Engagements with Generative AI and Personal Health Informatics: Opportunities for Planning, Tracking, Reflecting, and Acting around Personal Health Data

SHAAN CHOPRA*, University of Washington, USA
KATHERINE JUAREZ, University of Washington, USA
JAMES FOGARTY, University of Washington, USA
SEAN A. MUNSON, University of Washington, USA

Personal informatics processes require navigating distinct challenges across stages of tracking, but the range of data, goals, expertise, and context that individuals bring to self-tracking often presents barriers that undermine those processes. We investigate the potential of Generative AI (GAI) to support people across stages of pursuing self-tracking for health. We conducted interview and observation sessions with 19 participants from the United States who self-track for health, examining how they interact with GAI around their personal health data. Participants formulated and refined queries, reflected on recommendations, and abandoned queries that did not meet their needs and health goals. They further identified opportunities for GAI support across stages of self-tracking, including in deciding what data to track and how, in defining and modifying tracking plans, and in interpreting data-driven insights. We discuss GAI opportunities in accounting for a range of health goals, in providing support for self-tracking processes across planning, reflection, and action, and in consideration of limitations of embedding GAI in health self-tracking tools.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **Empirical studies in ubiquitous and mobile computing**.

Additional Key Words and Phrases: health tracking; personal informatics; self-tracking; reflection; personal health data; sense-making; personalization; actionability; generative AI; qualitative research

ACM Reference Format:
Shaan Chopra, Katherine Juarez, James Fogarty, and Sean A. Munson. 2025. Engagements with Generative AI and Personal Health Informatics: Opportunities for Planning, Tracking, Reflecting, and Acting around Personal Health Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 75 (September 2025), 33 pages. https://doi.org/10.1145/3749503

1 INTRODUCTION

People commonly engage in self-tracking around aspects of health (e.g., diet, exercise, sleep, biometrics, symptoms) [78], enabled in part by mobile and wearable support [47]. Personal health informatics is thus an important approach to understanding and pursuing personal health [26], often based in tools that *"help people collect personally relevant information for the purpose of self-reflection and gaining self-knowledge"* [50]. Distilling common challenges across varying settings, researchers have developed models of personal informatics that

*Corresponding author.

Authors' Contact Information: Shaan Chopra, schopra7@cs.washington.edu, University of Washington, Seattle, Washington, USA; Katherine Juarez, kajuarez@cs.washington.edu, University of Washington, Seattle, Washington, USA; James Fogarty, jfogarty@cs.washington.edu, University of Washington, Seattle, Washington, USA; Sean A. Munson, smunson@uw.edu, University of Washington, Seattle, Washington, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
© 2025 Copyright held by the owner/author(s).
ACM 2474-9567/2025/9-ART75
https://doi.org/10.1145/3749503

characterize challenges and opportunities in stages of deciding what to track, selecting tools, tracking, integrating data, reflecting, and acting [29, 50]. However, heterogeneity in the range of data, goals, expertise, and context that people bring to self-tracking often leaves people needing individualized support beyond what most tools provide, ultimately preventing people from meeting their goals [12, 26–28, 61].

With the rapid rise of Generative AI (GAI), researchers have noted its potential to support health, including improving patient-provider collaboration [3, 15], addressing healthcare needs requiring clinical validation [8, 15, 67, 81, 87], and supporting self-tracking and sense-making of personal health data [42, 44, 55, 59]. Within the context of our focus on self-tracking, commercial tools have begun adding GAI-based support for analyzing data and generating insights around sleep, recovery, and activity [1, 48, 85]. However, these remain limited in the data they integrate and insights they can provide [10, 42]. Recent work has identified additional opportunities for GAI in self-tracking contexts, including to support sense-making and reflection on personal health data [55], to motivate behavior change [42], and to provide personalized lifestyle recommendations [62]. Research has also examined benefits and risks of querying general-purpose GAI tools (e.g., ChatGPT) about personal health [89]. Building on such explorations of GAI within individual stages of self-tracking, our work investigates the potential for GAI to support people in navigating complexities introduced by heterogeneous tracking data, goals, expertise, and context throughout the multiple stages of personal health informatics.

We investigate the following research questions:

- **RQ1**: How do people who self-track for health goals interact with GAI for reflecting on, interpreting, and potentially planning action on their personal health data?
- **RQ2**: What are needs and opportunities for GAI in supporting varying individual needs and goals across different stages of personal health informatics?

To examine these questions, we conducted study sessions with 19 people from the United States who track varying aspects of their health, using GAI conversational agents (e.g., ChatGPT, Copilot) as technology probes [40]. We examined (i) how participants interact with GAI to make sense of their health data and (ii) what support they want from GAI in their self-tracking. Each session consisted of a semi-structured interview, an interactive session in which participants explored GAI around their health data and goals, and a case scenario presented after the interactive session to elicit consideration of GAI capabilities participants may not have encountered in their own exploration. We analyzed data drawing upon (i) models of personal informatics [29, 50] in analyzing participant tracking goals, stages of self-tracking, and barriers in each stage, and (ii) the bi-directional human-AI alignment framework [75] in analyzing participant interactions with and perceptions of GAI.

We make three contributions. First, we provide an account of participant interactions with GAI, using personal information and self-tracked data to ask questions related to their health goals. Using GAI conversational agents as technology probes in interactive sessions, we describe participant processes of forming queries and reflecting on responses, as well as challenges they encountered. We further detail how participants (i) refined queries toward more aligned responses or (ii) abandoned querying when unable to obtain support they were seeking. Second, we identify additional support participants desired from GAI, informed by their long-term experiences with personal informatics, by what they tried in their interactive sessions, and by GAI capabilities explored through the case scenario. Participants described GAI's potential to support varied health goals and across multiple stages of self-tracking, such as deciding what and how to track, setting up and modifying tracking plans, integrating data, interpreting recommendations, and acting on data-based insights. Finally, we discuss prior work and our findings, identifying GAI opportunities (i) to account for a range of health goals and questions, (ii) to scaffold self-tracking and support people across planning, reflecting, and acting upon personal health data, and (iii) concerns and implications for embedding GAI support in health self-tracking tools.

2 RELATED WORK

Our research is motivated and informed by past literature on personal informatics. This includes motivations for self-tracking, barriers people encounter throughout different stages of self-tracking, and past efforts to support people in navigating these challenges. Our work is also inspired by examinations of GAI in broader health contexts, specifically concerns about actionability and credibility raised in those studies. Finally, we draw upon recent work using GAI to support specific personal health informatics processes, such as reflection and action based in health data.

2.1 Self-tracking for Health

2.1.1 Models of and Motivations for Self-tracking. Researchers have proposed models characterizing self-tracking processes. The stage-based model of personal informatics describes five stages (i.e., preparation, collection, integration, reflection, action) along with cascading barriers at each stage [50]. The lived informatics model expands this to include deciding what and how to track, emphasizes an ongoing process of collecting, integrating, and reflecting, and highlights lapses, resumption, and changing goals [29]. The goal-directed self-tracking framework emphasizes the need to design tracking around individual goals, scaffolding processes of deciding what, when, and how to track in alignment with those goals [58, 73, 74]. We leverage these models of personal informatics processes to explore potential GAI support throughout such stages of self-tracking.

People also have varying motivations for engaging in self-tracking, such as improving health, finding new life experiences, behavior change, maintaining a record, or curiosity [14, 29, 54]. Past work identified broad categories of goals associated with self-tracking, including (i) tracking goals (e.g., "tracking hours of sleep"), (ii) information goals (e.g., "does lack of sleep make migraines more likely") and (iii) management goals (e.g., "improve symptoms") [73, 74]. Participants in our study similarly report diverse tracking motivations, within which we examine how they interact with GAI around their health goals and data.

2.1.2 Barriers and Challenges in Self-tracking Processes. Researchers have highlighted barriers people encounter in and across stages of self-tracking [4, 50, 70]. Entire regimens can become ineffective when not centered around an individual's goals [23, 58]. In reflection and sense-making, barriers can be created by data itself (e.g., data that is sparse, lacks context, or is not appropriate for a person's question) or by an individual's abilities to make sense of data (e.g., difficulty in integrating, interpreting, or visualizing data to draw insights) [5, 6, 17, 51]. Even expert self-trackers often lack rigor in their practices around tracking, reflection, and analysis [14], further highlighting the need to support sense-making and reflection [7]. During the action stage, a lack of specific guidance is a major barrier [24, 50, 53]. The design of self-tracking tools is often misaligned to individual goals, hindering planning and action [37, 49, 73]. *Fit4Life*, a hypothetical design intended to provoke critical reflection about the ethics of persuasive technologies, exemplifies complexities of aligning technology to individual goals [69]. Although the design purports to support individual health goals, it also depicted challenges of surveillance, coercion, and overreliance on technology.

Researchers have attempted to address such challenges by designing tools that support end-to-end self-tracking processes and guide people throughout stages of tracking. For example, OmniTrack supports the collection and integration stages via construction of customized trackers meeting individual needs and goals [45]. MigraineTracker [74] implements goal-directed tracking [73], guiding people's expression of self-tracking goals and providing integrated support across subsequent stages of tracking. Goal-directed tracking is also informed by challenges characterized by the Tracker Goal Evolution Model [61], including in translating qualitative goals (e.g., reduce migraines) into quantitative goals (e.g., drink 2-4 gallons water daily). Although qualitative context can support reflection and action [13, 33, 39], extracting and using contextual information can be challenging. Our research is informed by these past studies of self-tracking processes and by challenges of developing tools that provide domain knowledge and support across stages of planning, tracking, reflecting, and acting. For

example, OmniTrack's flexbility can be combined with templates co-designed with clinicians or provided by experts [45] to give users clear guidance on what and how to track. The design of MigraineTracker [74] similarly incorporated knowledge about common migraine-related goals (e.g., common symptom contributors, reasonable tracking frequencies and granularity). A key finding in such work has been that flexible tools must be paired with significant domain and tracking knowledge to result in success. This knowledge varies across goals and health contexts, making it difficult for designers and tools to provide. This motivates us to examine whether GAI might provide support and access to this knowledge without requiring expert creation and maintenance of different templates or tools for each health goal, context, or question.

2.2 Generative AI and Personal Health

Generative AI is increasingly leveraged in personal health [3, 8, 15, 67, 81, 87], with an emphasis on its potential to enhance personalization, including making recommendations (e.g., exercise, diet, sleep) based on individual background (e.g., genetics, lifestyle, family history) [62]. COVID-19 saw increased adoption of agents for answering questions, recommending care, checking symptoms, and performing administrative tasks [67], leading researchers to examine patient perceptions of GAI. Studies have found GAI responses to be empathetic, often attributed to long responses as an indicator of more time spent contemplating [81] and of higher information quality [3]. We use the bi-directional human-AI alignment framework [75] to analyze interactions with GAI around self-tracked health data, examining participant query formation, evolution, and refinement as well as perception of responses. This framework guides examination of both (i) aligning AI to humans (i.e., ensuring AI generates desired individualized responses), and (ii) aligning humans to AI (i.e., how individuals perceive AI-generated responses and potentially adjust to current AI capabilities).

However, GAI also presents concerns. It can give assertive responses despite uncertainty [52] and hallucinate [9] plausible yet incorrect content [88], potentially causing harm and jeopardizing trust in critical contexts like health [84]. Familiarity with prompting techniques is essential, as certain methods (e.g., chain-of-thought [83]) are less susceptible to hallucinations than others (e.g., ReAct [86]). Researchers have also highlighted concerns about widespread adoption of AI-based technologies in health (e.g., trust, clinical safety and reliability, regulatory oversight, privacy) [56, 79, 87]. This includes concerns about liability for AI-generated health information [20] and the quality of online health information [34, 77], warning that decisions based on low-quality information can lead to unintended harm [77]. Consumers of online health information struggle with assessing trustworthiness and credibility [34], a challenge that persists with GAI, as tools lack surface-level credibility cues [21]. Studies have also identified cultural, racial, and gender disparities in provider-patient communication, warranting concerns that GAI might perpetuate or amplify these [32, 41, 71, 81]. Researchers have highlighted opportunities for designing inclusive, personalized tools that capture diverse patient backgrounds, personalities, and motivations [81] and integrate personal characteristics to offer tailored suggestions in health and wellbeing [18].

Researchers also remain interested in making GAI responses more actionable [11], suggesting not just *what* can be done but also *how*. However, actionability can have multiple dimensions across individuals and contexts (e.g., understanding how to, being willing to, or being able to act on recommendations). Our work examines opportunities for supporting participants in reflecting and acting on insights from self-tracked health data, while also discussing appropriate integration of GAI into health tracking tools.

2.3 Generative AI Applications in Personal Informatics

Recent research examines GAI potential in specific stages of personal informatics, such as for collecting and reflecting on health-related data. GAI chatbots have been found effective for collecting self-reported sleep, exercise, and productivity data, with prompt design playing a significant role (i.e., information format and personality modifiers) [82]. Researchers have also developed GAI applications supporting collection and reflection. For

example, MindfulDiary is a GAI-enabled chatbot that encourages documenting daily experiences and improved journaling adherence through natural conversations [44]. MindScape similarly enhanced data collection and self-reflection through GAI-powered journaling that integrates passively-collected behavioral health data (e.g., conversational engagement, sleep, location) for personalized and context-aware prompting [59]. Research has also identified potential for GAI-generated narratives and summaries to supplement quantitative data and support self-reflection [76] and clinician reflection on patient data [25].

Research has also examined perceptions of GAI responses and insights. Some commercial personal informatics tools (e.g., Whoop [85], Onvy [1]) now include GAI-based coaches which use conversational interfaces to support question-asking around data in sleep, recovery, and activity. However, these do not extract or incorporate contextual information, limiting their ability to support reflection and action on derived insights [42]. To examine better analysis and interpretation of such personal health data, Merrill et al. developed a GAI-based Personal Health Insights Agent (PHIA) [55]. When compared against numerical reasoning and code generation, PHIA showed higher accuracy for objective health queries (e.g., How many times did I exercise today?) and received higher ratings on overall reasoning, relevance, and clarity (e.g., What is the total time I spent swimming for sessions lasting 40 minutes or less?). Similarly, GPTCoach addressed personalization challenges by using multimodal reasoning to integrate personal data with contextual factors [42]. This enabled reasoning through both semantic data captured in conversations (e.g., goals, preferences) and personal quantitative data (e.g., step count from wearable). Both PHIA and GPTCoach demonstrate potential of GAI-based health agents for supporting sense-making and reflection and for providing personalized and actionable insights. As a complement to such fine-tuned tools, our research investigates self-tracker interactions with general-purpose GAI around their personal health data and their perspectives on GAI's potential for supporting self-tracking journeys around diverse health goals and needs.

3 METHODS

Our IRB-approved study took place from February 2024 through January 2025 in the United States. We conducted individual study sessions with 19 participants and employed a hybrid process of inductive and deductive thematic analysis [31] to understand: (RQ1) how people interact with GAI around their health self-tracking data, and (RQ2) what more support they would like for their personal health goals and related self-tracking.

3.1 Participant Recruitment

Participants were recruited using word-of-mouth and purposive sampling [80], from social media platforms (i.e., Twitter, LinkedIn), the Quantified Self forum[1], and author personal and professional networks. Participants completed a screening survey (Appendix A) to ensure they: (i) are 18 years of age, and (ii) track data to understand and answer questions about their health. To prevent fraudulent participants [64], the survey required detailing types of data they track, how they track (e.g., mobile apps, wearables, paper), and examples of questions they want to answer with data. The survey also asked about familiarity with GAI tools (e.g., ChatGPT, Copilot) and work expertise in health and/or technology. These were not inclusion or exclusion criteria, but we used them to seek a range of experiences across participants. Table 1 details participant demographics, including health data they tracked.

3.2 Study Session Overview

We conducted individual study sessions with 19 participants. Each session lasted 60 to 90 minutes, and was compensated with a $20 or $40 gift card depending on the length of the session. All sessions were conducted over Zoom, except P2 used the researcher's laptop during an in-person session. The researcher shared their screen

[1]https://forum.quantifiedself.com/

Table 1. Demographic details of participants, including the types of health data they tracked and devices/modes of tracking. All except P19 reported familiarity with using GAI-based conversational agents (e.g., ChatGPT, Copilot, Gemini, Claude). Four reported prior use of GAI related to health and self-tracking: P7 used ChatGPT for meal plan recommendations meeting his dietary goals, P12 used a GAI-enabled fitness tracker called Whoop [85], P17 used ChatGPT to answer questions related to her menstrual health and tracking, and P18 used ChatGPT to get recommendations around diet and exercise for fat loss.

P#	Gender	Age Range (in years)	Race/Ethnicity	Highest Education	Work in Health or Tech?	Type of Health Data Tracked	Device/Mode of Tracking
P1	Cis-Woman	25-34	Asian / Pacific Islander	PhD	None	Cardiovascular Health, Diet, Exercise, Sleep, Menstruation/Ovulation	Wearables, Mobile or Web Apps
P2	Cis-Woman	25-34	Asian / Pacific Islander	Master's	None	Cardiovascular Health, Exercise, Sleep	Wearables
P3	Cis-Woman	25-34	Multi-Racial / Bi-Racial / Japanese American / Asian American	Bachelor's	Technology	Body Temperature, Cardiovascular Health, Exercise, Mental Health, Menstruation/Ovulation	Wearables, Mobile or Web Apps, Paper-based Mediums
P4	Cis-Woman	25-34	Asian / Pacific Islander	Master's	None	Blood Pressure, Exercise, Sleep Menstruation/Ovulation	Wearables, Mobile or Web Apps
P5	Cis-Woman	18-24	Asian / Pacific Islander	Bachelor's	Technology	Cardiovascular Health, Diet, Exercise	Wearables, Mobile or Web Apps, Physical Devices
P6	Cis-Woman	18-24	White / Caucasian	Bachelor's	None	Exercise, Medications/Supplements, Menstruation/Ovulation	Wearables, Mobile or Web Apps
P7	Cis-Man	25-34	Asian / Pacific Islander	Master's	None	Diet, Exercise	Wearables, Mobile or Web Apps, Paper-based Mediums, Mental Notekeeping
P8	Cis-Woman	25-34	Asian / Pacific Islander	Master's	None	Cardiovascular Health, Exercise, Sleep, Hydration, Nausea	Wearables, Mobile or Web Apps
P9	Cis-Woman	25-34	Asian / Pacific Islander	Bachelor's	Both	Exercise, Diet, Medications/Supplements, Menstruation/Ovulation	Wearables, Mobile or Web Apps, Mental Notekeeping
P10	Cis-Man	45-54	Prefer not to say	Master's	Technology	Air Pressure, Blood Biomarkers, Blood Pressure, Blood Sugar, Weight, Medications/Supplements, Body Temperature, Diet, Exercise	Wearables, Mobile or Web Apps, Physical devices, Comprehensive Blood Tests
P11	Prefer not to say	25-34	Asian / Pacific Islander	Bachelor's	Technology	Weight, Diet, Exercise, Sleep, Migraines, Menstruation/Ovulation	Wearables, Mobile or Web Apps, Paper-based Mediums, Physical Devices, Mental Notekeeping
P12	Cis-Man	25-34	Asian / Pacific Islander	Master's	Technology	Cardiovascular Health, Diet, Exercise, Sleep, Hydration, Allergens, Medications/Supplements	Wearables
P13	Cis-Woman	25-34	Asian / Pacific Islander	Bachelor's	None	Diet, Exercise, Sleep, Habits, Mental Health, Medications/Supplements	Mobile or Web Apps, Other-Spreadsheet
P14	Cis-Man	18-24	White / Caucasian	College Credit, No Degree	Technology	Body Composition, Weight, Teeth Health, Bowel Movements, Stool, Cardiovascular Health, Exercise, Sleep, Skincare, Medications/Supplements	Wearables, Mobile or Web Apps, Other-Progress Photos
P15	Cis-Woman	25-34	Hispanic	High School Graduate, Diploma or Equivalent	Technology	Menstruation/Ovulation	Mobile or Web Apps
P16	Cis-Woman	25-34	Hispanic	Master's	None	Diet, Exercise	Wearables, Mobile or Web Apps
P17	Cis-Woman	25-34	White / Caucasian	Professional	Health	Exercise, Menstruation	Wearables, Mobile or Web Apps
P18	Cis-Woman	25-34	Hispanic	Vocational Training	Technology	Diet, Exercise, Weight Sleep, Menstruation	Wearables, Physical Devices, Mobile or Web Apps
P19	Cis-Woman	35-44	White / Caucasian	Bachelor's	None	Exercise, Sleep, Menstruation	Mobile or Web Apps, Paper-based Mediums

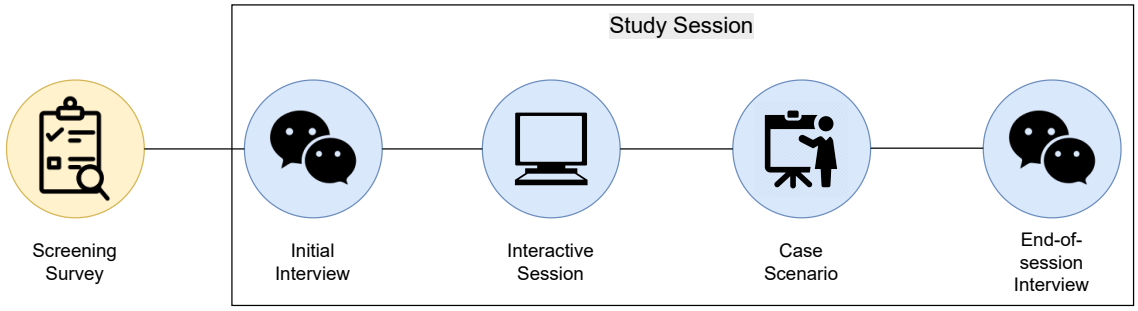


Fig. 1. Study procedure. Potential participants filled out a screening survey to determine their eligibility. Those eligible were invited to participate in a single study session comprising (i) an initial interview to learn about participant health tracking practices, data, and questions; (ii) an interactive session where participants engaged with GAI around their health data; (iii) a case scenario to demonstrate a broader range of GAI capabilities and elicit participant considerations; (iv) an end-of-session interview to learn about participant overall experiences in the study. For the interactive sessions, participants P1-P4 used the free version of ChatGPT (GPT 3.5x), while P5-P19 used a version of Copilot with commercial data protection (built upon GPT-4). Participants P1-P4 were not presented the case scenario.

and participants used Zoom's remote control option to directly interact with GAI. Informed consent, including permission to record, was obtained before each session. All recordings were transcribed for analysis.

As in Figure 1, study sessions consisted of four phases: an initial interview, an interactive session, a case scenario[2], and an end-of-session interview. We asked participants to bring their self-tracked health data. We did not ask them give us their data but instead to describe and/or use it to ask questions. Consistent with the iterative nature of semi-structured qualitative interviewing [35], we evolved and expanded the protocol through the course of the study. This section further details our study protocol, including additions/changes and their rationale. The complete study protocol is also provided in Appendix B.

3.2.1 Initial Interview. The initial semi-structured interview focused on participant data, health goals, and tracking motivations. We asked participants to describe health data they brought and their self-tracking practices (e.g., "What is the data about?", "How and why did you collect it?", "Have you used it or drawn any conclusions from it?"), including questions they wanted to answer using self-tracking (e.g., "What questions do you want to answer based on your tracking and why?", "How have you tried to answer this question, if at all?", "If you have been unable to answer this question, what has been the challenge?"). The initial interview therefore elicited both participant goals and prior long-term experiences pursuing those goals with self-tracking.

3.2.2 Interactive Session. This portion of the session used a GAI conversational agent (either ChatGPT or Copilot) as a technology probe [40] with goals to (i) collect information about participants and their experiences using GAI around their self-tracking and (ii) inspire participants to consider GAI opportunities for diverse self-tracking needs and desires. Participants used GAI to ask specific questions based on their self-tracked data. To facilitate the interactive session, we asked about reactions to GAI responses (e.g., "How do you feel about this response?", "In what ways is it useful versus not?", "Do you feel there is anything missing in the response?"). When needed, we prompted participants to consider refining or reframing questions, asking follow-ups or new questions, or providing more detail for a more personalized or appropriate response.

[2]As detailed in Section 3.2.3, participants P1-P4 were not presented the case scenario.

We initially used the free version of ChatGPT (GPT 3.5.x) as the technology probe. To ensure anonymity, all interactions with ChatGPT were conducted using an account created by the researcher. Although participants brought their self-tracking data to the session, we asked them to refrain from uploading it or sharing private or identifying information. Instead, we asked them to use their data to frame specific questions. We encouraged them to talk aloud throughout the session.

We observed participants had varying comfort levels sharing data in ChatGPT queries, describing frustrations with forming queries that would support personalized responses without divulging too much personal information. Although this was informative regarding participant privacy perceptions around GAI, it created a frustrating experience for participants and limited our ability to fully examine our research questions. After observing this in the first four sessions (P1-P4), we switched to Microsoft Copilot with commercial data protection, built upon OpenAI's GPT-4. Together with being the most recent version of GPT, this ensured participant data, queries, and responses were not stored or used by Microsoft or OpenAI, thus making it safer to share sensitive information. We updated study scripts to provide context around commercial data protection, removed guidance to refrain from uploading self-tracked data or sharing private or identifying information, and encouraged participants to be as specific as desired.

3.2.3 Case Scenario. The interactive session emphasized varying participant data and varying participant questions around that data. Although this supported learning about ways in which participants approached GAI support, it sometimes masked opportunities to gather participant perspectives (e.g., when participant data did not support certain types of questions or when participants otherwise did not explore GAI capabilities), thus limiting our ability to examine RQ2. After observing this in early sessions (P1-P4), we decided to introduce a case scenario showcasing a range of GAI capabilities for supporting self-tracking and sense-making. We created the case scenario using real health tracking data contributed by a participant in a prior study, demonstrating potential GAI support for self-tracking related to chronic shoulder pain. Detailed in Appendix C, the case scenario illustrates relevant GAI capabilities, including receiving uploaded images or datasheets, summarizing uploaded data, recommending new self-tracking variables, and creating a tracking template according to individual health goals. After using slides to present the case scenario, we asked participants for their thoughts on (i) prompting GAI by directly uploading health data and then formulating related questions, and (ii) the nature of GAI responses, specifically in terms of personalization, actionability, and trustworthiness. We further asked about perceived usefulness of GAI capabilities demonstrated in the case scenario and what additional capabilities would be useful.

To avoid biasing or leading participant interactions, we presented the case scenario only after participants had already interacted with GAI around their own health. This prevented it from influencing how participants used GAI for their own questions and data, while increasing their awareness of capabilities and potential uses of GAI to reflect on during the end-of-session interview. Because P1-P4 were not presented this case scenario, their reflections in the end-of-session interview were instead anchored only in their prior self-tracking experiences and experiences with GAI in the interactive session.

3.2.4 End-of-Session Interview. In this semi-structured interview, we asked participants to reflect on their overall experience as well as the usefulness, personalization, actionability, and trustworthiness of GAI responses (e.g., including factors that influenced these perceptions).

We then explored participant perceptions of GAI-supported reflection and sense-making as well as any actions participants might take based on understanding developed in the interactive session. Because initial participants described wanting GAI support with adjusting tracking (e.g., deciding what to track, how to track new variables, how to modify existing tracking), we expanded interview questions to prompt on how participants anticipated GAI responses might influence their tracking (e.g., "Do you anticipate changing anything in your health tracking or everyday life based on the responses you received?", "While this study focused on supporting reflection and

sense-making of already collected health data, would you also appreciate support of such AI tools in other parts of the process such as helping decide what data to track or change the type of health data to track?”).

We concluded by reviewing some limitations of GAI (e.g., it can give factually incorrect or biased answers, it can be coaxed into desired or potentially harmful answers [9]) and prompted for reactions. We used this debrief to caution participants, especially those unfamiliar with these limitations, against following GAI recommendations without seeking expert advice or cross referencing the recommendations with reliable sources.

3.3 Data Analysis

We analyzed collected data (session transcripts, session notes, screenshots of ChatGPT / Copilot chat) using a hybrid process of inductive and deductive thematic analysis [31]. This allowed us to capture codes anticipated by prior work in personal health informatics and human-AI interaction while preserving our ability to detect new data-driven codes. Where needed, we referred back to video recordings (e.g., to capture specific reactions).

We first identified two existing frameworks to guide deductive coding for (i) stages of tracking and (ii) aspects of human-AI alignment. We used the Lived Informatics model [29] to code for stages of tracking (e.g., deciding and planning to track; tracking and action based in a process of collection, integration, and reflection) and participant motivations for tracking (e.g., behavior change, maintaining a continuous record, curiosity). Because our research goal is to examine GAI potential across different stages of tracking, this model was integral for identifying where and how in the self-tracking process participants used or wanted to use GAI. We used the Bi-Directional Human-AI Alignment model [75] to code for human-AI alignment in the interactive session. More specifically, we used its *Align AI to Humans* direction to analyze how participants form, evolve, and refine queries, and its *Align Human to AI* direction to unpack how participants perceive responses (e.g., types of GAI support they anticipated or wanted, what worked, what challenges occurred), examining how documented challenges of GAI manifest in their application to self-tracking and personal health informatics. We developed a preliminary codebook with analysis guided but not confined by codes from the two frameworks. We also assigned inductive codes to data segments describing new patterns aligned with our research questions but not encompassed by deductive codes (e.g., “Not knowing how to integrate data into queries”, “querying GAI to validate current understanding or practices”). The first two authors coded the 19 sessions, regularly meeting to discuss analysis and to evolve the codebook to align with study goals and capture additional nuance (e.g., dropping irrelevant codes, adding new codes, splitting or combining codes, extending code descriptions). The codebook is provided in supplementary materials.

Next, we reviewed this codebook and clustered codes based on patterns in the data to construct higher-level themes around what participants did in their sessions (Section 4.1) and GAI support they anticipated and wanted in self-tracking for health (Section 4.2). We performed this iteratively, identifying relationships across categories of codes to formulate and revise themes, documenting them, and discussing with other authors as themes developed. All authors remained in constant touch with the data through collection and analysis, to continue developing shared understanding and to resolve disagreements.

Lastly, we analyzed GAI chat transcripts, coding for question topics, questions asked per topic, evolution of questions around a topic, format change, wording refinement, and types and extent of provided personal data.

3.4 Study Focus and Method Limitations

Because our study focused on examining opportunities and challenges for GAI support throughout stages of self-tracking, we recruited participants who already engaged in self-tracking and could bring their existing data, questions, and long-term self-tracking experiences to the study. Our study design focused on participant interactions with GAI during a single session, including their experiences before that session and attitudes about potential future use. We also captured participant intent to act on GAI insights around their self-tracked data.

However, we were unable to observe any future use or actual action. Our research questions and recruitment criteria likely affected participant demographics. All participants reported living in the United States and having at least high school education or similar. Participants skew under 35 years of age (17 of 19 participants), consistent with research finding greater interest in digital health tools among younger people [36, 57]. Consistent with our recruitment criteria, participants were technology savvy, health enthusiasts, and early adopters of new technologies [14]. Eight worked in technology, one in healthcare, one in technology and healthcare, and all but P19 noted being familiar with using GAI conversational agents (e.g., ChatGPT, Copilot, Gemini, Claude). Participant age, education, and professional expertise in technology and health likely influenced findings (e.g., perceptions about reliability and trustworthiness, ability to analyze data and formulate queries). We also explicitly recruited some participants from self-tracking communities (e.g., Quantified Self forum), and their tracking expertise likely affected their experience and our results. People new to self-tracking may have additional or different needs beyond those reflected in our results. As will be further discussed in Section 5.4, it remains important for future work to examine perspectives of a broad range of populations who were not represented in our study (e.g., lacking expertise in technology or health, less educated, of different identities and cultural backgrounds). Additional research should explore follow-up sessions and longitudinal deployment of GAI tools. Our work aims to inform such longitudinal research through uncovering a range of opportunities for GAI to provide support across various stages of self-tracking.

3.5 Ethical Considerations

This study was reviewed and approved as exempt by our institutional review board. Informed consent and permission to record was obtained before each participant session. All transcripts were de-identified before analysis and only de-identified quotes were reported. The free version of ChatGPT (GPT 3.5.x) was used for the first four sessions (i.e., P1-P4), with participants using an account created by the researcher to ensure anonymity. To make it safer to share sensitive information, the remaining sessions (i.e., P5-P19) switched to using Microsoft Copilot with commercial data protection (GPT 4). This ensured data, queries, and responses were not stored or used by Microsoft or OpenAI, potentially reducing participant privacy concerns (discussed in Section 5.4).

4 FINDINGS

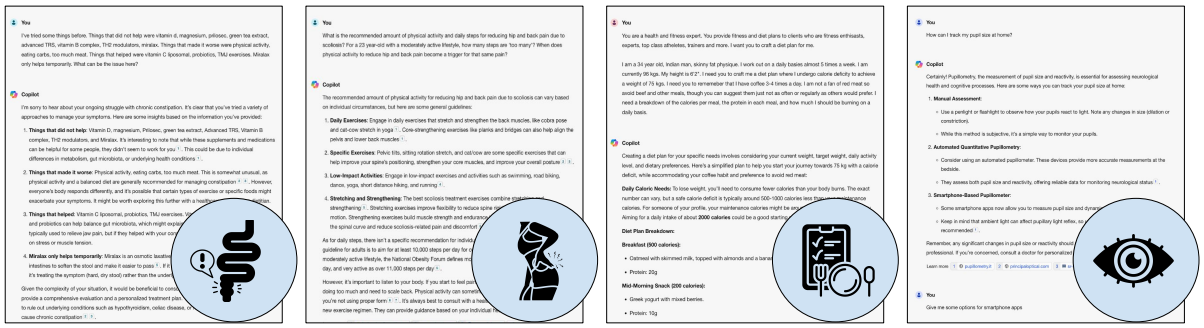
We first describe participant interactions with GAI during interactive sessions as they asked questions related to their health goals and made sense of self-tracked data (Section 4.1). Next, we describe types of support participants wanted GAI to provide for reflection and other stages of self-tracking (e.g., deciding what and how to track, setting up and modifying tracking, interpreting and acting on health data insights) (Section 4.2). Participants identified opportunities through reflecting upon their prior long-term self-tracking experiences, their interactive session, and GAI capabilities explored in the case scenario. Throughout our findings, we apply both quotation marks and italics to *"participant spoken quotes"* and only italics to *participant query text*.

4.1 Participant Interactions and Experiences with GAI

In interactive sessions, participants used GAI for a range of goals, including analyzing heterogeneous health data (P13), identifying correlations between health variables and lifestyle behaviors (P6, P12, P15, P16, P17), seeking more creative and individualized recommendations (e.g., P7 asked for high-protein meal options personalized to his demographics of *Indian male in early 30s with skinny fat physique* and preference to exclude tofu, P18 asked for workout and recipe recommendations to aid in losing fat around her midsection), validating current practices (P8, P10, P16), and identifying new health goals and how to start tracking for them (P10, P11, P14). Participants also prompted GAI in different ways, asking a variety of questions, sharing different amounts of self-tracked data, and refining, evolving, and abandoning queries based on responses. Table 2 presents participant motivations for

Table 2. Participants interacted with GAI to ask questions related to their health goals, on topics ranging from general wellness to specific health conditions. Motivations included working towards behavior change, obtaining a record of a behavior, validating current practices or understanding, curiosity, and social sharing. Participants shared personal data in their questions, including demographics, self-tracked values, and other observations or experiences. Supplementary materials include additional topics in which participants expressed interest but but did not explore in the interactive session.

P#	Motivations for Tracking & Questions	Question Topic	# of Qs Asked	Data Provided During GAI Interaction
P1	Behavior change Obtain a record Curiosity Validation	Sleep	4	age, cardiovascular fitness level, stress level
		Cardio levels	4	age, current cardiovascular fitness level, exercise/activity, target cardiovascular fitness level
		Weight loss	3	target heart rate zone, target weight loss
P2	Behavior change Obtain a record Validation Curiosity	Relation between various parameters (e.g., sleep, heart rate, cardiovascular activity)	6	age
P3	Behavior change Curiosity Validation	Exercise/Active energy	2	age, calories burned, gender
		Menstruation	5	first day of last menstruation, avg. daily exercise time
		Heart health	5	age
P4	Behavior change Obtain a record Validation	Sleep	5	age, gender, class end time, sleep hours, student status
		Blood pressure	3	chronic health condition
P5	Behavior change Curiosity	Running/ Fatigue	8	runs per week, running dist. & pace, avg. daily steps, step count, foot condition, exercise-related issues
		Sleep	1	current sleep schedule, nap times, target sleep schedule
P6	Behavior change Obtain a record Validation Curiosity	Magnesium supplements & migraines	5	–
		Physical activity & pain	4	age, activity level, chronic health condition,pain triggers, profession
		Menstruation	4	–
P7	Behavior change Social motivations	Diet plan	5	age, gender, current weight, target weight, height, ethnicity, dietary preferences, physique type, workouts per week
P8	Behavior change Obtain a record Validation	Exercise for fat loss/muscle gain	5	preferred exercises
		Pilates	10	preferred exercises, workout cadence, target weight loss time, location
P9	Behavior change Curiosity	Premenstrual syndrome	3	–
		Discontinuing birth control/hormones	6	time when stopped birth control, symptoms
		Running	2	daily running distance, avg. heart rate, max. heart rate
P10	Behavior change Obtain a record Curiosity Validation	Pupil size	3	–
		Light sensitivity	4	observations from personal experience
		Fatigue & its impact on the nervous system	1	–
P11	Behavior change Obtain a record Curiosity	Irritable bowel syndrome	5	symptoms
		Running	4	amount of time mile run increased by
		Sleep	4	observations from personal experience
P12	Obtain a record Validation Curiosity	Sleep	2	–
P13	Behavior change Curiosity	Constipation	9	chronic health condition, diet, activity levels, supplements, exercises
P14	Behavior change Curiosity	Dental health	5	–
		Sleep	5	daily wake-up time, bedtime, nap times, sleep quality
P15	Obtain a record Curiosity	Menstruation / PCOS	5	age, gender, height, ethnicity, period irregularity
P16	Behavior change Curiosity	Weight loss	4	weight loss goal, interactions with tracking app
		Weight loss & mind	1	mood
P17	Obtain a record Curiosity	Vasovagal syncope	2	symptom, activity
		Heart rate/running	1	heart rate, frequency
		Menstruation	1	medication, medication stop date, bleeding frequency, blood condition
P18	Behavior change	BMI	1	gender, height, weight
		Diet & exercise plan	3	fat loss goal, target body location, number exercise days, experience level, family medical history
		Tracking app recommendations	2	–
P19	Obtain a record Curiosity Validation	Sleep	4	–
		Physical activity & pain	4	–



P13: Used GAI to analyze heterogeneous self-tracked health data, such as to identify causes and solutions for her chronic constipation issues.

P6: Used GAI to identify correlations between her health variables and behaviors / practices, such as to understand the relationship between physical activity and hip / back pain.

P7: Used GAI to seek more creative and individualized meal recommendations toward personal health goals of maintaining a good physique and working towards his ideal body.

P10: Used GAI to identifying new health goals / questions and how to start tracking for them, such as to understand the science behind pupil dilation and how to measure pupil size.

Fig. 2. Participants interacted with GAI around their health data in a variety of ways. Questions were shaped by their diverse health contexts and goals, while specific interactions with and support desired from GAI were shaped by a range of expertise (e.g., in tracking, in prompting). Vignettes detailing this range of experiences are included in supplementary materials.

tracking, question topics, number of questions asked, and different types of provided data. Figure 2 shows some of the ways in which participants interacted with GAI during their interactive sessions.Supplementary materials also include four vignettes illustrating the variety of ways participants interacted with GAI, and how expertise levels (e.g., in tracking, in prompting) shaped that interaction and desired support. This section reports on participant interactions with GAI during the interactive session and challenges encountered in asking questions and reflecting on self-tracked health.

4.1.1 Shaping queries and deciding what data / personal information to include. Most participants shaped queries as questions. A notable exception was P2, whose queries only included topics (e.g., queried *blood pressure and menstrual cycle* expecting a response explaining any relationships between them). P1 and P18 asked GAI to do *"mathematical"* (P1) tasks and P14 asked it to analyze uploaded datasheets. P7 had prior experience with prompting GAI around his personal health goals and explicitly assigned the agent a role as a *"health data tracker or health and fitness expert"* before providing personal details and asking GAI to craft a custom diet plan. However, other participants (P5, P6, P10, P13) described challenges with not knowing how to formulate questions:

> *"The struggle is how do I phrase the question?… 'I'd be interested to know what level of physical activity [I should do]…' I just don't know how to phrase my own tracking into like a question"* (P6)

After asking questions, some participants remained uncertain they were *"asking the right questions"* (P11). Although Copilot provided suggested questions after each response, only P8 clicked a suggestion, likely because suggestions were not specific enough to what participants were seeking.

Participants provided different types of data or context in hope of more aligned or personalized responses. Some provided numerical data, often with an expectation of numerical responses. P3, P9, and P17 intentionally specified dates or timelines while seeking responses predicting the date of their next period or symptoms they might experience according to the stage of their menstrual and ovulation cycle. When specifying numerical data, participants often accessed their self-tracking devices and picked stats to include in queries (e.g., cardiovascular fitness levels (P1), heart rate and running stats (P9)). However, participants also noted this was tedious and they did not always know where to find desired values. Multiple participants provided *"identifying"* (P7) information, such as demographics and body-related details (e.g., age, gender, race/ethnicity, occupation, lifestyle, body type, height, weight), as they wanted responses personalized to those contexts. This also included information about

existing health issues or known symptoms (e.g., hip / back pain due to scoliosis (P6), IBS symptoms (P11)) and current practices (e.g., types of workouts and pilates movements (P8), prior treatments and their effectiveness for relieving chronic constipation (P13)). Participant motivation for providing such information was often to avoid general or harmful recommendations (e.g., that might not be appropriate for people with specific chronic issues) and redundant recommendations (e.g., things they have already tried). Participants often combined these different types of personal information when shaping queries, seeking even more tailored responses (e.g., *age: 26; gender: female; role: master's student with night classes end at 10 pm* (P4); a lupus patient (P4)). Despite providing demographics and other relevant context, participants often did not find GAI responses appropriately tailored, as we report in Section 4.1.4.

Participants also expressed challenges or discomfort sharing certain types of personal data. For example, P11 felt *"weird"* sharing their weight and described privacy concerns around uploading *"specific numbers"* because they would not want *"random people on the internet"* to have access to their personal health details. However, P11 also acknowledged *"other companies already have all my data anyways."* Others (e.g., P5, P14) were less concerned about privacy, describing being *"liberal"* (P5) about sharing self-tracked health data (e.g., exercise, diet) and not feeling a need to *"protect"* (P14) it as *"you cannot really do anything dangerous with it"* (P5). P8 further noted being open to sharing her personal health data to improve GAI models *"if it would help everyone collectively"* by making GAI models more *"trustworthy"* and based in *"people's real life scenarios"* instead of *"whatever [GAI] can find on the internet."* Several described challenges providing custom data or asking related questions, including how to convey data and insights from health tracking applications. P12 wanted to upload his self-tracking data from Whoop (a GAI-enabled fitness tracker), but did not know how to export it in csv format, and then ultimately did not ask many questions in the interactive session. P3 described potential for error if she tried to use data from Apple Health to ask questions, doubting her *"ability to relay my data to the computer."*

> *"As soon as I take [health tracking data] out of the Apple Health app, there's a chance for the information to be misrepresented in a [ChatGPT] prompt. It's not even that I don't trust the machine, but I don't trust myself to deliver the appropriate prompt… like to interpret the data and then feed it to a machine that again, is going to utilize patterns and recognition…If I use just the wrong words, it [ChatGPT] will spit out garbage, which could mislead me…"* (P3)

Similarly, P6 described not knowing how to *"input"* her maximum heart rate, because she did not know what it was, but *"Fitbit, it probably just knows that without my knowledge."*

4.1.2 Evolving and refining queries based on responses. Participants **asked follow-up questions and progressively added more personal information** to *"improve"* or better align queries to their needs. P5 wanted recommendations on how to *"switch up my routines to start getting into longer runs without overexerting myself."* After several iterations, she provided a week of step count data to give GAI *"a better idea"* about her activity levels. Participants also followed up when they felt responses were incorrect or contained information they already knew. Three (P1, P2, P4) included credibility prompts in follow-up queries, explicitly asking for recommendations *based on research* (P4) and *peer-reviewed resources* (P1), with P1 further checking quality of resulting references and noting she would better trust responses that *"cite real data."*

Participants evolved queries to **connect two or more forms of health goals** for which they tracked. P1's queries progressed from understanding what her sleep-related goals should be (*How much deep sleep do I need each night?*) and what her cardiovascular fitness reading means (*What does a cardio fitness level of 26.5 mean for a female aged 30?*), to connecting the two and adding a third variable (*high stress*) to ask *How much deep sleep does a 30-year old female with low levels of cardiovascular activity and working under high stress need?*

Participants also **refined wording** to describe health goals (e.g., from *losing weight* to *losing body fat*). They described that being precise with query formulation and wording was especially important in a context like

health: *"the consequences for getting bad health information are greater than the consequences of something, say, not summarizing my email correctly"* (P3).

After forming queries in different ways, some participants described that GAI perhaps could not break down complicated or multi-step prompts, so they **provided more scaffolding in queries** by breaking down and re-framing them. P10 noted that *"instead of trying to ask clever questions,"* he could try describing his *"observations in detail"* and *"be one level less abstract,"* asking questions like *"Hey, I noticed when I do this, this is the reaction. Can you explain?"* After using this approach, P10 felt more satisfied with the response. P14 described breaking down and re-framing his queries in a similar manner to *"coax"* GAI to a response he found useful.

4.1.3 Ongoing reflection and sense-making. Reflection happened throughout participant interactions with GAI, sometimes leading them to evolve information goals. P10's information goal evolved from measuring / tracking pupil size and understanding what factors affect pupil size, to understanding what other factors affect light sensitivity. P10's expertise in self-tracking and analyzing his data for a variety of short-term and long-term health goals likely informed his ability to reflect on and evolve his goal based on GAI responses. P6 evolved from querying to understand whether magnesium helps with her migraines and what magnesium supplements to use, to querying about other lifestyle changes that might help prevent migraines and which she might also consider.

Participants were motivated to act on recommendations they considered both aligned with their goals and appropriate to follow:

> *"I would like to try this out…Because this seems doable…Because 150 minutes of exercise per week means they're safe. I could just work out for five days a week. So that's 30 minutes each day…it's not like, it recommended me taking some sort of medication or something. So I don't think I would need to consult someone for increasing physical activity…I know I can increase physical activity, and that's fine."* (P1)

> *"I don't feel that it is necessary to ask [a professional before experimenting with a change in my sleep schedule]…because this time [11pm-3am suggested by ChatGPT] already makes more sense than my original sleeping time like 4pm to 8pm or something. So I think I can give it a try first…I think I will take action to try the suggestions ChatGPT gave me."* (P4)

Both participants felt recommendations about exercise and sleep were appropriate, but may have taken these recommendations based only on face validity (i.e., fiting what participants might expect is appropriate). Although multiple participants received recommendations related to tracking changes, few explicitly expressed interest in implementing those. P11 said they might *"start just noting down if I had like chickpeas or like other foods that day and then how I felt later on or the next day"*, and P10 and P14 said they would try the approaches GAI recommended for tracking pupil size and pocket depth. P14 felt he could trust the GAI-generated recommendation to take a photo of his teeth to measure and track pocket depth because it seemed benign—*"I do not see how this could go super wrong"*—and was not *"sketchy"* like *"stab a toothpick in your gums or something."* Lastly, multiple participants used responses to validate things they already knew or activities in which they already engaged. This meant sometimes *"ignor[ing] the answers"* they *"didn't really vibe with"* and using other parts of those answers to *"reaffirm the activities that I already do"* (P8).

Although most participants asked follow-up questions, sought clarifications, or refined queries, some (P2, P6, P8, P10, P13, P17, P19) tried to align responses to their needs without follow-up questions. This included correlating and making sense of responses relative to their own experiences and understanding. P13 discussed whether different parts of a response about potential GI tract dysfunctions applied to her:

> *"Celiac I've heard of that before…that it can cause constipation. I actually experienced that a lot with carbs. That's interesting…Electrolyte deficiencies? I've tried electrolytes before so I don't know if that's what*

> *I'm looking for…Fibromyalgia, that could be a possibility. I do have certain myalgic-like symptoms…"* (P13)

Some further tried to correlate responses to their self-tracking. For example, P6 reflected *"it was good for me to look at my data and look at what this information is telling…Like this information is telling me that I need to be gradual with my increases and decreases, and I obviously have not done that."*

Participant processes of reflection also included re-calibrating perceptions of what support GAI can provide. Some were surprised, like P6 who noted *"I think I underestimate the power of AI and what it can do."* On the other hand, P11 stopped querying after being satisfied with responses: *"I would need to enter in very specific numbers to get anything super personalized. So yes, as for how specific my questions, the responses are on par with my expectations."*

4.1.4 Abandoning querying goals. When not satisfied, participants either evolved queries (as in Section 4.1.2) or stopped because they felt unable to get support they were seeking.

Multiple participants described struggling to get GAI to integrate provided data (e.g., numerical data). P5 described not knowing if GAI *"actually uses the numbers"* and feeling responses providing calculations or numerical recommendations were not consistent. Similarly, P7 provided all variables required to perform calculations, but GAI responded with formulas for calculating BMI, energy expenditure, and calorie deficit without performing the calculations, leaving P7 to do the math himself, even though he felt it "*should have been done*": "*Why didn't it do the math when it had the input?*" P7 further attempted to refine his query and sought recommendations tailored to his demographics and personal context, but remained unsatisfied. He felt GAI *"completely butchers the whole context"* by becoming *"fixated"* on one aspect of his prompt (providing *Indian meal options*) while ignoring others in previous prompts (*high-protein meal options*). P7 was also unsure whether GAI was using context he had provided about his *"physique," "culture,"* and *"genetics."*

Other reasons for abandonment included when responses were too generic or *"abstract"* (P8), *"repeated back to me the things that I said"* (P13), provided already known or unwanted insights (e.g., advice for temporary pain relief (P13)), or lacked *"empathy", "reminiscent of the old man doctor who just tells me my problems…not directly addressing your issue, or even acknowledging that you are in discomfort"* (P5). Participants explained responses like these left them unsure of next steps. P13 noted *"It's like a good generic list, not really sure where to go from here."* For P8, advice about what to track was too vague to be actionable: (e.g., *"nutrition, variety, and progress tracking feel really vague. So like progress tracking, monitor your progress and adjust as needed…in my head I'm kind of like is that tracking my weight? Or do I get those body scans and I track my BMI, and body fat percentage?"*

4.2 Participant-Identified Opportunities for GAI Support in Planning, Tracking, Reflecting, and Acting

We report types of support participants desired from GAI around personal informatics processes, including planning for self-tracking, tracking, reflecting on tracked data, and acting upon resulting insights. Participant-identified opportunities draw upon experiences in the interactive session, their exploration of the case scenario, and their long-term experiences with health and self-tracking.

4.2.1 Supporting people in deciding what and how to track for health goals. **Identifying questions to answer and variables to track.** Participants described wanting GAI support in identifying questions related to their health goals and variables to track toward answering those questions. P10 wanted to learn about measuring and tracking pupil size because he was not sure what questions he might have. Although an expert at self-tracking and analyzing his data, P10 was unsure if pupil size was a valuable measurement, how to track it, or if it aligned with his new goal. He highlighted that many people *"don't even know what to track or what they should be worried about"* and considered this an opportunity for GAI. By the end of his interactive session, P10 was convinced pupil size is relevant and had a better understanding of how he might track it (e.g., using a recommended smartphone

app), though he needed to collect data before he could identify concrete health goals or questions. Similarly, P6 expressed wanting to know *"what am I not tracking now that I should track"* and indicated that might help her identify questions or health goals:

> *"I only look at the things I'm tracking, but there are probably things that I should track that I don't. . . . I think Copilot has more capacity to tell me like 'here are things that are helpful for knowing when trying to understand your side effects. . . But we notice that you're not tracking this. Why don't you consider doing that?'"* (P6)

P6 also described she would be interested in leveraging GAI to *"narrow down"* factors to track to better understand and act on hip and back pain.

Some found GAI support demonstrated in the case scenario to be less proactive than what they desired. For example, P12 felt that *"all the tracking ideas"* in the case scenario were from the character interacting with GAI and GAI itself *"was more assistive where it tried to just draw the patterns and create a worksheet"* (P12). P12 already had experience using a GAI-enabled fitness tracker and would have appreciated if *"AI could go one step forward, and I could just tell it my problem that I have shoulder pain, and then it helps me figure out what I should track, like work out, work stress, and whatever."* However, P12 acknowledged difficulty in identifying *"general [tracking] metrics"* that work well for everyone and speculated GAI might help individuals *"ideate"* contributors to their specific health concerns. Others indicated potential for GAI to be more interactive in recommending tracking (e.g., *"Hey, do you happen to have a step tracker?"* (P10)). P9 reflected that GAI might already be doing that in some capacity:

> *"Copilot is helping [the character in the case scenario] focus more on what to ask Copilot, which, in a way, it was also doing for me. . . Because based on the responses I would be. . . thinking, 'oh, these are things that I would like to ask a little bit more about, to learn a bit more about'. . . which did not occur to him like 10 min ago when asking the original question. . . "* (P9)

Others also appreciated GAI support in *"brainstorm[ing]"* (P17, P18, P19) and *"bouncing off ideas"* (P17) for tracking.

Deciding how to measure and track. Even when participants know what to track, they described challenges defining metrics or measuring variables. One reason was they did not know how to quantify concepts. P8 queried to learn how to assess whether pilates is an effective exercise for her. Although GAI recommended tracking strength and endurance *"metrics"* to self-assess progress, P8 desired more explanation of such metrics and how to track them:

> *"I feel like I would have to come up with some way of tracking like strength and endurance. . . I feel like the last one [recording any pain or discomfort during pilates sessions] is easy for me, like if I feel discomfort, it's like I know of it immediately, but, like the rest of it. . . I don't know if I am improving for flexibility and range of motion. . . maybe the first day of every month I to try to touch my toes or do yoga. So I have something to compare it to. I feel like in isolation, I don't really know how I can keep track of this. . . "* (P8)

After further prompting (e.g. *How do you suggest you can track strength? What are concrete activities I can do to get a better sense of how pilates is helping me?*), P8 received recommendations such as maintaining a workout journal, taking progress photos, and self-assessing how she feels after each session. She felt progress photos were a good suggestion, but she was still unsure how to self-assess after each session.

The case scenario included a GAI-created tracking template. Participants generally appreciated its definitions of tracking with recommendations for format (e.g., scales, describing location of pain) and granularity (e.g., scale from 1-10), but some wanted further explanations of how to measure variables, such as how and when to measure stress levels. P7 had concerns about how to input data, wanting an explanation of *"here's how you input this kind of data or track this"* or *"how to track these over the next few days."* Participants also wanted more descriptions of

how to quantify variables. For example, although the scale's endpoints were labeled—0 means no pain and 10 means severe pain—participants wanted more guidance about intermediate levels. P12 further described how scales could be subjective, with individual biases impacting tracking (e.g., individual perception of stress could be influenced by factors such as workload and physical exertion). Rather than subjective or ambiguous-to-interpret, P12 wanted *"very concrete objective metrics."*

4.2.2 Assistance with setting up and evolving tracking regimes. Multiple participants described wanting support creating and maintaining *"robust"* (P5, P8) tracking plans. P8 described tracking setup as the *"hardest part"*, but that maintenance was a challenge if *"a new metric to track"* meant she had to *"start over"*. P8 thus liked how Copilot could generate a tracking template, as that could *"reduce the barrier to keeping track of all these new metrics."* P5 wanted a worksheet *"like a doctor will give you… and tell you to fill it out and bring it in next week."* P13 and P18 wanted to upload their tracking spreadsheet or data log from a tracking application, as explored in the case scenario, and to try using GAI to generate a better spreadsheet (P13) or to identify *"inefficiencies"* (P18) in their current tracking. P18 also appreciated the case scenario showing how Copilot helped modify existing tracking regimes. However, some others found worksheets tedious to manage, *"inaccessible"* (P6) when needed for tracking, and of unclear utility even once filled out (e.g., *"Do you upload it to the chatbot or show this to someone.… show to your doctor? Maybe just show next steps…* (P15)).

One suggestion was to integrate data from tools a person already uses and support configuration of custom tracking based on that:

> *"I would input all the ways that I keep track of health already, and then see if they can integrate it or make it better… Maybe, like the calendar, Strava, Notion template… I think if I input that, then it closes the bridge… taking the information I provide and like recreating a template with the edits. So it would just give it to me… that would help with actually implementing it."* (P8)

4.2.3 Supporting direct integration of self-tracking data with GAI. Although only P14 attempted to directly upload their self-tracking datasheet in the interactive session, multiple participants said they wanted to directly integrate self-tracked data into GAI, by either uploading spreadsheets or integrating tracking devices. This could reduce errors in relaying information (P3), provide integration across applications (P7), and address challenges around formulating precise prompts (P6). The case scenario demonstrated uploading an image of a datasheet, in part to elicit reactions and in part because early participants had not explored uploading multimodal data. Participants appreciated GAI ability to use multimodal data (e.g., *"[Directly uploading datasheets and images as in the case scenario] just kind of blew my mind. I'm like, Oh, my gosh! Why didn't I think of this? This is like so useful!"* (P13)). Participants then also anticipated additional opportunities for GAI to integrate and interpret non-textual data such as photos or videos (P8) and handwritten notes or data tracked on paper (P11).

4.2.4 Support for understanding and interpreting insights enabled by health data and GAI. To better understand and act on health data insights, participants wanted additional support for interpretation. This included for *"parsing qualitative[ly tracked] data"* (P14), condensing long responses (P3), and providing meaningful explanations (e.g., *"most people don't want data. They want to know how you came to the data, like, what is the magic you did? So then it doesn't seem like magic"* (P7)). Participants described that GAI could help organize data into easy-to-understand formats (e.g., summaries, visualizations, lists). Although some preferred visualizations over text (e.g., P11, P12, P16, P19), others (e.g., P1, P9) wanted further explanations to understand visualizations. Participants also asked for data about other people and their experiences. P9 felt responses lacked a *"human voice aspect"* and wanted *"testimonies of other people"* against which to compare her own: *"people saying, 'Oh, this is the experience I went through… ' I guess that's kind of what I was trying to get at when I was asking it like 'what percentage of people have PMS for 2 weeks or so?'… like what do people actually experience?"* P6 and P8 wanted examples of how other people were implementing similar recommendations.

In wanting support with interpreting data-based insights, participants remained concerned about hallucinations:

> *"If it [a data-based insight] is coming from an AI assistant, I would not trust it a lot. I would personally check the graph and then go and verify with my data to ensure that it did not hallucinate…If it gives me a pie chart, I'm going to verify if the data it is using is not…It is just hallucinating that data. Or if it [GAI] is given me a correlation, I would go and see 2 or 3 data points just to verify, just to do a sanity check that it is not been hallucinated."* (P12)

P4 and P9 raised similar concerns and wanted to verify GAI-generated insights with a trusted medical provider.

4.2.5 Needing GAI to take initiative, be interactive, and provide in-the-moment support. Participants described opportunities for GAI to be *"proactive"* (P14), seek clarifications, and share data-based insights without explicit querying. P1 wanted GAI to *"not just spew out the calculation of how much calorie deficit you need every day"* but to ask follow-up questions such as *"Why do you want to lose X pounds in a week? What is your motivation?…Have you been working out regularly?…Are you practicing any form of diet or calorie deficit programmes?"*, feeling this could provide better context, support *"better answers"*, and reduce risk of potentially harmful suggestions. P5 similarly shared that GAI *"should also ask me questions in return"* to get a more *"holistic"* understanding instead of *"operating in a vacuum."*

Participants described how mixed-initiative approaches could support navigating challenges around formulating questions. A *"prompt generator"* (P3, P7) could provide support by asking for specific health data (P6), for *"key pieces of information"* (P19), or by sharing potential queries so that interactions do not feel like *"throwing things out into the universe and getting back like canned answers"* (P3). Others described how GAI should analyze their data and *"send notifications"* (P7, P14) to share insights:

> *"you didn't have to have a sit down session to be able to analyze the data. It would just ping you a notification and say something like. 'Oh, I noticed that on days where you have caffeine after 3 pm, you haven't been getting great sleep.'"* (P14)

They described system-initiated insights as useful because they may not *"seek it out unless something is wrong"* (P3). P9 described that in-the-moment recommendations based on prior sessions might better support taking action:

> *"I think part of it is just the timing of it, like reading about it beforehand and executing it 5 hours later is probably going to be a little difficult for me in terms of just retaining that information and reminding myself when I'm running. I am in the zone, and I don't really have capacity to think about anything else, so I doubt there would be a moment where I think, oh, this is a moderate pace, or this is high intensity interval training…I don't think I have the capacity to categorize my running into those spaces while I am doing it"* (P9)

P9 further described Fitbit's analysis of her heart rate while running (e.g., *"you can do like a kilometer at the heart rate of like 155"*) as a model for opportunities for GAI.

5 DISCUSSION

We reflect on our findings and discuss key considerations around: (i) GAI in self-tracking tools to account for a range of health goals and questions, (ii) GAI opportunities for providing scaffolding at different stages of self-tracking, and (iii) concerns and implications in embedding GAI in personal health informatics tools.

5.1 Accounting for a Range of Goals and Questions when Providing GAI Support for Personal Health Informatics

Consistent with prior research [14, 29, 72], participants self-tracked for a variety of goals (Table 2) and queried GAI for a range of needs (e.g., summarizing heterogeneous health data, identifying correlations and new variables to track, validating current practices, seeking recommendations) (Section 4.1). However, participants grappled

with formulating and aligning queries to their diverse goals and needs, often leading to query abandonment or lack of satisfaction with responses (e.g., Section 4.1.1 and 4.1.4).

In interactive sessions, participants employed varying prompt engineering strategies (e.g., assigning GAI a role (P7), detailing context (P6, P11, P13, P17), providing numerical data (P3, P9, P18)). Despite using prompts aligned with best practices [19, 68], participants encountered challenges obtaining responses aligned with their goals. As described in Section 4.1.2, participants broke down queries by describing observations, providing context, and then asking about specific relationships between health variables or practices. For some, this resulted in useful insight, while others did not receive a response they found useful. Such variation may be explained by differences in whether participant data contained sufficient information to generate useful insights. However, we observed that GAI tools used in this study seemed to handle some questions (e.g., identifying presence of a relationship) better than others (e.g., identifying interactions between different potential causes of a problem). Some participants were also unsure what context might be useful to describe or how to best describe it, and may have benefited from tools asking more questions to elicit this information. Additionally, not all participants knew how to break down queries or about the capabilities of GAI tools used in the study. P13 and P18 learned about the ability to upload a tracking spreadsheet from the case scenario, too late to apply it in their interactive session.

Some of these challenges are common in general-purpose GAI agents [22]: without explicit cues about a system's capabilities, people are unsure of what to try or how to use it most effectively. Even the conversational agents used in this study may have been more usable for participant health goals if they were presented in an interface that offered examples and highlighted some of the ways participants could use the tool to access support. Combining structured inputs (e.g., prompts for people to describe and categorize a goal [2, 74]) or query-templates for a variety of goals may have resulted in greater success for study participants.

5.2 Opportunities for GAI to Provide Scaffolding in Personal Health Informatics

We discuss opportunities for GAI to better scaffold support for challenges across stages of self-tracking, including extending participant-identified opportunities with additional context from prior self-tracking research. Although our study design means we primarily observed participants using GAI to reflect on their health data, participants identified opportunities for support across deciding what and how to track, setting up and evolving tracking regimes, and interpreting and acting upon health data insights (Section 4.2). We directly observed some of these, while other participant perspectives were more speculative (e.g., a stated intent to act based on insights from an interactive session still needs be translated into actual action). Our findings can guide future research that examines longer-term experiences with GAI-integrated health tracking tools.

5.2.1 Supporting Planning for and Evolution of Self-tracking. Participants described barriers in defining what and how to track, together with opportunities for GAI support (Section 4.2.1). This included seeking recommendations of variables to track towards their goals, how to measure those variables (e.g., scales to use), and how to operationalize measurements that felt *"subjective"* (P12). These challenges of quantifying health variables that can be measured or perceived in different ways correspond to prior work in transforming qualitative goals into quantitative goals [61] and in examining how people align health goals, information goals, and tracking goals [74].

One opportunity is to support planning of tracking according to established best practices in personal health informatics. P12 emphasized that a scale for tracking stress could feel subjective in part because stress can be affected by other factors (e.g., physical exertion, workload) (Section 4.2.1). This is consistent with research that people who track to understand an outcome (e.g., stress) often fail to track potential contributors (e.g., physical exertion, workload) and therefore struggle to understand their relationship [14]. Recent research has also highlighted potential for GAI-enabled journaling [44] and personalized prompting [59] to encourage self-tracker reflection on contributing factors. GAI could thus be leveraged in prompting individuals to consider tracking

such context and contributors alongside a primary variable of interest, consistent with participant suggestions that GAI could help individuals *"ideate"* (P12) and *"brainstorm"* (P17, P18, P19).

Participants also described being unsure how to use scales with high granularity (e.g., a 10-point scale for pain), which prior work supports interpreting as participants lacking an information goal warranting that level of granularity [74]. GAI could support better aligning tracking to goals (e.g., recommending tracking a Boolean presence or absence of pain), which could be informed by recent explorations of using GAI to structure self-reporting prompts according to information goals [82]. Another opportunity arises in revising tracking over time (e.g., changing format or granularity). Sefidgar et al. [74] found that tracking plans need to be revisited and re-aligned when information goals evolve, and P8 expressed concern about *"start[ing] over"* if her tracking needs changed. GAI could perhaps support P8's desire for *"robust"* tracking plans by supporting evolution in terms of both new tracking and preserving value of an individual's previously-collected data.

5.2.2 Providing Knowledge and Context for Supporting Individual Goals and Reflection. In our study sessions, participants used GAI-generated content to support their understanding in reflection (Section 4.1.3) and they described wanting more support like this (Section 4.2.4). This content could validate existing understanding and support inferences around how content might apply to individual goals and questions.

Prior research emphasizes a lack of effective ways to obtain knowledge and integrate contextual information to supplement self-tracked health data [13, 51]. In practice, current tools address this by supporting specific goals (e.g., food tracking tools often support weight loss goals). A system's designers can then provide knowledge, context, and scaffolding to support tracking and reflecting toward this set of goals (e.g., impact of diet, activity, or other factors on weight loss). Although effective, this breaks down when people bring different goals (e.g., tracking food for goals distinct from weight loss [16], tracking physical activity without a goal of increasing activity [38]). Research has proposed highly-configurable tools in which individuals define their own tracking (e.g., OmniTrack [45]), but, by not constraining the supported goals, designers lack the constraints that help them know what relevant knowledge, context, and scaffolding to provide [29, 50]. Recent research in goal-directed tracking explores configuration at the level of individual goals [74]. This allows designers to consider and account for a variety of goals, but is limited by requiring designers explicitly anticipate and account for the range of goals that will be supported.

Our findings motivate exploring how GAI might be integrated into configurable tools, what types of knowledge and context GAI can provide, and how scaffolding across self-tracking stages could support goals and reflection. Other research has found that incorporating textual narratives alongside quantitative fitness data encouraged deeper reflection, increased participant engagement, and was perceived as rewarding [76]. Future work could investigate incorporating narratives to complement quantitative data and design GAI response formats to advance individual goals. However, it will also be necessary to navigate potential harms where GAI-generated knowledge and context is inappropriate, incomplete, or incorrect.

5.2.3 Supporting Action on Appropriate Insights. Clarity in specific actions to take is a barrier to acting on health data insights [50]. Participants stated intent to act on GAI-based insights, including making changes to tracking (e.g., to begin tracking data related to stress levels and contributors, tracking foods that potentially contribute to IBS symptoms) and health behaviors (e.g., changing diet, changing workout routine). GAI advised participants in both *what* actions to take (e.g., do physical activity to reduce pain) and *how* to take action (e.g., gradual increases or decreases in physical activity). We also observed participants changing information goals through reflection (Section 4.1.3).

Actionability can include many dimensions that vary across individuals (e.g., understanding how to pursue an action, being willing to pursue an action, being able to pursue that action). Motivated by concerns for GAI being assertive despite lacking necessary context, prior research developed an interactive system that asks follow-up questions and abstains from guessing when unsure, aiming to make GAI recommendations safer [52]. Informed

by such approaches, support for identifying actionable recommendations amidst above-noted dimensions should also evaluate recommendations for appropriateness (e.g., safety). An action of beginning to track an existing activity likely does not require additional considerations (e.g., beginning to track consumption of certain foods and their effect on symptoms (P11), tracking pocket depth by taking teeth photos (P14)). In contrast, some GAI recommendations may require more care. P1 and P4 found recommendations on lifestyle changes to be benign (e.g., exercise more, better sleep at specific hours), explicitly stating they did not feel a need to consult an expert before taking action (Section 4.1.3). However, such assessment of appropriateness might be based on face validity, and so individuals might inappropriately trust recommendations that correspond to their existing beliefs or expectations and might distrust recommendations that do not correspond. P1 reached a specific exercise recommendation through calculations of her calorie burn rate and desired calorie deficit, and P1 seemed well-informed about her body and appropriate levels of exercise and calorie control. We did not observe participants considering any recommendations we would consider dangerous, and Section 3.2 notes our procedure emphasized seeking expert advice prior to following GAI recommendations. Nevertheless, it may not be appropriate to assume that even well-informed individuals will always seek expertise or otherwise know what behavior change recommendations are appropriate.

Many safety concerns can depend on other personal context (e.g., existing medications), so approaches to this challenge are likely complicated by challenges participants experienced in obtaining recommendations they were confident accounted for their personal context (Section 4.1).

5.3 Concerns and Implications at the Intersection of GAI and Personal Health Informatics

Our findings underscore a need for designers and researchers to be mindful of potential harms at the intersection of GAI and personal health informatics.

5.3.1 GAI Breakdowns and Risks of Improper Personalization. Participants noted shortcomings of GAI, including shallow integration of personal data, repeating back queries without providing new insights, and failing to provide actionable next steps (Section 4.1.4). Other known limitations of GAI include hallucinating, factually incorrect or biased responses, and gullibility in being coaxed into desired or over-tailored responses [9]. Given such breakdowns, the integration of GAI in personal health settings presents important potential for harmful consequences. Participants raised similar concerns and described a need to verify GAI insights (Section 4.2.4). For example, P7's experience with GAI losing the context of his food-related queries was frustrating (Section 4.1.4), but analogous breakdowns losing context of an allergy or an eating disorder could be harmful. Designers of GAI in personal informatics tools could leverage techniques like chain-of-thought prompting to support reasoning through a series of intermediate steps [83], but designer intent to effectively break down queries can instead become coaxing of GAI, potentially introducing designer biases and assumptions that personal health informatics research has found problematic [46].

Merrill et al. similarly highlighted risks of improper personalization in personal health settings [55], and participants expressing frustrations with shallow integration of provided data and context. Integrations of GAI that offer surface-level personalization without effectively tailoring their underlying recommendations could be burdensome or problematically cultivate trust that is ultimately unwarranted. One approach could be through direct integration of self-tracking data into GAI (e.g., directly uploading datasheets, integrating data from tracking devices), which may help address challenges of relaying information and forming precise prompts (Section 4.2.3). Recent research has explored direct integration of self-tracked health data (e.g., wearable data, survey data) for personalizing the reflection and sense-making experience and for supporting actionable steps toward personal health goals [30, 60, 66, 76], but seamless integration remains challenging. For example, an exploration of combining wearable-triggered journaling with GAI support in a stress management context found a misalignment in detection of stress events relative to when a person might actually want to engage with an

intervention [60]. Effective integration with GAI will also require addressing challenges around synchronizing, interpreting, and analyzing data from varying types of tracking devices. Designers and researchers pursuing opportunities at the intersection of GAI and personal health data should therefore be mindful of multiple challenges, from lower-level challenges of data integration, to challenges of human-AI interaction design in specific personal health contexts, to risks of improper personalization and a need for safety measures to prevent unintended consequences.

5.3.2 Careful What We Wish for: Implications for Real-time GAI Support in Personal Health Informatics. Participants extrapolated from their experiences to express a desire for GAI agents that could access real-time data, take initiative, and provide in-the-moment advice (Section 4.2.5). Although some raised privacy concerns around sharing personal health details, others were more open to sharing self-tracked data to improve GAI models that could benefit self-trackers (Section 4.1.1). Situating participant perspectives within personal health informatics research, we note similarities between the future some participants imagined and Fit4Life, a speculative sensor-based health tracking system that employed principles from persuasive computing to provide real-time, personalized feedback promoting healthier behaviors [69]. However, the authors of Fit4Life imagined it to illustrate the potential for such in-the-moment feedback to create coercive and invasive designs that reduce human experience to algorithmic input, undermine human agency, and raise concerns for surveillance and privacy. A key question is therefore how to reconcile participant visions of future technology with this cautionary and dystopian tale. One possibility is that support which is desirable in explicit and intentional engagements—such as participants experienced in our study—would quickly become unappealing in ubiquitous and longitudinal use. Another is that participants may have been implicitly expressing confidence in the ability of designers to create systems that offer support that aligns with their holistic goals. Although some participants in our study assessed reliability of GAI through explicit credibility checks and asking for research-based recommendations (Section 4.1.2), further research in personal health informatics contexts is required to examine what activities and roles people want GAI to engage in, potential concerns for overreliance, and what activities and roles people want to keep for themselves. As our communities advance GAI-enabled systems for personal health, this is a challenge that researchers and system designers must meet.

5.4 Limitations and Future Research

As noted in Section 3.4, our focus on people who already self-track meant participants could bring their existing data, questions, and long-term experiences to the study. This supported us in learning about opportunities and challenges for GAI to support various stages of self-tracking across a range of data, goals, expertise, and health contexts, as summarized in Table 1 and Table 2. Although informative of experiences of people who are experienced in personal health tracking and/or early adopters of new technologies [14], this choice also comes with limitations and future research opportunities. First, new self-trackers may experience additional or different challenges (e.g., not collecting key data needed for their questions [14]), need more suggestions [70], or not be convinced about the value of tracking [43, 70]. People new to self-tracking may benefit more from the support GAI can provide, while people with more experience may have already learned relevant lessons. Lack of familiarity with tracking processes or with GAI may also lead to difficulty in prompting GAI for desired support and limit meaningful interactions around personal health data. Inexperience may also make it more difficult to detect responses that do not align with a person's goals, increasing the risk of GAI creating harm. Future research should engage with people new to tracking and with a broader range of technology experience (e.g., people unfamiliar with using GAI), although ideally after designs have been updated to address opportunities identified in this research. Second, participant demographics likely influenced their experiences in the study, but participants generally did not address such influence. An example was P7, who used GAI to seek recommendations personalized to his age, sex, and ethnicity but was unsatisfied with responses (Section 4.1.4). People with marginalized identities

or those not well-represented in GAI data may encounter biases [63, 65] not reflected in our study or experienced by our participants. It remains important for research to support a broad range of populations and to explore potential differences in experiences based on factors such as demographics and identity, with goals for informing more inclusive design of GAI-enabled health tracking technologies and preventing exacerbation of existing health inequities (e.g., by disproportionately benefiting people who are more educated, technology savvy, or of specific identities and cultural backgrounds).

Additionally, our single-session study design focused on participant experiences with GAI during the interactive session and participant attitudes about potential future use after interacting with GAI and seeing the case scenario. Developing systems or further technology probes that address opportunities highlighted by our findings (e.g., deeper integration of GAI tools with personal data libraries, helping people identify opportunities for GAI support and craft effective prompts toward that support, helping people translate query sessions into effective action) may be an important prerequisite to successful future longitudinal studies of GAI tools that support individuals across stages of personal informatics processes. Although our study did not limit participants to text-based interactions with GAI, participants generally did not upload and analyze multimodal data (e.g., images, datasheets). Participants appreciated demonstration of this capability in the case scenario (Section 4.2.3), motivating future research that examines multimodal GAI interactions around diverse forms of self-tracked health data (e.g., wearable data, images from physical journals, progress videos). We used existing, real-world GAI tools (i.e., ChatGPT, Copilot) as design probes, and those tools continued to evolve over the course of our study. Finally, so that participants could comfortably share their personal data with the GAI tool during the study session, for all but the first four participants, we used a version of Copilot that included commercial data protection. Although this commercial data protection may have reduced participant expressions of concerns around privacy, some still noted unwillingness to share specific health details (Section 4.1.1). Future research should further examine data ownership and privacy concerns around GAI in personal health informatics.

6 CONCLUSION

We presented a qualitative study examining self-tracker interactions with GAI as they worked to understand, interpret, and plan action upon self-tracked health data, including data collected using wearable and mobile devices. Participants formulated questions based on their data and reflected on GAI responses in a variety of ways. They refined and evolved queries, both according to information goals and as information goals evolved in ongoing reflection and sense-making. They progressively included more personal context and data, and they broke down and re-framed instructions toward desired responses, but also ultimately abandoned queries where GAI did not meet their needs. Reflecting on both their experiences in the study session, their long-term experiences with personal informatics, and past experiences using GAI, participants described opportunities for GAI support in self-tracking for health. This included support in deciding what data to track and how, setting up and modifying tracking regimes, and interpreting and acting on data-based insights. Finally, we further discussed these findings in terms of the potential integration of GAI in self-tracking tools to support a range of health goals and questions, GAI opportunities for scaffolding planning, reflecting, and acting on health data insights, and concerns and implications in embedding GAI in personal health informatics tools.

Acknowledgments

This research is supported in part by the National Institutes for Health, through the National Library of Medicine under award R01LM012810 and through the Institute of Translational Health Sciences, funded by the National Center for Advancing Translational Sciences under award number UL1TR002319. The content is solely the responsibility of the authors and does not necessarily represent the official views of funders. We are also grateful

for the valuable feedback provided by our reviewers. We thank all our participants for contributing their time and perspectives.

References

[1] [n. d.]. ONVY. https://www.onvy.health

[2] Elena Agapie, Patricia A Areán, Gary Hsieh, and Sean A Munson. 2022. A longitudinal goal setting model for addressing complex personal problems in mental health. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.

[3] John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine* 183, 6 (June 2023), 589–596. doi:10.1001/jamainternmed.2023.1838

[4] Eric P.S. Baumer. 2015. Reflective Informatics: Conceptual Dimensions for Designing Technologies of Reflection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 585–594. doi:10.1145/2702123.2702234

[5] Marit Bentvelzen, Jasmin Niess, Mikołaj P. Woźniak, and Paweł W. Woźniak. 2021. The Development and Validation of the Technology-Supported Reflection Inventory. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 366, 8 pages. doi:10.1145/3411764.3445673

[6] Marit Bentvelzen, Paweł W. Woźniak, Pia S.F. Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting Reflection in HCI: Four Design Resources for Technologies that Support Reflection. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 2 (March 2022), 27 pages. doi:10.1145/3517233

[7] Andrew B.L. Berry, Catherine Y. Lim, Calvin A. Liang, Andrea L. Hartzler, Tad Hirsch, Dawn M. Ferguson, Zoë A. Bermet, and James D. Ralston. 2021. Supporting Collaborative Reflection on Personal Values and Health. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 299 (Oct. 2021), 39 pages. doi:10.1145/3476040

[8] Shashank Bhasker, Damien Bruce, Jessica Lamb, and George Stein. 2023. Tackling healthcare's biggest burdens with generative AI. https://www.mckinsey.com/industries/healthcare/our-insights/tackling-healthcares-biggest-burdens-with-generative-ai

[9] David C and Paul J. 2023. ChatGPT and large language models: what's the risk? https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk#:~:text=they%20can%20get%20things%20wrong,are%20prone%20to%20'injection%20attacks

[10] Chelsea. [n. d.]. Ouraring please update your advisor so that it can analyze cyclical data over time. Massive oversight of the needs of your female users here, especially since you're classified as a cycle tracker for HSA purposes. https://www.threads.net/@techchatchelsea/post/DCrb07Iv9Xn?xmt=AQGzNxmp0pUyVI_wGiQygULT1NiV9NwwKJe7PCs50zuLZg

[11] Zhaoqi Cheng. 2024. Interpretable and generative AI for actionable insights from textual data. https://open.bu.edu/handle/2144/48748

[12] Janghee Cho, Tian Xu, Abigail Zimmermann-Niefield, and Stephen Voida. 2022. Reflection in Theory and Reflection in Practice: An Exploration of the Gaps in Reflection Support among Personal Informatics Apps. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM. doi:10.1145/3491102.3501991

[13] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding self-reflection: how people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare* (Barcelona, Spain) *(PervasiveHealth '17)*. Association for Computing Machinery, New York, NY, USA, 173–182. doi:10.1145/3154862.3154881

[14] Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1143–1152. doi:10.1145/2556288.2557372

[15] Shaan Chopra, Jeanne Carroll, and Jessica Pater. 2024. Providing Context to the "Unknown": Patient and Provider Reflections on Connecting Personal Tracking, Patient-Reported Insights, and EHR Data within a Post-COVID Clinic. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 449 (Nov. 2024), 34 pages. doi:10.1145/3686988

[16] Felicia Cordeiro, Daniel A. Epstein, Edison Thomaz, Elizabeth Bales, Arvind K. Jagannathan, Gregory D. Abowd, and James Fogarty. 2015. Barriers and Negative Nudges: Exploring Challenges in Food Journaling. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1159–1162. doi:10.1145/2702123.2702155

[17] Aykut Coşkun and Armağan Karahanoğlu. 2023. Data Sensemaking in Self-Tracking: Towards a New Generation of Self-Tracking Tools. *International Journal of Human–Computer Interaction* 39, 12 (2023), 2339–2360. doi:10.1080/10447318.2022.2075637

[18] Luigi De Russis, Alberto Monge Roffarello, and Luca Scibetta. 2024. Dialogues with Digital Wisdom: Can LLMs Help Us Put Down the Phone?. In *Proceedings of the 2024 International Conference on Information Technology for Social Good* (Bremen, Germany) *(GoodIT '24)*. Association for Computing Machinery, New York, NY, USA, 56–61. doi:10.1145/3677525.3678640

[19] DigitalOcean. [n. d.]. Prompt Engineering Best Practices: Tips, Tricks, and Tools. https://www.digitalocean.com/resources/articles/prompt-engineering-best-practices

[20] Mindy Duffourc and Sara Gerke. 2023. Generative AI in Health Care and Liability Risks for Physicians and Safety Concerns for Patients. *JAMA* 330, 4 (July 2023), 313–314. doi:10.1001/jama.2023.9630

[21] Adam G. Dunn, Ivy Shih, Julie Ayre, and Heiko Spallek. 2023. What generative AI means for trust in health communications. *Journal of Communication in Healthcare* 16, 4 (Oct. 2023), 385–388. doi:10.1080/17538068.2023.2277489

[22] Yogesh K. Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M. Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, Hanaa Albanna, Mousa Ahmad Albashrawi, Adil S. Al-Busaidi, Janarthanan Balakrishnan, Yves Barlette, Sriparna Basu, Indranil Bose, Laurence Brooks, Dimitrios Buhalis, Lemuria Carter, Soumyadeb Chowdhury, Tom Crick, Scott W. Cunningham, Gareth H. Davies, Robert M. Davison, Rahul Dé, Denis Dennehy, Yanqing Duan, Rameshwar Dubey, Rohita Dwivedi, John S. Edwards, Carlos Flavián, Robin Gauld, Varun Grover, Mei-Chih Hu, Marijn Janssen, Paul Jones, Iris Junglas, Sangeeta Khorana, Sascha Kraus, Kai R. Larsen, Paul Latreille, Sven Laumer, F. Tegwen Malik, Abbas Mardani, Marcello Mariani, Sunil Mithas, Emmanuel Mogaji, Jeretta Horn Nord, Siobhan O'Connor, Fevzi Okumus, Margherita Pagani, Neeraj Pandey, Savvas Papagiannidis, Ilias O. Pappas, Nishith Pathak, Jan Pries-Heje, Ramakrishnan Raman, Nripendra P. Rana, Sven-Volker Rehm, Samuel Ribeiro-Navarrete, Alexander Richter, Frantz Rowe, Suprateek Sarker, Bernd Carsten Stahl, Manoj Kumar Tiwari, Wil van der Aalst, Viswanath Venkatesh, Giampaolo Viglia, Michael Wade, Paul Walton, Jochen Wirtz, and Ryan Wright. 2023. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* 71 (2023), 102642. doi:10.1016/j.ijinfomgt.2023.102642

[23] Tina Ekhtiar, Armağan Karahanoğlu, Rúben Gouveia, and Geke Ludden. 2023. Goals for Goal Setting: A Scoping Review on Personal Informatics. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 2625–2641. doi:10.1145/3563657.3596087

[24] Femke Beute Elisabeth T. Kersten-van Dijk, Joyce H.D.M. Westerink and Wijnand A. IJsselsteijn. 2017. Personal Informatics, Self-Insight, and Behavior Change: A Critical Review of Current Literature. *Human–Computer Interaction* 32, 5-6 (2017), 268–296. doi:10.1080/07370024.2016.1276456

[25] Zachary Englhardt, Chengqian Ma, Margaret E. Morris, Chun-Cheng Chang, Xuhai "Orson" Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. 2024. From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 56 (May 2024), 25 pages. doi:10.1145/3659604

[26] Daniel A. Epstein, Clara Caldeira, Mayara Costa Figueiredo, Xi Lu, Lucas M. Silva, Lucretia Williams, Jong Ho Lee, Qingyang Li, Simran Ahuja, Qiuer Chen, Payam Dowlatyari, Craig Hilby, Sazeda Sultana, Elizabeth V. Eikey, and Yunan Chen. 2020. Mapping and Taking Stock of the Personal Informatics Literature. *PACM Interactive Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 4, 4, Article 126 (Dec. 2020), 38 pages. doi:10.1145/3432231

[27] Daniel A. Epstein, Monica Caraway, Chuck Johnston, An Ping, James Fogarty, and Sean A. Munson. 2016. Beyond Abandonment to Next Steps: Understanding and Designing for Life after Personal Informatics Tool Use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1109–1113. doi:10.1145/2858036.2858045

[28] Daniel A. Epstein, Jennifer H. Kang, Laura R. Pina, James Fogarty, and Sean A. Munson. 2016. Reconsidering the device in the drawer: lapses as a design opportunity in personal informatics. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 829–840. doi:10.1145/2971648.2971656

[29] Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. 2015. A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 731–742. doi:10.1145/2750858.2804250

[30] Cathy Mengying Fang, Valdemar Danry, Nathan Whitmore, Andria Bao, Andrew Hutchison, Cayden Pierce, and Pattie Maes. 2024. PhysioLLM: Supporting Personalized Health Insights with Wearables and Large Language Models. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. 1–8. doi:10.1109/BHI62660.2024.10913781

[31] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods* 5, 1 (2006), 80–92. doi:10.1177/160940690600500107

[32] Warren J. Ferguson and Lucy M. Candib. 2002. Culture, language, and the doctor-patient relationship. *Family Medicine* 34, 5 (May 2002), 353–361.

[33] Luis Fernandez-Luque, Randi Karlsen, and Jason Bonander. 2011. Review of extracting information from the Social Web for health personalization. *Journal of medical Internet research* 13, 1 (2011), e15. doi:10.2196/jmir.1432

[34] Markus A. Feufel and S. Frederica Stahl. 2012. What do Web-Use Skill Differences Imply for Online Health Information Searches? *Journal of Medical Internet Research* 14, 3 (June 2012), e2051. doi:10.2196/jmir.2051

[35] Anne Galletta and William E Cross. 2013. *Mastering the Semi-Structured Interview and Beyond: From Research Design to Analysis and Publication.* NYU Press. http://www.jstor.org/stable/j.ctt9qgh5x

[36] Asees Kaur Gulati, Rachel Edna Lobo, N Nihala, Vishweshwara Bhat, Neha Bora, K Vaishali, and Mukesh Kumar Sinha. 2024. Young Adults Journey with Digital Fitness Tools-A Qualitative Study on Use of Fitness Tracking Device. *F1000Research* 13 (2024), 1296. doi:10.12688/f1000research.158037.1

[37] Rebecca Gulotta, Jodi Forlizzi, Rayoung Yang, and Mark Wah Newman. 2016. Fostering Engagement with Personal Informatics Systems. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16)*. Association for Computing Machinery, New York, NY, USA, 286–300. doi:10.1145/2901790.2901803

[38] Sarah Homewood. 2023. Self-Tracking to Do Less: An Autoethnography of Long COVID That Informs the Design of Pacing Technologies. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 656, 14 pages. doi:10.1145/3544548.3581505

[39] Sarah Homewood and Anna Vallgårda. 2020. Putting Phenomenological Theories to Work in the Design of Self-Tracking Technologies. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 1833–1846. doi:10.1145/3357236.3395550

[40] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 17–24. doi:10.1145/642611.642616

[41] Rachel L. Johnson, Debra Roter, Neil R. Powe, and Lisa A. Cooper. 2004. Patient race/ethnicity and quality of patient-physician communication during medical visits. *American Journal of Public Health* 94, 12 (Dec. 2004), 2084–2090. doi:10.2105/ajph.94.12.2084

[42] Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James Landay. 2024. Supporting Physical Activity Behavior Change with LLM-Based Conversational Agents. doi:10.48550/arXiv.2405.06061

[43] Da-jung Kim, Yeoreum Lee, Saeyoung Rho, and Youn-kyung Lim. 2016. Design Opportunities in Three Stages of Relationship Development between Users and Self-Tracking Devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 699–703. doi:10.1145/2858036.2858148

[44] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-Woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3613904.3642937

[45] Young-Ho Kim, Jae Ho Jeon, Bongshin Lee, Eun Kyoung Choe, and Jinwook Seo. 2017. OmniTrack: A Flexible Self-Tracking Approach Leveraging Semi-Automated Tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 67 (sep 2017), 28 pages. doi:10.1145/3130930

[46] Susanne Kirchner, Jessica Schroeder, James Fogarty, and Sean A. Munson. 2021. "They don't always think about that": Translational Needs in the Design of Personal Health Informatics Applications. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM. doi:10.1145/3411764.3445587

[47] Predrag Klasnja and Wanda Pratt. 2012. Healthcare in the pocket: Mapping the space of mobile-phone health interventions. *Journal of Biomedical Informatics* 45, 1 (2012), 184–198. doi:10.1016/j.jbi.2011.08.017

[48] Oura Labs. [n. d.]. Oura. https://support.ouraring.com/hc/en-us/articles/26055991315859-Oura-Labs

[49] Jong Ho Lee, Jessica Schroeder, and Daniel A. Epstein. 2022. Understanding and Supporting Self-Tracking App Selection. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 166 (dec 2022), 25 pages. doi:10.1145/3494980

[50] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 557–566. doi:10.1145/1753326.1753409

[51] Ian Li, Anind K. Dey, and Jodi Forlizzi. 2011. Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th International Conference on Ubiquitous Computing* (Beijing, China) *(UbiComp '11)*. Association for Computing Machinery, New York, NY, USA, 405–414. doi:10.1145/2030112.2030166

[52] Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning. arXiv:2406.00922 [cs.CL]

[53] Lena Mamykina, Daniel A. Epstein, Predrag Klasnja, Donna Sprujt-Metz, Jochen Meyer, Mary Czerwinski, Tim Althoff, Eun Kyoung Choe, Munmun De Choudhury, and Brian Lim. 2022. Grand Challenges for Personal Informatics and AI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 76, 6 pages. doi:10.1145/3491101.3503718

[54] Lena Mamykina, Elizabeth M. Heitkemper, Arlene M. Smaldone, Rita Kukafka, Heather J. Cole-Lewis, Patricia G. Davidson, Elizabeth D. Mynatt, Andrea Cassells, Jonathan N. Tobin, and George Hripcsak. 2017. Personal discovery in diabetes self-management: Discovering cause and effect using self-monitoring data. *Journal of Biomedical Informatics* 76 (2017), 1–8. doi:10.1016/j.jbi.2017.09.013

[55] Mike A. Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, Kumar Ayush, Hao-Wei Su, Qian He, Cory Y. McLean, Mark Malhotra, Shwetak Patel, Jiening Zhan, Tim Althoff, Daniel McDuff, and Xin Liu. 2024. Transforming Wearable Data into Health Insights using Large Language Model Agents. doi:10.48550/arXiv.2406.06464

[56] Bertalan Meskó and Eric J. Topol. 2023. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine* 6, 1 (July 2023), 1–6. doi:10.1038/s41746-023-00873-0

[57] Neoteryx Microsampling. 2020. younger generations want patient-centric, preventative healthcare at home. https://www.neoteryx.com/microsampling-blog/younger-generations-want-patient-centric-preventative-healthcare-at-home#:~:text=The%20younger%20generations%2C%20Y%20and,to%20correct%20any%20health%20deficits.

[58] Sean A Munson, Jessica Schroeder, Ravi Karkar, Julie A Kientz, Chia-Fang Chung, and James Fogarty. 2020. The importance of starting with goals in N-of-1 studies. *Frontiers in digital health* 2 (2020), 3. doi:10.3389/fdgth.2020.00003

[59] Subigya Nepal, Arvind Pillai, William Campbell, Talie Massachi, Michael V. Heinz, Ashmita Kunwar, Eunsol Soul Choi, Xuhai Xu, Joanna Kuc, Jeremy F. Huckins, Jason Holden, Sarah M. Preum, Colin Depp, Nicholas Jacobson, Mary P. Czerwinski, Eric Granholm, and Andrew T. Campbell. 2024. MindScape Study: Integrating LLM and Behavioral Sensing for Personalized AI-Driven Journaling Experiences. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4 (Nov. 2024), 186:1–186:44. doi:10.1145/3699761

[60] Sameer Neupane, Poorvesh Dongre, Denis Gracanin, and Santosh Kumar. 2025. Wearable Meets LLM for Stress Management: A Duoethnographic Study Integrating Wearable-Triggered Stressors and LLM Chatbots for Personalized Interventions. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 588, 8 pages. doi:10.1145/3706599.3720197

[61] Jasmin Niess and Paweł W. Woźniak. 2018. Supporting Meaningful Personal Fitness: the Tracker Goal Evolution Model. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173745

[62] Kannan Nova. 2023. Generative AI in Healthcare: Advancements in Electronic Health Records, facilitating Medical Languages, and Personalized Patient Care. *Journal of Advanced Analytics in Healthcare Management* 7, 1 (April 2023), 115–131. https://research.tensorgate.org/index.php/JAAHM/article/view/43

[63] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[64] Aswati Panicker, Novia Nurain, Zaidat Ibrahim, Chun-Han (Ariel) Wang, Seung Wan Ha, Yuxing Wu, Kay Connelly, Katie A. Siek, and Chia-Fang Chung. 2024. Understanding fraudulence in online qualitative studies: From the researcher's perspective. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 824, 17 pages. doi:10.1145/3613904.3642732

[65] Jinkyung Park, Ramanathan Arunachalam, Vincent Silenzio, Vivek K Singh, et al. 2022. Fairness in mobile phone–based mental health assessment algorithms: Exploratory study. *JMIR formative research* 6, 6 (2022), e34366.

[66] Soobin Park, Hankyung Kim, and Youn-kyung Lim. 2025. Reimagining Personal Data: Unlocking the Potential of AI-Generated Images in Personal Data Meaning-Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 545, 25 pages. doi:10.1145/3706598.3713722

[67] Jaana Parviainen and Juho Rantala. 2022. Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care. *Medicine, Health Care and Philosophy* 25, 1 (March 2022), 61–71. doi:10.1007/s11019-021-10049-w

[68] OpenAI Platform. [n. d.]. Prompt engineering. https://platform.openai.com/docs/guides/prompt-engineering

[69] Stephen Purpura, Victoria Schwanda, Kaiton Williams, William Stubler, and Phoebe Sengers. 2011. Fit4life: the design of a persuasive technology promoting healthy behavior and ideal weight. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 423–432. doi:10.1145/1978942.1979003

[70] Amon Rapp and Federica Cena. 2016. Personal informatics for everyday life: How users without prior self-tracking experience engage with personal data. *International Journal of Human-Computer Studies* 94 (Oct. 2016), 1–17. doi:10.1016/j.ijhcs.2016.05.006

[71] Debra L. Roter, Judith A. Hall, and Yutaka Aoki. 2002. Physician gender effects in medical communication: a meta-analytic review. *JAMA* 288, 6 (Aug. 2002), 756–764. doi:10.1001/jama.288.6.756

[72] Jessica Schroeder, Ravi Karkar, James Fogarty, Julie A Kientz, Sean A Munson, and Matthew Kay. 2019. A Patient-Centered Proposal for Bayesian Analysis of Self-Experiments for Health. *Journal of healthcare informatics research* 3 (2019), 124–155. doi:10.1007/s41666-018-0033-x

[73] Jessica Schroeder, Ravi Karkar, Natalia Murinova, James Fogarty, and Sean A. Munson. 2020. Examining Opportunities for Goal-Directed Self-Tracking to Support Chronic Condition Management. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4 (Sept. 2020), 151:1–151:26. doi:10.1145/3369809

[74] Yasaman S. Sefidgar, Carla L. Castillo, Shaan Chopra, Liwei Jiang, Tae Jones, Anant Mittal, Hyeyoung Ryu, Jessica Schroeder, Allison Cole, Natalia Murinova, Sean A. Munson, and James Fogarty. 2024. MigraineTracker: Examining Patient Experiences with Goal-Directed Self-Tracking for a Chronic Health Condition. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3613904.3642075

[75] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. *arXiv preprint arXiv:2406.09264* (2024).
[76] Konstantin R. Strömel, Stanislas Henry, Tim Johansson, Jasmin Niess, and Paweł W. Woźniak. 2024. Narrating Fitness: Leveraging Large Language Models for Reflective Fitness Tracker Data Interpretation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3613904.3642032
[77] Yalin Sun, Yan Zhang, Jacek Gwizdka, and Ciaran B. Trace. 2019. Consumer Evaluation of the Quality of Online Health Information: Systematic Literature Review of Relevant Criteria and Indicators. *Journal of Medical Internet Research* 21, 5 (May 2019), e12522. doi:10.2196/12522
[78] Fox Susannah and Duggan Maeve. 2013. Tracking for Health. https://www.pewresearch.org/internet/2013/01/28/tracking-for-health/
[79] Tara Templin, Monika W. Perez, Sean Sylvia, Jeff Leek, and Nasa Sinnott-Armstrong. 2024. Addressing 6 challenges in generative AI for digital health: A scoping review. *PLOS Digital Health* 3, 5 (May 2024), e0000503. doi:10.1371/journal.pdig.0000503
[80] Maria Dolores C Tongco. 2007. Purposive sampling as a tool for informant selection. (2007).
[81] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. 2024. Towards Conversational Diagnostic AI. arXiv:2401.05654 [cs.AI] https://arxiv.org/abs/2401.05654
[82] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 87:1–87:35. doi:10.1145/3637364
[83] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] https://arxiv.org/abs/2201.11903
[84] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. doi:10.48550/arXiv.2112.04359 arXiv:2112.04359 [cs].
[85] Whoop. [n. d.]. Whoop. https://www.whoop.com/us/en/
[86] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629 [cs.CL] https://arxiv.org/abs/2210.03629
[87] Peng Zhang and Maged N. Kamel Boulos. 2023. Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges. *Future Internet* 15, 9 (2023). doi:10.3390/fi15090286
[88] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. https://arxiv.org/abs/2309.01219v2
[89] Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 156, 26 pages. doi:10.1145/3613904.3642385

A Screening Survey

Checking Eligibility

(1) Are you 18 years or older? [Yes -> Continue; No -> END SURVEY]
(2) Do you track data to better understand and answer questions about your health?

Questions about Health Tracking Data

(1) What kind of data do you track? (e.g., exercise, diet, sleep, chronic health conditions, pain level, etc?) What do you use to track data about your health? (multiple options can be selected) [Paper-based mediums (e.g., Diaries, journals, calendars); Mobile or web app; Wearables (e.g., Fitbit, Oura ring, Apple watch); Physical devices (e.g., BP monitoring machine, cough monitor device); Other _____]
(2) Have you had any recent questions you wish to answer using self-tracking? (e.g., Does caffeine consumption cause migraines? What kind of exercises trigger knee pain? Does lack of sleep correlate to raised blood pressure?). Please give a few examples of specific questions. _____

Contact Information

(1) Please provide your email address. We will use this to schedule a session with you. _____

Demographic Information

(1) What is your highest education? [No schooling completed; Nursery school to 8th grade; Some high school, no diploma; High school graduate, diploma or the equivalent (for example: GED); Some college credit, no degree; Trade/technical/vocational training; Associate degree; Bachelor's degree; Master's degree; Professional degree; Doctorate degree]
(2) Do you work in healthcare or/and technology? [Healthcare; Technology; Both; None]
(3) What is your age? _____
(4) What is your sex? [Male; Female; Other _____; Prefer not to say]
(5) What is your gender? (multiple options can be selected) [Woman; Man; Self-describe _____; Prefer not to say]
(6) What are your preferred pronouns? _____
(7) Which race or ethnicity best describes you? (Please choose only one) [American Indian or Alaskan Native; Asian / Pacific Islander; Black or African American; Hispanic; White / Caucasian; Multiple ethnicity/ Other (please specify) _____]
(8) What country do you currently live in? _____
(9) Have you used a Generative AI-based tool or chatbot (e.g., ChatGPT, Microsoft Co-Pilot) before? (You will not be disqualified if you answer No to this question) [Yes; No]

B Study Protocol

Hello. I am _____. I am conducting research to understand how people may use AI-based technology to answer questions about their health and draw insights from their health tracking data. Thank you for agreeing to participate in our study. I would like to reiterate that you may decline to answer any questions or withdraw from the study at any time.

I plan for the study session to last a maximum of 90 minutes. The session will be conducted in 3 parts [P1-P4] / 4 parts [P5-P19]:

(1) First, I will ask a series of general questions related to your health tracking practices and data.
(2) Second, you will interact with an AI-based technology, namely ChatGPT [P1-P4] / Copilot [P5-P19], to ask specific questions you have from your health tracking data and to reflect on the responses you receive.
(3) Third, I will also show you a case example created using real health tracking data to demonstrate different capabilities of these technologies and get your thoughts on that. [Only for P5-P19, not presented to P1-P4]
(4) At the end, I will ask questions about your experience with the overall study and of interacting with these different technologies for drawing insights from your health tracking data.

B.1 Initial Interview

I would like to start by giving a short introduction to our study and why we are doing it. Self-tracking is a popular approach for understanding and keeping tabs on one's health. With the coming of tools, like ChatGPT and Microsoft Copilot, there may be an opportunity to use them to provide personalized care and recommendations. So I am doing this study to better understand the potential of the underlying AI of such tools for supporting people in drawing more actionable and useful insights from their health tracking data. I am curious whether and how such tools can support people like you in making sense of and acting on health tracking data. When I say people like you, I mean those who already track their health data and want to answer specific questions based on it or use the data to make decisions about their health and lifestyle. So I am interested in seeing how you prompt

such tools using your tracked data and would also like to learn about your perceptions about the responses you receive. Any questions before we begin?

To start with, in the survey, you indicated that you track your health and want to understand it better/answer specific questions. Could you tell me a little more about what you track, how you track, and what questions you want to answer based on your tracking?

- Why do you want to answer this question?
- Have you been able to answer this question? If not, why not / what has been the challenge? If yes, could you tell me more about how you were able to answer the question? Have you tried to use your self-tracking data to answer this question? (Why/Why not? / If yes, describe how?)
- Have you taken assistance or guidance from someone (e.g., doctors) on how you can answer these questions?
- Anything else you want to share about your health tracking practice or data?

B.2 Interactive Session

Now, we will move on to the second phase of our session. In this, you will use a tool called ChatGPT [P1-P4] / Copilot [P5-P19] to ask the questions about your health that you just described to me.
First, I would like you to open and explore the interface. Feel free to type and ask it questions.
SHARE SCREEN AND GIVE SCREEN CONTROL
Now, I would like you to ask the health related questions you have here.

- For ChatGPT participants (P1-P4): Please refrain from trying to upload your health data or share too much private information. However, please do ask the specific questions you described earlier.
- For Copilot participants (P5-P19). Feel free to use the Chatbot version or the bigger/note-form questions + different modes. Unlike commercial softwares like OpenAI's ChatGPT, we are using a [organization name] licensed version of Microsoft Copilot. While Copilot has the same underlying software (GPT4) as ChatGPT, this version of Copilot has commercial data protection, which means it will not store any questions you ask or data you decide to enter or upload here and Microsoft will not use that data to improve its models. This makes it safer to upload and share sensitive information, including health data. So, please ask the specific questions you described earlier.

And as you go about it, think out loud about what is going on in your head.

[For each question: ask participants how they feel about the answer and prompt them to add more details/specifics (if needed) to get a more appropriate answer.]

- How do you feel about this answer? Positives? Is there anything missing in the answer? Can you add something more to your query/question to make it more specific or clear so as to get a more appropriate answer?
- How are you deciding what specifics/details to provide in your query? What kinds of data are you comfortable putting into this tool? Are there things/data you are not comfortable inputting? Why?
- What are your thoughts on how useful the answers/recommendations are? In what ways is it useful? In what ways is it not useful?
- Do you think this response is personalized to you? Why/ why not? Is there something (e.g., more data/details, specific prompts) you could add in your query to improve the response?
- Do you trust the answer/recommendations? Why/why not? If not, what is it about the response you do not trust? Is there something (e.g., more data/details, specific prompts) you could add in your query to improve the response?
- What will you do with the information? Do you feel you learned something new or useful? [If needed] Did the responses change your understanding of something about your health that you already knew about?

[If needed] Do you anticipate changing anything in: your lifestyle or health practices, what/how you track your health?
- If you think you can use this information/recommendations, how will you go about doing that? Do you plan to consult someone (e.g., medical providers) before taking any steps?

B.3 Case Scenario [Only presented to P5-P19]

Thank you for interacting with the interface and asking questions based on your health data. What I'd like to do next is walk you through a scenario we developed. It uses real data that a person tracked, but the series of prompts I'll show you are ones that we developed. Because different people I talk with try different things with their data in this interview, my goal is to make sure everybody has seen some of the capabilities that we want to be sure to talk about. So after I show this to you, we'll talk more about your thoughts on using these tools for health questions, based on your own data and this scenario.
PRESENTATION OF CASE SCENARIO (described in Appendix C and slides included in supplementary material)

- Do you have any initial thoughts about this scenario?
- What do you think about how John prompts the interface, including how he directly shares/upload health data to the interface and forms questions?
- What are your thoughts on the responses/insights provided by the interface based on the uploaded health data? Usefulness / appropriateness? Personalization / specificity? Actionability? Trustworthy / accuracy / credibility / potential harms?
- Did you feel the interface made assumptions about John's data or health? If so, what are your thoughts about that?
- Are there any other capabilities of this interface demonstrated by this case example that you thought are useful?
- Are there any other capabilities which might not be represented in this case example but you feel would be good to have/useful for supporting reflection and sense-making of health data?

B.4 End-of-Study Interview

I would love to know your thoughts about our study and your experience in this session. Any general thoughts on the interface you saw and interacted with today, as well as the case examples created using real health data?

- Thoughts on usefulness? Appropriateness of responses?
- What are your thoughts on how personalized the responses were to your needs and questions? Are there any potential improvements to the interface which could have supported you in receiving more personalized responses/insights from your health data? Is there something you would have wanted GAI to do differently?
- How actionable did you find the advice for your context and preferences? What are the next steps for or challenges in taking action on recommendations/responses you received? Do you anticipate this info/reponses changing the way you assess the things you already do? Do you anticipate changing anything in your health tracking or lifestyle based on the responses you received?
- Thoughts on trustworthiness/accuracy/credibility of responses/recommendations? [For P5-P19] While the version of the software we used today keeps your data private, the free versions of Copilot and ChatGPT may store your queries, including any data you upload, and use it to improve their models. If, say, your data was being stored and used in such a way, what would you have done differently in that case? Would that change what data you enter or types of questions you ask?
- While in this study we focused on supporting reflection and sense-making of already collected health data, do you feel you would appreciate support of such AI-based tools in other parts of the process such as helping decide what data to track or change the type of health data to track?

Thank you and before we conclude, I do want to tell you a little more about ChatGPT / Copilot and certain limitations of its underlying software or AI . Large Language Models or LLMs, which are the underlying magic here, are undoubtedly impressive for their ability to generate convincing content about a lot of questions we ask. However, they do contain some serious flaws. So for example, LLMs can get answers wrong and just 'hallucinate' incorrect facts. After all, they use information that is widely available on the internet. They can also give biased responses and are often gullible. For example, if you ask Copilot or ChatGPT a lot of leading questions, you might be able to coax it into giving an answer that you want, which could also be potentially harmful content. This is to say that one should not just do whatever this technology says you should. So do you have any reaction or thoughts about this? Does this information change your perceptions of the responses you received? Anything you would do differently in inputting your data or interpreting the responses?

C Case Scenario Description

To more fully characterize our method, we describe steps in the presented case scenario, including our rationale, the prompts used, and Copilot responses. The complete storyboard slide deck used to present the case scenario is also provided in supplementary materials.

- **Introduction (slides 1, 2, 3)**: The scenario introduces John, a person who experiences chronic shoulder pain and has been tracking his health data for about 2 months, trying to identify (i) correlations between pain level and potential pain triggers, and (ii) what helps with his pain. He uses a spreadsheet (shown in slide 2) to track his pain level on a scale of 0-10, activities he thinks help with the pain, and activities that might be worsening it.
- **Asking for summarization and high level analysis of uploaded data (slides 4, 5)**: John uses Copilot to make sense of his health data. He starts by uploading a screenshot of his datasheet, prompting the tool to analyze his data and summarize what activities are most likely to trigger his shoulder pain.
- **Seeking more specific insights and implementable recommendations (slides 6, 7)**: After receiving some high level insights, John modifies his queries to ask more specific questions. These include activities that trigger the highest level of pain, recommendations for pain relief once it is triggered, and whether relief methods differ based on different triggers of shoulder pain. He specifies in his queries that he is looking for recommendations that he can realistically implement.
- **Attempting to identify correlations between triggers and symptoms (slides 8, 9)**: John is further interested in drawing correlations between symptoms and triggers (e.g., between his stress levels and intensity of shoulder pain), and queries Copilot to do this with his uploaded data. Copilot recommends doing a systematic analysis of his data to determine correlations, and it mirrors John's curiosity about if stress could be related to his pain, but stress is not something John had tracked (except for a single note about potential association of stress to pain).
- **Seeking recommendations on making tracking changes (slides 10, 11, 12)**: Although John is curious about potential correlations between stress and pain intensity, he is unsure about what he should track to identify this correlation. John prompts Copilot to learn about how he might change his tracking in order to draw these insights in the future. After being recommended tracking of various factors (listed in slide 11), John further probes Copilot to create a worksheet for him to track and identify relationships between some of these factors (i.e., stress levels, stress triggers, exact location of shoulder pain). Co-Pilot creates a worksheet, also providing descriptions for how John should track the factors (e.g., scales to use). Note that slide 12 (i.e., creation of the worksheet) was not shown to P5 or P9 (i.e., the first two participants to complete the revised protocol).
- **Conclusion (slide 13)**: Based on responses from Copilot, John decides to change two things. First, he decides to try changing his activities, specifically his workouts, such as by consulting a physical therapist

and trying to find alternatives to ab and upper body workouts which he notes as most likely to cause shoulder pain. Second, John decides to also start tracking data related to his stress levels, stress triggers, and location of pain in hopes of identifying correlations between stress and shoulder pain in the future.