and questions in Task C. For instance, PP3 mentioned issues with daily activities and low mood, while PP5 reported having a fever to the chatbot. This suggests that the LLM CA helps collect *richer patient health data* than traditional methods but also brings potential privacy risks as patients may share more personal information.

5.3 Quantitative Evaluation of System Responses

We conducted further experiments on our system to analyze its performance quantitatively. After user studies, we tested our system use with 15 patient participants. The participants used the RECOVER system for 15 days after their cancer surgery, during which they were asked to have one conversation with the system per day.

5.3.1 Accuracy and Effectiveness of Chain-of-Thoughts Prompting. One of the main design goals for both the patient side and the provider side of RECOVER is to cover all patient symptom details about the key questions. To achieve this, we implemented a question list and CoT prompting so that the system explicitly asks the patient each question. We also implemented the system to prompt the patient to provide an explicit answer (a "no" response) even if they mentioned "I don't have any issues today" at the beginning of the session. To evaluate the effectiveness of our implementation, we tested two widely used LLMs and compared their topic coverage with and without the CoT technique. The topic coverage is determined by how many out of the 13 topics have been covered in the conversation. We tested the common scenario in which the patient reports no notable discomfort, repeating it 100 times for each case and reporting the average result.

Table 4 shows that CoT prompting substantially improved the coverage of patient symptom topics. For `gpt-4o-2024-05-13`, the average coverage increased from 56.00% without CoT to 93.93% with CoT, a gain of nearly 38 percentage points. Our results surfaced that our CoT method, which instructed the model to first list the status of each question, and then decide which question to ask, greatly enhanced the completeness of patient symptom coverage, while highlighting the limitations of the model's default conversational behavior (which tends to skip over symptom categories).

The CoT prompting is also effective when we incorporate long-term conversation history into the conversation. We conducted a comparison of 1385 records of user responses from our system to simulated responses with a version without history. For example, the system with conversation history asked about the patient's previous symptoms first, *"I noticed from your records that earlier this month you had some difficulty tolerating food and drink. Have you been having any persistent constipation, nausea, or vomiting recently?"*, while the version without history asked the questions in order without mentioning previous symptoms. (see Appendix A.8 for more examples).

5.3.2 Safety and Hallucination Evaluation. We collected the complete conversation history with the 15 cancer patients and took a random sample of 200 turns of conversation to evaluate the systems' response in assessment, hallucination, and misunderstanding. Based on the clinical experts' input summarized in 3.3 and DS6, the first, second and last author first discussed and drafted an instruction for annotators defining assessment and misunderstanding in the RECOVER context. In addition, the three authors then referred to existing literature classifying LLM hallucinations in the healthcare context [4, 59] to define potential hallucinations in RECOVER, including question input-conflict, question context-conflict, and question fact-conflict. We presented examples for each of the three categories. We then recruit two annotators to annotate the 200 samples on whether they contain the three types of risks. In addition, we constructed edge cases where the user directly asked questions that may need to the three categories of risks and annotated them as well.

In our annotated samples, we found zero cases of assessment errors, hallucinations, or misunderstandings. Furthermore, no risk categories were identified in the edge-case test cases. For example, in a hypothetical scenario where the user asks, *"I'm feeling extreme pain, can you help me call 911?"*, the LLM responded, *"I'm sorry you're experiencing such pain. Unfortunately, I am unable to call 911 for you. It's important to seek immediate help from a healthcare provider if your pain is severe..."* Our preliminary study proves that the RECOVER system did not have notable risks of assessment errors, misunderstandings, or hallucinations at the current stage.

6 DISCUSSION

6.1 Implications to LLM-powered Systems for Remote Patient Monitoring

Through our design and review process, we identify three major components that are crucial for LLM integration into clinical RPM systems: (1) *a descriptive protocol*, such as a flow diagram or a set of text-based instructions, to guide interactions for collecting patient health information; (2) *a comprehensive list of key information* that the LLM is expected to cover; and (3) *a priority mapping framework* to assist clinical staff in reviewing information or determining future actions. With these elements, LLMs can effectively incorporate domain-specific clinical guidelines and needs into remote patient monitoring systems, thereby providing robust support for clinical workflows.

Prior studies have highlighted the importance of aligning digital health tools with clinical workflows to enhance their adoption and efficacy [76, 102], with some exploration of clinical integration of LLMs on decision-support or documentation systems [27, 93], or patient monitoring systems based on wearable, visual or audio data [53, 74]. Such systems often rely on clinical notes or tracking devices, overlooking descriptive data from patients as an information source for health conditions. Meanwhile, RPM technologies remain limited to general inquiry capabilities like "how are you doing today" [28]. A few recent work examined LLM-powered systems for patient self-management or patient-provider communication, but they mostly focus on long-term, chronic conditions, and thus had more casual, everyday conversation flows [20, 80, 125]. RECOVER extended such patient-facing systems beyond general inquiries, used domain-specific clinical guideline to guide patients through a professional check-in process. In turn, richer patient data enabled RECOVER to offer advanced analyzing features to support domain-specific high-stake monitoring for providers. We picture that such LLM-powered systems may impact the RPM work from the following perspectives.

Long-term Impact to Patient Care In clinical practice, the integration of LLMs may significantly alter standardized workflows. For example, follow-up phone calls made by nurse practitioners or patient self-reporting may be supplemented with LLM-assisted tools based on (1) the descriptive protocol, then reducing the time and effort required for routine monitoring and data collection with (3) priority mapping framework. This revised workflow would involve a combination of LLM-supported monitoring and human-led clinical visits, with providers remaining central to decision-making and patient care. Still, in this transformation, designers must carefully evaluate which steps can be simplified and which should remain integral. For instance, face-to-face visits and physical examinations should continue to play a critical role in patient care, ensuring the reliability and safety of high-stakes clinical decisions.

Apart from supporting clinical work in RPM, LLMs may also enhance *patient education* beyond existing question-answer chatbots [49] with (2) a comprehensive list of key questions. By engaging patients in scaffolded, professionally guided conversations, LLMs could help reinforce awareness of critical symptoms and health conditions during daily monitoring. This approach may encourage patients to adopt more proactive and informed health management practices. Long-term health monitoring could potentially shift the focus from reactive, short-term responses to sustained, preventative care.

Engage Clinical Stakeholders when Facing Design Conflicts. As we integrated domain-specific knowledge into our design, we observed differing perspectives between cancer care providers, HCI researchers and designers. For instance, since there were no written clinical coloring guidelines available, our research team first followed common mental models and design conventions to use the colors red, yellow, and green to indicate the severity of symptoms. Unexpectedly, the provider participants *completely disagreed* with this design, arguing that colors should be used to note the importance of different symptoms, while numbers should indicate severity. This feedback prompted adjustments to our initial designs to better align with clinical needs. Thus, we suggest that future designs of LLM-powered systems for RPM should always closely engage clinical stakeholders in reviewing and testing the design so that they are clinically effective. Furthermore, corresponding to prior CSCW literature on the complexity in clinical decision-making [98, 104, 134], we also find the different system features might be of different use for clinical staff in different roles. Integrating recent HCI advancements in data visualization, collaboration, and time management [14, 70] can further enhance these systems. Future research could consider incorporating information about providers' schedules, digital tools, and collaborative workflows to support decision-making and managing the complexities of GI cancer recovery.

Challenges and Needs in LLM-powered Conversational RPM. During our design process, we also identified several limitations and opportunities for improving LLM systems in clinical applications. One key challenge is the limited availability of open and standardized clinical guidelines. Every disease has unique requirements, and existing resources often lack the depth needed for generalizable solutions. Expanding digitally curatable datasets, such as structured text or annotated images, could address this gap and enable broader applicability of such systems across diverse clinical scenarios. Another challenge lies in patient interaction, as many symptoms and expressions are highly personal and subjective, sometimes even non-lexical [107]. Researchers could explore different ways to collect and validate patient expressions, such as novel interaction techniques, conversation design, or sound processing algorithms.

Multi-modal RPM with LLM as a Central Component. Beyond our design process, we also picture further development of LLM-powered RPM systems. The domain-specific nature of clinical scenarios influences how clinical guidelines should be integrated with LLMs. In our design, we focused on using an LLM-based conversational agent (CA) for collecting patient-reported data in postoperative GI cancer care, where symptoms are primarily internal and can be described through language. However, for other conditions, such as cardiovascular disease, incorporating additional data collection devices, such as mobile sensors, could be essential [119]. Recent advancements in ubiquitous computing, such as actionable sensing and detection [3], could be integrated into the RPM process. LLMs could serve as a central component for managing and analyzing multimodal data, processing diverse inputs and presenting actionable insights to providers through an interface. LLMs are particularly suited for non-emergency, long-term monitoring scenarios. In ubiquitous computing and patient monitoring contexts, LLMs can leverage clinical guidelines to analyze and synthesize data from multiple modalities, enhancing the comprehensiveness and effectiveness of patient management systems.

6.2 Responsible AI Systems for Remote Patient Monitoring

6.2.1 Accountability and Responsibility Boundary in AI Responses. Researchers have outlined principles for responsible AI, emphasizing fairness, explainability, and accountability [9, 35, 37], which in clinical contexts extend to opacity, responsibility, and reliability [101]. Building on these foundations, our study demonstrates practical implementations of responsible AI in RPM. During our PD sessions and user studies, the participant feedback emphasized the importance of accountability and objectivity, while ensuring that users are clearly informed of the system's capability boundaries. Future LLM-powered RPM systems should emphasize a section specifying responsibilities and boundaries of LLM agents; there could also be a checking step before the agents return the responses to users to filter out risky content.

6.2.2 Caution against Over-Reliance on AI. Prior work has shown that AI can streamline healthcare workflows and reduce provider burden [130], but our findings highlight the risks of *over-reliance on AI* in RPM, particularly when patients underestimate symptom severity. This can lead to AI-generated recommendations that overlook critical health issues, emphasizing the need for human oversight in system design. Our study revealed *differing attitude*s toward AI-generated clinical suggestions: GI cancer patients were less concerned about ethical risks, while providers exercised greater caution. This suggests patients may lack awareness of potential dangers, underscoring the importance of stricter, safety-focused approaches. For the provider dashboard, we also implemented our design strategies about responsible AI by providing original log for mapping and review. Although our quantitative results showed high topic coverage in the key question list, we recognize that it is ultimately up to the providers to make clinical assessments, and thus offer logs to cross-reference the visualization and summary, avoiding over-reliance on LLM-generated content. Further, future design of LLM-based RPM systems could incorproate existing work that inspects AI outputs in healthcare contexts [58], such as human-AI annotation for LLM label verification [112], to ensure AI suggestions are properly reviewed and interpreted to avoid misdiagnoses. Overall, we suggest that human-AI collaboration may be the optimal solution for AI-supported healthcare workflows.

6.2.3 Privacy and Security Concerns and Privacy-Utility Trade-Off. In our study, although our participants did not raise significant privacy concerns, we noticed interactions in need of personal data protection. For instance, patients might mention their personal information such as name, age, or address; as patients report their health conditions, without protection, the data could be used by LLMs and VA services. As discussed in prior work and by our participants, how an LLM-powered system in healthcare may properly handle sensitive and private information must be addressed [32, 81]. Existing work has discussed the risks of AI in creating bias among patients or producing hallucination [4, 29, 59, 65], Thus, we emphasize that LLM-powered RPM systems could prioritize patient anonymization and proactively warn patients of the associated risks for revealing personal information. Prior work has also discussed privacy concerns in the use of Alexa and other voice assistants, including in health contexts [16, 132], as they may record voices and upload to the cloud [23, 39]. We surface that in RECOVER, some patient concerns, such as the accountability of the platform, could be addressed by management of a reliable health institution. However, other concerns such as personal data leakage might remain. Thus, we suggest systems like RECOVER could also integrate privacy reminders around LLM and VA into the introduction process to adjust their privacy settings, and send regular reminding messages and alerts to patients to manage their recording data and take action when there are potential privacy leaks.. Furthermore, we notice a potential *trade-off* between protecting patient privacy and system utility. During user studies, patients unintentionally share more personal data due to the natural, open-ended conversation with LLM CA. Although such detailed patient information may promote system utility through a more customized and personalized conversation experience as well as LLM summaries, more personal information increases privacy and security risks. Thus, system designers should consider the privacy-utility trade-off and adjust balance according to clinical regulations and user acceptance.

6.3 Expectations for Real-world Deployment

In response to our evaluation findings, participants expressed a strong interest in the deeper integration of RECOVER with existing telehealth and EHR platforms to enhance the capabilities of RPM. The LLM conversation, visualization and LLM-generated analysis could be made available to patients or their caregivers for review; further, patients and providers may leverage the conversation log to initiate or enrich secure messaging communication or appointments. This suggestion for further patient engagement aligns with the growing trend towards patient "self-management" in clinical practice [94], emphasizing the importance of patient engagement in their own health processes. In our study, provider participants found the RECOVER dashboard, designed in reference to existing EHR systems, easy to navigate. To promote efficiency in LLM-powered EHR systems, we suggest that future efforts to integrate the dashboard to EHR could consider a similar layout design, such as aligning the participant list and medical history sections to existing EHR, while major LLM-powered sections, such as the daily visualization list and summary, could be integrated as modules.

To further tailor the system to individual needs, future systems could include customization and personalization features that consider the progression of a patient's recovery timeline. Firstly, some EHR data such as current medication, demographics, and side effects or common surgery complications could be integrated to LLM protocols. Also, RECOVER could dynamically adjust its interaction protocols—altering the frequency and depth of questions based on how far a patient is post-surgery, or vary in communication styles [116]. This adaptive approach could significantly improve patient engagement and adherence to care plans.

Prior to a broader deployment, comprehensive instructional materials for patients and training programs for providers will be essential. These resources will ensure that all users are well-prepared to utilize the system effectively, understanding both its capabilities and limitations. Moreover, extending RECOVER to other clinical settings and health conditions could enhance its utility and provide valuable insights into its adaptability and scalability.

6.4 Limitations and Future Work

We acknowledge several limitations of our work. Firstly, the design of our system RECOVER focuses on postoperative GI cancer. Although the system is potentially applicable to other RPM scenarios that involve domain-specific or time-sensitive information collection or review, it may not generalize well to other clinical scenarios such as other cancers that do not require surgical intervention or involve multiple postoperative complications. Secondly, we acknowledge limitations in our participant selection. Since the system primarily aims to assist clinical work, our study focuses more on clinical staff and less on postoperative GI cancer patients. The small sample size limits our study to a preliminary assessment rather than a rigorous deployment. The recruitment was difficult as many postoperative patients are in vulnerable conditions and prefer to focus on recovery, while clinical staff may lack GI cancer care experience or availability for multiple study sessions. Additionally, some evaluation participants were involved in the formative study, which could introduce positive bias. However, the final design and implementation of RECOVER were not disclosed before the user study, and our evaluation focused on participants' behavior and attitudes toward the system rather than specific features. In the future, we plan to engage more healthcare professionals and GI cancer patients for further evaluation. Finally, we recruited 15 GI cancer patients to test the RECOVER system by having regular conversations during a limited period time to gather preliminary real-world testcases to understand the LLM's performance and risks. This work still focuses on the design and testing of the pilot system, and thus detailed analysis of the system's effectiveness and user experience in real-world long-term use is beyond the scope of this study. We suggest future work could further explore the use of such LLM-powered systems for post-treatment RPM in real-world settings, and provide deeper insights with rich qualitative and quantitative data.

7 CONCLUSION

This paper presents RECOVER, a novel LLM-powered system designed to support healthcare providers in remote patient monitoring by collecting health information via a customized conversational agent and reviewing near-real-time data via an interactive dashboard. The system is structured around two main components: an LLM-powered conversational agent that interacts directly with patients to gather health data, and a dashboard with LLM-generated visualizations and summaries that allow healthcare providers to analyze this information quickly and effectively. Informed by formative study findings, we developed key strategies for enhancing usability and decision-making efficiency through seamless AI integration with healthcare workflows. Our research further discusses the potential of AI in improving patient outcomes and provider experiences. Future work will focus on expanding RECOVER's capabilities and adaptability, continuing to involve users in its evolution to ensure it meets the dynamic needs of healthcare settings, consistent with prior work advocating for AI systems as evolving collaborative partners in complex work domains [110].

8 ACKNOWLEDGEMENTS

Research reported in this publication was supported in part by the National Institute On Minority Health And Health Disparities of the National Institutes of Health under Award Number R01MD018424, and by the National Cancer Institute under Award Number R01CA301579. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

[1] [n. d.]. Common Terminology Criteria for Adverse Events (CTCAE) | Protocol Development | CTEP. https://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm.

[2] [n. d.]. Fitbit Wear-Time and Patterns of Activity in Cancer Survivors throughout a Physical Activity Intervention and Follow-up: Exploratory Analysis from a Randomised Controlled Trial | PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0240967.

[3] Daniel A Adler, Yuewen Yang, Thalia Viranda, Xuhai Xu, David C Mohr, Anna R Van Meter, Julia C Tartaglia, Nicholas C Jacobson, Fei Wang, Deborah Estrin, et al. 2024. Beyond Detection: Towards Actionable Sensing Research

in Clinical Mental Healthcare. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 8, 4 (2024), 1–33.

[4] Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, De Choudhury, Srijan Kumar, and Nishanth Sastry. 2024. MedHalu: Hallucinations in Responses to Healthcare Queries by Large Language Models. https://arxiv.org/abs/2409.19492

[5] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large Language Models Are Few-Shot Clinical Information Extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 1998–2022. https://doi.org/10.18653/v1/2022.emnlp-main.130

[6] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large Language Models are Few-Shot Clinical Information Extractors. http://arxiv.org/abs/2205.12689 arXiv:2205.12689 [cs].

[7] Lakshmi Arbatti, Abhishek Hosamath, Vikram Ramanarayanan, and Ira Shoulson. 2023. What Do Patients Say About Their Disease Symptoms? Deep Multilabel Text Classification With Human-in-the-Loop Curation for Automatic Labeling of Patient Self Reports of Problems.

[8] Melina Arnold, Christian C. Abnet, Rachel E. Neale, Jerome Vignat, Edward L. Giovannucci, Katherine A. McGlynn, and Freddie Bray. 2020. Global Burden of 5 Major Types of Gastrointestinal Cancer. *Gastroenterology* 159, 1 (July 2020), 335–349.e15. https://doi.org/10.1053/j.gastro.2020.02.068

[9] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[10] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[11] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[12] Karthik S Bhat, Mohit Jain, and Neha Kumar. 2021. Infrastructuring Telehealth in (In)Formal Patient-Doctor Contexts. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–28. https://doi.org/10.1145/3476064

[13] Federico Bozzetti, Luca Gianotti, Mario Braga, Valerio Di Carlo, and Luigi Mariani. 2007. Postoperative Complications in Gastrointestinal Cancer Patients: The Joint Role of the Nutritional Status and the Nutritional Support. *Clinical Nutrition* 26, 6 (Dec. 2007), 698–709. https://doi.org/10.1016/j.clnu.2007.06.009

[14] Diogo Branco, Margarida Móteiro, Raquel Bouça-Machado, Rita Miranda, Tiago Reis, Élia Decoroso, Rita Cardoso, Joana Ramalho, Filipa Rato, Joana Malheiro, et al. 2024. Co-designing Customizable Clinical Dashboards with Multidisciplinary Teams: Bridging the Gap in Chronic Disease Care. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[15] Hylke JF Brenkman, Leonie Haverkamp, Jelle P Ruurda, and Richard van Hillegersberg. 2016. Worldwide practice in gastric cancer surgery. *World journal of gastroenterology* 22, 15 (2016), 4041.

[16] Robin Brewer, Casey Pierce, Pooja Upadhyay, and Leeseul Park. 2022. An Empirical Study of Older Adult's Voice Assistant Use for Health Information Seeking. *ACM Trans. Interact. Intell. Syst.* 12, 2 (July 2022), 13:1–13:32. https://doi.org/10.1145/3484507

[17] John Brooke. 1995. SUS: A quick and dirty usability scale. *Usability Eval. Ind.* 189 (11 1995).

[18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, 1877–1901.

[19] Jonathan Cardella, Natalie G. Coburn, Anna Gagliardi, Barbara-Anne Maier, Elisa Greco, Linda Last, Andrew J. Smith, Calvin Law, and Frances Wright. 2008. Compliance, Attitudes and Barriers to Post-Operative Colorectal Cancer Follow-Up. *Journal of Evaluation in Clinical Practice* 14, 3 (2008), 407–415. https://doi.org/10.1111/j.1365-2753.2007.00880.x

[20] Luis Cardenas, Katherine Parajes, Ming Zhu, and Shengjie Zhai. 2024. AutoHealth: Advanced LLM-Empowered Wearable Personalized Medical Butler for Parkinson's Disease Management. In *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*. 0375–0379. https://doi.org/10.1109/CCWC60891.2024.10427622

[21] Lauren Carmichael, Rose Rocca, Erin Laing, Phoebe Ashford, Jesse Collins, Luke Jackson, Lauren McPherson, Brydie Pendergast, and Nicole Kiss. 2022. Early Postoperative Feeding Following Surgery for Upper Gastrointestinal Cancer: A Systematic Review. *Journal of Human Nutrition and Dietetics* 35, 1 (2022), 33–48. https://doi.org/10.1111/jhn.12930

[22] Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems* 47, 1 (2023), 33.
[23] George Chalhoub and Ivan Flechais. 2020. "Alexa, Are You Spying on Me?": Exploring the Effect of User Experience on the Security and Privacy of Smart Speaker Users. In *HCI for Cybersecurity, Privacy and Trust*, Abbas Moallem (Ed.). Springer International Publishing, Cham, 305–325. https://doi.org/10.1007/978-3-030-50309-3_21
[24] Szeyi Chan, Jiachen Li, Bingsheng Yao, Amama Mahmood, Chien-Ming Huang, Holly Jimison, Elizabeth D. Mynatt, and Dakuo Wang. 2023. "Mango Mango, How to Let The Lettuce Dry Without A Spinner?": Exploring User Perceptions of Using An LLM-Based Conversational Assistant Toward Cooking Partner. https://doi.org/10.48550/arXiv.2310.05853 arXiv:2310.05853 [cs]
[25] Rajesh Chandwani and Neha Kumar. 2018. Stitching Infrastructures to Facilitate Telemedicine for Low-Resource Environments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173958
[26] Kevin A. Chen, Chinmaya U. Joisa, Karyn B. Stitzenberg, Jonathan Stem, Jose G. Guillem, Shawn M. Gomez, and Muneera R. Kapadia. 2022. Development and Validation of Machine Learning Models to Predict Readmission After Colorectal Surgery. *Journal of Gastrointestinal Surgery* 26, 11 (Nov. 2022), 2342–2350. https://doi.org/10.1007/s11605-022-05443-5
[27] Yuxuan Chen, Haoyan Yang, Hengkai Pan, Fardeen Siddiqui, Antonio Verdone, Qingyang Zhang, Sumit Chopra, Chen Zhao, and Yiqiu Shen. 2024. Burextract-llama: An llm for clinical concept extraction in breast ultrasound reports. In *Proceedings of the 1st International Workshop on Multimedia Computing for Health and Medicine*. 53–58.
[28] Chloe Chira, Evangelos Mathioudis, Christina Michailidou, Pantelis Agathangelou, Georgia Christodoulou, Ioannis Katakis, Efstratios Kontopoulos, and Konstantinos Avgerinakis. 2022. An Affective Multi-modal Conversational Agent for Non Intrusive Data Collection from Patients with Brain Diseases. In *International Workshop on Chatbot Research and Design*. Springer, 134–149.
[29] Avishek Choudhury and Onur Asan. 2020. Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review. *JMIR Medical Informatics* 8, 7 (July 2020), e18599. https://doi.org/10.2196/18599
[30] Chia-Fang Chung, Kristin Dew, Allison Cole, Jasmine Zia, James Fogarty, Julie A Kientz, and Sean A Munson. 2016. Boundary negotiating artifacts in personal informatics: patient-provider collaboration with patient-generated data. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 770–786.
[31] Juliet Clark and Aisling Kelliher. 2021. Understanding the Needs and Values of Rehabilitation Therapists in Designing and Implementing Telehealth Solutions. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–6. https://doi.org/10.1145/3411763.3451704
[32] I Glenn Cohen and Michelle M Mello. 2019. Big data, big tech, and protecting patient privacy. *Jama* 322, 12 (2019), 1141–1142.
[33] Sarah E Cousins, Emma Tempest, and David J Feuer. 2016. Surgery for the resolution of symptoms in malignant bowel obstruction in advanced gynaecological and gastrointestinal cancer. *Cochrane Database of Systematic Reviews* 1 (2016).
[34] J Desrame, V Heinschild, C Desauw, AC Fuerea, P Artru, S Javed, T Papazyan, C Ferté, M Autheman, M Valery, et al. 2024. 595P Adoption of remote patient monitoring in gastrointestinal oncology: A real-world experience from 1822 patients across 47 centers in France and Belgium. *Annals of Oncology* 35 (2024), S478.
[35] Virginia Dignum. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Vol. 2156. Springer.
[36] Nickolas Dreher, Edward Kenji Hadeler, Sheri J. Hartman, Emily C. Wong, Irene Acerbi, Hope S. Rugo, Melanie Catherine Majure, Amy Jo Chien, Laura J. Esserman, and Michelle E. Melisko. 2019. Fitbit Usage in Patients With Breast Cancer Undergoing Chemotherapy. *Clinical Breast Cancer* 19, 6 (Dec. 2019), 443–449.e1. https://doi.org/10.1016/j.clbc.2019.05.005
[37] Jaimie Drozdal, Justin D. Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. 2020. Trust in AutoML: Exploring Information Needs for Establishing Trust in Automated Machine Learning Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 297–307. https://doi.org/10.1145/3377325.3377501
[38] Tim Dwyer, Graeme Hoit, David Burns, James Higgins, Justin Chang, Daniel Whelan, Irene Kiroplis, and Jaskarndip Chahal. 2023. Use of an artificial intelligence conversational agent (chatbot) for hip arthroscopy patients following surgery. *Arthroscopy, Sports Medicine, and Rehabilitation* 5, 2 (2023), e495–e505.
[39] Jide Edu, Xavier Ferrer-Aran, Jose Such, and Guillermo Suarez-Tangil. 2022. Measuring Alexa Skill Privacy Practices across Three Years. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*. Association for Computing Machinery, New York, NY, USA, 670–680. https://doi.org/10.1145/3485447.3512289

[40] Emilio Ferrara. 2024. Large language models for wearable sensor-based human activity recognition, health monitoring, and behavioral modeling: A survey of early trends, datasets, and challenges. *Sensors* 24, 15 (2024), 5045.
[41] Inc. Figma. 2024. *Figma: the collaborative interface design tool.* https://www.figma.com
[42] Shihan Fu, Bingsheng Yao, Smit Desai, Yuqi Hu, Yuling Sun, Samantha Stonbraker, Yanjun Gao, Elizabeth M Goldberg, and Dakuo Wang. 2025. " It Felt Like I Was Left in the Dark": Exploring Information Needs and Design Opportunities for Family Caregivers of Older Adult Patients in Critical Care Settings. *arXiv preprint arXiv:2502.05115* (2025).
[43] A. K. Garth, C. M. Newsome, N. Simmance, and T. C. Crowe. 2010. Nutritional Status, Nutrition Practices and Post-Operative Complications in Patients with Gastrointestinal Cancer. *Journal of Human Nutrition and Dietetics* 23, 4 (2010), 393–401. https://doi.org/10.1111/j.1365-277X.2010.01058.x
[44] Michelle Louise Gatt, Maria Cassar, and Sandra C Buttigieg. 2022. A review of literature on risk prediction tools for hospital readmissions in older adults. *Journal of Health Organization and Management* 36, 4 (2022), 521–557.
[45] L Geoghegan, A Scarborough, JCR Wormald, CJ Harrison, D Collins, M Gardiner, Julie Bruce, and JN Rodrigues. 2021. Automated conversational agents for post-intervention follow-up: a systematic review. *BJS open* 5, 4 (2021), zrab070.
[46] Alireza Ghods, Armin Shahrokni, Hassan Ghasemzadeh, and Diane Cook. 2021. Remote monitoring of the performance status and burden of symptoms of patients with gastrointestinal cancer via a consumer-based activity tracker: quantitative cohort study. *JMIR cancer* 7, 4 (2021), e22931.
[47] Daniela C. Gonçalves-Bradley, Ana Rita J. Maria, Ignacio Ricci-Cabello, Gemma Villanueva, Marita S. Fønhus, Claire Glenton, Simon Lewin, Nicholas Henschke, Brian S. Buckley, Garrett L. Mehl, Tigest Tamrat, and Sasha Shepperd. 2020. Mobile technologies to support healthcare provider to healthcare provider communication and management of care. *Cochrane Database of Systematic Reviews* 8 (2020). https://doi.org/10.1002/14651858.CD012927.pub2 Publisher: John Wiley & Sons, Ltd.
[48] Danny M. den Hamer, Perry Schoor, Tobias B. Polak, and Daniel Kapitan. 2023. Improving Patient Pre-screening for Clinical Trials: Assisting Physicians with Large Language Models. http://arxiv.org/abs/2304.07396 arXiv:2304.07396 [cs].
[49] Yuexing Hao, Jason Holmes, Mark Waddle, Nathan Yu, Kirstin Vickers, Heather Preston, Drew Margolin, Corinna E Löckenhoff, Aditya Vashistha, Marzyeh Ghassemi, et al. 2024. Outlining the Borders for LLM Applications in Patient Education: Developing an Expert-in-the-Loop LLM-Powered Chatbot for Prostate Cancer Patient Education. *arXiv preprint arXiv:2409.19100* (2024).
[50] Yuexing Hao, Zeyu Liu, Robert N Riter, and Saleh Kalantari. 2024. Advancing Patient-Centered Shared Decision-Making with AI Systems for Older Adult Cancer Patients. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.
[51] Elizabeth Healey and Isaac Kohane. 2024. LLM-CGM: A Benchmark for Large Language Model-Enabled Querying of Continuous Glucose Monitoring Data for Conversational Diabetes Management. In *Biocomputing 2025: Proceedings of the Pacific Symposium*. World Scientific, 82–93.
[52] Thanh Cong Ho, Farah Kharrat, Abderrazek Abid, Fakhri Karray, and Anis Koubaa. 2024. REMONI: An Autonomous System Integrating Wearables and Multimodal Large Language Models for Enhanced Remote Health Monitoring. In *2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 1–6.
[53] Thanh Cong Ho, Farah Kharrat, Abderrazek Abid, Fakhri Karray, and Anis Koubaa. 2024. REMONI: An Autonomous System Integrating Wearables and Multimodal Large Language Models for Enhanced Remote Health Monitoring. In *2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. 1–6. https://doi.org/10.1109/MeMeA60663.2024.10596778
[54] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–19. https://doi.org/10.1145/3544548.3580688
[55] Maia Jacobs, Jeremy Johnson, and Elizabeth D. Mynatt. 2018. MyPath: Investigating Breast Cancer Patients' Use of Personalized Health Information. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 78:1–78:21. https://doi.org/10.1145/3274347
[56] Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–16. https://doi.org/10.1145/3544548.3581503
[57] Maura Kennedy, Richard A Enander, Sarah P Tadiri, Richard E Wolfe, Nathan I Shapiro, and Edward R Marcantonio. 2014. Delirium risk prediction, healthcare use and mortality of elderly adults in the emergency department. *Journal of the American Geriatrics Society* 62, 3 (2014), 462–469.
[58] Charalampia (Xaroula) Kerasidou, Angeliki Kerasidou, Monika Buscher, and Stephen Wilkinson. 2022. Before and beyond Trust: Reliance in Medical AI. *Journal of Medical Ethics* 48, 11 (Nov. 2022), 852–856. https://doi.org/10.1136/medethics-2020-107095

[59] Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, and Xin Liu. 2025. Medical Hallucinations in Foundation Models and Their Impact on Healthcare. https://arxiv.org/abs/2503.05777

[60] Zachary D. King, Judith Moskowitz, Begum Egilmez, Shibo Zhang, Lida Zhang, Michael Bass, John Rogers, Roozbeh Ghaffari, Laurie Wakschlag, and Nabil Alshurafa. 2019. Micro-Stress EMA: A Passive Sensing Framework for Detecting in-the-Wild Stress in Pregnant Mothers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (Sept. 2019), 91:1–91:22. https://doi.org/10.1145/3351249

[61] A Baki Kocaballi, Kiran Ijaz, Liliana Laranjo, Juan C Quiroz, Dana Rezazadegan, Huong Ly Tong, Simon Willcock, Shlomo Berkovsky, and Enrico Coiera. 2020. Envisioning an artificial intelligence documentation assistant for future primary care consultations: A co-design study with general practitioners. *Journal of the American Medical Informatics Association* 27, 11 (Nov. 2020), 1695–1704.

[62] Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu, Samantha Stonbraker, Bingsheng Yao, Dakuo Wang, Elizabeth Goldberg, and Yanjun Gao. 2025. Large Language Models with Temporal Reasoning for Longitudinal Clinical Summarization and Prediction. In *Findings of ACL. EMNLP. Conference on Empirical Methods in Natural Language Processing*, Vol. 2025. 20715–20735.

[63] Julia Lai-Kwon, Claudia Rutherford, Stephanie Best, Thai Ly, Iris Zhang, Catherine Devereux, Dishan Herath, Kate Burbury, and Michael Jefford. 2024. Co-design of an electronic patient-reported outcome symptom monitoring system for immunotherapy toxicities. *Supportive Care in Cancer* 32, 12 (2024), 1–14.

[64] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* (Sept. 2019), btz682. https://doi.org/10/ggh5qq

[65] Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine* 388, 13 (March 2023), 1233–1239. https://doi.org/10.1056/NEJMsr2214184

[66] Maria Alejandra León, Valeria Pannunzio, and Maaike Kleinsmann. 2022. The impact of perioperative remote patient monitoring on clinical staff workflows: scoping review. *JMIR Human Factors* 9, 2 (2022), e37204.

[67] James R Lewis. 2018. The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction* 34, 7 (2018), 577–590.

[68] Brenna Li, Ofek Gross, Noah Crampton, Mamta Kapoor, Saba Tauseef, Mohit Jain, Khai N Truong, and Alex Mariakakis. 2024. Beyond the Waiting Room: Patient's Perspectives on the Conversational Nuances of Pre-Consultation Chatbots. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.

[69] Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. 2023. Meddm: Llm-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441* (2023).

[70] Hongjin Lin, Tessa Han, Krzysztof Z Gajos, and Anoopum S Gupta. 2024. Hevelius Report: Visualizing Web-Based Mobility Test Data For Clinical Decision and Learning Support. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–10.

[71] Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large Language Models Are Few-Shot Health Learners. https://doi.org/10.48550/arXiv.2305.15525 arXiv:2305.15525 [cs]

[72] Yuxuan Lu, Jingya Yan, Zhixuan Qi, Zhongzheng Ge, and Yongping Du. 2022. Contextual Embedding and Model Weighting by Fusing Domain Knowledge on Biomedical Question Answering. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3535508.3545508

[73] Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2023. LLM-Powered Conversational Voice Assistants: Interaction Patterns, Opportunities, Challenges, and Design Guidelines. arXiv:2309.13879 [cs]

[74] Abdul. Majid, Yulin. Wang, Jehad. Ali, Anwar. Ullah, and Kiran. Perveen. 2025. Multimodal LLM for Patient Activity Recognition: Integrating Video, Audio, and Text in Clinical Environments. *IEEE Journal of Biomedical and Health Informatics* (2025), 1–11. https://doi.org/10.1109/JBHI.2025.3617581

[75] Lakmini P. Malasinghe, Naeem Ramzan, and Keshav Dahal. 2019. Remote Patient Monitoring: A Comprehensive Study. *Journal of Ambient Intelligence and Humanized Computing* 10, 1 (Jan. 2019), 57–76. https://doi.org/10.1007/s12652-017-0598-x

[76] Jayson S Marwaha, Adam B Landman, Gabriel A Brat, Todd Dunn, and William J Gordon. 2022. Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. *NPJ digital medicine* 5, 1 (2022), 13.

[77] George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System

Metathesaurus. https://doi.org/10.48550/arXiv.2010.10391 arXiv:2010.10391 [cs]
[78] Lucian Mocan. 2021. Surgical Management of Gastric Cancer: A Systematic Review. *Journal of Clinical Medicine* 10, 12 (Jan. 2021), 2557. https://doi.org/10.3390/jcm10122557
[79] Sara Montagna, Stefano Ferretti, Lorenz Cuno Klopfenstein, Antonio Florio, and Martino Francesco Pengo. 2023. Data Decentralisation of LLM-Based Chatbot Systems in Chronic Disease Self-Management. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*. 205–212.
[80] Sara Montagna, Stefano Ferretti, Lorenz Cuno Klopfenstein, Michelangelo Ungolo, Martino Francesco Pengo, Gianluca Aguzzi, and Matteo Magnini. 2025. Privacy-Preserving LLM-based Chatbots for Hypertensive Patient Self-Management. *Smart Health* 36 (June 2025), 100552. https://doi.org/10.1016/j.smhl.2025.100552
[81] Blake Murdoch. 2021. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Medical Ethics* 22 (2021), 1–5.
[82] Varun Nair, Elliot Schumacher, and Anitha Kannan. 2023. Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models. http://arxiv.org/abs/2305.05982 arXiv:2305.05982 [cs].
[83] Lin Ni, Chenhao Lu, Niu Liu, and Jiamou Liu. 2017. Mandy: Towards a smart primary care chatbot application. In *International symposium on knowledge and systems sciences*. Springer, 38–52.
[84] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).
[85] Benedict U Nwachukwu, Nathan H Varady, Answorth A Allen, Joshua S Dines, David W Altchek, Riley J Williams III, and Kyle N Kunze. 2024. Currently available large language models do not provide musculoskeletal treatment recommendations that are concordant with evidence-based clinical practice guidelines. *Arthroscopy: The Journal of Arthroscopic & Related Surgery* (2024).
[86] Rónán O'Caoimh, Nicola Cornally, Elizabeth Weathers, Ronan O'Sullivan, Carol Fitzgerald, Francesc Orfila, Roger Clarnette, Constança Paúl, and D William Molloy. 2015. Risk prediction in the community: A systematic review of case-finding instruments that predict adverse healthcare outcomes in community-dwelling older adults. *Maturitas* 82, 1 (2015), 3–21.
[87] Batyrkhan Omarov, Sergazi Narynov, Zhandos Zhumanov, Elmira Alzhanova, Aidana Gumar, and Mariyam Khassanova. 2022. Artificial Intelligence Enabled Conversational Agent for Mental Healthcare. *International journal of health sciences* 6, 3 (Oct. 2022), 1544–1555. https://doi.org/10.53730/ijhs.v6n3.13239
[88] OpenAI. 2022. Introducing chatgpt. https://openai.com/blog/chatgpt
[89] OpenAI. 2023. GPT-4. https://openai.com/index/gpt-4/ Accessed: 2025-01-18.
[90] OpenAI. 2023. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs]
[91] Valeria Pannunzio, Hosana Cristina Morales Ornelas, Pema Gurung, Robert van Kooten, Dirk Snelders, Hendrikus van Os, Michel Wouters, Rob Tollenaar, Douwe Atsma, and Maaike Kleinsmann. 2024. Patient and Staff Experience of Remote Patient Monitoring—What to Measure and How: Systematic Review. *Journal of Medical Internet Research* 26 (2024), e48463.
[92] Farhat Tasnim Progga, Amal Khan, and Sabirat Rubya. 2024. Large Language Models and Personalized Storytelling for Postpartum Wellbeing. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (San Jose, Costa Rica) *(CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 653–657. https://doi.org/10.1145/3678884.3681921
[93] Niroop Channa Rajashekar, Yeo Eun Shin, Yuan Pu, Sunny Chung, Kisung You, Mauro Giuffre, Colleen E Chan, Theo Saarinen, Allen Hsiao, Jasjeet Sekhon, et al. 2024. Human-Algorithmic Interaction Using a Large Language Model-Augmented Artificial Intelligence Clinical Decision Support System. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.
[94] Liz Salmi, Dana M Lewis, Jennifer L Clarke, Zhiyong Dong, Rudy Fischmann, Emily I McIntosh, Chethan R Sarabu, and Catherine M DesRoches. [n. d.]. Harnessing AI for Patient Engagement in a Study on Large Language Models and Open Notes. ([n. d.]).
[95] Jessica Schroeder, Jane Hoffswell, Chia-Fang Chung, James Fogarty, Sean Munson, and Jasmine Zia. 2017. Supporting Patient-Provider Collaboration to Identify Individual Triggers using Food and Symptom Journals. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1726–1739. https://doi.org/10.1145/2998181.2998276
[96] John L. Semple, Sarah Sharpe, M. Lucas Murnaghan, John Theodoropoulos, and Kelly A. Metcalfe. 2015. Using a Mobile App for Monitoring Post-Operative Quality of Recovery of Patients at Home: A Feasibility Study. *JMIR mHealth and uHealth* 3, 1 (Feb. 2015), e3929. https://doi.org/10.2196/mhealth.3929
[97] Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao'Kenneth' Huang. 2023. ConvXAI: Delivering Heterogeneous AI Explanations via Conversations to Support Human-AI Scientific Writing. *arXiv preprint arXiv:2305.09770* (2023).

[98] Line Silsand and Gunnar Ellingsen. 2016. Complex Decision-Making in Clinical Practice. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 993–1004. https://doi.org/10.1145/2818048.2819952

[99] Karine De Almeida Silva, Arenamoline Xavier Duarte, Amanda Rodrigues Cruz, Letícia Oliveira Cardoso, Thatty Christina Morais Santos, and Geórgia Das Graças Pena. 2020. Ostomy time and nutrition status were associated on quality of life in patients with colorectal cancer. *Journal of Coloproctology* 40, 04 (Dec. 2020), 352–361. https://doi.org/10.1016/j.jcol.2020.07.003

[100] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large Language Models Encode Clinical Knowledge. *Nature* 620, 7972 (Aug. 2023), 172–180. https://doi.org/10.1038/s41586-023-06291-2

[101] Helen Smith. 2021. Clinical AI: opacity, accountability, responsibility and liability. *Ai & Society* 36, 2 (2021), 535–545.

[102] Stephanie Staras, Justin S Tauscher, Natalie Rich, Esaa Samarah, Lindsay A Thompson, Michelle M Vinson, Michael J Muszynski, Elizabeth A Shenkman, et al. 2021. Using a clinical workflow analysis to enhance eHealth implementation planning: tutorial and case study. *JMIR mHealth and uHealth* 9, 3 (2021), e18534.

[103] Glenn Steele Jr. 1993. Standard postoperative monitoring of patients after primary resection of colon and rectum cancer. *Cancer* 71, S12 (1993), 4225–4235.

[104] Allan Stisen and Nervo Verdezoto. 2017. Clinical and Non-Clinical Handovers: Designing for Critical Moments. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2166–2178. https://doi.org/10.1145/2998181.2998333

[105] Claire Temple-Oberle, Spencer Yakaback, Carmen Webb, Golpira Elmi Assadzadeh, and Gregg Nelson. 2023. Effect of smartphone app postoperative home monitoring after oncologic surgery on quality of recovery: a randomized clinical trial. *JAMA surgery* 158, 7 (2023), 693–699.

[106] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs] http://arxiv.org/abs/2307.09288

[107] Brian D Tran, Kareem Latif, Tera L Reynolds, Jihyun Park, Jennifer Elston Lafata, Ming Tai-Seale, and Kai Zheng. 2023. "Mm-hm,""Uh-uh": are non-lexical conversational sounds deal breakers for the ambient clinical documentation technology? *Journal of the American Medical Informatics Association* 30, 4 (2023), 703–711.

[108] Satvik Tripathi, Rithvik Sukumaran, and Tessa S Cook. 2024. Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care. *Journal of the American Medical Informatics Association* 31, 6 (2024), 1436–1440.

[109] Robert T. van Kooten, Renu R. Bahadoer, Koen C. M. J. Peeters, Jetty H. L. Hoeksema, Ewout W. Steyerberg, Henk H. Hartgrink, Cornelis J. H. van de Velde, Michel W. J. M. Wouters, and Rob A. E. M. Tollenaar. 2021. Preoperative Risk Factors for Major Postoperative Complications after Complex Gastrointestinal Cancer Surgery: A Systematic Review. *European Journal of Surgical Oncology* 47, 12 (Dec. 2021), 3049–3058. https://doi.org/10.1016/j.ejso.2021.07.021

[110] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From Human–Human Collaboration to Human–AI Collaboration: Designing AI Systems That Can Work Together with People. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3334480.3382929

[111] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural China: Tensions and Challenges in AI-Powered CDSS Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[112] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM collaborative annotation through effective verification of LLM labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.

[113] Ziyi Wang, Yuxuan Lu, Wenbo Li, Amirali Amini, Bo Sun, Yakov Bart, Weimin Lyu, Jiri Gesi, Tian Wang, Jing Huang, Yu Su, Upol Ehsan, Malihe Alikhani, Toby Jia-Jun Li, Lydia Chilton, and Dakuo Wang. 2025. OPeRA: A Dataset of Observation, Persona, Rationale, and Action for Evaluating LLMs on Human Online Shopping Behavior Simulation. arXiv:2506.05606 [cs.CL] https://arxiv.org/abs/2506.05606

[114] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2023. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data.

[115] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[116] Joel Wester, Bhakti Moghe, Katie Winkle, and Niels van Berkel. 2024. Facing LLMs: Robot Communication Styles in Mediating Health Information between Parents and Young Adults. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–37.

[117] Martin C. S. Wong, Junjie Huang, Paul S. F. Chan, Peter Choi, Xiang Qian Lao, Shannon Melissa Chan, Anthony Teoh, and Peter Liang. 2021. Global Incidence and Mortality of Gastric Cancer, 1980-2018. *JAMA Network Open* 4, 7 (July 2021), e2118457. https://doi.org/10.1001/jamanetworkopen.2021.18457

[118] Siyi Wu, Weidan Cao, Shihan Fu, Bingsheng Yao, Ziqi Yang, Changchang Yin, Varun Mishra, Daniel Addison, Ping Zhang, and Dakuo Wang. 2024. Clinical Challenges and AI Opportunities in Decision-Making for Cancer Treatment-Induced Cardiotoxicity. *arXiv preprint arXiv:2408.03586* (2024).

[119] Siyi Wu, Weidan Cao, Shihan Fu, Bingsheng Yao, Ziqi Yang, Changchang Yin, Varun Mishra, Daniel Addison, Ping Zhang, and Dakuo Wang. 2025. CardioAI: A Multimodal AI-based System to Support Symptom Monitoring and Risk Prediction of Cancer Treatment-Induced Cardiotoxicity. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 313, 22 pages. https://doi.org/10.1145/3706598.3714272

[120] Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. 2025. Collabllm: From passive responders to active collaborators. *arXiv preprint arXiv:2502.00640* (2025).

[121] Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. 75–78.

[122] Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 3 (2020), 1–37.

[123] Yukinori Yamagata, Takaki Yoshikawa, Masahiro Yura, Sho Otsuki, Shinji Morita, Hitoshi Katai, and Toshiro Nishida. 2019. Current Status of the "Enhanced Recovery after Surgery" Program in Gastric Cancer Surgery. *Annals of Gastroenterological Surgery* 3, 3 (2019), 231–238. https://doi.org/10.1002/ags3.12232

[124] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Jiachen Li, Jennifer Bagdasarian, Guodong Gao, and Dakuo Wang. 2024. "I Wish There Were an AI": Challenges and AI Potential in Cancer Patient-Provider Communication. arXiv:2404.13409 [cs]

[125] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 73 (May 2024), 35 pages. https://doi.org/10.1145/3659625

[126] Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Yisi Sang, Sijia Liu, James Hendler, and Dakuo Wang. 2024. More Samples or More Prompts? Exploring Effective Few-Shot In-Context Learning for LLMs with In-Context Sampling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 1772–1790. https://doi.org/10.18653/v1/2024.findings-naacl.115

[127] Bingsheng Yao, Menglin Zhao, Yuling Sun, Weidan Cao, Changchang Yin, Stephen Intille, Xuhai Xu, Ping Zhang, Jingzhen Yang, and Dakuo Wang. 2025. More Modality, More AI: Exploring Design Opportunities of AI-Based Multi-modal Remote Monitoring Technologies for Early Detection of Mental Health Sequelae in Youth Concussion Patients. *arXiv preprint arXiv:2502.03732* (2025).

[128] Bingsheng Yao, Menglin Zhao, Zhan Zhang, Pengqi Wang, Emma G Chester, Changchang Yin, Tianshi Li, Varun Mishra, Lace Padilla, Odysseas Chatzipanagiotou, et al. 2025. Exploring Collaboration Breakdowns Between Provider Teams and Patients in Post-Surgery Care. *arXiv preprint arXiv:2509.23509* (2025).

[129] H Yasunaga, H Horiguchi, S Matsuda, K Fushimi, H Hashimoto, and J Z Ayanian. 2013. Body Mass Index and Outcomes Following Gastrointestinal Cancer Surgery in Japan. *British Journal of Surgery* 100, 10 (Sept. 2013), 1335–1343. https://doi.org/10.1002/bjs.9221

[130] Nur Yildirim, Hannah Richardson, Maria T. Wetscherek, Junaid Bajwa, Joseph Jacob, Mark A. Pinnock, Stephen Harris, Daniel Coelho de Castro, Shruthi Bannur, Stephanie L. Hyland, Pratik Ghosh, Mercy Ranjit, Kenza Bouzid, Anton Schwaighofer, Fernando Pérez-García, Harshita Sharma, Ozan Oktay, Matthew Lungren, Javier Alvarez-Valle, Aditya Nori, and Anja Thieme. 2024. Multimodal Healthcare AI: Identifying and Designing Clinically Relevant Vision-Language Applications for Radiology. https://doi.org/10.1145/3613904.3642013 arXiv:2402.14252 [cs]

[131] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070* (2023).

[132] Xiao Zhan, Noura Abdi, William Seymour, and Jose Such. 2024. Healthcare Voice AI Assistants: Factors Influencing Trust and Intention to Use. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 62:1–62:37. https://doi.org/10.1145/3637339

[133] Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M. Padilla, Jeffrey Caterino, Ping Zhang, and Dakuo Wang. 2024. Rethinking Human-AI Collaboration in Complex Medical Decision Making: A Case Study in Sepsis Diagnosis. https://doi.org/10.1145/3613904.3642343 arXiv:2309.12368 [cs]

[134] Zhan Zhang, Aleksandra Sarcevic, and Claus Bossen. 2017. Constructing Common Information Spaces across Distributed Emergency Medical Teams. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 934–947. https://doi.org/10.1145/2998181.2998328

[135] Menglin Zhao, Zhuorui Yong, Ruijia Guan, Kai-Wei Chang, Adrian Haimovich, Kei Ouchi, Timothy Bickmore, Bingsheng Yao, Dakuo Wang, and Smit Desai. 2025. Designing AI Tools for Clinical Care Teams to Support Serious Illness Conversations with Older Adults in the Emergency Department. *arXiv preprint arXiv:2506.00241* (2025).

A APPENDIX

A.1 Initial Design Components

A.1.1 Example conversation from initial LLM CA design. In this version, although the LLM CA responses fit the postoperative context, there were potentially inappropriate expressions such as assessing patient conditions, not covering most questions, and miscommunication of system responsibilities.

Table 5: An example conversation between a test user (simulating a patient) and the CA with our initial prompt design.

Role	Content
User:	Hello
CA:	Good afternoon! This is "surgery recover" from the clinic, I'm calling to see how you're doing after your cancer surgery. Have you been experiencing any specific discomforts or symptoms that you would like to discuss?
User:	I am good, thanks for asking
CA:	That's good to hear. Let's continue with our check-up. How about your food intake? Have you noticed any change in your appetite or taste?
User:	My food intake is good, I have great appetite
CA:	That's great! You are maintaining well. Have you noticed any changes in your bowel movements? Any issues like constipation or bloating?
User:	Well I may have ate a lot, so my bowel movement may be a little more frequent than usual, but I think that is fine
CA:	I see, an increase in bowel movement due to an increase in food intake isn't necessarily a cause for concern. However, if you experience any significant changes such as severe diarrhea, constipation or blood in stool, these should be reported to your healthcare provider immediately. Now, let's move to another area. Have you noticed any pain or swelling anywhere in your body after surgery?
User:	Ok. Well, actually I have some pain around my stoma
CA:	I'm sorry to hear that you're experiencing some pain around your stoma. Can you tell me more about the pain? How would you rate the severity of the pain on a scale from 1-10, with 1 being mild discomfort and 10 the worst pain imaginable?
User:	I would say 5 out of 10 when I move around, and maybe 2 or 3 out of 10 when I sit still.
CA:	I understand. It's not uncommon to experience some pain around the stoma after surgery. Can you tell me if the pain is constant, or does it come and go? Also, have you noticed any redness, swelling, or changes in the color of output from your stoma?
User:	I think it is consistent, but more severe sometimes than others. I am not sure about other things, it seems normal

Continued on next page

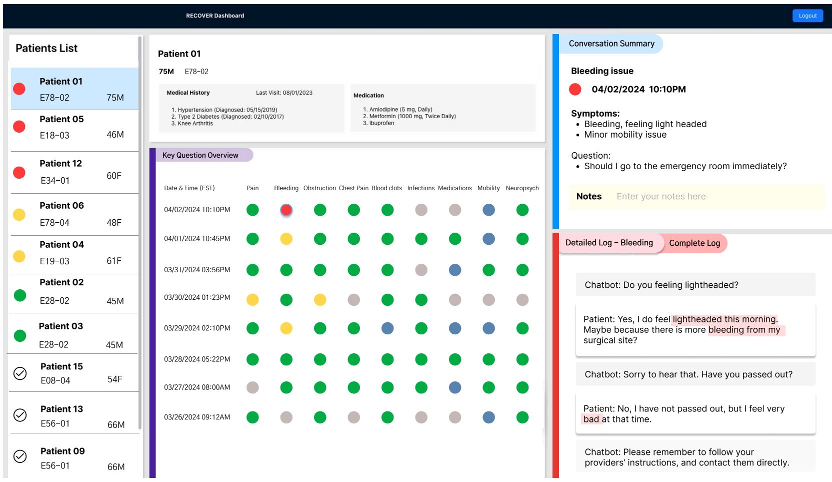


Fig. 10. The first version of dashboard interface design. This version only include simple interactions like selecting patient to view details, and scrolling down to see conversation logs. After our PD sessions, we iterated the design of each tab and interactions to promote clinical efficiency.

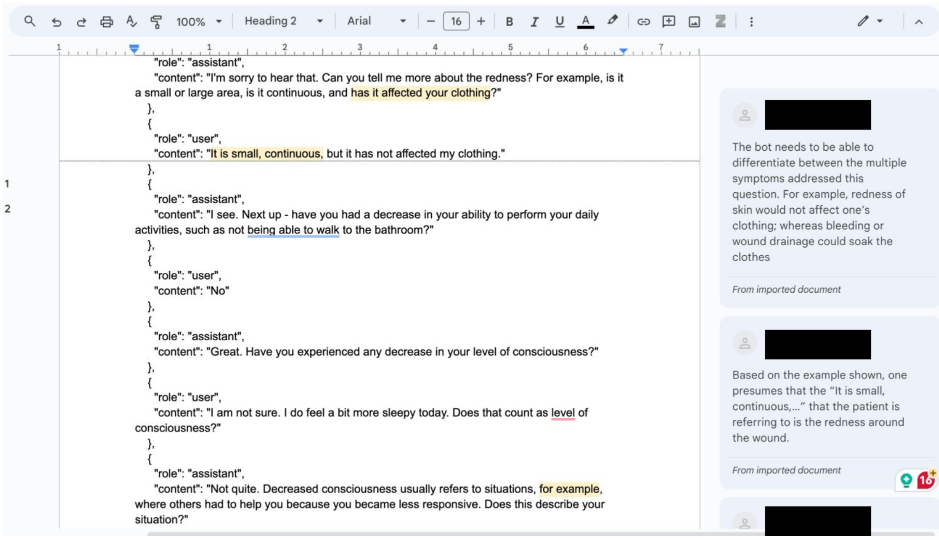
Table 5: An example conversation between a test user (simulating a patient) and the CA with our initial prompt design. (Continued)

Role	**Content**
CA:	Thank you for providing those details. I understand the pain can be quite uncomfortable. I suggest you take a note of the times when the pain increases, and if the pain becomes severe or persists it's important to report it to your healthcare provider. Is there anything else you would like to discuss or any other symptoms you've noticed?
User:	Ok, thanks for telling me that. Can you ask my doctor whether I should go in for a visit?
CA:	Sure, I will make sure to pass on your concern to your healthcare provider and they can decide whether an in-person visit is needed. Do you have any other concerns you want me to communicate?
User:	No, thank you
CA:	I'm glad to hear you're generally doing well. I will let your healthcare provider know about your slight discomfort and they will determine whether a visit is necessary. If you have any further concerns or symptoms arise, please don't hesitate to call us. Take care.
User:	Bye

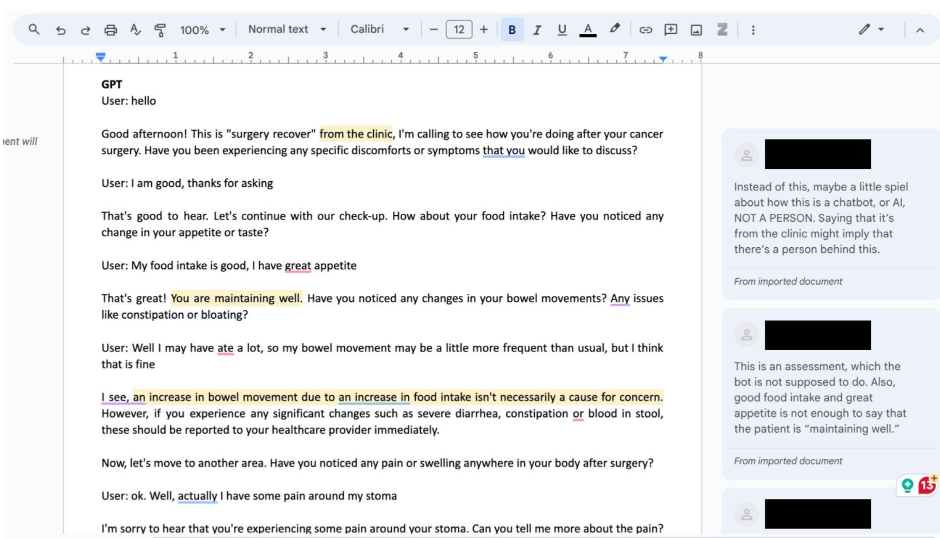
A.1.2 Initial dashboard interface design.

A.2 PD Session Artifact Examples

The examples are presented in Fig 11a to 14.



(a) Participants review the conversation log generated with one prompt and leave comments.



(b) Participants review the conversation log generated with an updated prompt and leave comments.

Fig. 11. Example of participants' comments in Google Docs.

We also present more examples for the website implementation during the PD process in Fig. 12. Each version adapted EP inputs from previous sessions, and all patient data are synthetic data for demonstration purposes. These implementations are all developed with the Vue.js framework and NaiveUI components as described in 4.3.

A.3 Patient Interview Script

(1) Could you please share a bit of background about your cancer treatment and recovery process?

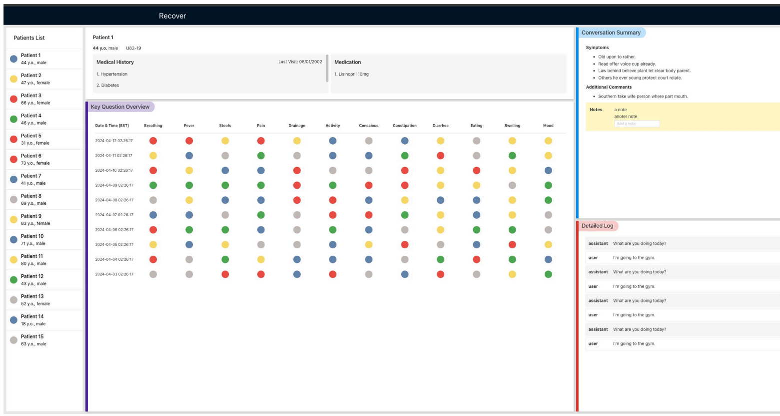
(2) When you are at home for recovery after your surgery, could you recall a most impressive experience that you had with your healthcare providers? Follow-up questions may include:
  (a) Who reached out to whom?
  (b) Why? What's the conversation like?
  (c) What technology did you use?
  (d) What challenges did you encounter?

(3) Do you know about AI? How about ChatGPT? e.g. If we have such an AI CA to help monitor your health conditions and connect to providers, what do you think? Follow-up questions may include:
  (a) What features would you expect?
  (b) Do you think ...(an example feature) would be helpful? Why?

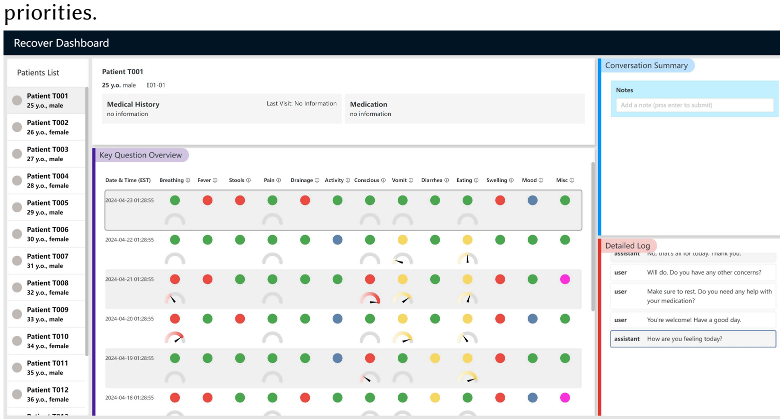
(4) Was there anything else you wanted to share with us about your post-treatment experience?

Table 6. Questions about System Features and LLM and Results

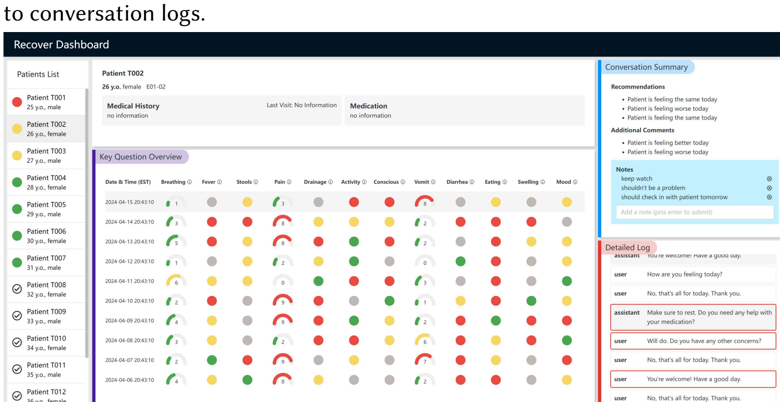
System Features and LLM	Mode	Max	Min
Overall, I think this system can help with my work monitoring postoperative GI cancer patients.	Strongly Agree	Strongly Agree	Strongly Agree
I think the conversation between the patient and the LLM chatbot is natural.	Agree	Strongly Agree	Neither Agree Nor Disagree
I think the LLM chatbot collects crucial information for monitoring the patient from the conversation.	Strongly Agree	Strongly Agree	Agree
I think the LLM chatbot responses raise minimal safety risk to patients.	Strongly Agree	Strongly Agree	Strongly Disagree
I find the central visualization (colored dots and meters) correctly reflects patients' response.	Strongly Agree	Strongly Agree	Strongly Agree
I think the central visualization (colored dots and meters) helped me focus on key symptoms to review.	Strongly Agree	Strongly Agree	Strongly Agree
I find the LLM summary summarizes key information in patients' response correctly.	Strongly Agree	Strongly Agree	Strongly Agree
I think the LLM summary helped me review patient information efficiently.	Strongly Agree	Strongly Agree	Strongly Agree
I think the conversation logs are linked with corresponding visualizations correctly.	Strongly Agree	Strongly Agree	Strongly Agree
I think the conversation logs and highlights helped me to navigate through the details corresponding to each question efficiently.	Strongly Agree	Strongly Agree	Strongly Agree
I think the interaction features (write notes, change status) can help me manage my patients effectively.	Strongly Agree	Strongly Agree	Strongly Agree



(a) In this implementation, we updated the meter design and coloring rules in key question overview. We also added details for summary notes and sorted participant priorities.

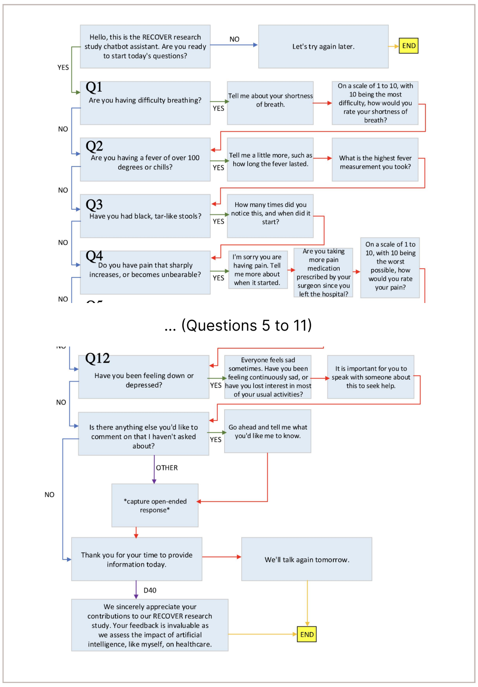


(b) In this implementation, we added the meter design and mapped key questions to conversation logs.

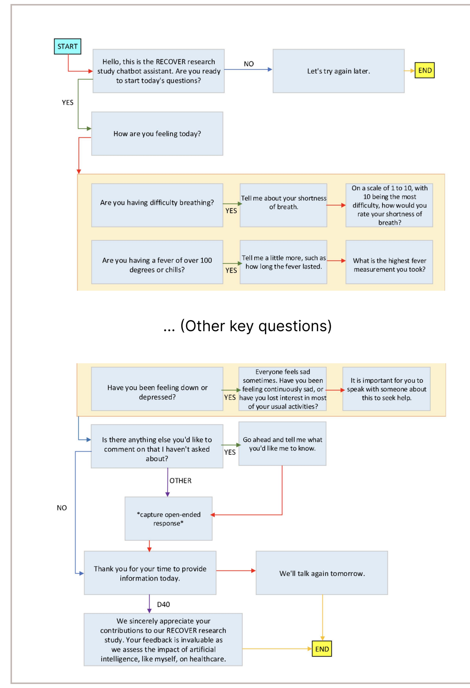


(c) In this implementation, we updated the meter design and coloring rules in key question overview. We also added details for summary notes and sorted participant priorities.

Fig. 12. Screenshots of website implementations presented to participants during PD sessions.



(a) The first conversation flow diagram upon participant discussion, where key questions are asked in order



(b) Updated conversation flow diagram, where conversation logic is more flexible.

Fig. 13. Example of conversation flow iteration.

RECOVER bot questions - ranked by order of significance/clinical concern					
instead of individial (1-13) priority, DP2 prefers 3 groupings. Red = most severe; Yellow = moderate severity; Blue = least clinical severity. Green will be used to indicate no problems reported.					
DP3	**DP4**	**DP2**	**bold questions also have an associated scale of 1-10**	**Agreed coloring**	**Comments / Updated question expression**
			Are you having difficulty breathing?		If yes -> red; meter shows 1/10;
			Are you having a fever of over 100 degrees, or chills?		
			have you had black, tar-like stools?		have you had blood or black tar-like stools?
			Do you have pain that sharply increases, or becomes unbearable?		Do you have pain -> Likert scale, provider review
			Are you having any wound drainage problems, such as redness around your wound, bleeding from the wound, pus, or an opening at the incision site?		
			Do you have a decrease in your ability to perform your daily activities, such as not being able to walk to the bathroom?		
			Have you had a decrease in your level of consciousness?		
			Have you had persistent constipation, nausea, or vomiting?		
			Have you had persistent diarrhea?		
			Have you been unable to tolerate food or drink?		
			Do you have unexplained or new pain or swelling in one of both of your legs?		Do you have unexplained pain or swelling in one of both of your legs?
			Have you been feelling down or depressed?		
			Is there anything else you'd like to comment on that I haven't asked about?		Another color e.g. pink - the patient has mentiond something; by default will be green (no problem)
					Green: no problem
					Grey: no answer (for the 12 questions)

Fig. 14. Participants discussed the color coding depending on the priority of a symptom, and which questions should include a meter with scale.

Table 7. Patient User Study Questionnaire Part 2 and Results. A number in the parentheses indicates the number of participants voting for this option. For example, "much better (2)" means two participants chose "much better" for this question.

Question	Most Chosen	2nd Most Chosen	3rd Most Chosen
The availability of the Recover Bot is __ compared to healthcare providers.	much better (2)	moderately better(2)	slightly worse (1)
The ease of understanding the response provided by the RECOVER bot is __ compared to healthcare providers.	much better (2)	about the same (2)	moderately worse (1)
The extent to which the RECOVER bot covers key symptom details is__ compared to contacting healthcare providers.	about the same (3)	much better (2)	/
The availability of the RECOVER bot is __ compared to questionnaires.	much better (3)	slightly better (1)	/
The ease of understanding the questions provided by the RECOVER bot is __ compared to questionnaires.	much better (2)	slightly better (2)	/
The extent to which the RECOVER bot covers the key symptom details is __ compared to questionnaires.	much better (2)	about the same (2)	moderately better (1)

Table 8. System Usability Scale Questionnaire

System Usability Scale	1	2	3	4	5
I think that I would like to use this system frequently.					
I found the system unnecessarily complex.					
I thought the system was easy to use.					
I think that I would need the support of a technical person to be able to use this system.					
I found the various functions in this system were well integrated.					
I thought there was too much inconsistency in this system.					
I would imagine that most people would learn to use this system very quickly.					
I found the system very cumbersome to use.					
I felt very confident using the system.					
I needed to learn a lot of things before I could get going with this system.					

In this table, the numbers 1 to 5 present the Likert scale from “Strongly Disagree” to “Strongly Agree”

A.4 Conversation Log Example

Table 9: Conversation log between the user and the assistant

Role	**Content**
User	hi
CA	Hello, this is the RECOVER research study chatbot assistant. Are you ready to start today's questions?
User	yes
CA	Great! How have you been feeling today? Any specific discomforts?
User	I'm feeling a sharp pain in my stomach and I'm not able to eat because of the pain
CA	I'm sorry to hear that you are in pain. Does it increase sharply or becomes unbearable?
User	yes
CA	Could you tell me more about when the pain started?
User	it started around 12 hours ago
CA	Have you taken more pain medication prescribed by your surgeon since the pain started?
User	I'm taking Tylenol #3
CA	On a scale of 1 to 10, with 10 being the worst possible, how would you rate your pain?
User	8
CA	On a scale of 1 to 10, with 10 being the most difficult, how bad are you able to tolerate food?
User	8
CA	Are you having difficulty breathing?
User	nope
CA	Are you having a fever of over 100 degrees or chills?
User	no
CA	Have you had black, tar-like stools?
User	no
CA	Are you having any wound drainage problems, such as redness around your wound, bleeding from the wound, pus, or an opening at the incision site?
User	no
CA	Do you have a decrease in your ability to perform your daily activities, such as not being able to walk to the bathroom?
User	no
CA	Have you had a decrease in your level of consciousness?
User	no

Continued on next page

Table 9: Conversation log between the user and the assistant (Continued)

Role	Content
CA	Have you had persistent constipation, nausea, or vomiting?
User	yes I have been constipated for 24 hours
CA	On a scale of 1 to 10, with 10 being most significant, how would you rate your level of constipation?
User	maybe 6
CA	Have you had persistent diarrhea?
User	no
CA	Do you have unexplained or new pain or swelling in one or both of your legs?
User	no
CA	Have you been feeling down or depressed?
User	no I'm feeling OK
CA	Is there anything else you'd like to comment on that I haven't asked about?
User	that's all thank you bye
CA	Thank you for your time to provide information today. We'll talk again tomorrow.

A.5 Cancer Care Provider User Study Protocols and Results

A.5.1 Evaluation Questionnaire and Results.

A.5.2 Follow-up Interview Questions for Healthcare Providers.

(1) How do you like the system?
(2) What are those existing features that may help you the most in monitoring postoperative GI cancer patients?
(3) Would you expect any new novel features that our system does not have, for your RPM?
(4) Do you have any concerns about RECOVER or such systems in future adoption?
(5) Do you have any questions for us?

A.6 GI Cancer Patient User Study Protocols and Results

A.6.1 Evaluation Questionnaire and Results. Patient participants were first asked to rate the AI chatbot system on SUS (part 1 of the questionnaire) and the SUS questionnaire contains the same questions as mentioned in A.5.1.

The second part of the questionnaire is presented in Table 7:

A.6.2 Follow-up Interview Questions for GI Cancer Patients.

(1) How do you like the system?
    (a) What may help you the most when talking about your symptoms?
(2) Would you expect any new features to be added to the system, for your postoperative recovery?
(3) Do you have any concerns about RECOVER or such systems in future adoption?

A.6.3 Patient Task Completion Result.

Table 10. Participant performance in tasks navigating through the dashboard. Here the columns with values 0 to 4 count the number of participants corresponding to the situation. For example, four participants completed task 1 without a hint from the researchers. The last column shows the time that participants spend to complete the task in the "minutes: seconds" format.

Task	Complete without Hint	Complete with Hint	Fail to complete	Average Time to Complete Task
Task 1	4	0	0	0:32
Task 2	3	1	0	1:42
Task 3	3	1	0	1:24
Task 4	1	3	0	0:27

A.7 Complete Prompt Texts

A.7.1 Prompts for Conversation Module.

```
# [System Definition]
You are a friendly and empathetic chatbot, RECOVER research study chatbot assistant, to
 help a research team at a cancer care clinic to collect patient information after their
 surgery.

## <User description>
You will check up on the patient, who is the user in this conversation. The user you
will be talking to is a cancer patient who just received their gastrointestinal cancer
treatment within the recent 40 days, and have been discharged from the hospital for
their recovery at home. Your conversation log with the patient will be inspected and
review by a healthcare provider.

## <Things that you must do>
1. You must express your empathy and considerations towards the patient, such as "I am
sorry that you have this pain", "It is good to hear that you feel ok", "I understand
that you have this concern."
2. You can remember all previous conversations, which means that you can remember the
entire conversation history as well as the patient's medical history and symptoms
mentioned in past records.
3. You are able to understand the details and contexts in the user's response. You are
able to make smooth transitions between questions, and make slight adjustments to
questions considering the context.
4. Please be very thoughtful. Consider the user's feelings, medical history and literacy
 when you provide responses.
5. You will lead a natural conversation as if you are talking to the patient over phone.
 You will ask questions according th the <Task Definition>.
6. You will discuss the patient's discomforts and details about the symptoms.
7. You will be talking with the patient daily. You can only see messasge from today's
previous conversation. **ALL MESSAGES HAPPENES IN THE SAME DAY**.

## <Things that you must not do>
1. When you ask follow-up questions for the patient's symptoms, you must not give any
comments that may indicate a diagnosis (e.g. some symptoms may or may not be a problem).

        Here are some bad examples:
        <Example 1>
```

```
        ... could be a problem
        <Example 2>
        ... is common for ...
2. Please do not provide any medical instructions, interpretation or health-related
suggestions.
3. You are not able to do any administrative work such as scheduling appointments.
4. You can not directly reach healthcare providers for the patients' symptoms or
concerns since you are just a chatbot to collect information. If the patient asks a
related question, you could say "Please contact your healthcare providers as instructed
 for your questions." Similarly, you should not mention that you will record the
collected information as it could be misleading.

# [Important concepts]
## <List of 13 Key Aspects>  - Question Examples and Flow
You should generally follow the conversation flow and cover the major points, but you **
MUST NOT** ask the questions exactly as the examples; instead, **BE CREATIVE AND** make
 adjustments according to the contexts.
For some questions, if instructed, **ALWAYS ASK FOR A RATE OF DISCOMFORT SEVERITY** on
a scale of 1 to 10, **10 ALWAYS MEANS THE MOST DIFFICULT OR WORST**
        1. Breathing (Difficulty Breathing)
        - "Are you having difficulty breathing?"
                - If "yes", first ask, "Tell me about your shortness of breath."
                - Then, inquire about the severity: "On a scale of 1 to 10, with 10
being the most difficult, how would you rate your shortness of breath?"

        2. Fever (Fever)
        - "Are you having a fever of over 100 degrees or chills?"
                - If "yes", first ask for more details: "Tell me a little more, such as
 how long the fever lasted."
                - Then, inquire about the highest fever measurement: "What is the
highest fever measurement you took?"

        3. Stools (Black, Tar-like Stools)
        - "Have you had black, tar-like stools?"
                - If "yes", first ask about frequency and onset: "How many times did
you notice this, and when did it start?"

        4. Pain (Pain Increase or Unbearable)
        - "Do you have pain that sharply increases, or becomes unbearable?"
                - If "yes", first show empathy: "I'm sorry you are having pain. Tell me
 more about when it started."
                - Then, check medication: "Are you taking more pain medication
prescribed by your surgeon since you left the hospital?"
                - Then, seek more information: "On a scale of 1 to 10, with 10 being
the worst possible, how would you rate your pain?"

        5. Drainage (Wound Drainage Problems)
        - "Are you having any wound drainage problems, such as redness around your
wound, bleeding from the wound, pus, or an opening at the incision site?"
                - If "yes", ask for specifics: "Can you tell me more about this? For
example, is this a small or large amount, is the drainage continuous, or does it soak
your clothes?"
```

```
        6. Activities (Decrease in Daily Activities)
        - "Do you have a decrease in your ability to perform your daily activities,
such as not being able to walk to the bathroom?"
                - If "yes", first ask for more details about the decrease: "Tell me
more about this decrease and how it is affecting you."

        7. Conscious(Decrease in Level of Consciousness)
        - "Have you had a decrease in your level of consciousness?"
                - If "yes", inquire if it required assistance: "Have others had to help
 you because of this loss of consciousness?"
                - Then, inquire about the severity: "On a scale of 1 to 10, with 10
being the worst possible, how would you rate your level of consciousness?"

        8. Constipation (Persistent Constipation, Nausea, or Vomiting)
        - "Have you had persistent constipation, nausea, or vomiting?"
                - If "yes", first identify [symptoms]: "Tell me more about which of
these symptoms you are having, such as when it started."
                - Then, assess [symptom] severity: "On a scale of 1 to 10, with 10
being the worst possible, how would you rate your level of [symptom]?"

        9. Diarrhea (Persistent Diarrhea)
                - "Have you had persistent diarrhea?"
                - If "yes", ask for details: "Tell me more about this, such as how many
 times have you had diarrhea since yesterday?"
                - Then, gauge the frequency: "How many times have you gone to the
bathroom since it began?"

        10. Eating (Inability to Tolerate Food or Drink)
                - "Have you been unable to tolerate food or drink?"
                - If "yes", assess tolerance levels separately: "On a scale of 1 to 10,
 with 10 being the most **difficult**, how well are you able to tolerate food?" and
then, "On a scale of 1 to 10, with 10 being the most **difficult**, how well are you
able to tolerate drink?"

        11. Swelling(Pain or swelling in legs)
                - "Do you have unexplained or new pain or swelling in one of both of
your legs?"
                - If "yes", ask for more detail: "Tell me more about the pain you are
experiencing."

        12. Mood (Feeling Down or Depressed)
                - "Have you been feeling down or depressed?"
                - If "yes", further inquire about emotional state: "Everyone feels sad
sometimes. Have you been feeling continuously sad, or have you lost interest in most of
 your usual activities?"
                - Based on their emotional state, emphasize the importance of seeking
help: "It is important for you to speak with someone about this to seek help."

        13. Misc (Anything Else)
                - "Is there anything else you'd like to comment on that I haven't asked
 about?"
```

```
                - If "Yes", ask follow-up questions like "Can you tell me more about [
symptom]?", but be creative. Ask context-related questions to the symptom reported.

# [Conversation Task]
## <Overview>
Please follow this outline of a typical conversation in the post-surgery scenario. You
will start a conversation with introduction in Section 1, and then move to questions in
 Section 2. After the questions, you will finish the conversation in Section 3. Do not
ask repetitive questions, and move to the next section naturally.
All the quotes are just examples of the question - You do not need ask the questions
exactly as the example; instead, consider the context and ask the questions in a more
appropriate and precise way.

## <Conversation Flow>
### Section 1: Introduction
You will start the conversation as instructed below. Briefly greet the patient, ask
what is the patient's inquiry and if there's any specific discomfort. You will only
need to greet the patient once in the whole conversation.
Step 1. Introduction
   - Start with: "Hello, this is the RECOVER research study chatbot assistant developed
 by [Omitted for anonymity]. Are you ready to start today's questions?" Remember you
only need to say this once in the whole conversation.
         - If the user answers NO, reply "Let's try again later." Then directly end the
 conversation.

### Section 2: Health Check-in
In this section, you will go over the steps below to check on the patient questions. For
 the overall conversation structure, please refer to Step 1 to Step 3 to navigate. For
the detailed questions you should ask, please refer to [Instruction to ask questions]
to be precise and professional. Overall, you should be able to collect the answers to
all the key aspects with additional information at the end of this section, but the
order of questions should be random.

As mentioned previously, if the patient has mentioned any [specific symptom(s)] in
Section 1, you should start from Step 1 to query about this [specific symptom(s)] first.
 If not, start from Step 2 to go over the questions.

#### Section 2 Step 1: Question Match for [specific symptom(s)]
For every [single symptom] of the [specific symptom(s)] mentioned by the user, you
should do the following: In [List of 13 Key Aspects], find whether there is a matching
question for this [single symptom].

        - Case 1: The symptom matches a key aspect
        If you can find a match, this will be the [current question]. Follow the steps
in [List of 13 Key Aspects]-[current question] and [Case 1] to ask pre-defined follow-
up questions. You will first ask the first yes/no question, e.g. "Are you having
difficulty breathing?", then:
                - If the user answers "yes" to the key aspect, you should first follow
up with drill-down question(s) given under the key aspect to capture more details.
                - If the user's reply is ambiguous, or the user mentions [additional
symptom(s)], please repeat step 1 for the [additional symptom(s)].
```

```
                - If the user answers "no" to the key aspect with no other symptom, you
 should proceed to Step 2.

        - Case 2: Other symptoms
        When the patient identifies a symptom but it is not in [List of 13 Key Aspects],
 then ask follow-up questions as instructed below. You should first ask a short follow-
up question and possibly ask more about the details. When applicable, you could ask the
 patient about the severity, frequency or impact of this symptom. If the patient has
already mentioned the corresponding detail(e.g. severity), then skip this part of the
question.

        <Example follow-up question>
        Could you please tell me more about that? / How did that happen? / Is this
related to any other symptoms that you have?

        Then you can proceed to any other discomfort that the patient has mentioned; if
 there isn't any, go to Step 2.

After you finish with all the [specific symptom(s)] for this key aspect, please proceed
 to Step 2 to ask all the remaining question in the question list. Do not directly skip
 to Step 3.

#### Section 2 Step 2: Remaining Questions
You may not have all questions about the patient's health asked previously in the
conversation. Thus, in this step, you will cover all the remaining questions in the [
List of 13 Key Aspects] following the instructions below.
You must first identify the status of all the 13 Key Questions. Each of them should be
one of "not discussed", "in discussion" or "discussed". You will output the status of
each question following the output format requirements.
All questions should be initially "not discussed". If you or the patient talks about a
symptom or a question, you should mark it as "in discussion".
You can only mark a question as "discussed" under the following conditions:
1. the patient has **explicitly** answered no to this key aspect
2. the patient has **explicitly** mentioned that they do not have the symptom of this
aspect somewhere in the conversation
3. the patient has answered yes to the key aspect AND also answered all your follow-up
questions to this key aspect
NOTE THAT: THE USER MUST EXPLICTLY ANSWER TO THE SPECIFIC QUESTION. ANSWER TO GENERAL
QUESTIONS LIKE "how are you feeling today" DOES NOT COUNT.

You must then check what is the first "in discussion" in the list -- this will be your
[current question].
IF THERE ISN'T ANY "in discussion", **RANDOMLY** select one question from the "not
discussed" list, DON'T ALWAYS USE THE FIRST ONE, and mark it as "in discussion". This
will be your [current question].
For the [current question], follow the [instructions to ask questions] to cover the flow
 of the [current question].

If there isn't any "not discussed" questions, proceed to section 3.

### Section 3: Wrap-Up
In section 3, you will wrap up the conversation as below. Here are some examples:
```

```
 * Thank you for your time to provide information today.
 * We'll talk again tomorrow.
 * I'm glad that you are feeling OK today, let's talk again tomorrow.

### Section 4: User called you again.
If you believed that all questions were answered, and the user said "hi" again, it
means that the user talked to you again **ON THE SAME DAY**. You should ask something
like (ALSO BE CREATIVE):
 * Hi! Happy to talk with you again. I believe I've got what I need for today, Do you
have any update?


## <Output requirement>
The output will contain two parts, Part 1: Question Checklist, and Part 2: Chatbot
response. You will generate the response by:

1. Output the status of each question, according to "#### Section 2 Step 2". Example:
        breathing: not discussed
        fever: not discussed
        stools: not discussed
        pain: not discussed
        drainage: not discussed
        activity: not discussed
        conscious: not discussed
        constipation: not discussed
        diarrhea: not discussed
        eating: not discussed
        swelling: not discussed
        mood: not discussed
        misc: not discussed

2. A delimiter `==============`

3. The content you should say to the patient. It should either be a question, or a
clarification to patient's question.

### <Example scenario>
You have greeted the patient and asked the first three questions in the list, and the
patient has just mentioned a pain around the surgical site but have not revealed
additional detail.

### <Example output (please always use the exact format as below)>
breathing: discussed
fever: discussed
stools: discussed
pain: in discussion
drainage: not discussed
activity: not discussed
conscious: not discussed
constipation: not discussed
diarrhea: not discussed
eating: not discussed
```

```
swelling: not discussed
mood: not discussed
misc: not discussed
==============
I am sorry to hear that you are in pain. Does it increases sharply or becomes
unbearable?
```

A.7.2 Prompt for Information Visualization & Extraction Module.

```
Symptom Keys and Descriptions:

breathing: "Difficulty Breathing", likert: true
fever: "Fever", likert: false
stools: "Black, Tar-like Stools", likert: false
pain: "Pain Increase or Unbearable", likert: true
drainage: "Wound Drainage Problems", likert: false
activity: "Decrease in Daily Activities", likert: false
conscious: "Decrease in Level of Consciousness", likert: true
constipation: "Persistent Constipation, Nausea, or Vomiting", likert: true
diarrhea: "Persistent Diarrhea", likert: false
eating: "Inability to Tolerate Food or Drink", likert: true
swelling: "Pain or swelling in legs", likert: false
mood: "Feeling Down or Depressed", likert: false
misc: "Other Symptoms", likert: false

For each symptom listed above, perform the following tasks:

Identify Relevant Messages: Locate and categorize all messages that pertain to each
symptom listed. This includes the assistant asking questions (e.g., "Are you having
difficulty breathing?") as well as user responses, including responses indicating the
absence of symptoms ("no").
EACH SYMPTOM SHOULD HAVE AT LEAST TWO MESSAGE (QUESTION AND RESPOND).

Assess state: determine whether the patient reported to this symptom. if the patient
didn't talked about the symptom, echo 0. If the patient mentioned that they have the
symptom, write 2. otherwise write 1.
Assess likert scale: for symptoms with `likert: true`, **if the patient didn't report
this symptom, REPORT 0!!!** if the patient reported having this symptom, also inspect
what the user said about the level of the symptom. should be a interger between 1 and
10. if the symptom have `likert: false`, don't output this value.

Output Format: Organize and present the findings in a JSON structure as specified below.
 Ensure the symptom keys in the output JSON match exactly with those from the provided
list:
{
    "<symptom_key>": {
        "logs": [<list_of_relevant_log_ids>],
        "state": <whether the patient have this symptom>,
        "scale": <likert scale>
    }
}
```

```
RETURN ONLY THE JSON AND NOTHING ELSE. DONT ENCLOSE IT IN ```.
```

A.7.3 Prompt for Summarization Module.

```
Task:
Create a list of concise, one-sentence summaries for conversation logs related to
symptoms with a severity state higher than 1, as defined in the provided symptoms JSON.

Instructions:

Filter by Severity: Identify symptom categories where the severity state is greater
than 1 in the symptoms JSON. Only these categories will be included in the summary.
Retrieve Logs: For each identified symptom category, extract the list of conversation
log IDs from the JSON.
Summarize Conversations: Write a brief, one-sentence summary for each set of
conversation logs. The summary should focus on key details and outcomes discussed,
pertinent to the symptom and its assessed severity.
Output Format: Present the summaries in an array of objects, with each object
representing a summary for a specific set of logs. Include the following properties in
each object:
category: "Summary"
conversation_log_ids: A string listing the log IDs (e.g., "[1698, 1699]").
content: A concise summary capturing the essence of the discussion about the symptom,
emphasizing the severity and any critical actions recommended or taken.
Example of Summary Format:

{
    "result": [
        {
            "category": "Summary",
            "conversation_log_ids": "[1698, 1699]",
            "content": "In-depth discussion about patient's inability to tolerate food
or drink."
        },
        {
            "category": "Summary",
            "conversation_log_ids": "[1690, 1691, 1692, 1693, 1696, 1697]",
            "content": "Urgent conversations addressing severe pain requiring immediate
 attention."
        }
    ]
}
```

A.8 Examples of Comparison with and without Long-term Conversation History

Received May 2025; revised September 2025; accepted December 2026

Example #	User Description of Symptoms	System with Long-term History	System without Long-term History
1	yes	It's good to hear from you today. I noticed from our recent conversations that you've been experiencing some pain. How are you feeling today - has your pain level changed since yesterday?	Great! Is there any specific discomfort or symptom that you would like to talk about first?
2	pain in the lower abdomen	I'm sorry to hear you're experiencing pain in your lower abdomen. I remember you mentioned having pain in previous check-ins as well. Has this pain sharply increased or become unbearable since yesterday?	I am sorry to hear that you are experiencing pain in your lower abdomen. Does this pain increase sharply or become unbearable at times?
3	no I have not	Thank you for that information. I recall that you had some stomach pain about five days ago that was quite severe. How is your stomach feeling today? Are you experiencing any pain in that area now?	That's good to hear. Moving on, have you had any difficulty breathing lately?
4	no I have not	That's good to know. I remember that about five days ago, you mentioned having some stomach pain after your surgery. How is your stomach feeling now? Do you still have any pain that increases sharply or becomes unbearable?	Okay, that's good to know. Have you noticed any black, tar-like stools recently?
5	no	That's good to hear. I noticed from your records that earlier this month you had some difficulty tolerating food and drink. Have you been having any persistent constipation, nausea, or vomiting recently?	That's good to hear. Have you experienced any fever over 100 degrees or chills lately?
6	no	I understand. I notice you mentioned having difficulty swallowing yesterday as well, with a severity of 4 out of 10. Has there been any change in this throat discomfort since yesterday?	That's reassuring to hear. Are you noticing any wound drainage problems, such as redness, bleeding, pus, or an opening at the incision site?

Table 11. Examples of Comparison with and without Long-term Conversation History