Supporting the Work of Patients and Providers in Complex Chronic Illness

Adrienne Pichon

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2025

Abstract

Supporting the Work of Patients and Providers in Complex Chronic Illness

Adrienne Pichon

Chronic illnesses, particularly poorly understood and complex conditions like endometriosis, present significant challenges for patients and healthcare providers. Endometriosis is a systemic, multifactorial condition affecting approximately 6–10% of women of reproductive age. It is characterized by highly variable and unpredictable symptoms, a lack of biomarkers, no cure, and individualized responses to treatment. This requires significant patient-provider collaboration and ongoing self-management. Despite advances in artificial intelligence (AI) and personal informatics systems for managing chronic illness, limited support exists for complex conditions like endometriosis, where significant uncertainty and variation impede care and management.

This dissertation seeks to understand the needs of individuals faced with the complex chronic illness that is endometriosis and address those needs, particularly through leveraging patient-generated data and personal informatics tools to support care. Throughout this work, we take a human-centered AI (HAI) approach to promote the perspectives of individuals and align their needs and priorities with the capabilities of the technologies we use. In this research, we first document the work of patients and providers in caring for such a complex chronic illness, and elicit the data and technology needs of patients and providers (Aim 1). There, qualitative research methods, including focus groups and interviews, reveal that patients and providers face barriers in synthesizing complex health data, aligning perspectives, and navigating individualized

management pathways. Next, we develop and evaluate interpretable temporal phenotypes of health status (Aim 2). There, we use a probabilistic modeling approach to generate interpretable, temporal phenotypes of health status from self-tracked data, then validate these health status representations through user feedback and a real-world computational task. Finally, we identify human and technical specifications for an intelligent system that provides adaptive self-management recommendations using reinforcement learning (RL) (Aim 3). There, we propose and implement a novel HAI framework, which facilitates conducting a mixed-methods study where we map and align human needs and values with technical capabilities and requirements. This research can inform the development of adaptive, explainable, and personalized self-management recommendations. Findings from this dissertation demonstrate the potential of computational approaches and novel intelligent systems to empower individuals with endometriosis by augmenting their understanding and use of their health-related data and self-management efforts with data-driven insights and AI-enabled intelligent systems.

This dissertation has several important contributions. *First*, this work both leverages and advances human-centered AI. The introduction of the Multi-Perspective Directed Analysis (MPDA) framework provides an approach to bridge human and technical needs in the design of AI-enabled systems. By aligning insights from end-users with the specifications of data science, MPDA operationalizes an HAI approach to design, offering a reproducible approach for other researchers seeking to address similar interdisciplinary challenges. This framework highlights the potential of HAI in translating patient needs into actionable computational design requirements and provides a blueprint for tackling open questions in health and other domains through human-centered AI. We also elaborate on several HAI principles, for example, how to empower patients through control of intelligent systems. *Second*, this dissertation contributes to advancing personal informatics technologies. We have documented a range of technology gaps and opportunities to innovate solutions to address these gaps. The development of interpretable, temporal representations of health status and the requirements gathering for an AI-enabled personal informatics tool for individualized recommendations are both novel contributions. These innovations expand the literature on chronic illness support, particularly by demonstrating the potential of AI in addressing a particular real-world, complex health scenario. We also highlight and articulate several sociotechnical gaps where technologies cannot meet the complex needs of users at this time. Despite the barriers in translating these technologies into practical systems, this work explores how AI can enhance chronic disease management through pragmatic, user-centered solutions. *Finally*, we advance illness-specific endometriosis research. This dissertation addresses an illness context with significant gaps in research and technologies to support care and management. By identifying the complex work undertaken by patients and providers in care and the associated needs of patients and their care teams, this research provides a foundation for designing tools and systems tailored to this context. While we propose HAI solutions for some of these gaps, we also highlight additional opportunities for computational and interactive systems that could better support individuals and care teams managing endometriosis.

As an ultimate goal, this thesis seeks to understand and address the needs of individuals caring for endometriosis, a complex and poorly understood chronic illness. In particular, we aim to develop HAI technologies to support patients and their care teams. Through the alignment of human needs and values with technological constraints, this research facilitates patient-centered care, empowers individuals with their data, and contributes to the broader fields of biomedical informatics, human-computer interactions, and human-centered AI.

Table of Contents

List of Figures

List of Tables

Acknowledgments

There are many people to whom I owe my deepest gratitude. While the words that follow offer only a brief acknowledgment, they can never truly capture the depth of my appreciation.

First, I would like to thank my advisor, Dr. Noémie Elhadad, who has shepherded me on this incredibly challenging yet rewarding journey. Your unwavering guidance, wisdom, and encouragement have been instrumental in shaping not only this dissertation but also my growth as a researcher and scholar. I am profoundly grateful for your mentorship, patience, and support.

Next, I would like to extend my heartfelt thanks to my dissertation committee. Dr. Bakken, you have been a mentor of my research, and champion of my leadership and advocacy for justice-related topics in informatics. Dr. Mamykina, you gave language to my passion for working with patients, you opened the door to the field of human-computer interactions for me, and that has deeply impacted my career trajectory. Your thoughtful insights have supported my academic and personal growth. To all of my advisors, you inspired and motivated me.

To the participants of all endometriosis research, and specifically those who contribute to Phendo and Citizen Endo studies: I am deeply grateful for your engagement and generosity in sharing your experiences and data. Your contributions are invaluable, and they have been the foundation for this work.

To Sharon, Emma, and Iñigo: Thank you all so much for your efforts in supporting my research. I truly could not have produced this work without you. Collaborating with you has been both a

privilege and a joy.

To my wonderful peers, in particular Harry, Pooja, Oliver, Apara, Tara, and the Justice Informatics Collaborative: Thank you for being alongside me throughout this journey. Your personal and professional support has been essential, allowing me to thrive as a whole person in this challenging environment. I cannot articulate how meaningful your friendship is to me.

To Jess: My most dedicated student, you have taught me how to be a mentor. Working together has been a wonderful experience, and I have learned so much from you. I wish you the greatest future and look forward to seeing all that you will accomplish.

To my mentors and colleagues across the years: Somjen, you were my first mentor, and you have become my family. Melissa, your friendship means the world, I am grateful to have you to look up to. David Norman, I am grateful for the opportunity to develop alongside you as a scholar, your early friendship shaped my worldview. Dr. Schnall, you gave me a crash-course introduction to informatics and a decade of experience in just a few short years — I am so glad to remain connected to you.

To my family — my wife, mom, dad, siblings, and grandparents — I love you all so very much. I hope I have made you proud. Your love, encouragement, and support have been the foundation upon which I've built my life and work. I would not be the person or scholar I am today without you all. To my baby and pup, your joy has given me the fuel to power through hard days.

Finally, I want to acknowledge the land where I live, work, and conduct my research, which is the ancestral home and unceded territory of the Lenape people. I also want to recognize the Mohawk people, whose land rejuvenates and inspires me whenever I spend time in the outdoors. I give thanks for the connection and grounding that this land provides.

This dissertation is the result of the collective support, inspiration, and generosity of many incredible individuals. To each of you, I offer my deepest gratitude.

Dedication

To my family — Breandra, Wavy, and Rio — you are my whole entire world and the reason my work matters.

Chapter 1: Introduction

1.1 The need for supporting the work of patients and providers in complex chronic illness

Chronic conditions burden those who must live with and manage them [1, 2, 3, 4, 5]. Different chronic diseases have a range of short- and long-term impacts on individuals, and many vary significantly even among individuals with the same disease [6]. Further, gaps in medical knowledge persist, complicating diagnosis, management, and treatment [7]. Chronic illness leads to a significant amount of work for patients and their care teams to undertake [8, 9]. Patient-centered care — where patients actively participate in their own care alongside their providers and are considered essential members of the care team — is important for providing care for individuals with chronic illness [10].

To mitigate these burdens and facilitate patient-centered care, researchers have developed various personal informatics tools — technologies designed to collect, integrate, and analyze personal data to support reflection and decision-making — to support care for different diseases, support treatment, help reach patient goals, and improve quality of life [11, 12, 13]. With the proliferation of such personal informatics tools across health contexts and purposes, individuals managing chronic conditions often generate a considerable volume of personal health data about their illness experience that is available to support care [14]. With the rise of artificial intelligence (AI) — here, broadly considered as leveraging computational techniques to analyze data and provide actionable insights or perform useful tasks in support of human users — there are even more opportunities to support care of chronic disease [15].

There is a particular opportunity to apply AI in the context of poorly understood, complex, enigmatic conditions, especially where the experience of illness varies greatly from individual to individual. While there are some intelligent systems to capture illness data and reflect on health-related data in these contexts (e.g., tools for identifying triggers in enigmatic diseases, such as irritable bowel syndrome (IBS) [16, 17], rosacea [18], and migraine [19]), few go beyond logging and reflecting on data or provide computational features to use the data to support care. Further, barriers persist in developing AI-enabled technologies that can be implemented in real-world settings, with particular challenges to privacy, trust, and autonomy for technologies that support patients as end-users [20, 21].

The research in this thesis focuses on endometriosis, an inflammatory chronic, multi-factorial, and systemic condition estimated to affect 6-10% of women[1] of reproductive age [23]. In this burdensome chronic condition, the types and severity of symptoms vary greatly from individual to individual and over time. Despite recent research interest, endometriosis remains enigmatic [24]. There are substantial gaps in knowledge about the disease, leading to a lack of established medical guidelines [25]. Treatment and management options are not reliably effective, and the benefits vary between patients [26, 27]. Because endometriosis is poorly understood, has no biomarker for diagnosis, has no cure, and treatment is complex, individualized, and often ineffective, monitoring and care for the condition remains extremely challenging [28, 29].

In complex and poorly understood diseases such as endometriosis, it can be very difficult to characterize individuals' health status, understand what is going on with their health, and communicate about the experience of illness with a care team [30, 31, 32]. Further, without reliable treatments, individuals often rely on their own expertise and independent self-management to mitigate their symptoms [33]. Self-experimentation with self-management strategies — where people engage in a trial-and-error approach to find strategies that are effective for mitigating their symptoms — is a common approach for people with chronic conditions to develop effective personalized regimens [34, 35]. Tools have been developed to support self-management tasks [36, 37, 38, 39, 40, 41, 42, 43], facilitate self-experiments [44, 45, 46, 47, 19, 48], and provide just-in-time recom-

[1] In this dissertation, we reference endometriosis as a condition that impacts "women." While imperfect, the use of this term is important because "women's health" is under-studied and often stigmatized precisely *because* they are women's concerns, and stripping this label could obscure this problem (as also argued by Grimme et al. [22]). At the same time, we recognize that intersex people, non-binary individuals, and transgender men, for example, may also have endometriosis. In the same way that not all women menstruate and not all menstruators are women, we also acknowledge that not all people who are impacted by "women's health" issues are women either.

mendations [49] for individuals managing their illness. Still, at a basic level, people with complex, poorly understood conditions, such as endometriosis, lack support in using their tracked data to identify trends and changes in their health status, and further lack tools for computational support with their data [11]. Individuals with endometriosis often lack technological support in care tasks that may benefit from intelligent systems that could support making sense of their health status and facilitate self-experiments with self-management. This gap offers an area for research to address and improve monitoring and care delivery for individuals with varying and uncertain illness experiences.

1.2 Thesis approach

In this thesis, we rely on principles from the nascent and rapidly developing body of literature on human-centered AI (HAI) to guide our research questions and approach [50, 51, 52, 53, 54, 55, 56, 57, 58, 59]. Here, we consider HAI as the design and development of AI systems that prioritize user empowerment, ethical principles, and contextual adaptation to enhance human agency and well-being. This approach aims to enable reliable, transparent, and trustworthy technologies while supporting dynamic, personalized interactions that align with users' values and needs.

The research in this thesis relies on human perspectives to guide the design and development of technologies, grounding the process in the needs and experiences of end-users who are patients living with the challenges of such a burdensome chronic condition. To design computational mechanisms and intelligent systems that are practical and feasible and also align with HAI principles (e.g., autonomy, privacy, and trust), it is critical to harmonize human needs and values with the constraints and requirements of data and technology [60]. The work in this thesis leverages the input and self-tracked data of end-users to prioritize their experiences, values, and priorities while simultaneously addressing the specific demands of the computational mechanisms used. By doing so, the resulting technologies capitalize on human strengths (such as contextual understanding and intuition) while benefiting from the computational capabilities of AI [61]. Through this alignment of human needs and values with technological constraints, this research facilitates patient-centered care, empowers individuals with their data, and contributes to the broader field of human-centered AI.

As an ultimate goal, this thesis seeks to understand and address the needs of individuals caring for endometriosis, a complex and poorly understood chronic illness. With the lack of knowledge impeding the delivery of care for enigmatic conditions like endometriosis, patient-centered care is especially difficult and even more critical [28, 29]. Thus, we center the perspectives of patients, while also consulting endometriosis specialists. We rely on qualitative methods to document these needs and identify opportunities for technology to provide solutions. To address challenges in making sense of and characterizing health status, we construct meaningful and interpretable[2] representations of the illness experience (i.e., digital phenotypes — digital representations of health status derived from self-tracked illness data), which can enable the patient's experience of their health status to be captured and reconciled in a machine-readable format. These representations can support individuals in their own care and management and could potentially be used by personal informatics tools to tailor interventions or perform other computational analyses. To address challenges in self-management, we explore the potential for an intelligent system that could provide automated, individualized support for self-management. To do so, we propose and implement a novel HAI framework for mapping the possible design space in this context. We imagine that the interpretable, meaningful representation of health status over time might provide a machine-readable foundation for such an intelligent system.

The first studies in this thesis elicit the needs and priorities of users — **Aim 1** applies qualitative research methods (focus groups and interviews) to understand what patients and providers need when caring for a complex, enigmatic chronic condition. **Aim 2** focuses on using phenotyping methods to develop temporal phenotypes of health status to construct meaningful, interpretable representations of health experiences. Interpretable approaches are the focus (so that they can be useful to human users), and the phenotypes are evaluated by end-users and in a real-

[2] Throughout this thesis, we refer to both interpretability and explainability, which are related but distinct concepts. Interpretability means that a model is inherently understandable to users, where a person can understand how an input leads to an output. Explainability is related to describing or justifying a decision-making process, even if a model is not inherently interpretable.

world computational task. **Aim 3** identifies human and technical specifications for an intelligent system that provides individualized, adaptive self-management recommendations. In this aim, a novel human-centered AI framework enables triangulation of qualitative findings with quantitative findings for mapping these requirements for the individualized self-management recommendations within a specific machine learning (ML) approach — Reinforcement Learning (RL). RL is unique in its ability to make sequential recommendations that adapt to changes in a complex environment, which makes it an ideal candidate to power an intelligent system for individualized self-management. By following this framework, we conduct a mixed-methods study to map the human, data, and ML requirements and constraints for such an intelligent system, organized around the principles of RL, then use these to develop recommendations for design. Our HAI framework allows us to understand what goals and features are important to users, the design space for meeting these goals, and how the machine-readable phenotypes of health status (from Aim 2) might be used to facilitate individualized recommendations in the proposed intelligent system. In these studies, data from the Phendo app (described in section 2.2.1) are used. The Columbia University institutional review board (IRB) has reviewed and approved all study procedures for these research activities under protocols #AAAQ9812 and #AAAR8513.

1.2.1 Aim 1: Elicit patient and provider needs

Objective: Elicit the needs in caring for the complex, enigmatic chronic condition endometriosis and requirements for technology to support these needs.

Sub-Aim 1.1. Characterize the work of patients and providers in the management of an enigmatic condition.

Research Questions:

$RQ_{1.1}$: In the work of patients and providers when caring for endometriosis, what aspects of their collaborative work pertain specifically to such a complex condition?

$RQ_{1.2}$: What role does technology play in facilitating the partnership between endometriosis patients and providers and their collaborative work, and what opportunities are envisioned?

Sub-Aim 1.2. Characterize the work of patients on their own in the self-management of an enigmatic condition.

Research Questions:

$RQ_{1.3}$: In the work of endometriosis patients, what aspects of their independent work pertain specifically to such a complex condition?

$RQ_{1.4}$: What role does technology play in facilitating the success of endometriosis care and management for patients on their own, and what opportunities are envisioned?

Methods and Materials for Aim 1: To elicit the needs of patients and providers in the care of endometriosis — a complex enigmatic chronic condition — and opportunities to support these needs, the research for this aim relies on qualitative methods to engage both patients and providers to identify and characterize current experiences and practices both within and outside of the clinical context, unmet needs in care and self-management, and gaps where technology could provide support. We conduct focus groups with endometriosis patients and interviews with providers, then use thematic analysis to understand the work of patients, both on their own and in partnership with providers, and the information and design needs for intelligent systems to address these needs.

Primary Findings for Aim 1: This study revealed key needs of patients and providers in managing endometriosis, underscoring the complexities of data use and patient-provider collaboration, as well as the individualized challenges of self-management. Patients and providers face significant barriers in gathering and reflecting on comprehensive health data to support both pre-visit preparation and in-clinic discussions. Challenges include synthesizing fragmented data across symptoms and domains, navigating various temporal resolutions of data, and integrating personal insights to create a holistic view of health status. Additionally, both patients and providers experience challenges in facilitating effective partnerships due to misalignments in understanding, stigmatization, and the absence of standardized knowledge. Patients often take on the role of self-advocates, tailoring data presentations for different providers and working to convey a "story" that captures the nuances of their experience while still fitting into the clinical workflow.

Meanwhile, self-management is particularly demanding, with patients resorting to unsupported trial-and-error approaches to find strategies that alleviate symptoms, due to limited clinical guidelines and the disease's individualized nature. These findings highlight a pressing need for tools and support systems that can bridge data gaps, help make sense of health status, align patient-provider perspectives, and empower patients in crafting personalized management regimens.

1.2.2 Aim 2: Interpretable, temporal health status phenotypes

Objective: Develop and validate the phenotyping of users to represent temporal, individualized illness states that are interpretable and meaningful to users and suitable for use in intelligent systems.

Sub-Aim 2.1. Develop and validate the phenotyping of users to represent temporal, individualized illness states.

Research Questions:

$RQ_{2.1}$: Can a digital phenotyping approach aggregate individual-level data to enable interpretable representations of health status in the context of a complex enigmatic condition?

Sub-Aim 2.2. Evaluate the temporal phenotypes with human end-users.

Research Questions:

$RQ_{2.2}$: Can digital phenotyping methods construct meaningful representations of health status for individuals?

$RQ_{2.3}$: Can digital phenotyping methods construct meaningful representations of health status for individuals over time?

Sub-Aim 2.3. Evaluate the temporal phenotypes with a real-world computational task.

Research Questions:

> $RQ_{2.4}$: Can digital phenotyping methods construct temporal representations of health status that can be used in real-world tasks?

Methods and Materials for Aim 2: In Aim 1, we identified key needs and technology gaps in the work of caring for the complex, enigmatic chronic condition endometriosis. In this aim, we create computable representations of illness states to help individuals characterize their health status and that may be used in intelligent systems. This aim involves phenotyping user health states to develop representations of an illness state space that can be used an RL-based intelligent interactive system, which will be interpretable and meaningful to users.

To do so, we use a mixed-membership probabilistic model capable of accommodating self-tracked data across different domains, that has been previously validated for user-level phenotypes with self-tracking data. Here, we extend this model to generate phenotypes at the user-week level (i.e., the data from each user are aggregated by week). Phenotypes can be described as a mixture of characteristics across domains of illness experience. The set of learned phenotypes represent latent subgroups of health statuses experienced by individuals based on similar illness characteristics. The phenotypic profile (i.e., the mixture of phenotypes) provides a computable representation that characterizes the health status of an individual at a particular time. We also construct baseline phenotypes using rules based on the simple daily check-in field, “How was your day?” We use Phendo data for both learned and baseline phenotype models. We validate the learned model by comparing it to the baseline model in various ways. We also evaluate the learned phenotypes by seeking real-world feedback from Phendo users and through a real-world prediction task.

Primary Findings for Aim 2: We learned a model with four distinct phenotypes, representing severity-based health statuses. For validation, we compared this learned model with a simple baseline model, constructed from the frequently tracked “How was your day?” question in Phendo. We find the learned phenotype provides useful information about users’ health statuses across various domains of health. Compared to the baseline model, which is limited by missing data and over-reliance on a single-feature indicator, the learned model incorporates comprehensive self-tracked data, capturing diverse health experiences and minimizing missing assignments. Temporal analyses of phenotype transitions over time further highlight the learned model's ability to reflect individualized health patterns across an illness trajectory (the experience, impact, and work related to living with the illness). The learned model provides a more nuanced, holistic, and interpretable view of health status that accommodates complex, individualized experiences, establishing it as a more complete representation of user health dynamics.

We conducted a two-part evaluation with a computational task and a user study to assess the learned phenotype's performance and acceptability to users. In the computational task, the learned phenotypes performed better at the task of predicting flare-ups compared to the baseline phenotypes. When inspecting the key evaluation metrics, the baseline phenotypes performed poorly while the learned phenotypes performed better with forecasting flare-ups. This suggests that the learned temporal phenotypes are promising for use in intelligent systems that can meet the needs of individuals with complex chronic illness, especially when binarized. In the user study, both learned and baseline phenotypes had similar performance with matching between user and AI assessments, difficulty, and certainty. While there was only limited agreement between users and the AI-generated health status (phenotypes), there were some positive aspects that were uncovered in the user study. The learned phenotypes more closely aligned with the human users' process of determining health status — participants used much of the same data that the learned phenotypes relied on in making the health status assignments. Further, while neither the baseline nor the learned phenotypes neatly matched with the user's assessment, the baseline model overwhelmingly assigned a better health status to the data than users, while the learned model erred on the side of assigning a worse status than the users — which users prefer. This aligns with the results from the computational task, where the baseline model was unable to forecast bad health statuses and was outperformed by the learned model. Participants also reported that they felt there was value in the AI-generated health statuses, for example to create a personal baseline for them to work from or as a sort of "sounding board" when working through their own health assessments. They were also optimistic about bringing summaries to their providers.

1.2.3 Aim 3: Adaptive self-management recommendations

Objective: Identify and formulate human and technical specifications of adaptive self-management recommendations.

Research Question:

> $RQ_{3.1}$: What insights at the intersection of human needs and values, human self-tracking behaviors as evidenced by "in the wild" self-tracking data, and capabilities and constraints of RL, can inform the design of RL-based intelligent systems for self-management of endometriosis?

Methods and Materials for Aim 3: In Aim 1, we used qualitative methods to elicit the needs of users, both for patients and providers managing endometriosis together and for patients self-managing their illness independently. In Sub-Aim 1.2, a key problem was identified — developing individualized self-management regimens is an open problem that is significant and worth addressing. The frustrating and burdensome trial-and-error process could be augmented by adaptive recommendations, to tailor self-management strategies to individuals and their particular circumstances, within conditions of uncertainty. Consequently, Aim 3 focuses on identifying and operationalizing specifications of adaptive recommendations to support individuals in self-managing their endometriosis.

In this aim, we map the human and technological constraints and needs, which calls for an interdisciplinary effort between human-computer interactions (HCI) and data science. We propose and implement an HAI framework — Multi-Perspective Directed Analysis — to align human and technological requirements and constraints that can guide design of intelligent systems for self-management. We conduct a mixed-methods study — we use concepts from a particular ML technique, RL, to elicit user needs, through directed content analysis of user interviews, and uncover practical data constraints, through analysis of "in the wild" user engagement logs from the

Phendo app. We gather and triangulate human-machine-data requirements for a self-management tool for individuals with endometriosis and use these to develop recommendations for developing a system that aligns with needs, capabilities, and constraints from human user, data, and ML perspectives.

Primary Findings for Aim 3: This study leveraged a novel mixed-methods approach to map users' self-management needs to computational requirements for a proposed AI-enabled tool that uses RL. Findings confirm that users are interested in experimenting with self-management regimens using AI recommendations, emphasizing the importance of individualized user models. Considering the action space, users demonstrated a preference for a broad, flexible range of personalized self-management options. The state space requires holistic, contextual user data to provide recommendations aligned with daily routines and environmental influences, although capturing these real-world dynamics presents technical challenges. For the reward function, participants preferred short-term outcomes (e.g., daily pain reduction) as indicators of strategy success, aligning with RL's evaluative mechanisms. Finally, the agent/policy findings underscore the need for personalized, explainable recommendations, as users are heterogeneous in their responses to strategies and want insights into suggested actions. High engagement levels suggest that users are likely to interact with the tool frequently enough to meet RL's computational needs, supporting its feasibility and utility in real-world self-management scenarios.

1.3 Contributions

This dissertation makes several important contributions to research across the fields of human-centered AI, informatics, and endometriosis.

First, this work both leverages and advances human-centered AI. AI-enabled personal informatics tools have the potential to offer support in real-world, complex circumstances. But their benefits in supporting people with chronic illness has not yet been fully realized. Barriers exist in translating the technology that is being developed into pragmatic systems to enhance the delivery of care and the experience of self-managing chronic illness. Human-centered AI is emerging as a framework and research agenda that guides this research; while this work draws from the HAI literature and existing methods, it also contributes to HAI efforts.

The introduction of the Multi-Perspective Directed Analysis (MPDA) framework provides an approach to bridge human and technical needs in the design of AI-enabled systems. Thus, we contribute a framework and an example of applying this framework to identify and align both human needs and technical needs within the frame of HAI. By aligning insights from end-users with the specifications of data science, MPDA operationalizes the HAI agenda, offering a reproducible approach for other researchers seeking to address similar interdisciplinary challenges. This framework highlights the potential of HAI in translating patient needs into actionable computational design requirements and provides a blueprint for tackling open questions in health and other domains. We also offer insights to advance HAI through various principles, such as mechanisms to enhance autonomy and control over intelligent systems. We also call for innovation around representing embodied experiences of illness.

Second, this dissertation has contributions for the advancement of personal informatics technologies. We have documented a range of technology gaps and opportunities to innovate solutions to address these gaps, in the context of complex chronic illness. The development of interpretable, temporal representations of health status and the requirements gathering for an AI-enabled personal informatics tool for individualized recommendations are both novel contributions. These innovations expand the literature on chronic illness support, particularly by demonstrating the potential of AI in addressing a particular real-world, complex health scenario. We also highlight and articulate several sociotechnical gaps where technologies cannot meet the complex needs of users at this time. Despite the barriers in translating these technologies into practical systems, this work explores how AI can enhance chronic disease management through pragmatic, user-centered solutions.

Finally, we advance endometriosis research. Endometriosis is a burdensome condition that is under-funded in research and under-supported with interventions and technology [62, 63, 64]. The disease impacts a sizable subset of the population (about 10% of those who have ever menstruated [65]), placing a substantial burden on individuals with the disease and the healthcare system more broadly. It is poorly understood and very individualized, thus compounding the complexity of supporting patients and their care teams in this context [25]. These considerations have not been broadly addressed when designing personal informatics tools for supporting care of this chronic illness. Thus, this dissertation addresses an illness context with significant gaps in research and technologies to support care and management.

By identifying the complex work undertaken by patients and providers in care and the associated needs of patients and their care teams, this research provides a foundation for designing tools and systems tailored to this context. While we propose solutions for some of these gaps, we also highlight additional opportunities for computational and interactive systems that could better support individuals and care teams managing endometriosis. Thus, by articulating the needs of patients and their care teams in this particular illness context and opportunities for technologies to support these needs, we make an important contribution to the literature.

1.4 Guide for the reader

Chapter 2 presents background information from the literature to ground the thesis in existing work. **Chapter 3** addresses Aim 1 and reports on the qualitative work completed through engaging patients and providers in discussions about their health, management, and technological needs. This study sets up the rest of the thesis. **Chapter 4** addresses Aim 2 and describes the phenotyping experiments to construct and evaluate meaningful, temporal representations of health status. **Chapter 5** addresses Aim 3 and presents the human-centered AI work to map the human, data, and ML requirements of an interactive system for supporting self-management under conditions of uncertainty, and use these to develop design recommendations. **Chapter 6** discusses conclusions, contributions of this thesis, limitations of the research, and future work.

Chapter 2: Background

2.1 Care and management for chronic illness

Chronic illness burdens the healthcare system and the patients who suffer the symptoms of their illness, thus care for chronic conditions is a major health priority globally [1]. Persistent barriers to care further compound the stress of illness and often requires patients to engage in self-management outside of the clinic [2, 3, 4, 5]. Because the nature of chronic illness is persistent, and impacts the person every day, it necessitates a holistic approach to care, such as patient-centered care. **Patient-centered care** is a paradigm that moves beyond a biomedical-only frame of disease, which focuses primarily on biological factors of disease, to consider the patient holistically within the context of their illness [10]. This approach recognizes that illness is not just a physiological phenomenon but also involves emotional, psychological, social, and environmental dimensions that significantly impact a patient's health and well-being.

In patient-centered care, patients are considered essential members of the care team and take an active role in their care. This framing is responsive to individualized needs and aligns with patient priorities. Patient-centered care has emerged as a prominent framework for delivering care, especially in the context of chronic conditions. It establishes guidelines for providers to engage patients and their caregivers in accessible, coordinated, well-informed care [10]. **Shared decision making**, where patients and providers work together towards an approach to care and treatment, is key to patient-centered care [66]. Informatics solutions, and particularly AI-enabled intelligent systems, can enable patient-centered care that facilitates tailored treatment based on patient data, helps reach patient goals by providing insights and fostering communication that can enable shared-decision making, and improves quality of life by empowering patients to take an active role in their health monitoring and care [12, 13].

Patients with chronic illness and those who provide care for them dedicate substantial effort, time, and resources to care and management. The "illness trajectory" describes the experience of living with the illness, the related work, and the impact of the illness and work on those involved (e.g., patients, families, and care teams) across the course of illness [9], and involves various **lines of work** overlapping and interacting dynamically over time [8]. As a core aspect of patient work, **self-management** — the day-to-day activities individuals undertake outside of the clinic to cope with their chronic illness — plays a critical role in managing and preventing the progression of disease and has become a necessary part of caring for chronic illness [67, 68, 69, 70, 8, 71, 72]. However, self-management does not occur in isolation but is embedded within a broader sociotechnical system that shapes patients' ability to engage in care. Research by Carayon et al. [73] highlights how work system design — including people, tasks, tools, the environment, and organizational structures — impacts patient outcomes. This perspective underscores the need to consider the complexity of patient work within the broader healthcare system, ensuring that interventions support, rather than burden, individuals managing chronic illness.

Research in interactive technology has established the value of technology and data for supporting providers at the point of care [74, 75] and patients in self-managing their condition [76, 77, 78, 79, 18, 80, 81]. Patient-generated data can facilitate a data-driven workflow of clinical encounters [82, 83, 84]. Some research has focused on understanding and designing for patient-provider collaboration in care [85, 86, 17, 87], and recent work highlights the importance of reflection [88, 89, 90], context [91, 92, 84], and personal narrative [93, 94] in managing chronic illness, with some focus on collaborative reflection between patients and providers [95, 96, 97].

2.1.1 Enigmatic chronic illness - Endometriosis

Enigmatic conditions are poorly understood scientifically. Because of this uncertainty, care for these conditions has added challenges. One such enigmatic condition, endometriosis is an inflammatory, estrogen-dependent disorder defined by the presence of a tissue similar to uterine endometrium located in physiologically inappropriate body locations leading to chronic, cyclic, and persistent or progressive symptoms [65]. It is estimated to affect 6-10% of women of reproductive age [23]. Endometriosis is a debilitating chronic illness that has no biomarkers, no clear treatment guidelines, and no cure [24, 23, 27, 25, 26, 65]. Endometriosis symptoms and treatment responses vary widely between patients. Pelvic pain and infertility are hallmark symptoms of the disease. Other symptoms range widely in description and severity and include generalized pain and fatigue, gastrourinary symptoms, dysmenorrhea, and pain associated with sex. Symptoms are often unexplained, biologically undetectable, and seem to vary widely from one person to another [65].

With so much uncertainty, patient expertise can play a key role in patient-provider interactions. When we talk about patient expertise, we often mean that they are experts in their own lived experience of illness [98]. This is certainly true, but with enigmatic conditions, patients can develop expertise on their condition through independent research and keeping up to date with new formal knowledge [99]. This expertise has positive and negative aspects — it can both reduce and compound the stress of illness; while patient expertise empowers individuals with endometriosis to manage their care and advocate for better outcomes, it also imposes emotional and cognitive burdens, could worsen health disparities, and may lead to conflicts with healthcare providers [100].

Since a lack of knowledge impedes the delivery of care [28, 29], patient-centered care for enigmatic conditions like endometriosis is more difficult while at the same time critical. The unpredictability and long, diverse list of symptoms hinders communication and assessment of health status. Medical uncertainty compounds the effort required for patients and providers to work together and complicates the care partnership [101]. The lack of treatment options and clinical guidelines hinders decision-making, and self-management is often required for patients to manage their symptoms [102, 103, 100, 33]. Either independently or together with their providers, patients often engage in self-experimentation to come up with a personalized management regimen that allows them to control and mitigate their symptoms. While some simple tools for helping individuals experiment with self-management have been developed [16, 104, 47, 45, 105], the frustrating and burdensome trial-and-error process is not currently supported by personal informatics tools in the context of enigmatic chronic illness.

There may be a particular opportunity to leverage self-tracked data, that can represent patients' illness experiences from their point of view alongside computational methods that can address the need for rich, holistic representations and heterogeneous responses to treatments that are common with endometriosis. To design technologies to support the multifaceted efforts of caring for endometriosis, we must first understand the dynamics of the tasks, or work, of patients and providers in this enigmatic chronic condition.

2.2 Intelligent systems for health and management of chronic illness

A rich body of literature has been established related to developing intelligent systems and tools to use personal health data to support individuals in their health-related goals [11]. In particular, research in human-computer interactions (HCI) commonly explores **personal informatics** solutions, which aim to empower users in their health by facilitating engagement with personal data — to collect, reflect on, analyze, and use data in order to gain self-knowledge and support behavior change. Li et al. [106] proposed a stage-based model of personal informatics that outlines various stages of user engagement with their personal data (preparation, collection, integration, reflection, and action), which has been widely applied in HCI. Personal informatics tools help individuals understand their illnesses and guide health-related actions through reflection and analyses of historical and in-the-moment data.

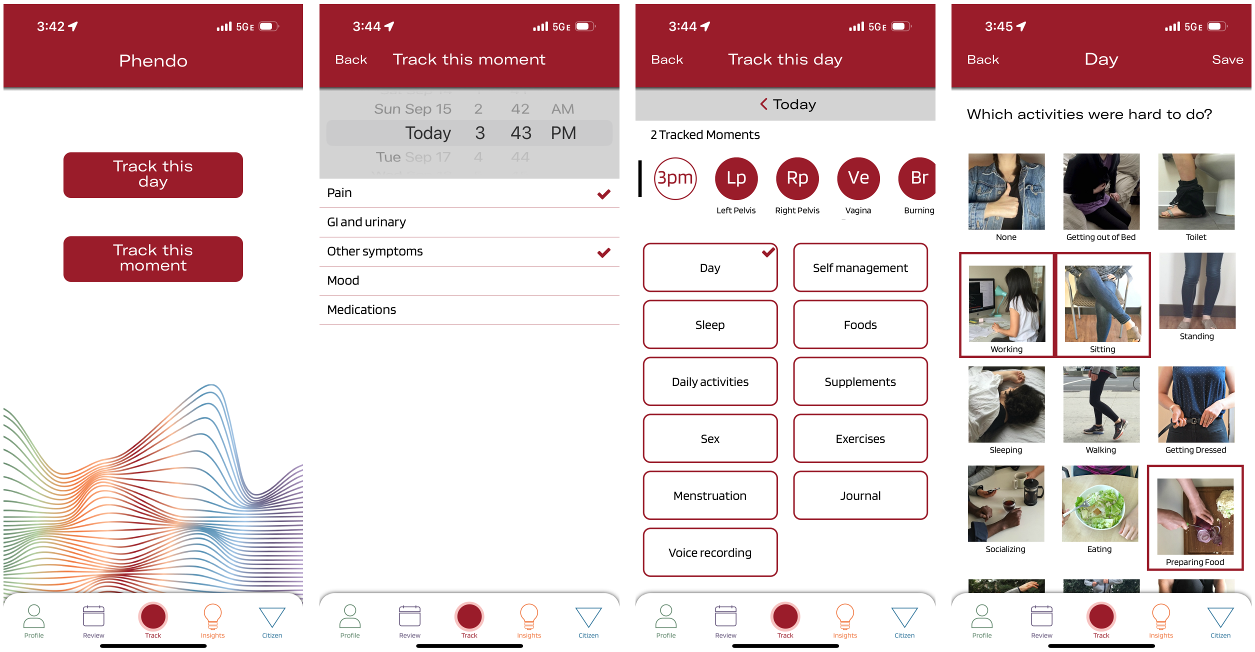
One of the primary functions of personal informatics technologies is to support individuals in the collection and aggregation of personal health data [106]. A variety of personal informatics systems that capture multi-modal information to generate personalized, health-related feedback have been developed. For example, Bentley et al. developed *Health Mashups*, a system that presents statistical patterns between wellbeing data and contextual information, aiming to make data more actionable for users and promote changes in behavior [107]. Similarly, Rabbi et al. introduced *MyBehavior*, a system that uses passive tracking of data, such as activity levels and location, to automatically provide personalized health feedback based on user behaviors and preferences [108]. Rather than using passive data from a smartphone, Cordeiro et al. integrated photo capture into mobile food journaling to engage users and capture more accurate information with less effort [109]. Finally, Karaturhan et al. combined real-time and retrospective self-reflection in a mobile photo-based journaling app to improve the quality of data and encourage users to extract meaning from their information [110].

Given the capture and aggregation of this data, personal informatics tools provide individuals the opportunity to reflect on patterns and garner insights that can inform their health-related actions. Research in personal informatics has explored how to facilitate this reflection, often through the use of visualization and pattern extraction through machine learning. For example, Mishra et al. documented ways that self-tracking technologies could help individuals with Parkinson's disease identify patterns in their symptoms and maintain a sense of agency [79]. For patients with IBS, flexible tracking tools that fit into users' routines were key for supporting individualized, actionable insights and facilitating patient-provider collaborations [111]. In the context of adolescents with autism spectrum disorder, adaptive, custom self-tracking tools were a promising approach to support self-awareness and communication with caregivers [112]. Similarly, some work has explored systems designed to promote reflection through shared health data, while also protecting privacy and trust when managing HIV collaboratively with caregivers and healthcare providers [92]. More recently, personal informatics systems have been augmented with ML to better support reflection and action, for example in the context of diabetes [113, 114].

Still, in practice, many gaps remain in the project of designing tools to support care for chronic conditions, for example to incorporate and focus on domains of everyday life beyond medical aspects, to design for collaborative care, and to create personalized interventions tailored to individual patients [115]. Furthermore, conditions that are enigmatic, or poorly understood scientifically, require substantial effort to understand a patient's illness, work in partnership with a care team, and make decisions about care and management. These considerations are critical in the design of solutions to support care.

2.2.1 Phendo: Personal informatics for endometriosis

The research in this thesis relies on the Phendo app, its engaged community of users, and the data they provide. Doing so allows us to foreground the perspectives of individuals with endometriosis and the experiences that they have painstakingly self-tracked with Phendo. The Phendo app [116, 117] is a mobile research app that was developed in partnership with endometriosis patients to capture the real-world experience of the disease by allowing users to catalog the day-to-day signs and symptoms, self-management activities, and other lived experiences of endometriosis outside of the clinic. A broader goal is to leverage citizen science (i.e., the involvement of volunteers from the general public in scientific research, where individuals contribute to data collection, analysis, or other research activities, generally in collaboration with professional scientists, see [118]) to facilitate ML from the collected data to discover new insights about the



(a) Home screen, with day and moment tracking options

(b) Moment-level questions

(c) Day-level questions

(d) One specific question with pre-set multiple choice answers (activities of daily living that were difficult to carry out)

Figure 2.1: Screenshots of the Phendo app

disease and support individuals in the care and management of their illness. A series of prior studies explored users' motivations for tracking and important dimensions to track, via interviews and focus groups [119], and further elicited variables (e.g., symptoms, self-management strategies) that people with endometriosis find important via online surveys and content analysis of an online endometriosis community [120]. This self-tracking app was designed alongside end-users — individuals with endometriosis — so that the collected data reflect the illness experience of individuals from the patient perspective. In this way, these data can be considered a type of "counter data" [121] that challenges assumptions, reveals biases, and captures perspectives that are commonly rendered invisible by mainstream or traditional data sources.

Participants of the research app have been recruited through patient advocacy networks, and interest and engagement have persisted over time. Since December 2016 when the app was launched, nearly 18,000 people have signed up to use the app. The community of Phendo users is active and engaged with self-tracking activities that provide a rich day-to-day picture of the disease [122]. An analysis of Phendo app usage patterns showed that long-term users of the app are more likely than short-term users to self-track their self-management activities [123].

During onboarding, participants complete the informed consent process in the Phendo app before proceeding to input profile information (e.g., demographics). Participants are also provided with a link to complete the WERF EPHect [124] questionnaire detailing their health history using a standardized instrument that was developed by the endometriosis research community. Once enrolled, users of the Phendo app self-track a wide range of information about the different dimensions of their illness, including signs and symptoms, quality of life, and treatments and self-management — screenshots from the Phendo app are shown in Figure 2.1. The particular domains, options, and structure for capturing the self-tracking data were identified and elaborated in early participatory design work. A visual overview of the main Phendo data capture screen is shown in Fig 2.2, and the full vocabulary can be found in Appendix B.1.

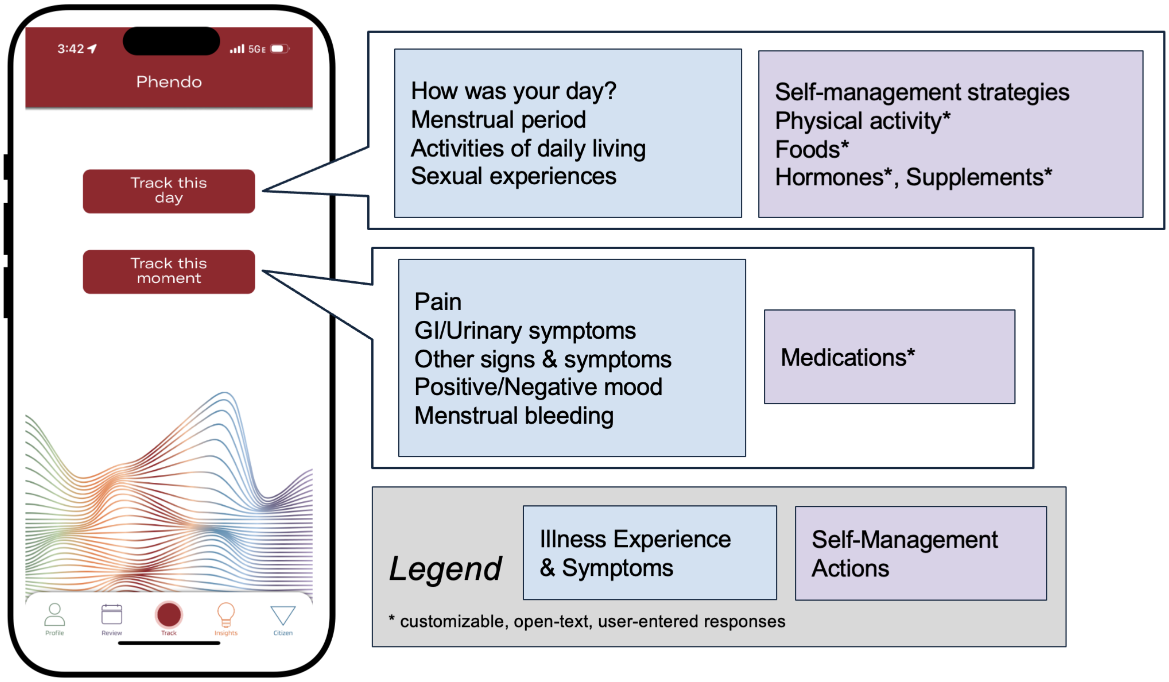


Figure 2.2: Overview of Phendo app tracking domains.

At the *moment level* (i.e., tracking as many or few times throughout the day as the user wants to log), participants can track details and severity of their illness experience. They can log: pain across very specific body locations along with modifiers (e.g., "aching" or "sharp") and severity; a wide range of GI/UI symptoms and other symptoms (e.g., "fatigue," "headache," or "touch sensitivity"); positive and negative moods; bleeding patterns ("clots," "breakthrough bleeding," "spotting"); and medication intake (that is customized by each user in the profile tab).

At the *day level* (i.e., tracking only once per day), users can track a functional assessment of their day — "How was your day?" — from "great" to "unbearable". They can also log their menstruation and flow levels, which activities of living were difficult to do, and experiences with sex. Participants can also track user-entered foods and exercises that may hurt or help, hormones, and supplements. Finally, they can enter free text into the unstructured daily journal.

Significant and novel research has been made possible by the gold-standard dataset created using the Phendo app. Prior work with Phendo data has focused on characterizing the enigmatic illness. An analysis of self-tracked Phendo data has helped to augment what is known about the disease and fill gaps in the medical literature [125]. Other analyses have detailed specific experiences of endometriosis self-management, in particular, the impacts of physical activity [126]. More recent work has sought to understand the needs of individuals and their providers in supporting care and management with personal informatics tools, which is the focus of this thesis. The research activities in this dissertation leverage the Phendo app, the community of users, and the available data from users of the Phendo app.

2.3 Human-centered AI (HAI)

To meet the real-world needs of users in designing such technologies, a human-centered approach is necessary [127]. In **human-centered AI**, humans are positioned at the core of the development lifecycle of intelligent systems in order to create systems that are effective and ethical [128]. It is critical to center the needs of users, and in particular to engage them as experts of their own experiences and data. Human-centered AI tools have the potential to empower human users, and should strive for high levels of human control alongside the automation that AI can provide [59]. In this way, HAI promotes the development of systems that augment human capabilities, rather than replacing or undermining them.

HAI is a topic that arises across various disciplines, and requires interdisciplinary focus to study and address [50, 51, 52, 53, 54, 55, 56, 57, 58, 59]. We consider HAI as the design and development of AI systems that prioritize user empowerment, ethical principles, and contextual adaptation to enhance human agency and well-being. This approach aims to enable reliable, transparent, and trustworthy technologies while supporting dynamic, personalized interactions that align with users' values and needs.

While there is no consensus among researchers on what HAI actually means, Andersen et al. [129] provide a useful framework with five "lenses" for how HAI is used and applied in healthcare: 1) human-centered as a characteristic of AI systems that align with human and societal values; 2) HAI as a design process where users are engaged in creating appropriate AI-based systems; 3) HAI as a focus on the interaction between users and an AI system, e.g., usability, efficiency, or explainability; 4) a focus on the sociotechnical aspects of HAI systems; and 5) a focus on the implementation of AI-enabled systems "in the wild." Various reports from professional societies, government agencies, consumer groups, and corporations have identified a wide range of key principles that characterize human-centered intelligent systems [130, 131], which we also draw from in this thesis. There is increased acknowledgment of the importance of human-centered interactive systems in health, however there are open and urgent questions about how these systems should be designed and developed [54].

In the pursuit of designing intelligent, data-powered, AI-enabled, human-driven systems to support care for this painful, poorly understood chronic condition, it is critical to take an HAI approach — both in terms of methodological approach and in design principles for proposed solutions. This work focuses on the user, their needs, understanding and representation of their illness, control over data and interactive systems, and maximizing benefits from their own personal health data alongside enhanced computational capabilities.

The research in this dissertation focuses on HAI as it relates to the design process, and in designing in alignment with key principles of HAI systems. In this work, we consult endometriosis patients and providers to understand their needs in care and opportunities for technology to support these needs. We then use self-tracked illness data to create temporal phenotypes that can represent health statuses over time. We also consult with individuals who have self-tracked their own illness data to evaluate the performance and acceptability of such tools. Finally, we innovate and implement a novel HAI approach to design that accounts for the human, data, and ML perspectives in developing intelligent systems for care in the context of the complex, enigmatic condition endometriosis. In this way, we seek to create tools that can improve the lives of individuals with endometriosis while also contributing to the informatics and HAI literature.

Chapter 3: Understanding the Needs of Patients and Providers in Complex Chronic Illness

3.1 Introduction and related work

To elicit the needs of patients and providers in the care of endometriosis[1] — a complex enigmatic chronic condition — and opportunities to support these needs, we engage both patients and providers to identify and characterize current experiences and practices both within and outside of the clinical context, unmet needs in care and self-management, and gaps where technology could provide support. To design tools to support the collaborative effort of caring for patients, we must understand the dynamics of the work of patients and providers in this enigmatic chronic condition. This study uses the framework of work and aligns with prior literature documenting the difficult and ongoing work of caring for chronic illness along with understanding the tools needed to support patients and providers. We both concentrate on understanding the work of patients and providers in caring for endometriosis and how breakdowns in this context could be supported by technology, while also focusing on the independent work of patients in self-managing endometriosis and identifying where there are opportunities to support individuals in their care with personal informatics tools.

This study addresses Aim 1 of the thesis. Here, we ask the following research questions:

$RQ_{1.1}$: In the work of patients and providers when caring for endometriosis, what aspects of their collaborative work pertain specifically to such a complex condition?

$RQ_{1.2}$: What role does technology play in facilitating the partnership between en-

[1] The manuscript detailing these results, titled "Divided We Stand: The Collaborative Work of Patients and Providers in an Enigmatic Chronic Disease" was published and presented at Computer Supported Collaborative Work (CSCW) [132]. In this chapter, we present an abbreviated version of the results and discussion.

dometriosis patients and providers and their collaborative work, and what opportunities are envisioned?

$RQ_{1.3}$: In the work of endometriosis patients, what aspects of their independent work pertain specifically to such a complex condition?

$RQ_{1.4}$: What role does technology play in facilitating the success of endometriosis care and management for patients on their own, and what opportunities are envisioned?

3.1.1 Patient-provider partnership and collaboration in care of chronic illness

Patient-centered care reconciles the traditional biomedical-only frame of disease with care "that is respectful of, and responsive to, individual patient preferences, needs, and values, and ensuring that patient values guide all clinical decisions" [10]. Key to patient-centered care is shared decision-making, which supports patient-provider collaboration and negotiation in approach to care, where both providers and patients have relevant expertise and perspectives [133, 66]. Providing access to clinical documentation has been linked to higher perceived shared decision-making among patients and insights into their own care and self-management practices [134], but patient portals still lack functionality to allow patients to fully participate in care decisions [135]. Patient-generated data can also facilitate shared decision-making, for example, one study with Parkinson's patients found that graphical summaries of sensor data helped to guide collaborative conversation [87].

Beyond the clinical encounter, **patient self-management** is a fundamental component of coping with chronic illness [69], especially in the context of an enigmatic illness, where treatments are not reliably effective at mitigating symptoms. While health professionals consider self-management a routine element of a patient's medical regimen, patients think about self-management as a process to facilitate day-to-day normalcy and structure, often through trial-and-error, working through the emotional toll of illness, and challenging the medical dominance over their illness experience [136]. Patient experience and expertise are not consistently acknowledged in the current medical model or the traditional role of the patient, but are in fact key to patient empowerment and enable pragmatic handling of uncertainty in the intricate day-to-day contingencies of self-management [137]. Patients engage in problem-solving to transfer insights from past experiences onto current self-management situations [138]. Self-tracking tools have been shown to support patients coping with incurable illness by facilitating problem-solving and coping with detrimental emotional reactions [79]. Furthermore, patient expertise allows patients to integrate clinical knowledge into self-care practices and enables finding common ground with providers, supported by incorporating data beyond clinical documentation into care tools [139].

Common ground, as introduced by Clark [140], is the process of building shared knowledge, beliefs, and assumptions which evolve through time in a partnership. The more common ground, the more successful the communication between actors and the better the collaboration; when common ground is lacking, misalignments abound. Coiera [141] expanded on the idea of grounding, suggesting that effort required to ground conversations could happen ahead of time or at the time of an interaction. **Pre-emptive grounding** is best suited for information-focused tasks where information is stable, repetitive, archival, or critical but rare and that may be worth formalizing. On the other hand, **just-in-time grounding** is best suited for communication-focused tasks where new knowledge is exchanged, information is informal, local, personal, or rare, and prediction of what needs to be shared is difficult. Conversations with substantial common ground shared between patients and providers can be succinct, but poorly grounded conversations must be supported with information exchange and often rely on artifacts to facilitate communication, align perspectives in care, and make sense of uncertainties together. For providers, technology has been explored to facilitate common ground in a clinical setting (handoffs in the ICU) with high complexity due to high data volume and coordinating across multidisciplinary care teams [142, 143], but these studies are limited to collaboration within clinical teams, rather than across patients and providers.

Care for chronic illness requires **collaborative and ongoing efforts** to align perspectives and attend to the embodied complexities of illness, beyond treatment and self-management, particularly in situations with substantial uncertainty. Empowering patients and including them as partners in their care is important, but autonomy and independence are not universal ideals. Patient choice does not always lead to better outcomes, with limits on the controllability of disease and the potential for this responsibility of choice imposing burden on patients [144, 145]. Relying on patients for self-management and involvement in medical care has both benefits and downsides [33]. While patients asserting their expertise and questioning medical dominance can help them understand and manage their condition, patients who are not doing everything they can to mitigate or actively seeking to cure their disease may be blamed or deemed a personal failure. Individual responsibility in the absence of a patient-provider partnership may do more harm than good. In fact, patients are not generally looking to be autonomous in their care [146, 147], but rather to partner with providers for decision-making [148]. The role patients play and want to play in their care fluctuates (within and across individuals) and depends on context [149] and trust in the patient-provider relationship [150].

Boundary negotiating artifacts can be used in complex knowledge-sharing tasks to exchange information, negotiate roles and expertise, and establish and align perspectives within multidisciplinary collaborative teams, like patients and providers [151]. These tools facilitate crossing and pushing boundaries in dynamic, context-dependent situations where expertise are shared and misalignments are common. Chung and colleagues [82, 152, 111] argue that using self-tracking in patient-provider collaboration can be conceptualized as a dynamic process of navigating tensions between the patient and provider scope of expertise through creating and using boundary negotiating artifacts. Piras and Miele [153] also explore applying self-tracking implemented by patients to set boundaries and reaffirm independence in self-management. Based on their own understanding of their bodies and patterns, prior research has shown, patients may utilize strategies of noncompliance to further their personal self-care goals or values, based on their own experience and expertise [67, 100]. Widespread use of mobile phone apps for health management suggests opportunities to leverage patient-generated data and health informatics tools to support clinical encounters and meet healthcare needs [69]. **Personal informatics tools**, that enable both collection and reflection of personal health data (e.g., behaviors, symptoms, treatment progress, and general health) [154, 155], can support patient self-discovery independent of their providers and when shared with providers can enable negotiation of roles and facilitate patient-provider communication, support diagnosis and personalized treatments, and enhance motivation, accountability, and engagement with tracking and the treatment plan [82].

3.1.2 Work across the illness trajectory

In designing for patient-centered, collaborative care, we apply the **patient work framework** that extends design focus beyond a singular biomedical lens of illness by "attending to the embeddedness of patients' health management in larger processes and contexts and prioritizing patients' perspectives on illness management" [156]. By understanding the efforts entailed in care and the dynamics of getting this work accomplished, we can design technology that is responsive to and supportive of the lived experience, local context, and work activities of caring for chronic illness.

Patients with chronic illness and those who provide care for them dedicate substantial effort, time, and resources to care and management. The "illness trajectory" describes the experience of living with the illness, the related work, and the impact of the illness and work on those involved across the course of illness [9], and involves various lines of work overlapping and interacting dynamically over time [8]. **Illness work** (diagnostic, treatment, and symptom management activities) and **everyday life work** (daily or regular tasks to keep up personal and home life) are regular, ongoing activities to facilitate patients' day-to-day lives with their illness [8]. Existing technology often supports illness work, and designs are beginning to incorporate context-aware solutions [157].

Biographical work entails understanding and reconstructing identity and life meaning in relation to one's illness and social history. This effort to understand one's life and identity across the illness trajectory overlaps with the largely implicit and unacknowledged **sentimental work**, necessary towards both humanistic and pragmatic ends [9]. Effort in communication and relationship-building can provide comfort, satisfy social norms (e.g., active listening), and help get back on track after a negative patient-provider interaction. Building trust between patients and providers is critical to establishing a collaborative partnership and can also help motivate patients in their care regimens. One study documented how providers' emotions are used to coordinate care in the ER (e.g., to facilitate a shared mental model about a situation, or to communicate concern as a call to action) and how they are represented (or not) using technology or in paper documentation [158].

Across the illness trajectory, projects or "arcs of work" comprised of tasks (or clusters of tasks) must be coordinated amongst actors varying in experience, skill, knowledge, training, and social location [159]. **Articulation work** is the complex interplay of mostly implicit work that organizes and coordinates tasks and actors, enables tasks to be carried out, and "gets things back on track" after unexpected contingencies [159, 8, 160, 161]. Articulation work has been extensively studied in CSCW, particularly with medical records [162, 163].

Information work is pervasive across the various arcs of work, including articulation work, wherever information is given, received, or exchanged [71, 164]. Patients engage in other forms of information work to facilitate care of chronic illness. To understand their illness experience, patients can produce and reflect on self-knowledge (e.g., symptom self-tracking, diabetes self-monitoring) to gain insights by searching for patterns and linking past information to inform current or anticipate future situations. Artifacts of this **reflexive work** could reduce the cognitive load required, and insights from reflection could mitigate the stress associated with managing illness. Self-monitoring may also enable a process of (re)discovery and (re)learning about one's illness experience through experimenting [165]. Personal informatics tools can facilitate reflection. Work in HCI suggests strategies and tools to support self-reflection and communication, which may allow patients to correct misalignments with providers and articulate their values, self-care approaches, and how these intertwine [89]. Tools with prompts to steer reflection are promising to minimize burden and enable control over disclosure [90].

Beyond the expertise gained from reflecting on one's own lived experience, patients with chronic illness also build lay expertise by consulting online resources, researching established literature, and connecting with online health communities to discuss their unique case and brainstorm broader solutions to fill in gaps in knowledge. This **work of becoming an expert patient** [33] and the responsibility to self-manage can empower patients and reduce stress associated with illness.

But while these self-tracking technologies have the potential to empower patients, they also risk adding stress related to the tracking or the illness and may magnify surveillance or pressure for patients to be "disciplined" in their involvement [165, 33].

3.2 Methods: A qualitative approach

Focus groups and interviews. We engage patients through focus groups, where semi-structured group discussions lasting around 90 minutes center around how patients assess their own health status, communicate with their care team, and self-manage their condition outside of the clinical context. We also consult with endometriosis specialists through semi-structured interviews lasting around an hour, where we ask providers about their approach to a typical visit with endometriosis patients, perspectives about shared decision-making in practice, attitudes towards using patient-generated data at the point of care, and use of technology to support care. In the interviews with providers, we show them two visualizations (see Appendix A.1) to facilitate discussion about how patient-generated data and associated technologies might impact their care practices. We conduct focus groups and interviews until we reach information saturation. These qualitative studies enable us to understand both patients' and providers' perspectives about the process of care and the patient-provider partnership, the use of data for care, communication, and self-management, and desires for tools that could help address unmet needs. The interview and focus group guides can be found in Appendix A.1 and Appendix A.2, respectively.

Recruitment and eligibility. We recruit endometriosis patients who used the Phendo app at the time of the study, and also via social media posts, flyers hung near clinics, and through partner organizations. Eligibility for participation includes: English-speaking adults with a diagnosis of endometriosis, having experienced symptoms in the past three months, and having received care for endometriosis in the past year. Patients are compensated with a $25 pre-paid card for participating in the focus groups. We recruit providers recruited from large institutions that provide endometriosis care and through recommendations from patient advocacy groups. When recruiting providers, we preference individuals from a diversity of clinical specialties and across a wide range of experience levels. Eligibility criteria include: self-reporting endometriosis expertise in their practice.

Data and analysis. Data are collected and stored in a secure, electronic format. Audio files are transcribed for analysis. Transcripts are checked against the audio recordings and all names and identifying information are removed. Thematic analysis is guided by our goals to elucidate the types of work entailed in endometriosis care and to identify opportunities for the design of technology to support this work. We follow the methodology as detailed in [166]. Coding is carried out iteratively, with initial codes generated broadly as margin notes, then organized to search for and generate themes. The codebook is revised until consensus is reached among coders that data are represented in proposed grouped codes. Themes related to our overall research questions of work of different actors are selected from the broad list generated. Transcripts are coded for themes by two independent coders and discrepancies are discussed. We use Cohen's Kappa to assess the level of agreement between two coders, across two transcripts, with a higher value indicating better reliability and a score of 0 being equivalent to chance and a score of 1 indicating perfect agreement [167]. After the coding is complete, a third coder addresses the extent to which coders apply the coding framework to the transcripts. Findings from patients and providers are synthesized, then compared and contrasted against each other. Finally, once the themes are identified, we share them with our participants for feedback and to assess their fidelity. This type of member checking provides further confidence in our findings.

3.3 Results: The needs of patients and providers

We conducted 5 focus groups with a total of N = 21 endometriosis patients. We conducted 10 interviews with specialists. We engaged individuals across a range of experiences, as detailed in Table 3.1. For the qualitative analysis, the final Kappa coefficient for two coders was 0.89, showing a high degree of agreement.

Through the thematic analysis, we found that the work to care for endometriosis, including reflecting on and making sense of the illness experience, preparing for clinical visits and planning

Providers *(N=10)*	***n (%)***	***Patients*** *(N=21)*	***n (%)***
Gender		Age	
Female	7 (70)	*Younger than 30*	7 (33)
Male	3 (30)	*30 or older*	14 (67)
Years Experience		Years Diagnosed	
Less than 5	3 (30)	*Less than 5*	12 (57)
5 to 10	3 (30)	*5 to 10*	6 (29)
10 or more	4 (40)	*10 or more*	3 (14)
Specialty		Race or Ethnicity	
Gynecologist	2 (20)	*White*	14 (67)
Surgeon	3 (30)	*Black*	5 (24)
Physiatrist	2 (20)	*Latina*	2 (10)
Pelvic Phys Therapist	2 (20)	*Asian*	1 (5)
Pain Specialist	1 (10)		

Table 3.1: Participant characteristics across provider interviews and patient focus groups. Patients could select more than one race; race and ethnicity were asked separately.

for care, and engaging in self-management, is significantly compounded by the complex nature of the disease. We distinguish four aspects that complicate the work of patients and providers; these themes and some select examples with quotes are presented in Tables 3.2, 3.3, 3.4, and 3.5. As we design solutions, the enigmatic nature of endometriosis calls for complementary approaches from human-centered computing and artificial intelligence, and thus opens a number of design opportunities and future research avenues — for supporting the work of both patients and providers in caring for and managing this complex condition and for patients on their own in self-managing their illness.

Theme 1: Enigmatic condition means uncertainty and frustration in care
Patients reported relying on their lived experiences and personal records, but mostly feeling uncertain when **reflecting and making sense of their own disease experience**. Because of the enigmatic nature of the condition, patients feel lost in **assessing their health status**. *"Things happen to my body and I don't know if it's related to endo." "I have the hardest time figuring out how I'm feeling. Is how I feel normal? I don't know what good is supposed to feel like." "I'm stage one... but the pain does not feel like stage one, you feel self-conscious like, oh gosh have I been really overreacting, is my pain tolerance really low?" "I've had doctors tell me, 'Well this shouldn't be happening'...well it is happening so what now?"* Patients also exhibited doubt about **assessing their treatments**: *"Is this even working? I don't know how to tell,"* said one patient, and another agrees, *"I suck at evaluating,"* and a third said, *"We're all just guessing."* This difficulty in assessing health status and treatment progress was echoed by providers across specialties.
When planning for care, patients and providers described a **trial-and-error approach to treatment** frequently taken in an enigmatic condition, often experimenting with multiple methods of *"hacking endo"* as one provider called it. Many providers spoke optimistically about this approach: *"I have no shortage of options and combinations of options that we can try until we find something,"* while others encouraged caution and close collaboration with the patient: *"We don't know enough about endo to dictate medications or procedures. I have to encourage patients to be involved with the process."* But patients talked about the toll it takes on them to fight a disease without standard care: *"It became this trial and error, [...] 'Let's try this because there is nothing else left to try.' It was exhausting."* They also perceive the providers' frustrations when treatment options are exhausted. One patient says, *"They are almost exasperated when you don't feel better. I do think that it comes from their own frustration with endo,"* and, *"They get frustrated when the textbook answer doesn't work for you."* With the lack of reliable success with treatments, care often falls to patients to **figure out a self-management regimen** that works for them, which often means a lengthy process of experimentation. They have a hard time figuring out what to try (*"I'm literally willing to try anything,"*) and evaluating if these strategies work to mitigate their symptoms, and could not find technology to support their experimentation.
Several providers suggested designing "self-tracking prescriptions" to support the trial-and-error approach to treatment and self-management, where patients and providers collaboratively select key symptoms, triggers, behaviors, treatment, or self-management strategies to track consistently for a specific period of time. One provider noted the potential for collaboration: *"These are the symptoms that are important to you. These are the measurements that are important to me. Let's see if we can narrow down."* Providers highlighted opportunities to garner buy-in from patients, *"I ask the patient 'Hey, I'm noticing a pattern here. Can you over the next three months track these four or five things really, really well for me? And focus less on these other things?'."*

Table 3.2: Overview of Theme 1 across provider interviews and patient focus groups for identifying the needs of patients and providers in their care tasks. We also present some selected examples and quotes.

Theme 2: Multi-factorial and systemic condition overwhelms patients and providers in working together for comprehensive care
Patients emphasized feeling overwhelmed when **reflecting on their health status** because *“there are so many facets of the disease and there are so many different systems of the body that it can impact.”* One patient asserted the value of keeping track of these different aspects and using them to gain insight: *“It’s easy to get caught in the different kinds of pain in the different organs. But I think assessing the different kinds of pain also helps linking things together.”*
Providers noted how inter-related causes of different symptoms added complexity to their own clinical work and when **reflecting with patients to figure out what is going on with their health**: e.g., *“Is that rectal bleeding because your endo has eroded through your rectum or is it because of constipation due to chronic pelvic pain?”* and, *“Is it chronic pain that was taken over by the central nervous system or is it pain due to a lesion?”* They also noted the systemic impact of endometriosis further challenged understanding a patient’s illness experience, sometimes even questioning the patients’ reports: *“Sometimes you see patients who come here and everything hurts. You’re wondering how much is real, related to the endo, or stuff I cannot help with,”* explained one surgeon, while a physical therapist wondered *“Endo is multi-system, they feel a lot of things in different spots and it’s hard to sort through: Did you feel a fleeting pain or was this really disabling?”* To adequately care for this complex disease, some providers underscore the importance of **approaching each patient holistically**, as a physiatrist explains, *“It’s not just their urinary stuff. It’s not their GI stuff. It’s not their neurologic. I mean, we are standing back and we’re putting them all together,”* and a physical therapist says, *“We want to engage not just the pelvis but the brain and the heart in all of it.”*
Providers highlighted the complex work that goes into **reviewing and interpreting clinical and self-tracked patient data** from across quite a few domains of health. Deriving insight from raw low-level, day-to-day self-tracking data is cognitively difficult and time-consuming, reconciling these insights with clinical history is complex, and incorporating the information into a patient’s medical record adds an administrative burden of manually inputting data due to lack of interoperability. *“I have to make a mental model of what I think is going on based on that data, interpret it, and write it down free-hand into my console note. [...] I have to rely on my ability to quickly process that information and make sure I’m not missing anything.”*
Providers and patients both expressed desire for synthesized, summary reports to facilitate at-a-glance assessment to get a quick overview of patient data (*“A report rather than raw data because this honestly takes a lot of work to go through all of this,”*) and analytics to support them in summarizing and identifying trends and in discussing these patterns together during the encounter. And while too many details can get in the way of constructing a full picture of the patient, sometimes these details can provide useful clues to disentangle what may be causing symptoms or how to address them: *“Sometimes there is too much information, then you get caught too much in the weeds [...] you just need things that help, you want some granular detail but, you don’t want too much that you miss the forest for the trees.”* Tools bringing clinical and self-tracked data together could mitigate the multi-factorial aspect of the disease by enabling providers to *“discuss symptoms that patients are feeling, but that we traditionally forget to ask about,”* or *“as a way to target my questions a little more. So, it might not save time but it might get me at more granular information.”*

Table 3.3: Overview of Theme 2 across provider interviews and patient focus groups for identifying the needs of patients and providers in their care tasks. We also present some selected examples and quotes.

Theme 3: Chronic condition with different temporal resolutions adds confusion for both patients and providers
The different temporal resolutions of endometriosis — its chronic aspect, its cyclical variations, and its rapid fluctuations in symptoms — emerged as an important complicating factor, specifically when **making sense of the disease and planning for the future.** Patients and providers understand these time frames much differently, which confuses patient-provider communication and prevents building a shared understanding of the patient's experience of illness. Patients often referred to their chronic illness as a journey through the years and life events. One patient framed her illness experience *“in stages of life,”* and others related reframing their life goals in the context of their conditions, along with the disappointment that their illness had caused: *“There's a big element of grief to the endo journey.”* In dealing with the emotion and uncertainty about the future, one patient said, *“I'm treating this as an adventure.”* In this **biographical work**, patients reported reflecting on their journey to understand their illness history, contextualize their current experience, and forecast what their illness trajectory might look like in the immediate and long-term future. When it came to **reporting their experiences of symptoms** to providers, however, patients felt they could realistically make sense of their symptoms short-term only. *“For me, it's such a daily thing. It's hour-by-hour, day-by-day [...] so when a physician asks me, ‘How have you been in the past three months,’ that's kind of a tricky answer.”*
Providers in contrast conceptualized the patients' health status at longer time intervals, like time between appointments. This resolution difference was frustrating to providers: *“It can be really hard to sort through, if they are like ‘and on this hour I felt this way and that way.’ Most of us think in terms of weeks instead of days,”* while another complained: *“People don't know how long their symptoms lasted. They don't know when they started. They don't know when they first felt them.”* To this end, providers imagined that technology could support patients in recalling details of symptoms and reflect on them to extract useful insights for their care with more synthesis and abstraction than what is currently offered in self-tracking technology. *“When I see them every six months or a year, it gets really, really hard I think from a recall perspective of remembering, 'How did I feel seven months ago? I have no idea.”'*

Table 3.4: Overview of Theme 3 across provider interviews and patient focus groups for identifying the needs of patients and providers in their care tasks. We also present some selected examples and quotes.

Theme 4: Patients and providers negotiate knowledge and expertise, attempting to align perspectives
Patients emphasized the considerable **work of becoming an expert patient** about the scientific nature of their disease across their trajectory of illness: *“I really had to do my own research, tons of hours put into. I feel like I’m a bigger endometriosis specialist than my own doctor.”* Providers were aware of their patients’ expertise: *“Compared to other conditions I take care of, I found most women that are affected with endo are very well versed in the condition,”* and in fact acknowledged that in contrast to most chronic conditions, *“there is not that much difference between what the physicians know and what the patients know.”*
Negotiating expertise to assess health status or determine approach to treatment was commonly considered an asset by providers: *“One of my favorite questions to ask is, ‘What do you think is going on?”’* or just part of the clinical dynamic: *“Sometimes, they are totally reasonable. Sometimes, they have the absolute right solution. Sometimes, they have one of many right solutions.”* But, patients pushing boundaries may threaten the established dynamic: *“It’s your reputation, and it’s how you deal with those expert patients. They know they are not gonna win against me,”* said a physical therapist, while a surgeon remarked, *“It is better to have a little bit of ignorance. Because sometimes they read too much, and then it’s very difficult to then guide them in their therapy.”*
Because of the inherent complexity and medical uncertainties, providers emphasized the need to **manage patients’ expectations** at the start of their partnership: *“It’s more difficult to manage if you don’t set those expectations, they are gonna expect you to cure them.”* A surgeon insists, *“It’s important to discuss that we are so behind with this condition. We just try to bandage, this may be a battle that we cannot win.”* This sentiment carried across specialties, when reflecting on realistic goals with no cure: *“Our goal is to help you take your pain from a 10 down to a 5,”* and, *“It’s important to let the patient know that there’s limits to what surgical management can do.”*
While patients acknowledged the limitations of treatment and the need to remain realistic with their expectations, they felt unheard and dismissed about their own **illness experience**, especially when it came to their experience outside of objective clinical measures. When preparing for the clinical encounter, patients talked about the **sentimental work** involved in reflecting on their own experiences. One patient asserted, *“It feels like a trial, I have seven minutes as a lawyer to prove my case, to present enough evidence, to hit the particular buzzwords.”* Patients commonly felt their provider’s understanding of their disease did not align with their lived illness experience.
Patients consider their self-tracking data and narratives as a critical component of their disease status and bring these artifacts to the clinical encounter because they want their providers to be aware of the whole picture and to reflect on the information together. Beyond the value of a holistic picture for care, patients felt dismissed when providers minimized the relevance or believability of this information. *“I’m not tracking just for myself. It’s going to be used in a way that’s gonna help my care. I’ve been doing ‘that’ for years now. I know ‘that,’ but does my provider know ‘that’?”* Providers acknowledged that sharing self-tracking data can act as *“a trust builder between [the patient] and I, so she can choose to show me this or not show me this,”* going on to explain, *“It’s a way for people that have been so minimized to say ‘My pain is very real.”’* Patients agreed, *“When I have something to show them, it’s not just me saying things to them, there’s actual records of it,”* hinting that self-tracking data can act as an objective metric, considered more acceptable to providers than a verbal narrative.

Table 3.5: Overview of Theme 4 across provider interviews and patient focus groups for identifying the needs of patients and providers in their care tasks. We also present some selected examples and quotes.

3.3.1 The needs of patients and providers in management

Patients and providers described a wide range of challenges and needs related to bringing together, reflecting on, and using patient data to prepare for clinical visits or within the clinical encounter, and various needs related to facilitating patient-provider partnerships and collaborative work.

Bringing together, reflecting on, and using patient data to prepare for clinical visits and within the clinical encounter. Patients and providers described challenges related to using personal health data to assess health status and construct holistic representations of their illness experiences. The enigmatic nature of the disease was described by patients and providers as a key feature that spurs a lot of uncertainty and frustration when trying to make sense of a person's illness experiences and health status. They also talked about how the many body systems involved and the wide variety of domains potentially relevant to a person's illness, both within and outside of healthcare domains, overwhelms them and makes it difficult to synthesize information. And the different temporal resolutions (i.e., the chronic nature, the cyclical variations, and the rapid fluctuations in systems) compound the complexity of assessment and using data for care. Patients and providers reported difficulties bringing together, organizing, and synthesizing their large volumes of complex, granular data across a wide breadth of relevant symptoms and domains. Disconnected data streams from various apps and other sources with no way to collect them in one place or export them to share with providers (including combining medical data from providers with patient contributed data) was one challenge discussed that makes it difficult to compile a comprehensive data profile. They also discussed how they are unable to reconcile temporal resolutions in data to align with both patient needs for tracking and understanding (patients conceptualize their illness experiences at either a moment-to-moment scale or as a journey across stages of life) and, at the same time, how providers need the data for clinical assessment and care (e.g., month-to-month or between appointments). Participants also described difficulties reconstructing a holistic picture of how patients are doing, especially since the illness experience cannot be captured fully in data, and patients need to add context, interpretation, and narrative to bring together a more ‘gestalt’ view. Patients also reported difficulties identifying changes, trends, and associations; they also reported having difficulty comparing their current health status to how they felt previously (e.g., from a personal baseline, or from a previous time point). Finally, both patients and providers explained how difficult they find it to develop insights from reflecting on changes, trends, associations, or comparisons, and then lack methods to document insights, update them, and then use them for care.

Facilitating patient-provider partnerships and collaborative work. Patients and providers also described needs related to facilitating patient-provider partnerships, negotiating expertise, and correcting misalignments. With a lack of established medical knowledge, patients often develop expertise around what is known scientifically about their disease. Patients and providers both asserted that patient expertise is legitimate and valuable. Some providers also warned that this type of expertise could lead to more challenging patient-provider dynamics, which is further threatened by the stigmatization of endometriosis, both as a disease associated with female reproductive organs and the menstrual cycle, and as a chronic condition with pain as a central feature. Providers described how the uncertainty inherent in endometriosis care also means more opportunities for misalignment. Another complication relates to the fact that many different types of providers and specialists care for endometriosis patients, and patients may have a large care team or no care team at all. Patients described feeling like they needed to prepare for their clinical visits, yet had difficulties in doing so and lament lacking technical support, especially since they explained that it is best to tailor their prep work and any artifacts (records, data, notes) to the specific provider, a particular specialty, and/or the circumstances. Patients told us that they have a hard time figuring out what information is critical to convey to a particular provider or for a particular visit, how to curate representations of their data, and how to present their information to providers. Participants also described a tension — balancing the need for rich, detailed representations of personal health data alongside personal narrative that “tells the story,” with artifacts that enable a quick overview “at a glance”. Patients and providers talked about how the rich, holistic representation aligns with the patient experience and can help providers structure the visit and target their questions, while the "at a glance" overview fits into the clinical workflow and is easily digestible. Patients emphasized that it is especially difficult for them to "tell the story" quickly with new providers and to align the provider's understanding of their experiences with their own. Finally, participants described challenges related to navigating interpersonal relationships, expertise, and misalignments in understanding, perspectives, or expectations, and lack supportive mechanisms to identify, negotiate, and explicitly reconcile misalignments and to fill gaps in knowledge.

3.3.2 The needs of patients in self-management

The patients and specialists we spoke to for this research described various, common challenges that patients encounter when trying to self-manage their condition.

Developing individualized self-management regimens through trial-and-error. Participants described self-management as an essential task for mitigating the effects of symptoms that patients experience as a result of their illness, and emphasized that self-management strategies can compliment formal care or address symptoms that other treatments have not been helpful at eliminating. Since endometriosis is so poorly understood and the experience of illness and what works for each patient is so individualized, research participants emphasized that self-management is critical to promoting quality of life and mitigating the burden of illness on patients' everyday lives. But, providers lamented that successful self-management in this context is impeded by a lack of established clinical guidelines to follow or reliable management options and no existing biomarkers or symptom profiles to rely on for evaluation. The multi-factorial and systemic features also mean that there are different body systems and non-specific symptoms, where patients talked about competing priorities in self-management.

Patients discussed their difficulties working to develop a personalized self-management regimen that works for them, since what works for one person may not work for other similar patients. Many described relying on trial-and-error with self-management strategies, but are not supported in their self-experimentation. They talked about lacking support for identifying strategies to try, figuring out what to try (i.e., what might work for a particular individual, and under what circumstances a particular strategy might work for the individual), implementing strategies, evaluating if strategies work to address the symptoms they are trying to manage, and keeping track of what strategies work in what contexts to implement in the future.

3.4 Discussion

Overarching themes suggest that the complex nature of caring for endometriosis does not create new work, but rather intensifies every aspect of patient and provider work, as well as complicates their relationship. While some technology solutions exist and are used by both patients and providers, they fall short of supporting them in dealing with a condition with no established medical guidelines nor enough knowledge to produce reliable treatment plans. We argue that the enigmatic, complex, and ambiguous nature of endometriosis necessitates complementary approaches from human-centered computing and artificial intelligence, opening numerous avenues for future research and design. In Sections 3.3.1 and 3.3.2, we identified a wide range of needs of patients and their care teams and gaps in technological support. In this section, we elaborate on two key opportunities that we focus on addressing in the subsequent studies of this thesis.

3.4.1 Opportunities to support the work of patients and providers in complex chronic illness

In this study, we identified a range of unmet needs of patients and providers in managing endometriosis. Across all contexts, patients and providers discussed the difficulties they face in making sense of an individual's health status over time, particularly given the broad variety of symptoms and domains that are often experienced with endometriosis. Further, all stakeholders conveyed a dire need for support in self-management of this complex condition. Because of the enigmatic and chronic nature of endometriosis, treatments are often ineffective and management regimens are highly personalized, commonly requiring lengthy and involved trial-and-error with various strategies. While patients emphasize that they want low-impact, restorative self-management strategies to help them live with and ameliorate symptoms, they are not well-supported in their lengthy experimentation and have difficulty identifying strategies, sticking with them, and evaluating if they are working. Participants imagined features of intelligent systems that could help them scaffold this trial-and-error self-experimentation process — helping them figure out what to try, determine if it is working, and structure the data collection process. A tool to help users experiment with self-management strategies to help develop effective individualized regimens can make use of large volumes of self-tracking data along with computational approaches and human-centered AI to help facilitate action and support care. We largely focus on these gaps in the following studies.

Chapter 4: Learning Interpretable, Temporal Health Status Phenotypes from Self-Tracked Patient Data

4.1 Introduction and related work

In the previous qualitative study, we identified key needs and technology gaps in the work of caring for the complex, enigmatic chronic condition endometriosis. Patients and providers described how challenging it is to care for and manage endometriosis, especially given how unpredictable and burdensome the illness experience is for patients. They described how challenging it is to characterize their illness state, at a particular time and over their illness trajectory, and to understand how their illness states change over time or in response to treatments and self-management. These challenges are due in part to the complexity of endometriosis and the systemic nature of its presentation, to the significant week-to-week variations in health status and needs, and to the unpredictable variations in health status that are inherent in endometriosis. While others have addressed this question at the population level (e.g., though symptom clustering [168]), we tackle the open question of characterizing individual-level health status.

There are new tools for self-tracking and management that could support individuals in their care and management. Thus, there is an opportunity to leverage self-tracked data and machine learning to support patients in reflecting on and understanding their health and to enable intelligent systems to support people in their care tasks. But, this will require meaningful and interpretable computable representations of illness states over time. In this work[1], we create a model capable of analyzing data from the Phendo app (described in Section 2.2.1), where users self-track their experiences of illness as they happen in day-to-day life. We use these data to generate temporal

[1] This study was presented at the American Medical Informatics Association (AMIA) Annual Symposium in 2024, in a podium talk titled: "Learning Interpretable, Temporal Health Status Phenotypes from Self-Tracked Patient Data."

health status phenotypes that might be useful for individuals in understanding their illness and that could provide valuable information to intelligent systems to support care and management of complex illness.

These studies address Aim 2 of the thesis. Here, we ask the following research questions:

$RQ_{2.1}$: Can a digital phenotyping approach aggregate individual-level data to enable interpretable representations of health status in the context of a complex enigmatic condition?

$RQ_{2.2}$: Can digital phenotyping methods construct meaningful representations of health status for individuals?

$RQ_{2.3}$: Can digital phenotyping methods construct meaningful representations of health status for individuals over time?

$RQ_{2.4}$: Can digital phenotyping methods construct temporal representations of health status that can be used in real-world tasks?

4.1.1 Digital phenotyping

Digital phenotyping is a promising approach for the task of constructing meaningful temporal representations of health status from self-tracked health data. Onnela and Torous have defined and operationalized digital phenotyping as the "moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices, in particular smartphones" [169, 170]. With so much data being generated by individuals online and through mobile devices, digital phenotyping has become a common approach to construct these complex representations [171, 172, 173, 174, 175, 176, 177, 178, 179]. Researchers have put forth various frameworks for constructing holistic representations from multi-modal data [180, 181, 182].

There are many benefits to using phenotyping approaches [183]. Phenotyping with smartphone data can enable naturalistic data capture, representing real-world patient experiences [184]. Phenotyping can handle large volumes of patient data and facilitate dimensionality reduction, while providing meaningful labels [185]. It can also generate temporal representations of patient data across periods of time [186], which can facilitate prediction tasks and other computational tasks of AI-enabled systems. Finally, it can use interpretable models to create health status representations that are meaningful for use by humans in understanding their illness and in supporting their care tasks [187], while also are machine-readable that can be used by intelligent systems to support computational tasks [184].

In this research, we create digital phenotypes of user health states to form computable representations of illness states. We aim to create digital phenotypes which can be meaningful and useful for individuals and that can also be used by AI-enabled intelligent interactive systems for chronic disease management. These studies focus on developing interpretable, meaningful representations of health status for individuals over time. Using digital phenotyping, we can leverage a variety of data from a person's smartphone to enable automated characterization of their phenotypic state. Combined with data science and machine learning techniques, this approach to analyzing patient-generated data has the potential to facilitate personal informatics tools and interventions to support care and management of enigmatic chronic illness.

Because endometriosis is poorly understood clinically and manifests heterogeneously among those with the disease, an unsupervised approach is ideal and interpretability of the model results is important. Further challenges arise when considering the characteristics of this corpus of self-tracked data: illness experiences vary drastically from individual to individual, so user patterns are heterogeneous between individuals and likely for individuals across their own timeline; thus, similar to other medical data, the Phendo self-tracked data are also heterogeneous, noisy, and sparse. A recent publication [123] details common engagement patterns for Phendo users and reveals potential biases in their self-tracking behaviors.

To address the particular challenges of these heterogeneous data and the complex, uncertain illness context, we rely on unsupervised probabilistic methods, similar to the approach taken by Urteaga and colleagues [27], who used data from the Phendo app for phenotyping individuals with endometriosis. This mixed-membership model is a specific type of generalized low-rank model, which is well-adapted to generating interpretable phenotypes. However, rather than aggregating an entire user's record for individual-level phenotyping, here, we extend this work to create temporal phenotypes, by aggregating each user's data by week. An interpretable, temporal representation (i.e., being able to represent an individual's timeline of health experiences as a dynamic mixture of phenotypes over different weeks) is likely to be suitable to the type of real-world interventions proposed, rather than aggregating all data from each user agnostic of time. In the context of this research, the phenotypic profile characterizes the health status or illness states (comprised of characteristics across domains of illness) at a particular time for a particular person.

4.2 Methods: Creating and evaluating health status phenotypes

4.2.1 Creating health status phenotypes

Data and cohort selection. We use data from the Phendo app, described in section 2.2.1. Our analysis is limited to those who meet specific criteria. Eligibility requirements include: self-reported diagnosis of endometriosis and at least one self-tracking entry (with a minimum of five data points). All available self-tracking data are aggregated by user-week. Each user-week is described by a vector of vectors, where domains/dimensions are represented as counts by specific item. A detailed description of the mapping is provided in Appendix section B.1 on page 185. Each domain represents a related set of tracking questions and responses, which maintains a distinct separation of illness experience variables and self-management variables.

Model. We use a mixed-membership probabilistic model, where a person's self-tracked data, aggregated by week, can be represented as a probabilistic mixture of phenotypes. Phenotypes can be described as a mixture of characteristics across domains of illness experience. The phenotypic profile (i.e., the mixture of phenotypes) provides a computable representation that characterizes the health status of an individual at a particular time.

Mixed-membership models are Bayesian generative models that have the ability to capture and model latent structures within collections of groups of data. A recognizable and commonly cited example of mixed-membership models is the topic model [188], which is useful for inferring latent topics within a corpora of documents using a probabilistic model. Topic modeling is a valuable tool to organize and summarize information, explore themes to discover new insights, and extract meaning from large volumes of text. The results are often intuitive, where the model commonly extracts concepts from a corpus that humans would identify, but does so without supervision. Results are interpretable and useful because the hidden structure that is inferred generally resembles the thematic structure of the collection of text. The generative statistical model enables sets of observations (words from documents in a corpora) to be explained by unobserved groups (topics) that explain why some parts of the data are similar. The model mathematically accounts for different words occurring in several topics, and allows for multiple topics to be represented in different proportions within a single document. Posterior inference with observed data is used to identify what are the likely generations that would have produced a particular document from a particular corpus.

Topic models represent a primary example of mixed-membership models — in our work, we reframe the components of the problem: instead of a corpora of text-based documents, our "documents" are a set of aggregate summaries of illness data self-tracked by an individual over the span of a week (each individual can contribute multiple "documents"), and our "words" are self-tracked signs and symptoms, treatments, and quality of life measures, which are generated from the "corpus" of data from endometriosis patients who self-track in Phendo. As in topic modeling, the mixture components (i.e., topics/phenotypes) are shared across the population, but the mixture proportions vary per user-week.

We extend the more basic type of mixed-membership model to accommodate multi-modal data, where each modality is a specific question with its own vocabulary, as in previous work [125, 27, 189]. Self-tracked variables from each user-week are used to train the mixture-model to learn a set of topics (phenotypes) that represent latent subgroups of health status/illness states experienced by individuals based on similar illness characteristics.

Model training and selection of hyperparameters. Development of the phenotypes is iterative. To determine the best hyperparameters to use for the phenotyping model, we experiment with held-out data that is split 80/20 train/test ratio with no crossover of participants between the training and test set, following the same method as [190]. To identify the optimal hyperparameters, we examine the models and compare the log-likelihoods (5-fold cross-validation). The hyperparameters are varied within these ranges: $K \in \{2, 3, 4, 5, 6, 7, 8, 9, 25\}$, $\alpha \in \{0.1, 0.01, 0.001\}$, and $\beta \in \{0.1, 0.01, 0.001\}$.

4.2.2 Phenotype validation

We validate the phenotypes in various ways, to ensure that they are appropriate and suitable for the intended tasks. We first examine the learned phenotype model to understand how it has characterized health status. We look at the vocabulary, visualized by heatmaps and wordclouds. We use these to make sense of what types of insights the mixed-membership model has learned and how it has characterized health status, e.g., has it learned differences in types of illness experience (gi-based vs pain-based) or has it learned differences based on severity of the illness experience that week. We compare the learned phenotype model to the baseline phenotype model, to see if there are differences in coverage. We also look at the correlation between the learned and baseline model, using the Pearson Correlation. Next, we look at the temporal dynamics of the learned phenotype model. We create a user timeline of phenotype assignments across user-weeks for each individual in the dataset. We look at the transitions between each of the phenotypes (i.e., how often a user-week stays the same phenotype from one week to the next vs how often it switches to a different phenotype), and create a state transition diagram to examine transitions between phenotypes across the population of users. We also plot each of the user timelines across time, separated by engagement levels (i.e., regular trackers, usual trackers, occasional trackers, and seldom trackers). Additionally, we plot the distribution of phenotypes across weeks where users tracked their menstrual cycle to see if there is alignment in phenotypes and menstruation.

The proposed phenotypes are then validated by researchers by examining associations between the different phenotype assignments and real-world self-tracked data. Each phenotype is visualized along with the data that were determined to be informative in the generation of the phenotypes. We use the Kruskal-Wallis test to identify if there are any significant differences between the phenotypes and the self-tracked data domains. We then use the Wilcoxon test to identify which pairs of phenotypes are significantly different from each other for each domain. This helps strengthen researcher confidence in the phenotypes before seeking evaluating them in a user study and with a task-based evaluation.

4.2.3 Phenotype evaluation

We evaluate the learned phenotype model in two ways, to assess how well the phenotypes might work in real-world scenarios. First, to ensure that these state space representations represent meaningful facets underlying the experience of illness, we consult with patients to validate the phenotypes. Second, to evaluate the performance of the phenotypes in real-world tasks, we also conduct a task-based evaluation with a computational model.

Evaluation metrics based on human user preferences. From other work[2] asking individuals with endometriosis about technologies to forecast flare-ups, participants expressed their preference for false positives (predicting a flare-up that does not happen) over false negatives (predicting no flare-up when in reality symptoms do worsen). As one person explained:

> *I'd rather have it tell me, 'I'm gonna have an awful day,' and I feel good, versus telling me 'I'm having a great day,' and then I don't. If it's telling me I'm gonna have an awful day, I'm not going to cancel everything for the day. It is more just trying to manage those symptoms and get the mental preparedness, whereas it could be a little discouraging, and maybe make you lose trust in the system a little bit, if you feel terrible, and it says, 'Nope, everything's great, things are fine.'*

There was broad consensus among participants of those focus groups that they would *"be more frustrated if it was like, 'You're fine,' and then I woke up and was flaring."* Participants in that

[2] These conversations are from focus groups following a pilot study with people with endometriosis, where we seek to develop voice-based technologies to forecast flare-ups. The manuscript detailing this work, titled "The Voice of Endo: Leveraging Speech for an Intelligent System That Can Forecast Illness Flare-ups" has been accepted for publication at CHI [191].

study also said that too many incorrect predictions, but especially too many false negatives, would diminish their trust in a system. According to conversations with those end-users, individuals are also more interested in when they will have flare-ups, rather than when they will have "good" days.

User evaluation of phenotypes

The most important evaluation of the learned phenotypes is through engaging human end-users and asking them to review, use, and provide feedback on the proposed phenotypes. For this, I conduct a two-part user study. In the first part, users are asked to evaluate the phenotypes using their real-world data. In the second part, participants are provided timelines with simulated data and are asked to evaluate the phenotypes through attempting several tasks. We refer to both the learned and baseline phenotypes as "AI-generated Health Status" when showing them to participants and discussing the phenotypes in the user study.

Cohort. For this study, we recruit and engage Phendo users with at least 6 weeks of self-tracked data. In order to enroll participants who have sufficient Phendo data to review for the study, we re-contact individuals who have consented to be recruited for research. All users are English-speaking, have been diagnosed with endometriosis, and are engaged in self-tracking their experience of illness (i.e., have logged at least 6 weeks of data, with at least 5 datapoints each, across at least 3 of the 4 phenotypes).

Study Design. The user evaluation consists of two parts — the first to evaluate phenotypes with users' real-world data to see if the health statuses align with their understanding of their illness; the second to evaluate the temporal component of the health statuses to see if they help individuals evaluate changes in health status over time. Participants are asked to think-aloud as they complete each task. Along with this, participants are asked to assess the difficulty and certainty, along with their agreement of the AI-generated health status. The survey used is printed in Appendix B.3.

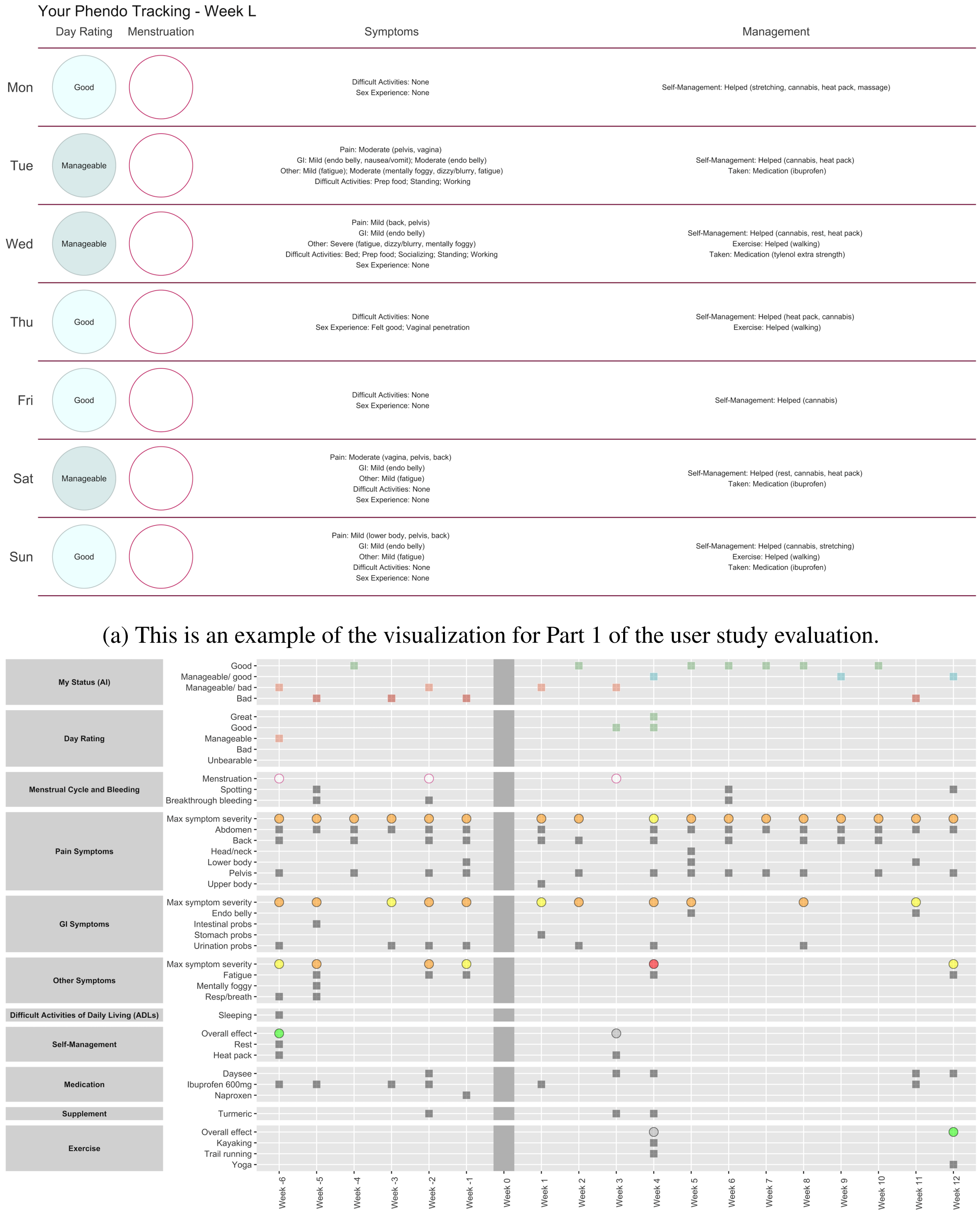
Part 1 — Static, non-temporal evaluation with user's real-world data. Individuals are presented with their data one week at a time (see an example in Fig 4.1a). First, they are asked to complete a primary assessment, where they assign the week to a health status among A (good), B (manageable/good), C (manageable/bad), and D (bad). We also ask them to assess (on a likert scale) how certain they are of their assignment and how easy or difficult it was to make the assignment. After they make their assignment and reflect on it, we will reveal the AI-generated health status and ask them about their agreement with the AI-generated health status (likert scale). We then ask them to complete a secondary assessment with that same data (now that they know the AI-generated health status), where they assign the week to a health status (A to D) and assess their level of certainty and perceived difficulty. We also collect qualitative data from the think-aloud component as participants complete their assessments.

For each of their 6 weeks of data, participants are asked to complete the primary and secondary assessments twice: (a) once with the learned phenotypes, and (b) once with the baseline phenotypes. The order of the instances (weeks of data) and the condition (learned/baseline phenotypes) shown to participants are randomized and counterbalanced. We consider patient perspectives of their own experiences as the "gold standard" in this evaluation.

Part 2 — Dynamic, temporal evaluation with simulated data and tasks. Participants are presented with simulated data and asked to evaluate health status over time for three different patient tasks. These tasks include: (1) evaluate health status pre/post surgery; (2) evaluate self-management strategies; and (3) prepare for a clinical visit. The data shown to users is based on real user data, but is modified so that the "ground truth" is known. E.g., for the case where users are asked to evaluate surgery, the data is augmented so that some symptoms are improved following surgery, see Fig 4.1b; in the case of preparing for a clinical visit we have extracted the patient's report of going to their provider from the journal entry on that day:

> *Had Dr. Appt with new Gyno, pelvic exam caused a lot of pain. Had low back pain. They want to do pelvic floor therapy soon and a laparoscopy. [...] So frustrated, not believed on pain because I wasn't having ovarian pain today and I guess I wasn't really sick today for the exam so was told their other patients are usually sicker and more chronic.*

This enables us to assess how well participants are able to complete tasks (i.e., we can evaluate if



Your Phendo Tracking - Week L

	Day Rating	Menstruation	Symptoms	Management
Mon	Good		Difficult Activities: None Sex Experience: None	Self-Management: Helped (stretching, cannabis, heat pack, massage)
Tue	Manageable		Pain: Moderate (pelvis, vagina) GI: Mild (endo belly, nausea/vomit); Moderate (endo belly) Other: Mild (fatigue); Moderate (mentally foggy, dizzy/blurry, fatigue) Difficult Activities: Prep food; Standing; Working	Self-Management: Helped (cannabis, heat pack) Taken: Medication (ibuprofen)
Wed	Manageable		Pain: Mild (back, pelvis) GI: Mild (endo belly) Other: Severe (fatigue, dizzy/blurry, mentally foggy) Difficult Activities: Bed; Prep food; Socializing; Standing; Working Sex Experience: None	Self-Management: Helped (cannabis, rest, heat pack) Exercise: Helped (walking) Taken: Medication (tylenol extra strength)
Thu	Good		Difficult Activities: None Sex Experience: Felt good; Vaginal penetration	Self-Management: Helped (heat pack, cannabis) Exercise: Helped (walking)
Fri	Good		Difficult Activities: None Sex Experience: None	Self-Management: Helped (cannabis)
Sat	Manageable		Pain: Moderate (vagina, pelvis, back) GI: Mild (endo belly) Other: Mild (fatigue) Difficult Activities: None Sex Experience: None	Self-Management: Helped (rest, cannabis, heat pack) Taken: Medication (ibuprofen)
Sun	Good		Pain: Mild (lower body, pelvis, back) GI: Mild (endo belly) Other: Mild (fatigue) Difficult Activities: None Sex Experience: None	Self-Management: Helped (cannabis, stretching) Exercise: Helped (walking) Taken: Medication (ibuprofen)

(a) This is an example of the visualization for Part 1 of the user study evaluation.

(b) This is an example of the visualization for Part 2 of the user study evaluation. This shows the task of evaluating if an individual has improved health status pre/post surgery. In this example, the AI-generated health statuses (i.e., learned phenotypes) are shown on the top row.

Figure 4.1: Example visualizations for the user study evaluation of the learned phenotypes.

they are “correct”).

For each of the three tasks (i.e., surgery, self-management, and clinical summary), participants are asked to complete their assessment task twice: (a) once with just the raw data without the AI-generated health status shown, and (b) once with both the raw data and the AI-generated health status over time (i.e., learned phenotypes) shown (see the top row in Fig 4.1b). Similar to part 1, in their primary assessment, participants evaluate the health status over time (this time, qualitatively), and assess their certainty and difficulty. Then, the row of AI-generated health statuses are revealed on the timeline and participants discuss and assess their agreement with the learned phenotypes. Finally, they are asked to complete their secondary assessment where they evaluate the health status over time and assess their certainty and difficulty in making their evaluations.

Task-based evaluation of phenotypes

We assess if the health status phenotypes are useful in a real-world computational task. Specifically, we see if we can use self-tracked data to forecast flare-ups (i.e., predict the health status phenotype in the next week, using this week’s data).

Data. The feature vector consists of each user’s self-tracked data, aggregated to the week-level. The outcome vector consists of the subsequent week’s phenotype, and we experiment with both the learned and baseline models, for both the 4-class problem and binary 2-class problem (grouping the two worst phenotypes as a flare-up, and the two best phenotypes as non-flare-up). For the final dataset, only the top 12 most dense user-weeks (according to both the feature and outcome vectors) are selected for each individual. We split the final dataset 80/20 train/test ration, with the same assignments as the phenotyping experiments (i.e., if an individual was in the training set for the phenotype experiments, they will also be in the training set for the computational evaluation task experiments). The computational task experiments are carried out across three datasets: (1) Baseline phenotypes; (2) Learned phenotypes, limited to only cases where that user-week also has a Baseline phenotype (i.e., the Baseline phenotype is not missing); and (3) Learned phenotypes with all data (i.e., even if the user-week has a missing Baseline phenotype).

Models. A variety of models are tested in the analysis, including Logistic Regression, Gradient Boosting, Random Forest, and Decision Tree to evaluate the effectiveness of these features in predicting future phenotypes. The goal is to assess how different models perform in capturing the relationships between self-tracked health data and subsequent phenotypic health states.

Evaluation. We examine multiple evaluation metrics across models. In selecting metrics to rely on for evaluation, we seek to align with user priorities — i.e., to prioritize detecting flare-ups and minimize false negatives. In this case, the best metric to use is the F2 score, which prioritizes recall over precision. Further, we will want to prioritize identifying the worst health status phenotypes. For this reason, we focus on the F2 score for the worst phenotype (i.e., "Flare-up"). We also look at the AUROC and AUPRC for each model.

4.3 Results: Temporal health status phenotypes

4.3.1 Dataset for phenotyping experiments

Data from the Phendo app were used for this analysis. All questions from all users were downloaded, and then data were preprocessed to facilitate the analysis. This analysis was conducted at the week-level, so the unit of analysis for this study has been set to: *user-week*. Each user's data were aggregated to the week-level and treated as independent records. Records with fewer than 5 data points have been excluded from the analysis. Self-tracked data were mapped to meaningful domains and specific responses. All symptoms (pain, GI, and other symptoms) have all been mapped to the domain of "symptoms," while the severity of these symptoms has been left separate. Foods and exercises have also been normalized to a discrete set of options. The full mapping is presented in Appendix B.1.

The final dataset for phenotyping includes data from a total of n = 11,852 users. Users in the dataset have tracked an average of 4.3 weeks. The dataset includes data from 51,187 user-weeks. User-weeks in the dataset include an average of 52.6 moments. Summary statistics for the user-weeks included in the analysis are shown in Table 4.1.

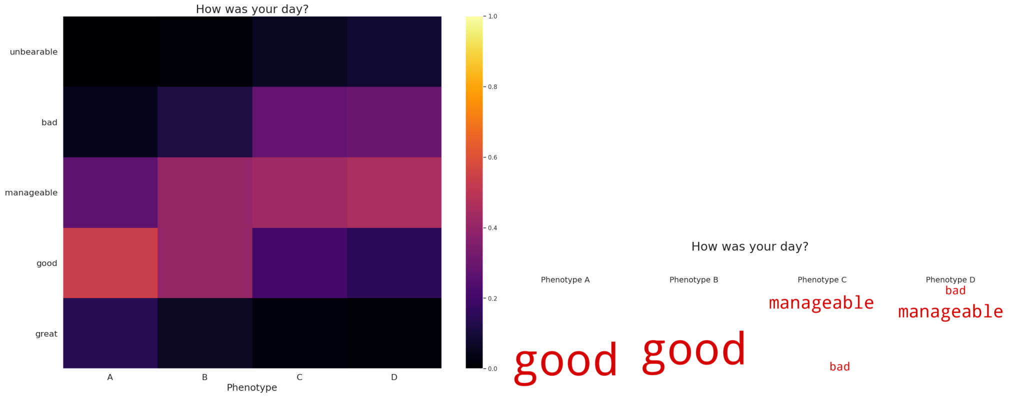
Question	**# of Obs**			**# Tracked Days**		
	mean	*std dev*	*max*	*mean*	*std dev*	*max*
How was your day?	3	2.5	26	3	2.3	7
What symptoms are you experiencing?	17	28.2	941	3	1.7	7
How severe is the pain?	4	5.9	244	2	1.6	7
How severe is the GI symptom?	3	3.1	58	2	1.5	7
How severe is the other symptom?	3	2.9	90	2	1.4	7
What exercise did you do?	6	5.0	54	3	2.2	7
Did you experience negative effects from exercise?	2	2.0	12	2	1.9	7
Did you experience positive effects from exercise?	3	1.9	12	2	1.9	7
What food did you eat?	8	9.1	123	3	2.2	7
Did you experience negative effects from food?	3	2.0	19	3	1.9	7
Did you experience positive effects from food?	3	2.1	14	3	2.0	7
What is your period flow?	4	3.5	37	3	2.2	7
What kind of bleeding?	2	2.3	47	2	1.2	7
What activities were hard to do?	12	14.7	249	3	2.0	7
How was sex?	4	3.4	27	2	1.8	7
What did you do to self-manage?	6	6.4	58	3	2.1	7
Was self-management effective?	3	2.2	16	3	2.1	7
Did you take any medication or supplements?	9	14.5	239	3	2.2	7
Total	**53**	**63.7**	**1199**	**4**	**2.3**	**7**

Table 4.1: Summary statistics for the user-weeks of self-tracked data included in the Phenotyping dataset.

4.3.2 Phenotyping experiments and final model

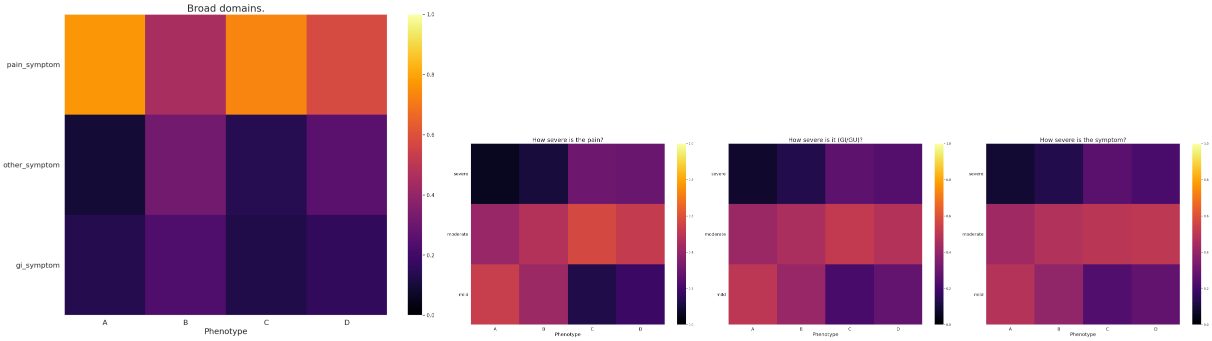
Selecting K. First, to select K, models were run across all candidate number of topics K using sparse parameters (a = b = 0.001). The best run from each of the experiments was inspected to evaluate how many topics appear to be represented in the data. K = 3, K = 4, and K = 5 resulted in the most meaningful models and were inspected in more detail. More than 5 topics had redundant phenotypes, and fewer than 3 lost nuance and detail. It was determined that K = 4 was the optimal number of topics to select for the final model, since more topics did not capture discriminating features or offer different insights.

Candidate models. Experiments were run across all hyperparameters at K = 4. A total of 9 experiments (with 10 runs each) were conducted and examined. The experiments run across different hyperparameters were robust at K = 4, with very consistent characterization of the four learned phenotypes.



(a) Heatmap — ‘How was your day?’ (b) Word Cloud — ‘How was your day?’

Figure 4.2: Learned Phenotype Model: ‘How was your day?’ (Phenotypes shown across the x-axis in order of A, B, C, D, and Daily Rating is shown on the y-axis with unbearable at the top and great on the bottom.)



(a) ‘What symptoms are you experiencing?’ (top-*pain*; middle-*other*; bottom-*GI*)

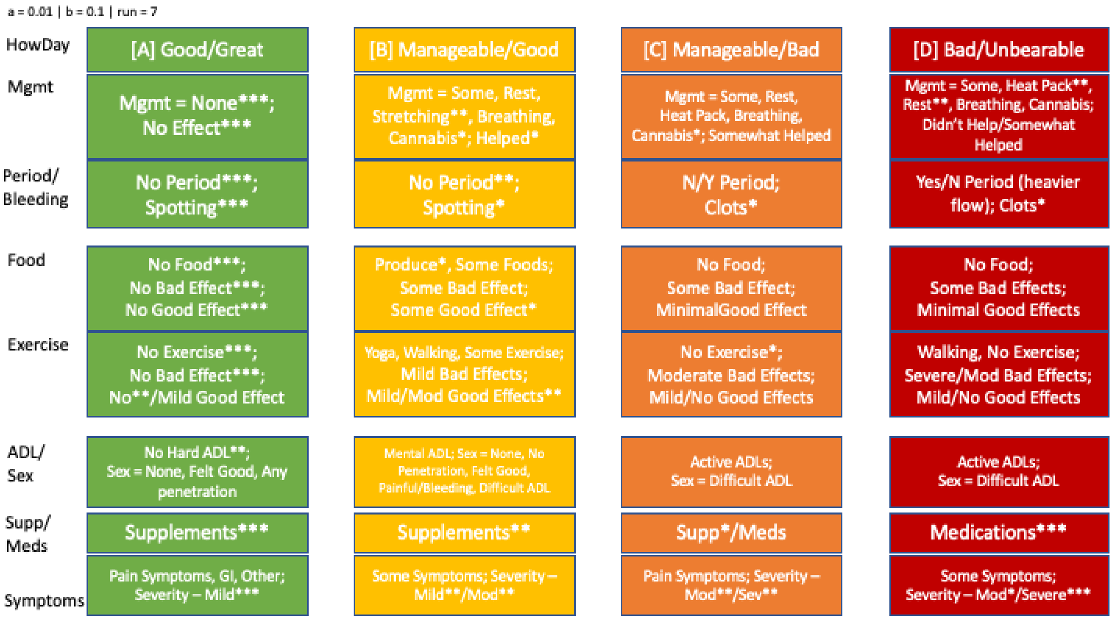
(b) ‘How severe is the pain?’ (top-*severe*; middle-*moderate*; bottom-*mild*)

(c) ‘How severe is the GI symptom?’ (top-*severe*; middle-*moderate*; bottom-*mild*)

(d) ‘How severe is the other symptom?’ (top-*severe*; middle-*moderate*; bottom-*mild*)

Figure 4.3: Learned Phenotype Model: Symptoms and Severity

Final model. The final model has K = 4 topics, with hyperparameters of a = 0.01 and b = 0.1. A heatmap and wordcloud of the 4 phenotypes for “How was your day?” is shown here in Fig 4.2 and just the heatmaps are shown for symptoms and symptom severity in Fig 4.3. The heatmap visualizations show the per-question probability distributions of each phenotype, with each heatmap representing the likelihood of the responses within each domain for each phenotype. For instance, the response ‘good’ is highly likely to be tracked under phenotype A and not likely to be tracked under phenotype D (pink versus purple) in Fig 4.2a. The wordcloud visualizations show an alternate representation, with the font size of each response reflecting its likelihood to be tracked within the phenotype; in these, only the most discriminative features are shown. Together, the heatmaps and wordclouds present a visual representation of the learned phenotype model. The heatmaps and wordclouds for all other domains are shown in Appendix B.2. An overview of discriminating features for this model is presented in Fig 4.4.



a = 0.01 | b = 0.1 | run = 7

HowDay	[A] Good/Great	[B] Manageable/Good	[C] Manageable/Bad	[D] Bad/Unbearable
Mgmt	Mgmt = None***; No Effect***	Mgmt = Some, Rest, Stretching**, Breathing, Cannabis*; Helped*	Mgmt = Some, Rest, Heat Pack, Breathing, Cannabis*; Somewhat Helped	Mgmt = Some, Heat Pack**, Rest**, Breathing, Cannabis; Didn't Help/Somewhat Helped
Period/ Bleeding	No Period***; Spotting***	No Period**; Spotting*	N/Y Period; Clots*	Yes/N Period (heavier flow); Clots*
Food	No Food***; No Bad Effect***; No Good Effect***	Produce*, Some Foods; Some Bad Effect; Some Good Effect*	No Food; Some Bad Effect; MinimalGood Effect	No Food; Some Bad Effects; Minimal Good Effects
Exercise	No Exercise***; No Bad Effect***; No**/Mild Good Effect	Yoga, Walking, Some Exercise; Mild Bad Effects; Mild/Mod Good Effects**	No Exercise*; Moderate Bad Effects; Mild/No Good Effects	Walking, No Exercise; Severe/Mod Bad Effects; Mild/No Good Effects
ADL/ Sex	No Hard ADL**; Sex = None, Felt Good, Any penetration	Mental ADL; Sex = None, No Penetration, Felt Good, Painful/Bleeding, Difficult ADL	Active ADLs; Sex = Difficult ADL	Active ADLs; Sex = Difficult ADL
Supp/ Meds	Supplements***	Supplements**	Supp*/Meds	Medications***
Symptoms	Pain Symptoms, GI, Other; Severity – Mild***	Some Symptoms; Severity – Mild**/Mod**	Pain Symptoms; Severity – Mod**/Sev**	Some Symptoms; Severity – Mod*/Severe***

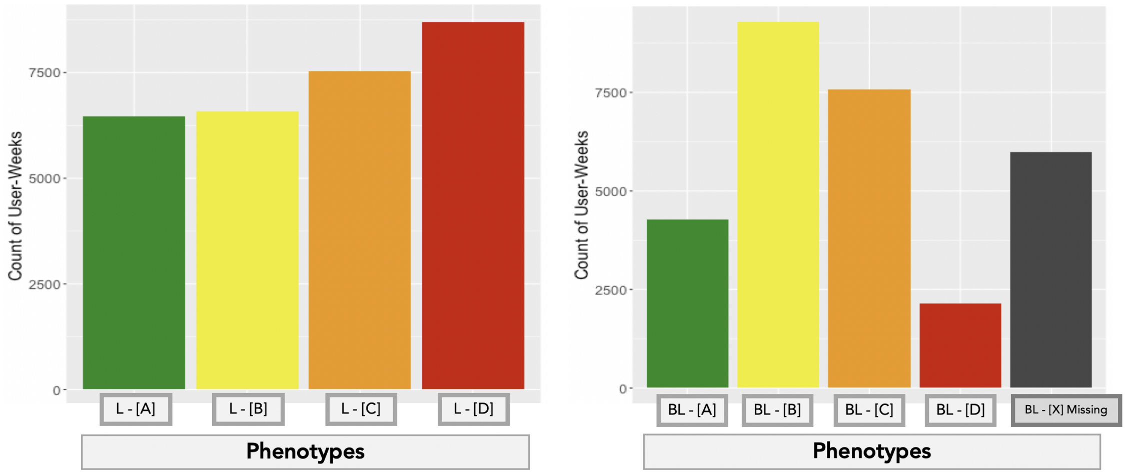
Figure 4.4: Learned Phenotype Model: Overview of Phenotypes

The mixed-membership model learned a largely severity-based representation of health statuses. This was not a given, and the model could have learned various other dynamics instead (e.g., based on body system or type of management that was helpful). This severity-based characterization was consistent across all K topics and a/b hyperparameters. This robustness provides confidence in the model, and that a severity-based phenotype is the most appropriate approach. It also helps us to learn about the underlying population and experience of illness, which will be explored in the following sub-section.

Comparing to baseline model. A baseline model was constructed using rules based on the 'How was your day?' Phendo question. Several sets of rules were explored, including using the most frequently tracked response and the 'worst' response of the user-week. The responses 'Good' and ‘Great’ were combined so that the baseline model resulted in 4 phenotypes, to align with the learned model. These rules allowed for soft assignments, similar to the learned model, as well as a single hard assignment from these soft assignments. This single indicator within the Phendo app is useful to broadly understand a user’s health condition. It is also a frequently tracked feature in the data. But, there are still substantial weeks missing this indicator data, resulting in substantial ‘missing’ assignments for the baseline model, as seen in Fig 4.5b (i.e., the NoHowDayData phenotype). For part of the validation, plots of the baseline model are presented alongside plots of the learned model.

4.3.3 Validating the learned phenotype model

Coverage and missingness. In the final learned model, each of the four phenotypes is well-represented in the data, with some variation. The baseline model, on the other hand, is more skewed towards some phenotypes with less representation of others (i.e., the best and worst phenotypes have fewer assignments in the dataset). The baseline model also has a lot of ‘missing’ assignments that are not missing in the learned model, due to the ‘How was your day?’ question not being answered. These ‘missing’ assignments could belong to any of the phenotypes, but since they did not track that single variable, they do not have a health status assignment according to the baseline



(a) Learned Phenotypes

(b) Baseline Phenotypes

Figure 4.5: Phenotype Model: Distribution of Phenotypes

model. Further, individuals that track any of the ‘How was your day?’ responses could be having a better or worse health experience than that one variable can capture. The learned phenotype takes all of the self-tracked variables into account, so it gives a richer and more meaningful picture of what is happening.

Comparing the learned and baseline models directly in Fig 4.6, it is clear that there is a slight correlation between the learned and baseline models (shown in the bold boxes), but the associations are not large. The highest correlation is between the ‘best’ learned and baseline phenotype, with a coefficient of 0.25. Each of the other phenotypes has a positive correlation between the learned and baseline phenotypes, and there is a negative association between the ‘best’ and ‘worst’ phenotypes, which is to be expected and a good sign. So while the models do correlate, the learned model does not neatly align with the ‘How was your day?’ variable. Thus, the learned phenotypes are useful to characterize what is going on with individuals and their health, and provide richness and

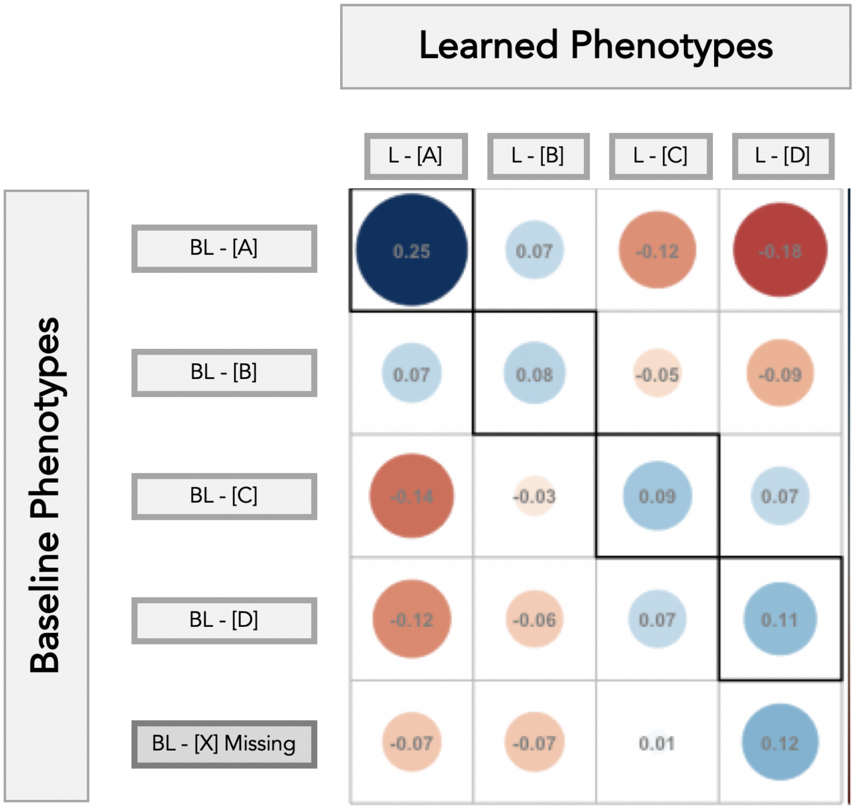


Figure 4.6: Correlation Between Learned and Baseline Phenotype Models

nuance. For the records without a baseline phenotype assignment (i.e., there was no 'How was your day?' tracked), we can see a slight negative association with the 'best' learned phenotype and a slight positive association with the 'worst' learned phenotype, suggesting that records with missing information may be having a flare-up or a bad week of symptoms. This information would be lost if only using the baseline phenotype assignments.

Temporal dynamics. We have also created user timelines of all user-weeks from an individual. Table 4.2 shows a summary of each user's record, with the proportion of how frequently users 'stay' or 'switch' among different phenotypes within their timeline and also the proportion that each phenotype is assigned across the timeline. This shows that individual users are not assigned to a single phenotype across their entire timeline (i.e., the model is not learning user-level dynamics), and that there is variation in phenotype assignments. This is true for both the learned and baseline model, and mirrors what is shown at the population-level in Fig 4.5. This table shows wide heterogeneity across each individual's timeline and across different users.

	Learned Phenotypes			**Baseline Phenotypes**		
	mean	*std dev*	*max*	*mean*	*std dev*	*max*
Stay	0.2	0.3	0.99	0.2	0.3	0.99
Switch	0.8	0.3	1.00	0.8	0.3	1.00
Phenotype A	0.2	0.3	1.00	0.2	0.3	1.00
Phenotype B	0.2	0.3	1.00	0.3	0.4	1.00
Phenotype C	0.2	0.4	1.00	0.2	0.3	1.00
Phenotype D	0.2	0.4	1.00	0.1	0.2	1.00
No *'How Was Your Day'* Data				0.2	0.9	1.00
No Tracking Data	0.1	0.3	0.99	0.1	0.2	0.99

Table 4.2: Summary of each user's temporal record. The top portion of the table shows how often an individual stays at the same phenotype from week-to-week vs switches to a different phenotype. The bottom of the table shows the proportion of assignments to each phenotype across an individual user's timeline.

The next visualization, Fig 4.7, plots the weekly assignment of learned phenotypes for users according to their engagement levels (based on prior research [123]). This plot suggests that the model is capturing temporal patterns, since each individual is not assigned to the same phenotype across their entire timeline. These different dynamics can help us to learn about the experience of disease, and can give us insights about our population. We can see heterogeneous patterns across different users' timelines. There are some cyclical patterns, some individuals swap back and forth between different phenotypes, and some have periods of good phenotypes followed by periods of not-so-good phenotypes. This shows something we already know, that the experience of illness is highly individualized. The learned phenotypes help us to capture and represent this, and could

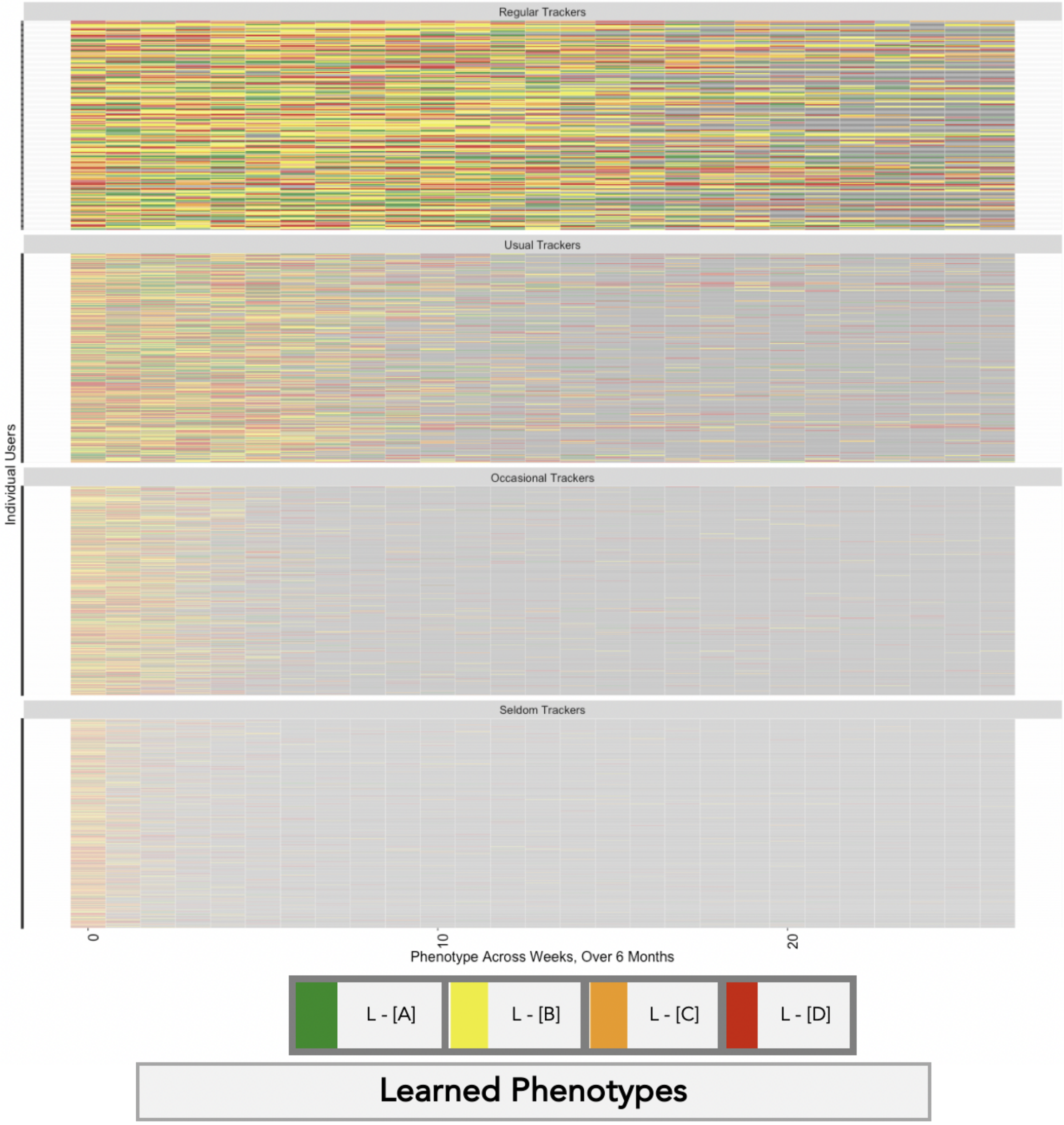


Figure 4.7: Learned Phenotype Model: Temporal Plot of Phenotypes Across Engagement Levels

potentially help us to tailor interventions or treatments to these different dynamics.

When examining what else the phenotypes might be associated with, to ensure that we are not just learning patterns that could be computed directly, we see in Fig 4.8 that the learned phenotype model does not align clearly with weeks where users have their menstrual period. This gives us further insights into the population of users — bad weeks are not only during menstrual periods.

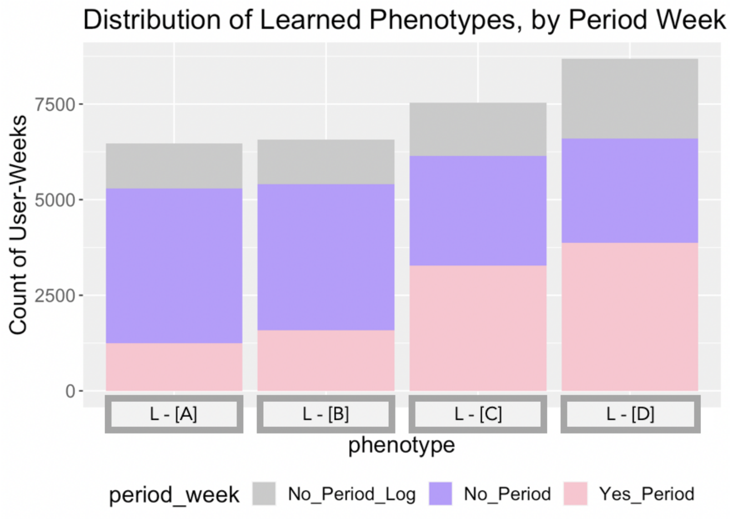
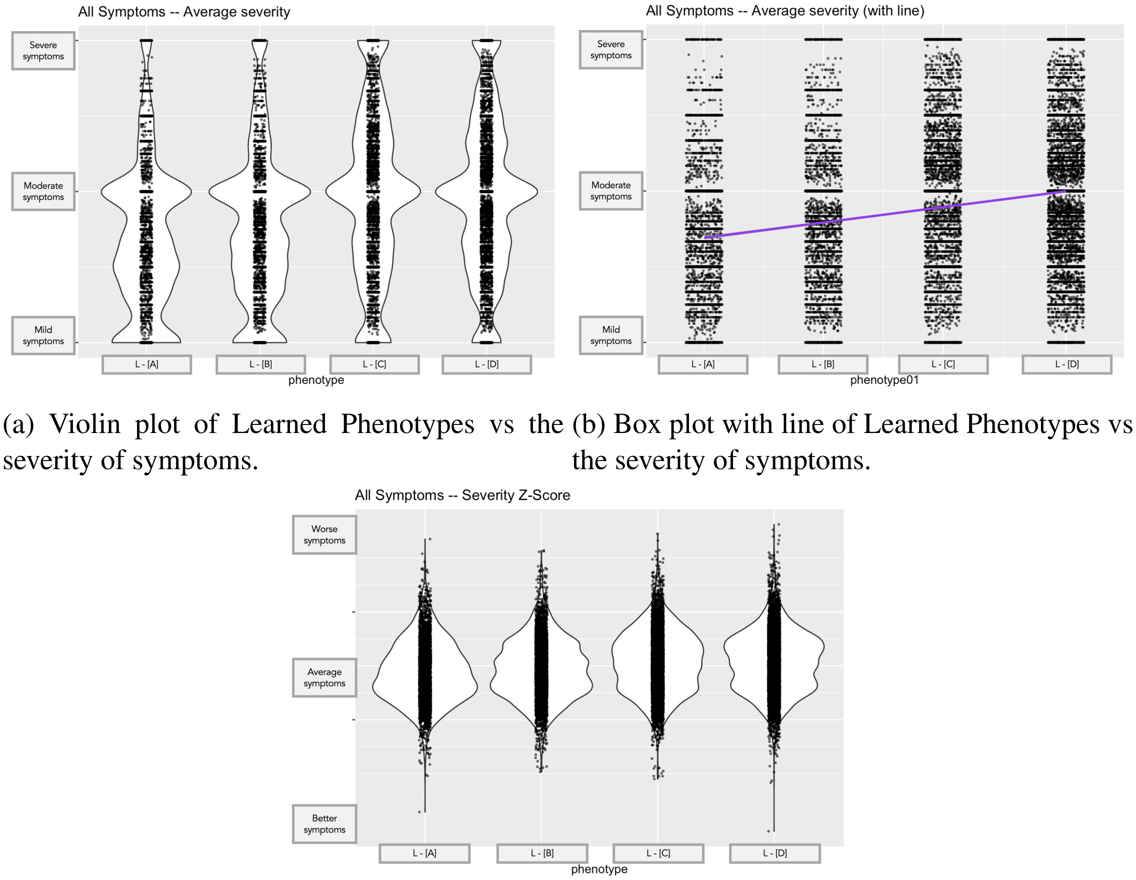


Figure 4.8: Phenotype Model: Distribution of Learned Phenotypes with Menstrual Cycles

Examining user-week phenotype assignments against self-tracked data. In order to strengthen researcher confidence in the proposed phenotype models, we have aggregated Phendo data across the phenotype assignments and assessed differences across phenotypes for each of the self-tracked domains. In this analysis, we have used the learned model to infer weekly phenotypes of the self-tracked data, and then compared how the phenotypes relate to the data tracked that week.



(a) Violin plot of Learned Phenotypes vs the severity of symptoms.

(b) Box plot with line of Learned Phenotypes vs the severity of symptoms.

(c) Z-scores of Learned Phenotypes vs the severity of symptoms.

Figure 4.9: Plotting the Learned Phenotypes vs the severity of symptoms across user-weeks.

	Phenotype A	Phenotype B	Phenotype C
Phenotype B	< 0.001	—	—
Phenotype C	< 0.001	< 0.001	—
Phenotype D	< 0.001	< 0.001	0.34

Table 4.3: Wilcoxon test for data shown in Fig 4.9a.

In Fig 4.9, we can see clear associations that the “Good” phenotype is associated with less severe symptoms (across pain, gi, and other symptoms), while the “Bad” phenotype is associated with more severe symptoms. Fig 4.9a and Fig 4.9b show the distribution of severity scores for user-weeks that have been assigned each of the phenotypes. The violin plot emphasizes the distributions, while the best-fit line in the center panel shows the trend of better phenotype and less severe symptoms vs worse phenotype and more severe symptoms. The Kruskal-Wallis test (chi squared value = 1424.6, df = 3, p-value < 0.001) shows that there is a significant difference across each of the phenotypes. The Wilcoxcon test (see Table 4.3) identifies which pairs of phenotypes are significantly different — all of them show a statistically significant difference from one another except for [C] and [D], which fail to show a statistically significant difference. Fig 4.9c shows the phenotypes vs severity z-scores for each individual. This plots the difference between an individual's average severity score and what severity was tracked that week. You can see that phenotype [A] (Good) has z-scores that indicate "better" severity scores, while phenotype [D] (Bad) has z-scores that indicate "worse" severity scores. We performed these comparisons across a range of tracked data, which all show similar patterns. This "gut check" helps us to confirm that the learned phenotypes are capturing variations in real-world data.

4.3.4 Evaluating the learned phenotype model

User evaluation

Cohort. We recruited N = 5 individuals who previously used the Phendo app to track their experience of illness (minimum 6 weeks of data), to evaluate the phenotypes. Their demographics are shown in Table 4.4. While the sample lacks diversity in race and ethnicity, there is a range of illness experiences represented among the research participants.

Here, we present the results of the user study. First, we present the quantitative results from part 1; next, we present the quantitative results from part 2; finally, we present the qualitative results from both parts of the study.

Quantitative results from part 1 - assignment of health status by week. In Part 1, participants reviewed their own self-tracked data and made assignments from "best" to "worst" across the phenotypes A - Good, B - Manageable/Good, C - Manageable/Bad, and D - Bad (for both the learned and baseline phenotypes — in this section, referred to as "AI-generated health statuses"), giving data for n = 30 instances for each learned and baseline phenotyping model. For this part of the study, participants evaluated their health status assignment, certainty, and difficulty, both before

Table 4.4: Participant demographics for evaluation user study (N = 5).

Age	
Mean (SD)	40 (8.5)
Median	41
Range	30-50
Gender	**n (%)**
Woman or Female	5 (100)
Race	**n (%)**
Hispanic	1 (20)
White	5 (100)
Relationship Status	**n (%)**
Married or domestic relationship	3 (60)
Single, never married	2 (40)
Highest Level of Education	**n (%)**
College +	5 (100)
Income	
Mean (SD)	72k (13k)
Median	65k
Range	58-90k
Employment Status	**n (%)**
Employed	4 (80)
Not employed	1 (20)
Living Environment	**n (%)**
Suburban	2 (40)
Urban	3 (60)
Years Diagnosed	**n (%)**
Less than 5	2 (40)
5 to 10	1 (20)
10 or more	2 (40)
Any periods, past 3 months	**n (%)**
Yes	4 (80)

Note: Participants could select more than one race/ethnicity; race and ethnicity were asked together. Categories with no responses have been omitted from the table (Black, Asian, Native American, Other).

and after knowing the AI-generated health status (i.e., in primary and secondary assessments).

Participants' assignments matched the Learned Phenotypes 23% of the time (n = 7), and matched the Baseline Phenotypes 33% of the time (n = 10). We performed significance testing to evaluate if there was a statistically significant difference in how participants matched or did not match the learned and baseline assignments. Neither a Pearson's Chi-squared test nor a Fisher's Exact test indicated there was any statistical difference in performance across the learned and baseline models. Figure 4.10 shows the count of instances where the phenotype assignment (AI-generated health status) matched the user's assignment, by learned vs baseline phenotype, for the primary assessment.

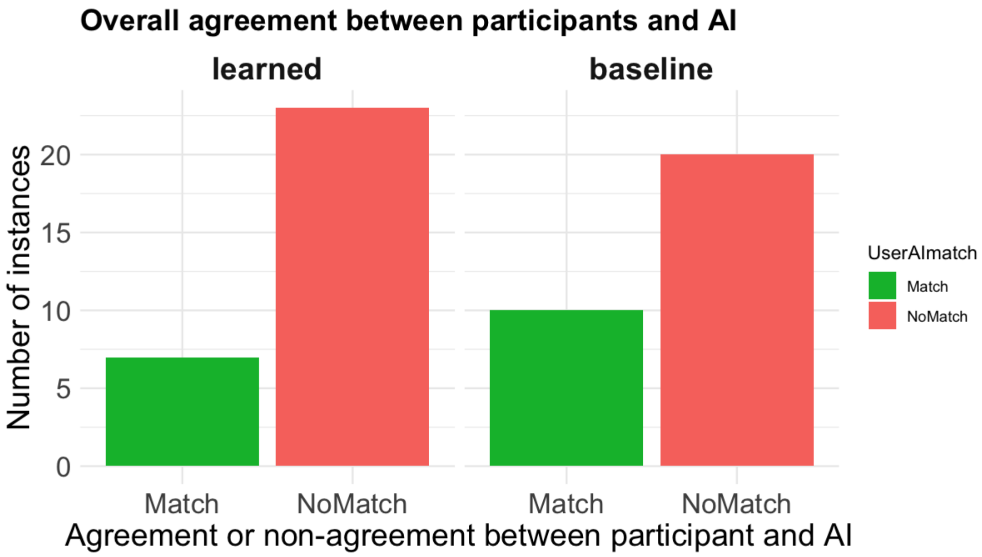
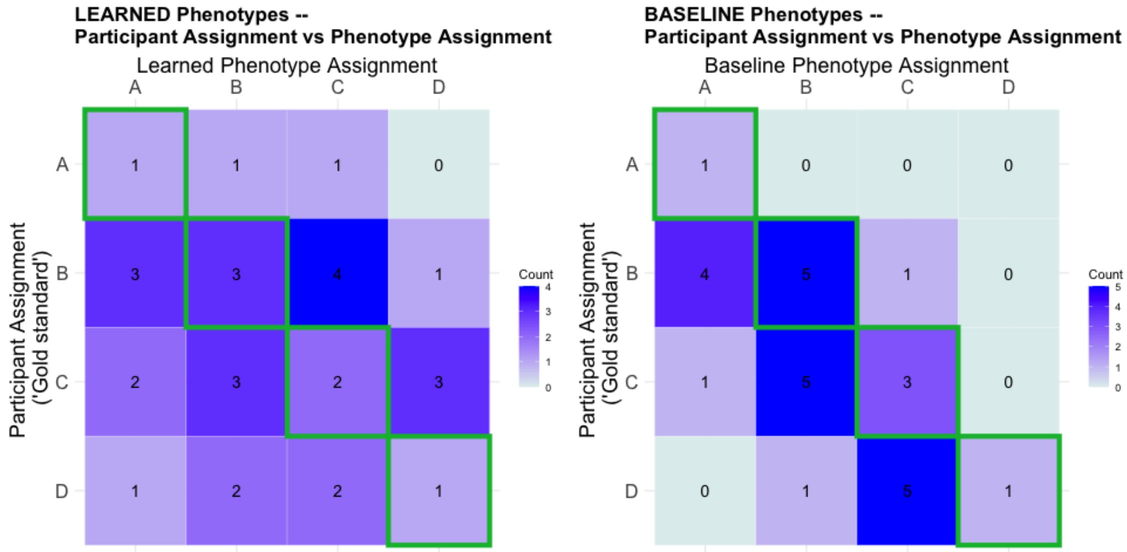


Figure 4.10: Bar chart depicting the frequency of matches between the AI-generated health status and user assignment (for the primary assessment), by the learned and baseline phenotype models.



(a) Heatmap of the **learned** phenotype assignments vs the user assignments

(b) Heatmap of the **baseline** phenotype assignments vs the user assignments

Figure 4.11: AI-generated health status vs user assignments (for the primary assessment), across the learned and baseline phenotypes. The “matched” boxes (where both the phenotype assignment and the user assignment match) are outlined in green.

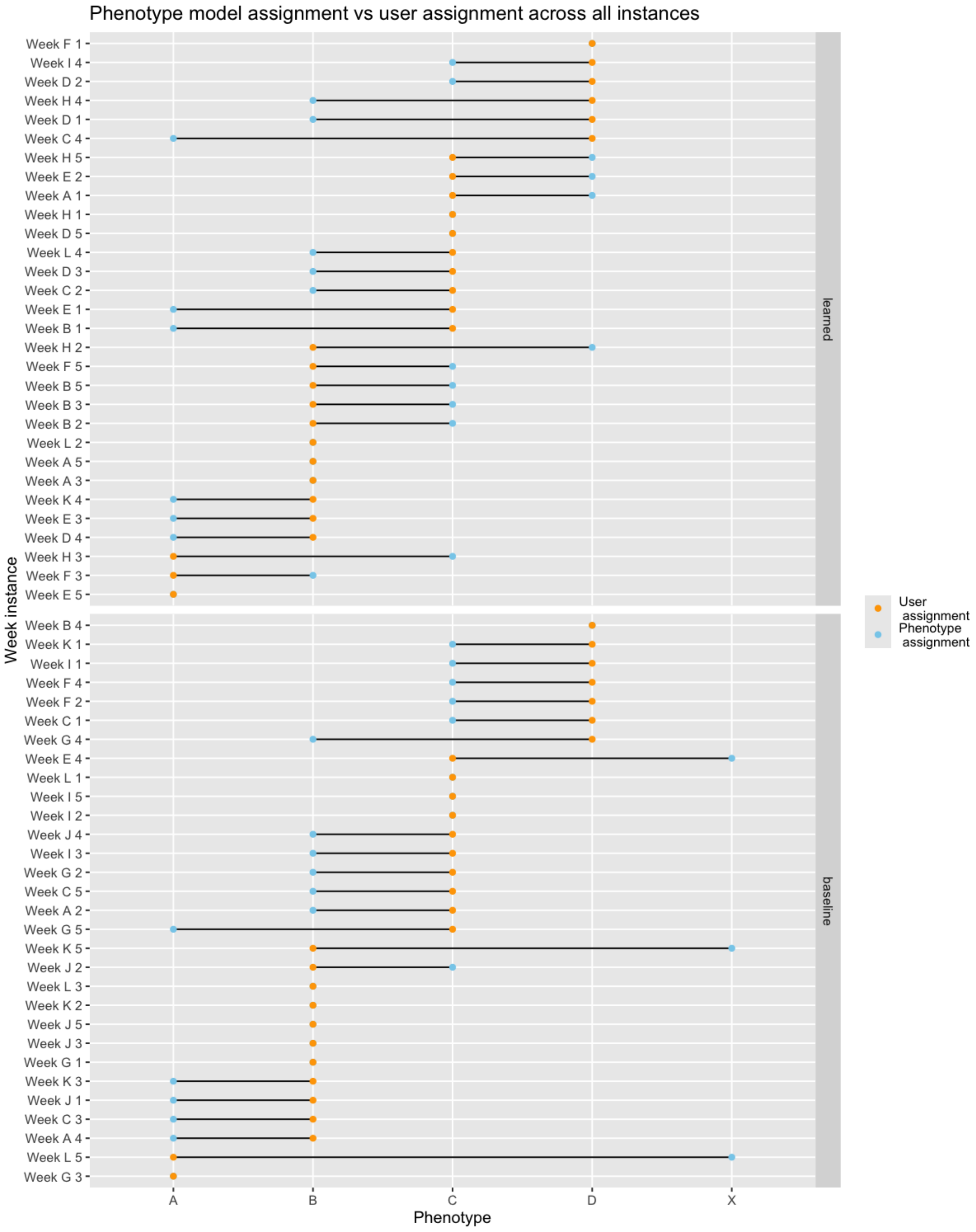


Figure 4.12: All AI-generated health status vs user assignments (for the primary assessment), with the learned on top and baseline on bottom. The line shows the difference between the model and user assignments; if there is no line, the assignment matched across the model and user assignments.

The plots in Figure 4.11 give details about the phenotype model assignments vs the user assignments across the learned and baseline phenotype models. With so many user assignments in Fig 4.11b under the black boxes (where assignments "match"), the baseline model tended to underestimate individual's health status (i.e., among disagreements, in 16 cases, the AI-generated health status was better than the user's rating, while in only 1 case was it worse). While the learned model sometimes under-estimated the severity of individuals' health statuses (see the assignments under the black boxes in Fig 4.11a, incorrect assignments made by the learned model were more likely to be more severe than less severe, when compared to the users' own assessments of their health status (i.e., in 13 cases the AI-generated health status was better than the user's rating, while in 10 cases the AI-generated health status was worse than the user's rated health status). This aligns with the user's preference for the AI to err on the side of assigning a status that is worse than reality, rather than assigning one that is better than reality.

Figure 4.12 shows the differences in each assignment between the AI-generated health status (phenotype assignments) and the assignments by users, where the user assignment is shown in orange and the AI-generated health status is shown in blue. If the orange dot is to the right of the blue dot, the participant rated the week as worse than the AI-generated health status. If the blue dot is to the right of the orange dot, the AI-generated health status rated the week as worse than the participant's assignment. If the orange dot is on top of the blue dot, the ratings matched. The learned phenotypes are shown on the top half, and the baseline phenotypes are shown on the bottom half of the plot.

Results from across the two assignments where we showed participants the same data suggest limited test-retest reliability. It is notable that individuals were somewhat inconsistent with their assignments. Across all users and instances, individuals made inconsistent assignments 33% of the time. Some of them even acknowledged that they had changed their mind from the first time they had seen that week of data. The consistent and inconsistent assignments are shown in Table 4.5.

Despite their lack of consistency, participants rated themselves as relatively certain and felt the task was somewhat easy. Across all of the participants, results on the perceived difficulty of

Consistent Assignments (n = 21)		
Category	**n**	**Percentage (%)**
Health Status A (Good)	1	5
Health Status B (Manageable/Good)	7	33
Health Status C (Manageable/Bad)	7	33
Health Status D (Bad)	6	29
Inconsistent Assignments (n = 9)		
Category	**n**	**Percentage (%)**
Good ↔ Manageable/Good	3	33
Manageable/Good ↔ Manageable/Bad	5	56
Manageable/Bad ↔ Bad	1	11

Table 4.5: Test-retest reliability. Consistent and inconsistent assignments made by participants (for the primary assessment) across the two times they saw the same week of data.

the task and their self-rated certainty range quite a bit. Their ratings for difficulty and certainty, after making their primary assessments is shown in Fig 4.13. Notably, no participants selected “Very Hard” or “Not at all Certain” for any of the data that they viewed, before the AI-generated health status was revealed (i.e., during the primary assessment). In general, the easier a participant reported an assignment was to make, the more certain they were about the assignment, and the harder they reported making the assignment, the less certain they were.

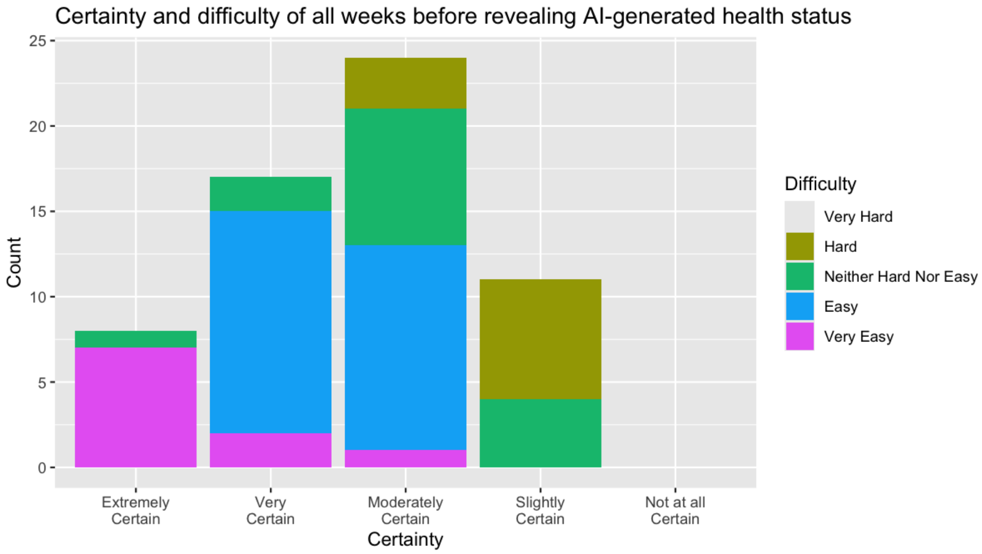
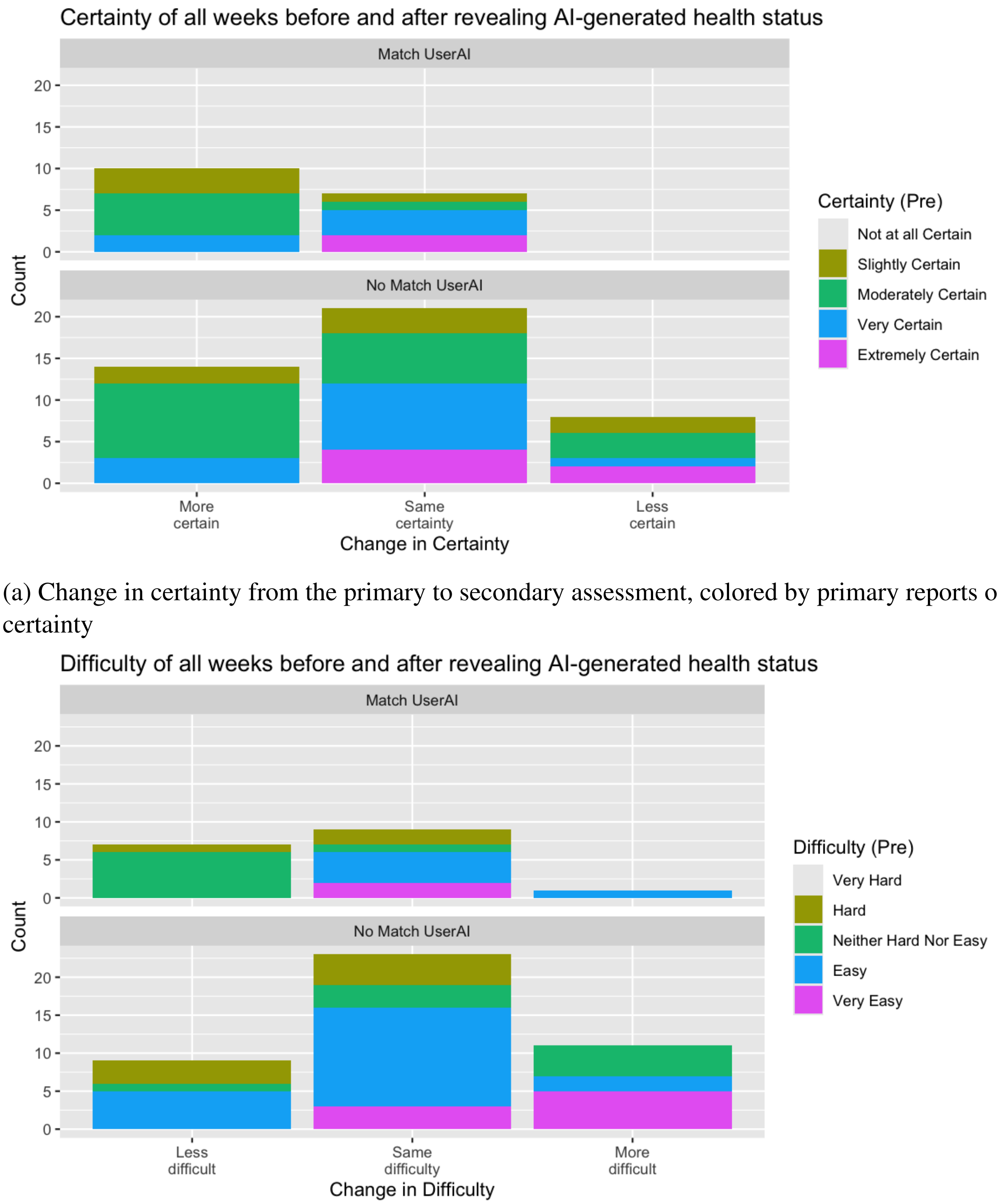


Figure 4.13: User-rated degree of difficulty and certainty of health status assignments during the primary assessment, before being told the AI-generated health status.



(a) Change in certainty from the primary to secondary assessment, colored by primary reports of certainty

(b) Change in difficulty from the primary to secondary assessment, colored by primary reports of difficulty

Figure 4.14: Change in certainty and difficulty from participants making their primary assessments to after they know the AI-generated health status (secondary assessments), colored by their primary reports. Results for the baseline model are shown on the left, and the learned model on the right. The top row of each plot shows the instances where users matched the AI, and the bottom row shows where users and the AI disagreed.

At the same time, very few participants changed their Health Status assignments after finding out what the AI-generated health assignment was (i.e., from the primary to the secondary assessment), despite the frequent discrepancy between their assignment and that of the phenotype model. In 93% of cases (n = 56), individuals kept the same assignment. In only 7% of cases (n = 4) did participants change their assignments — one from Health Status A (Good) to Health Status B (Manageable/Good); two from Health Status B (Manageable/Good) to Health Status C (Manageable/Bad), and one from Health Status D (Bad) to Health Status C (Manageable/Bad).

In Fig 4.14, we visualize the certainty and difficulty of users making their primary assignment and the change in certainty and difficulty after they know the AI-generated health status for part 1 of the user study. In these plots, there is no clear change in difficulty or certainty across instances. But, there are some interesting patterns. The plot on the top, Fig 4.14a, shows the change in certainty for users across all assignments. Across the baseline and learned models, participants generally maintained the same certainty or became more certain when their assessment matched the AI. When participants and the AI disagreed, certainty decreased for participants and self-reports shifted more towards "less certain," especially for those who initially reported high certainty. Alignment with AI predictions reinforces participant confidence, while disagreement reduces certainty. The plot on the bottom, Fig 4.14b, shows the change in difficulty for users across all assignments. Across the baseline and learned models, participants mostly reported no change in difficulty or found it less difficult when their assessments matched the AI. Participants tended to perceive the task as more difficult when their judgments conflicted with the AI, especially those who initially rated making that week's assignment as "very easy," which is more pronounced for the learned model. Participants reported the task as harder to assess health status when their judgment contradicted the AI's assignment of health status. Even still, sometimes when there was misalignment between user and AI assignment, users still found their assessment to be easier and with more certainty, which is somewhat surprising.

Quantitative results from part 2 - evaluation of health status over time. In Part 2, participants evaluated patient data to evaluate if they could assess any changes over time for patients who have (1) undergone surgery; (2) experimented with self-management; and (3) are preparing for a clinical visit. Just visualizing the data without the AI-generated health statuses (phenotypes), participants were somewhat successful in evaluating the patient's health status over time. They felt it was somewhat difficult, and where somewhat uncertain, see Fig 4.15. When they were provided with the AI-generated health status, their ability to assess changes over time improved, and they were quicker to make their assessments. However, it was not necessarily easier, and they were not necessarily more certain about their assessments, see Fig 4.16.

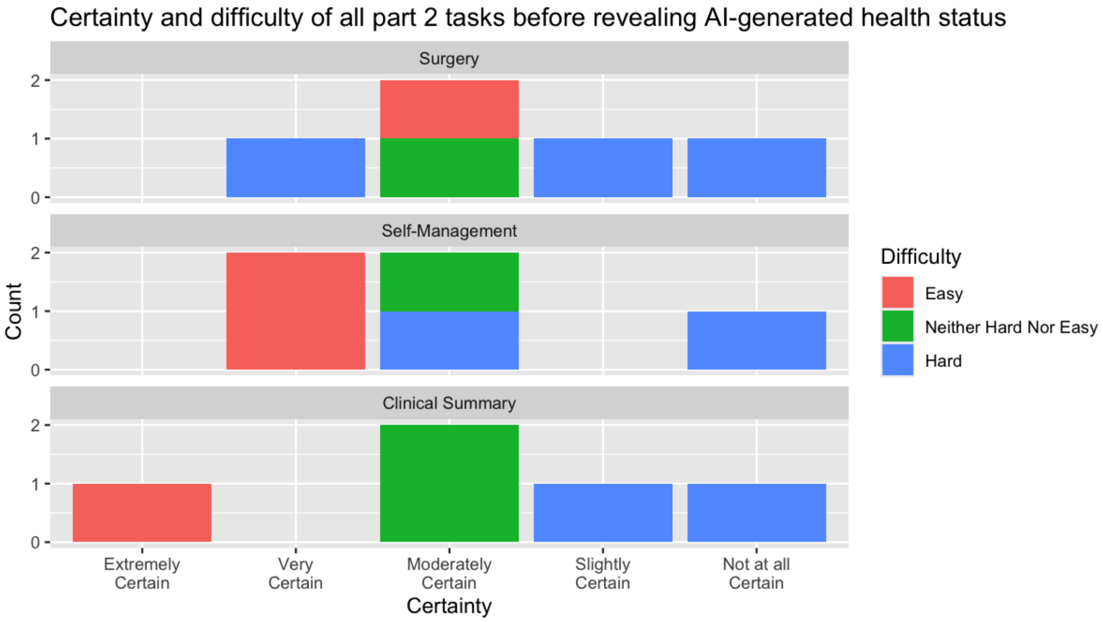
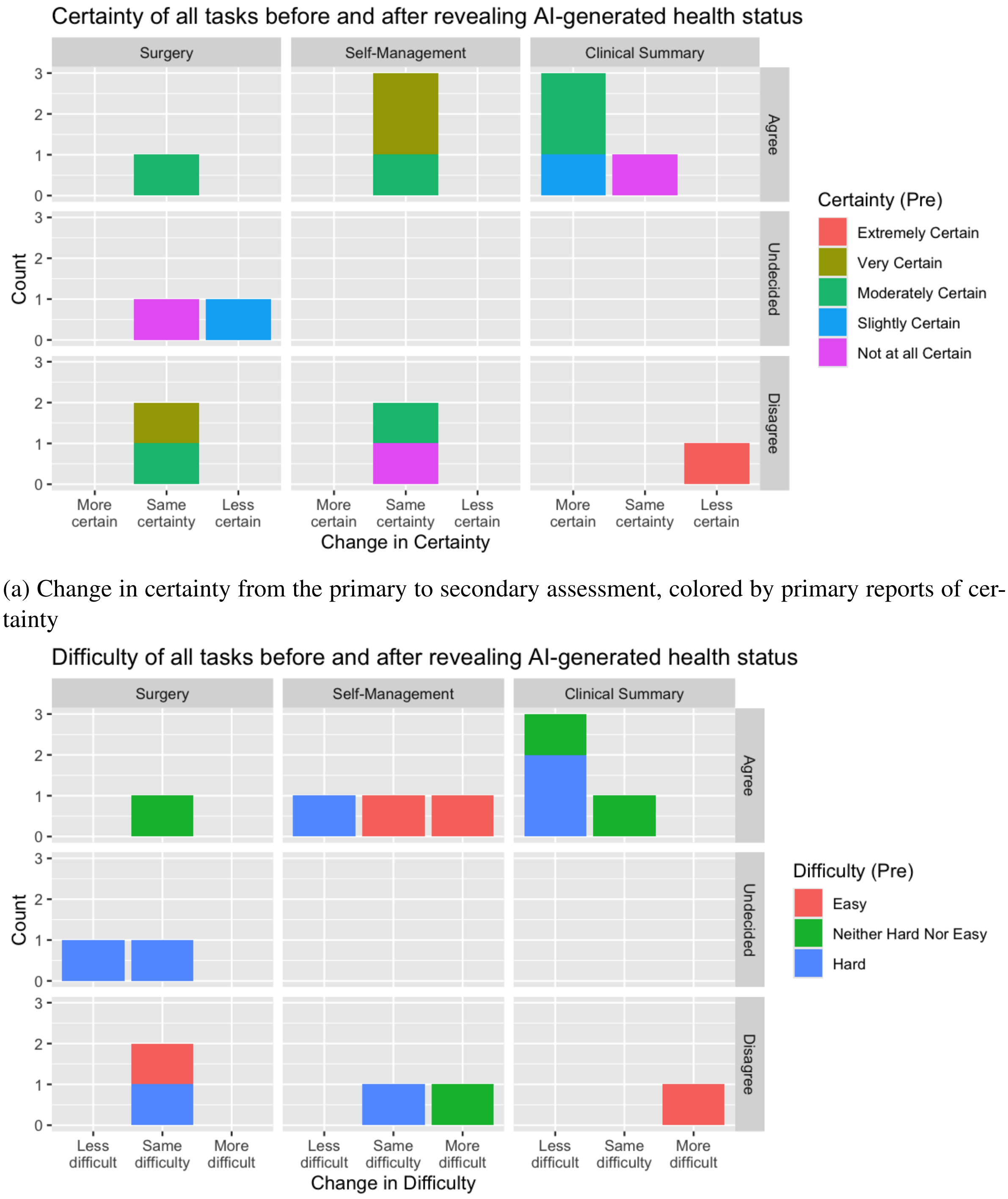


Figure 4.15: User-rated degree of difficulty and certainty of their assessments of health status over time for part 2 before being told the AI-generated health status (i.e., the primary assessment).



(a) Change in certainty from the primary to secondary assessment, colored by primary reports of certainty

(b) Change in difficulty from the primary to secondary assessment, colored by primary reports of difficulty

Figure 4.16: Change in certainty and difficulty from participants making their primary assessments to after they know the AI-generated health status (secondary assessments), for all tasks of part 2. The rows of the plot indicate the participant's level of agreement or disagreement with the AI-generated health status reports.

To contextualize all of these findings, we turn to the insights from the qualitative data captured during the study visits.

Qualitative results - insights from think-aloud discussions with participants across parts 1 and 2. Participants talked about their thought processes in making the health status assignments, how difficult or easy it was to make their assignments, how certain they were about their assignments, their agreement or disagreement with the AI-generated health status, and how their assignment, difficulty, and certainty changed after having access to the AI-generated health statuses. In Part 1, participants talked about using a range of indicators to assess their health status, in combination. This mirrors what the learned phenotype relied on to make the assignment, more than the baseline phenotype, which used only the single Day Rating indicator. In Part 2 participants described using the same cues, such as symptom severity and use of medication, but noted that it was difficult because each person with endometriosis experiences it differently. They commented on challenges they experienced because of the complexity of the data they were reviewing. When they were provided with the AI-generated health status, their ability to assess changes over time improved and they felt it was easier, and were more certain about their assignments. Across both parts, the results related to how hard or easy the task was, and how certain the person was on their assignment was not clear.

Indicators used to make health status assignment. Even though we only talked to a small sample of participants, the individuals that we spoke to described an extremely wide range of indicators that they used to assess their health status in making their assignments from the self-tracked data. At the same time, none of the indicators described were unique to any individual. Each participant described using some indicators more than others, weighted them differently, interpreted them differently, and used them to varying degrees to make their assignments. Participants used many of these things together to make their assessments of health status. They talked about each of these indicators helping to nudge their judgment of their health status towards better or towards worse, rather than relying on any indicator to make an absolute judgment.

All participants talked about using their self-tracked **Day Rating** to some degree — one person in particular used this to a strong degree (*"I'm just looking at my day ratings and I'm going to say the week was manageable/good. Gosh, it's like, that feels like such a rare evaluation of a week."* (HS3)), while others used it as one indicator in concert of the others. They also talked about their impressions of how they made those assessments of "How was your day?" to begin with. Multiple people talked about rating their days (and sometimes their symptom severity, too) as better than they probably were feeling at the time (e.g., *"Yeah, manageable [Day Rating]. I know for me, manageable is not necessarily good, or just, I can deal with it."* (HS4) and *"This is definitely during that time when I was using moderate when it was actually severe."* (HS2)).

Participants talked about using their **Symptoms** to make assessments — the presence/absence of symptoms, the specific symptoms logged, the frequency of symptoms, and the severity of symptoms. For participants, no symptoms logged were indicative of a better health status while the presence of symptoms was seen as contributing to a worse health status. If a participant perceived that they had a logged a lot of symptoms, especially a wide range of symptoms, then they interpreted it as a worse week. Sometimes, specific symptoms were seen as a worse health status. For example, pelvic pain was seen by one participant as a symptom indicative of a worse health status. The more severe the symptoms, the worse participants ranked their health status (e.g., *"This is only moderate pain versus severe pain. So that would be a huge difference. And I take pain well, so moderate isn't going to kill me. Where it's severe pain, standing up is hard."* (HS1). As a further indicator of the impact of these symptoms on their health status, many people talked about impaired activities of daily living (ADLs). Many participants talked about experiencing difficulties with activities as being an indicator that they were having a bad week (e.g., *"Where it says difficult activities, bed, dressing, prep, food, sitting, those are days where it takes me hours to get myself to the kitchen and make a little bit of food. So I know this was a very bad week"* (HS2)). At the same time, seeing evidence of difficult activities did not always mean it was considered a bad week by a participant, e.g., *"It's also interesting to see, okay, what's my baseline? You know, it's like sitting for a long time always hurts. And that doesn't necessarily mean that it's a terrible week because it always* *hurts."* (HS3). **Menstruation** and bleeding were also indicators used by a few individuals, one person relied strongly on her cycle to make judgments (e.g., *"I started my period, and it got very difficult, which is standard for me. The beginning of my period is usually worse than the end of it."* (HS5)). Others didn't menstruate or didn't use it as an indicator at all.

Participants also talked about aspects of **Self-management** as giving them insight into their health status. People talked about seeing logs of self-management strategies used often indicate that they were not having a good week (e.g., *"I'm seeing a lot of self-management, which I know, usually for me indicates that it's kind of a bad time."* (HS4)). A few people mentioned specific strategies giving some insight (e.g., *"So if I was using cannabis, that means I was really having a hard time."* (HS3) and *"I used a heat pack and breathing exercises and rest. So that means I didn't get out of bed."* (HS5)). Further, several participants talked about if they logged that the self-management strategies were helpful indicated a better (but not good) week, while "no effect" suggested a worse week (e.g., *"Seeing more self-management too but it looks like a lot of it helped so this one I would say is probably more of the manageable/good."* (HS4) vs *"I'm seeing a lot of the self-management where it had no effect at all doing those things — and taking more medication, but just not feeling much better so I'd say bad, health status D."* (HS4)). Additionally, logging pain medications was seen as a strong indicator of a bad week (*"I was doing the Tramadol, which is my primary rescue pain medication."* (HS5)). On the other hand, one participant explained that when she saw only her "regular medications and supplements," she knew it was a good week (e.g., *"This one should be good because it's just my normal medication."* (HS1). Very few people talked about foods, exercise, or sex experiences although a few participants mentioned seeing data related to having sex and doing exercises suggesting that they were having a pretty good week (e.g., *"I did core exercises. That's really good. I can see that I was very functional as a body."* (HS3); *"But I see that in exercise, I was able to do things so that that does align with the good. So I would say manageable/good."* (HS3); *"I was having sex in there. So to me, that's usually an indicator that it was somewhat manageable."* (HS4)), or food suggesting a bad week (e.g., *"Some of the food that I listed, while it might not have had any effect in the short term, I was charting some patterns* *of eating that I know are usually driven by my endo symptoms... Kind of like survival mode. So, because of that, I would put this at a D."* (HS5)).

Several people talked about using tracking patterns to infer how they were feeling that week, although they also talked about how complicated that was (e.g., *"Like, was I just feeling so good that I decided I didn't need to track or was I feeling so awful that I didn't track?"* (HS4)). Tracking something was seen as strongly indicative because they "felt the need to note" something good or bad (e.g., *"If I was feeling good enough to be like, I need to make a note of this, that was kind of rare as well."* (HS4) vs *"And I still was like, this is so bad that I need to record it."* (HS5)). Breaks in tracking were seen as potentially being due to them feeling either good or bad.

Aspects impacting the difficulty of making health assessments. Weeks with very little data were hard for participants to make assessments of. They also had a hard time when they felt that different days of the week would give different assessments, or may sway a week's health status towards better or worse than their overall assessment may otherwise be (e.g., *"It's hard when, one day is really bad, but the rest of the days are not"* (HS1)). Another explained:

> *There's a part of me that wants to say, C, because there's a majority of days where I was probably doing all right. But there's the other part of me that's like, Sunday was completely non-functional, like I couldn't human that day. So does that sort of cut it off at the legs for the health status of the week? This one is harder. I'm kind of torn between C and D.* (HS2)

One of the reasons individuals speculated that they likely report that their health status is worse than what would be assigned relying only on the Day Rating is that individuals tend to minimize their pain and also do not like to log that their experience was too bad. For example when asked why she clicked manageable in the app but then assessed her health status as worse than that, one participant explained:

> *Yeah I think I try not to but sometimes charting that things are really hard feels dramatic and so then I feel almost even though it's a private app, I shouldn't feel guilty*

> *about it, but I feel like, Oh it's not that hard I'm just being dramatic. There are other people that have it harder, I can't use the hardest part of the app, where it's like this is the absolute worst, like I can't touch that because what if it is worse, what if there is a worse and I just haven't been there yet and other people are there and I'm skewing the data by being dramatic.* (HS5)

Several people talked about checking their sense of reality, and report that they often under-report their experiences:

> *Reading the symptoms, difficult activities. The fact that I put manageable there is funny too, because I remember, over time, getting more and more used to Phendo being like, if I'm having trouble standing, that's not manageable, like stop putting that.* (HS2)
>
> *Just a lot of difficulty every day. I'm sure at that point I was like, it's fine. In hindsight, I'm like, you know, that doesn't really seem very fine.* (HS4)

Making sense of disagreements between participants and AI assessments. When there was a high degree of disagreement, we asked the individuals to explain. One person explained, *"I wonder if that [disagreement, where the model assigned a worse health status] isn't a reflection of what I'm used to, and I'm used to bad. So bad isn't that bad to me."* (HS2). Others would say that they could see where the AI was coming from, or what maybe made the AI make an assignment one way or another.

While many participants changed their assessment from one viewing of the data to the next, very few changed their minds when their assignments did not match the AI's. Most people relied on their own assessments. (e.g., *"I can see why [the AI made the assignment, that didn't match mine], especially, the fact that I was able to have sex in there, that's usually a good sign. But also, I was having issues with doing everything, it looks like, on some of those days. Yeah, the activities of daily living on Sunday and Monday. Yeah, I disagree."* (HS4). A few did question their own assessments:

> *It feels like calibrating your sense of reality. Like with the whole experience of Endo where you're like, oh, other people aren't experiencing this, I didn't realize that everyone wasn't in this kind of pain every month. So, I'm a little on the fence, I know it gets worse than this week. So in that sense, no, I'm not wrong. But in that other sense, if anybody else who isn't used to having endometriosis was experiencing the same thing, it would probably be bad for them. So that's hard. That's a hard one.* (HS2)

One person did knowingly change her health status assignment from the first time she saw the data to the second, explaining:

> *I'm actually kind of on the fence about this one. I think I might go with B and be a little closer to what the AI was thinking last time. [...] So we were the same this time. It's interesting. I mean, I'm not surprised, I know people aren't perfectly consistent, so I'm not surprised that I sort of adapted and changed my interpretation from the first time I saw it. So I guess I would say that I agree with the AI. I'm going to say agree just because I had some uncertainty about it. To strongly agree would be a little bit to me, like overconfident.* (HS2)

When they disagreed, participants often emphasized the aspects of their data that strongly suggested to them the health status that they assigned, like using pain medications or recording severe pain. When the AI over-estimated their health status, they explained: *"I don't think an A-good week rating should be given to a week with this many days where pain is noted."* (HS5) and *"I disagree, mostly because of the number of days that I had a level of pain at all that was pathologic in nature."* (HS5)

Assessing health status over time. Participants were largely able to provide accurate assessments of health status over time, even without access to the AI-generated health statuses. But they had some uncertainty about it and it took a long time for them to review the detailed self-tracked data. They also commented that it was harder to evaluate another patient, rather than using their own self-tracked data (e.g., *"It's really harder looking at somebody else's data, so that was hard."*

(HS5)). After the AI-generated health statuses were revealed to participants, they generally had an easier and faster time in making their evaluations.

Participants had the most difficulty with assessing the health status change pre-post surgery (e.g., *"That's tough to discern.... That's really tough to say if it's improved or not. I'm not sure."* (HS3)), which was the most ambiguous case. They also had the most difficulty and were the least certain making this assessment. Participants analyzed that there was not a ton of symptom improvement pre-post surgery (e.g., *"But then, the severity looks pretty much exactly the same. And still symptoms in all of the same areas."* (HS4)). A few participants commented on the reduction in medication usage (e.g., *"I see a gap in medication usage in a stretch after surgery"* (HS5)) and a few symptoms (e.g., urination problems), which could suggest improvement. One participant took nearly ten minutes to review the case and found very small details that explained the person's experience, for example there is a day with elevated fatigue symptoms, but it is also a day where the patient went kayaking and trail running. This person was uncertain about the AI's assessment of improvement, aligning with most of the other participant's assessments as well. Several people also commented on how difficult it is after surgery to notice improvement due to the challenges of recovering from the surgery itself (e.g., *"I'm using my own experience as after surgery, you don't feel great immediately, and you have to go through the healing process, which initially, the first couple months, you might not feel better."* (HS3)).

Participants had a relatively easy time assessing the impact of self-management (e.g., *"So without even looking at the actual meaning of the data yet, those pre- and post, all three of them, the post column, it looks a little lighter in terms of what's being reported. So that's just an early indication to me that maybe it's been helpful. So now I want to look at the details."* (HS2)), and found the AI-generated health status agreed with their assessments (e.g., *"So having those, I feel like this one actually is really helpful with the health statuses because the first two, I think I still agree, it was a pretty dramatic improvement. We have the major decrease in the difficulty with the daily activities, symptom severity has either gone down or they're just not having symptoms anymore. So those are a lot clearer. And then for the last one, I do think it's a slight improvement."* (HS4)).

Participants also benefited from having three examples to review to evaluate if the management strategy has a positive impact for the person under evaluation.

Participants had a pretty accurate assessment of the clinical summary case, when compared with the patient's own self-report from their journal entry. When asked what they would tell a doctor, participants said: *"I would say that I was having, obviously, more bad days than good days, but some of them were manageable."* (HS1) and, *"A lot of hard days, a lot of bad days. And some intermittent bleeding, symptom severity, kind of an increase of symptoms going into the visit, that would have been a really hard couple of days before that. And yeah, some good days where things were helping at the beginning, but then between week one and five, and then weeks eight, nine, ten, eleven, looks like a lot of really hard days and pain in a lot of places."* (HS5). The AI-generated health status gave participants confidence in their own assessments. After viewing the AI-generated health statuses, one participant said: *"Oh, Okay. Yeah, that looks like what I would think. I would definitely tell [the doctor] that this person is experiencing a very life-disrupting amount of pain and suffering."* (HS2) Another explained how having access to these health statuses would give her confidence in her own assessment:

> *If I had this, I would feel a little more confident being like, overall it's manageable, but in a bad way, or just maybe bad overall. I definitely have the tendency to be like, oh, it's totally fine because you're just kind of used to it because it's chronic. And then when you look back at the data and you're like, oh, I had four days last week I didn't get out of bed, and it kind of validates that. So with this, I think it would be easier to say, no, things were actually pretty bad.* (HS4).

Imagined real-world use of AI-generated health statuses. Participants were all optimistic about using such a technology if it were available to them. They felt it would help them in assessing their own health status, although they viewed themselves as the experts and would want control over the AI's behavior and outputs. For example, one participant explains that such a tool could help her get an overall sense of her health status and monitor her medication usage, potentially enabling her to link this data to the symptoms that prompted her to take the medications.

> *I think it would give me personally an overall sense of what my days were like and also weeks were like and also how much medication I took. Because for me, I have multiple medications that I take, and I can only take so many per certain amount of time. So it would be helpful for me to see how severe I was and what symptoms I had that made me take medication.* (HS1)

Another person described the tool as a sounding board to reflect and give a reference point for her own experiences.

> *I would definitely use it. And I think that it's a really good, kind of like a sounding board, it's some kind of reflection where you compare your own assessment to the AI and go, that's very wrong, but somehow it helps give a reference point.* (HS3)

Another person emphasized that she would not take the AI's health assessment over her own, but also explained that she imagined feeling validation from a tool helping her reflect on her burdensome symptoms and know that they are real.

> *I would definitely use them. I don't think I would necessarily take it over my work, like if it's telling me, you're doing great. And I'm like, no, I don't feel like I'm doing great. I'm not gonna be like, well, the AI told me I'm good. But I think the bigger part would be that validation piece of like, okay, I actually am not doing well or, kind of checking in with getting used to those things. And if your normal is feeling like you're constantly, I don't know, like at a seven on a pain scale out of 10, I think it's easy to get caught up in that and be like, well, that's just my life. So I would definitely use the tool to really just kind of validate that and say, okay, I'm not crazy. I really have been having bad symptoms or there have been changes in how I've been feeling.* (HS4)

Participants also described potential value in taking the health status reports to their healthcare providers. One person talks about providing her doctor a report, and especially sending data ahead of time so the doctor could review it before her appointment:

> *The migraine apps that I use, you can send a report of all the information to your doctor. Because I know my doctors appreciate if I write stuff down that I give them a copy as well as me having a copy. So I guess you could do that with the printout from the AI or whatever. But I think sometimes sending something ahead of time, they're able to glance through it before you even come in.* (HS1)

Another participant talks about benefits in helping her prepare for a doctor visit and figure out what to report. She talks about the value of a summary to save herself effort and frustration, and also to validate her experiences to affirm that her symptoms are not made up.

> *It would be a useful tool, especially for conveying to doctors the consistency of pain and life disrupting symptoms. I absolutely think it would be very helpful, especially because it's a lot of information and being able to boil it down to sort of a stoplight, kind of red-yellow-green. I've definitely found myself sitting in a doctor's office kind of having this same sort of thinking out loud, like okay pros and cons, this got a little bit better that got a little bit worse, how it all weighs together. Trying to grapple with worse or better and how bad is bad, I'd like if there's a health status summary that I can feel confident in its accuracy, that would be an incredibly valuable thing because it would just save me so much analysis and humming and hawing and trying to make sure that I'm reporting things accurately, and yeah I think it would be a really really useful tool. Also that little part that's like, oh no, it's not your imagination. This really does suck. It really is bad. You do need to be believed. That is just another piece of validation for that. I think would be great.* (HS2)

On the other hand, one person (HS5) explained that while she thought *"it could be helpful in seeing trends"*, she imagined reservations about interpreting the patterns *"until my charting improves"*. Thus, she reported that she would *"use it for my own personal knowledge"* but would want more consistent data before showing it to a provider. Thus participants still have reservations about if such a technology would be pragmatic and useful to them in their lives.

Task-based evaluation

Data. We created three variations of the dataset, based on the phenotype (baseline vs learned), and split the data into a training set and a test set. The sample sizes of the different datasets and outcomes represented within them are presented in Table 4.6 for the training data and in Table 4.7 for the testing data. As was mentioned above related to the distribution of the learned and baseline phenotypes, there is some class imbalance in the phenotypic outcome variables.

Training Data Sample Sizes	Baseline Phenotypes n = 8584	Learned Phenotypes (with matched BL) n = 8584	Learned Phenotypes (all, even no matched BL) n = 10452
4-class	A (n = 1574) B (n = 3346) C (n = 2892) D (n = 772)	A (n = 2241) B (n = 2017) C (n = 2162) D (n = 2164)	A (n = 2566) B (n = 2314) C (n = 2679) D (n = 2893)
Binary CD Flare	No Flare (n = 4920) Flare (n = 3664)	No Flare (n = 4258) Flare (n = 4326)	No Flare (n = 4880) Flare (n = 5572)
Binary D-only Flare	No Flare (n = 7812) Flare (n = 772)	No Flare (n = 6420) Flare (n = 2164)	No Flare (n = 7559) Flare (n = 2893)

Table 4.6: Sample size of training data, for evaluating the phenotypes on a real-world computational task.

Testing Data Sample Sizes	Baseline Phenotypes n = 4973	Learned Phenotypes (with matched BL) n = 4973	Learned Phenotypes (all, even no matched BL) n = 6490
4-class	A (n = 1172) B (n = 2097) C (n = 1395) D (n = 309)	A (n = 386) B (n = 973) C (n = 1952) D (n = 1662)	A (n = 667) B (n = 1232) C (n = 2383) D (n = 2283)
Binary CD Flare	No Flare (n = 3269) Flare (n = 1704)	No Flare (n = 1359) Flare (n = 3614)	No Flare (n = 1899) Flare (n = 4591)
Binary D-only Flare	No Flare (n = 4664) Flare (n = 309)	No Flare (n = 3311) Flare (n = 1662)	No Flare (n = 4207) Flare (n = 2283)

Table 4.7: Sample size of test data, for evaluating the phenotypes on a real-world computational task.

Model selection and evaluation. We built various models across a range of algorithms (i.e., logistic regression, gradient boosting, random forest, and decision tree), datasets (i.e., baseline phenotypes, learned phenotypes with matched baseline instances, and learned phenotypes with and without matched baseline instances), and outcome definitions (i.e., 4-class definition of phenotypes A vs B vs C vs D, and 2-class definition of phenotypes AB vs CD). A summary of the evaluation metrics across all candidate models and outcome definitions is presented in Table 4.8 for the 4-class problem, and in Table 4.9 for the 2-class problem, where a flare-up is defined as phenotype C or D. We focus on the F2 score as the primary evaluation metric, and specifically the F2 score for flare-ups. We also present the AUROC and AUPRC.

Model	Baseline Phenotypes	Learned Phenotypes (with matched BL)	Learned Phenotypes (all, even no matched BL)

Logistic Regression	F2 weighted = 0.38
	F2 Flare = 0.02
	AUROC = 0.601
	AUPRC = 0.311
		F2 weighted = 0.34
		F2 Flare = 0.53
		AUROC = 0.607
		AUPRC = 0.363
			F2 weighted = 0.36
			F2 Flare = 0.59
			AUROC = 0.603
			AUPRC = 0.355

Gradient Boosting	F2 weighted = 0.38
	F2 Flare = 0.24
	AUROC = 0.608
	AUPRC = 0.313
		F2 weighted = 0.35
		F2 Flare = 0.53
		AUROC = 0.604
		AUPRC = 0.358
			F2 weighted = 0.36
			F2 Flare = 0.59
			AUROC = 0.603
			AUPRC = 0.351

Random Forest	**F2 weighted = 0.39**
	F2 Flare = 0.31
	AUROC = 0.583
	AUPRC = 0.293
		F2 weighted = 0.33
		F2 Flare = 0.45
		AUROC = 0.580
		AUPRC = 0.358
			F2 weighted = 0.34
			F2 Flare = 0.48
			AUROC = 0.579
			AUPRC = 0.319

Decision Tree	F2 weighted = 0.32
	F2 Flare = 0.10
	AUROC = 0.509
	AUPRC = 0.256
		F2 weighted = 0.29
		F2 Flare = 0.33
		AUROC = 0.527
		AUPRC = 0.263
			F2 weighted = 0.30
			F2 Flare = 0.36
			AUROC = 0.526
			AUPRC = 0.265

Table 4.8: Evaluation metrics across candidate models for the 4-class problem. For metric "F2 Flare," the D phenotype is considered a flare-up. The top metrics are bolded.

Model performance varied across the algorithms, datasets, and outcome definitions that we tested. Based on the F2 flare-up score, Gradient Boosting and Logistic Regression outperformed the other models tested for both the 4-class model and the 2-class model.

The learned phenotypes outperformed the baseline phenotypes on the forecasting task. While the models with the baseline phenotypes generally have a higher weighted F2 score, we can see that the F2 "Flare-up" score is always lower than the models with the learned phenotypes, across all algorithms and outcome definitions. The learned phenotypes also have better AUPRC scores.

Model	Baseline Phenotypes	Learned Phenotypes (with matched BL)	Learned Phenotypes (all, even no matched BL)

Logistic Regression	**F2 weighted = 0.63**
	F2 Flare = 0.28
	AUROC = 0.620
	AUPRC = 0.450
		F2 weighted = 0.56
		F2 Flare = 0.61
		AUROC = 0.544
		AUPRC = 0.763
			F2 weighted = 0.59
			F2 Flare = 0.69
			AUROC = 0.534
			AUPRC = 0.745

Gradient Boosting	**F2 weighted = 0.63**
	F2 Flare = 0.32
	AUROC = 0.629
	AUPRC = 0.457
		F2 weighted = 0.57
		F2 Flare = 0.62
		AUROC = 0.554
		AUPRC = 0.767
			F2 weighted = 0.59
			F2 Flare = 0.68
			AUROC = 0.552
			AUPRC = 0.755

Random Forest	**F2 weighted = 0.63**
	F2 Flare = 0.39
	AUROC = 0.606
	AUPRC = 0.433
		F2 weighted = 0.54
		F2 Flare = 0.57
		AUROC = 0.539
		AUPRC = 0.753
			F2 weighted = 0.55
			F2 Flare = 0.61
			AUROC = 0.533
			AUPRC = 0.740

Decision Tree	F2 weighted = 0.57
	F2 Flare = 0.44
	AUROC = 0.539
	AUPRC = 0.360
		F2 weighted = 0.52
		F2 Flare = 0.54
		AUROC = 0.516
		AUPRC = 0.731
			F2 weighted = 0.54
			F2 Flare = 0.58
			AUROC = 0.526
			AUPRC = 0.716

Table 4.9: Evaluation metrics across candidate models for the 2-class problem, where a flare-up is defined as phenotype C or D. The top metrics are bolded.

This means that the learned phenotypes are much better for the real-world task of forecasting an upcoming symptom “flare-up.”

The 2-class outcome definition of flare-ups performed better than the 4-class flare-up outcome. This means that the binarized phenotype would be better for future intelligent systems in a prediction task.

4.4 Discussion

4.4.1 Preliminary benefits of the learned model compared to the baseline model

There are several key points that suggest the learned model is better than the baseline model. First, the learned phenotypes model represents the latent structure of user health status and does not rely on a single feature as an indicator. This means that all available self-tracked data can inform the phenotypes and therefore minimize missing phenotype assignments. The learned model is more complete, leverages all of the data available, and is a robust representation of health status. It is also a more nuanced and holistic representation of health status compared to a baseline model that only uses a single feature to represent health status. The heatmaps and wordclouds in Appendix B.2 showcase how the learned model enables health status to capture detailed human experience of illness. The learned model is also calibrated and interpretable, so it is more meaningful and robust compared to a single variable or a large volume of raw self-tracked illness data.

4.4.2 Performance and acceptability of learned phenotypes

In the user study, both learned and baseline phenotypes had similar performance across matching between user and AI assessments, difficulty, and certainty. While there was not very high agreement between users and the AI-generated health status, there were still some positive aspects that were uncovered in the user study. We found that the learned phenotypes more closely align with the human users' process of determining health status. Participants used much of the same data that the learned phenotypes relied on in making the health status assignments. Further, while both the baseline and learned phenotypes did not neatly align with the user's assessment, the learned model erred on the side of assigning a worse status than the users, while the baseline model overwhelmingly assigned a better health status to the data than users. This aligns with the results from the computational task, where the baseline model was unable to forecast bad health statuses and was outperformed by the learned model. Participants also reported that they felt there was value in the AI-generated health statuses, for example to create a personal baseline for them to work from or as a sort of "sounding board" when working through their own health assessments. They were also optimistic about bringing summaries to their providers.

In the computational task, the learned phenotypes performed much better at the task of predicting flare-ups compared to the baseline phenotypes. In some ways, the baseline phenotypes had better evaluation metrics. However, when we look at the metrics that matter (i.e., that prioritize identifying flare-ups, and preferencing false positives over false negatives), the baseline phenotypes performed very poorly while the learned phenotypes performed much better with forecasting flare-ups. This suggests that the learned temporal phenotypes are promising for use in intelligent systems that can meet the needs of individuals with complex chronic illness, especially when binarized.

4.4.3 Implications for the learned phenotype model

Results from the user study highlight the nuanced relationship between participants' self-assessments of health status and AI-generated phenotypes, offering insights into their thought processes and where the current phenotypes might be improved. Participants demonstrated a clear reliance on a wide range of indicators, including symptom severity, medication usage, and activities of daily living, to make health assessments. These indicators, while overlapping among participants, were often interpreted and weighted differently, reflecting the individualized and subjective nature of self-assessment. Interestingly, while participants valued the AI-generated health statuses as a reference point, they overwhelmingly maintained their own assessments, showcasing their role as the primary experts of their lived experiences. At the same time, there were significant discrepancies in repeated evaluations of the same data. This requires further study to fully understand and reconcile with a machine-readable health status phenotyping model.

Despite this, participants recognized the potential utility of AI-enabled tools in validating their experiences and assisting with summarization, particularly to be used as a resource for communicating with their care teams. Additionally, while the AI-generated health statuses enhanced participants' efficiency in assessing longitudinal health trends, they did not necessarily increase their confidence or make decision-making about their evaluations easier.

In line with HAI, these findings underscore the importance of designing intelligent systems that complement rather than override user expertise, emphasize transparency in AI reasoning, and account for the complexity of individual health narratives. These findings also call for innovation in how users can remain "in the loop" with these models, or otherwise enhance autonomy and control over how their data represent their experiences. Integrating AI as a supportive, participatory tool has the potential to enhance both self-awareness and patient-provider interactions, but first further work is required so that the outputs align with users' expectations and lived realities.

The evaluation of the phenotypes also gave insight into how the phenotypes could be improved. In the user study, many participants talked about using medication to make their assessments. This information was not fully incorporated into model training. In the current model, medication was included as a simple binary (took medication, yes), and future work could map and categorize the user-entered medications to identify pain medications. This is an insight directly garnered from individuals in the user study. The computational task evaluation provides further insight into the computational performance of such a technology. The results from the experiments to train various models to forecast flare-ups suggest that the binarized model is the best performing model at this time, but it may not be suitable for all tasks. However, there is considerable opportunity to improve upon the performance. While the learned phenotype model performed only marginally well in the evaluation, potential applications for summarization could be developed now (e.g., to help individuals aggregate their data by health status to look for trends in their data and prepare for clinical visits), and others could be developed after further work on the phenotypes. Future work on the phenotypes could include improving how current data are used (e.g., mapping pain medications), using more advanced ML techniques to use other existing data (e.g., NLP methods to use the open-ended journal text), or incorporating additional datatypes (e.g., passive sensing).

Chapter 5: Informing the Design of Individualized Self-management Regimens from the Human, Data, and Machine Learning Perspectives

5.1 Introduction and related work

In the first study, we elicited the needs of users, both for patients and providers managing endometriosis together and for patients self-managing their illness independently. In the previous studies, we built the computational foundation for using interpretable health state phenotypes to support individuals and for use in AI-enabled intelligent systems. This study[1] aims to elucidate a set of design criteria for a human-centered, RL-enabled intelligent system that will satisfy human user needs and values, and has the potential to be successfully implemented considering the characteristics of self-management data and the complex domain of enigmatic chronic illness. We propose a novel framework — Multi-Perspective Directed Analysis — to guide the analysis. We use MPDA to conduct a mixed-methods study to identify the needs of end-users by analyzing real-world self-tracking data from existing users of the Phendo app alongside conversations with patients about their health experiences, and to identify the requirements for an AI-enabled self-management tool through conversations with the data scientist about how these findings fit with and impact ML and computational decisions. We triangulate these results and map them onto concepts of RL to identify the boundaries and constraints, from human, data, and ML perspectives.

This study addresses Aim 3 of the thesis. Here, we ask the following research question:

$RQ_{3.1}$: What insights at the intersection of human needs and values, human self-tracking behaviors as evidenced by "in the wild" self-tracking data, and capabilities

[1] The manuscript detailing these results, titled "Informing the Design of Individualized Self-management Regimens from the Human, Data, and Machine Learning Perspectives" has been accepted for publication in: Transactions on Computer-Human Interaction (TOCHI) [192]. In this chapter, we present an abbreviated version of the results and discussion.

and constraints of RL, can inform the design of RL-based intelligent systems for self-management of endometriosis?

5.1.1 Human-centered AI

A human-centered approach is critical to designing intelligent, personalized systems that meet the real-world needs of individuals [127]. User-centered and participatory design have been established as critical pathways towards developing technological solutions, which have been increasingly applied in AI [193, 194, 195]. However, traditional user-centered design methods often prioritize user needs over technical capabilities and limitations [196]. While this approach has been shown to be widely successful in traditional software development [197, 198, 199, 200, 201], it may have limitations when applied to machine learning (ML) and AI. Given the need for human-centered design and an uptick in interactive systems that incorporate ML/AI [202, 203, 204], there is a need for new design methods that balance human needs with the technical capabilities of these systems.

Various human-centered AI (HAI) principles have been put forward to address this gap [205]. Chancellor [53] argues that human-centered machine learning practices must be applied throughout the whole ML pipeline of problem brainstorming, development, and deployment. Human-Centered Algorithm Design proposed by Baumer [127] outlines design strategies across theoretical, speculative, and participatory processes, with the focus on incorporating social interpretations into the design approach. Value-Sensitive Algorithm Design [206] focuses on the early elicitation of human insights to guide the abstract and analytical creation of algorithms, aiming to mitigate bias and avoid compromising user values; however, there is no way to explicitly account for the demands and specifications of specific ML techniques, and it does not account for the data perspective. By contrast, Stakeholder-Centered AI Design [207] accounts for both human and data perspectives, using co-design with stakeholders' own data to prototype with AI, but only addresses generic algorithm considerations. However, while these approaches include user, data, and/or ML perspectives, they do not account for all of these perspectives simultaneously and have no way to design for a specific ML algorithm. In fact, to a large degree, these new efforts continue to prioritize user needs and values and use them as a blueprint for designing new technologies. Since AI-enabled technologies often have unique capabilities and hard to change restrictions that must be considered when designing user-facing systems [208], neither the traditional user-centered design, nor the newer user-centered methods for AI account for these constraints. While these approaches offer some directions, a gap remains in practices to design HAI technologies, particularly for a given ML solution [209]. This requires new design approaches that place both users and AI on the same level and enable negotiation between them.

In this chapter, we propose and implement an HAI framework for designing intelligent personal informatics systems — Multi-Perspective Directed Analysis (MPDA) — that accounts for human, data, and machine learning requirements and constraints, concurrently. MPDA uses constructs extracted from an ML approach, RL, to elicit both user needs, through directed content analysis of user interviews, and practical data constrains, critical for ML and AI-driven systems, through analysis of user engagement logs with an app for collecting self-monitoring data. We applied the proposed framework to gathering and triangulating human-machine-data requirements for a self-management tool for individuals with endometriosis — a poorly understood, complex chronic condition with no cure or reliable treatment.

5.1.2 Self-management for complex chronic illness

Self-management is key in managing and preventing the progression of disease [68, 69, 70, 8, 71, 72]. However, establishing a self-management care regimen can be a major hurdle. Faced with often generic guidance, individuals are left with the burden of translating this information into their day-to-day lives [210, 211, 212, 213, 214]. For instance, the recommendation to "engage in regular exercise" is left up to individuals to determine how to implement (e.g., which exercises would make sense for them, how often, and how intense). Furthermore, it is not known a-priori which strategies will be successful for a given person. Thus, individuals have to experiment through trial-and-error, which can be a lengthy process. This approach becomes even more complex when individuals have to choose among multiple strategies that can be combined into a regimen. Finally, in conditions with limited scientific knowledge and no established self-management guidelines, there is an additional burden on the individual to identify candidate strategies [215, 102, 216] and effective personalized regimens [81, 216].

5.1.3 Personal informatics for self-management

Personal health informatics solutions have been proposed to support self-management for individuals [36, 37, 38, 39, 40, 41, 42, 43], scaffold for problem solving activities [217, 218], and promote experimentation to identify triggers of disease flares [44, 45, 46]. Tools for self-tracking and reflection have helped in multiple contexts like migraine [19], IBS [16, 111], autism spectrum disorder [112], HIV [92], diabetes [91, 219], and for those with multiple chronic conditions [89, 90]. However, most solutions still leave a lot of the analytical work to individuals, e.g., which strategies to experiment with (either alone or in concert of each other), how to go about experimenting with them, and determining if they work. As such, designing intelligent systems to support management in the context of complex chronic illness represents an example of Ackerman's sociotechnical gap [220] — i.e., there is a known discrepancy between the nuanced, flexible, and contextual real-world task of self-management, and the rigid and brittle capabilities of technology.

In the earlier studies of this thesis, we documented a need for solutions that can provide individualized self-management recommendations and help to identify strategies with a positive health impact. These recommendations should take individual factors and context into consideration, help discover and select strategies to try out, and learn which ones are effective for each person. Given the structure of the problem identified in this work, a promising research direction for individualized, adaptive recommendations for self-management is the use of artificial intelligence (AI) methods in general, and reinforcement learning (RL), in particular. RL is unique in its ability to make sequential recommendations that adapt to changes in a complex environment. With RL, an agent learns a policy, i.e., a mapping from states to recommended actions (e.g., for a person living with diabetes, blood glucose level measurements mapped to recommended insulin doses), to maximize a pre-defined reward (e.g., decrease in HbA1c levels in a person living with diabetes) while simultaneously adapting to changes in the state, i.e., the environment resulting from the actions (e.g., the health condition of the person with diabetes). This ability to adapt in decision-making under uncertainty makes RL a plausible candidate to power an intelligent personal informatics system to support the trial-and-error process of self-management, by providing individualized recommendations of strategies for users to try and evaluating their success — over time and through interactions with the system — for learning individualized self-management regimens that work for each person [221]. However, RL also has a somewhat rigid conceptual model and requires structuring of a problem space into its framework of action space, state space, reward, and agent/policy. Furthermore, RL is notoriously data hungry and requires large training datasets to enable learning. As a result, while RL has many unique benefits, the introduction of RL may widen the sociotechnical gap [220] that already exists in intelligent solutions for personal health.

5.1.4 Personal informatics for self-experimentation

One common approach to supporting increased self-knowledge using personal health data is through self-experimentation. Traditional personal informatics tools, even without the use of AI or statistical analysis, can help individuals identify trends and patterns in their health data to support personal discovery and behavior change [47, 19, 48]. Personal informatics interventions have also been developed specifically for this purpose. These systems are frequently designed to facilitate conducting n-of-1 trials (also called single case designs), where users act as their own controls to highlight an individual's response to a treatment rather than a group's [46, 44]. While many of these systems have shown promise in supporting experimentation, they have also been constrained by rigid options and limited personalization [16, 104, 47, 45, 105]. But users with complex health conditions are still in need of personalized, customizable solutions to support self-experimentation towards developing effective long-term regimens.

5.1.5 Approaches for automated, individualized self-experimentation

Some research has explored the use of advanced computational frameworks that use ML to develop more flexible, person-centered systems for self-experimentation [222, 223]. Other research has investigated automated, individualized intervention approaches such as adaptive treatment strategies [224, 225], micro-randomized trials (MRTs) [226], and just-in-time adaptive interventions (JITAIs) [49], for a variety of health-related outcomes [227]. These tools exist in domains that have more well-defined parameters and outcomes, but could provide guidance in the context of complex chronic illness self-management.

Some of these existing approaches to adaptive intervention fit into the RL paradigm [228, 229]: a computational approach to understanding, automating, and optimizing a sequenced set of actions that maximize an outcome of interest. An RL agent learns which intervention is best to suggest from a pre-determined (continuous or discrete) set of *actions*, given a *state*, towards maximizing a total *reward*. JITAIs adjust the type, timing, and framing of support provided based on a user's dynamic internal and contextual state, aimed at maximizing the positive impact of an intervention. JITAIs, which sometimes use RL approaches to learn the best intervention for users, result in (expert or learned) decision rules that map an individual's current state to a particular intervention or treatment at each decision time point. Therefore, an RL approach to an intelligent system for supporting the trial-and-error process in self-management of a complex chronic condition is a promising solution to investigate.

5.1.6 Reinforcement learning (RL)

RL agents make sequential decisions as they interact with the environment, i.e., as the world changes, toward achievement of a specified goal. An RL agent learns how to update its recommendation policy from previous actions and evaluated rewards, based on the variables it observes. The *policy* dictates the behavior of the *agent* (system) and uses observations to map from the *state space* to the *action space* when in particular observed states. The state, i.e., the information used for individualization, helps decide when and/or how to intervene with particular actions. A *reward signal* defines the goal of the RL problem, which the agent maximizes, while the *value function* (which the RL agent updates as it interacts with the environment) quantifies the long-term desirability of states (after taking into account likely subsequent states and corresponding rewards).

RL as a candidate for automated, individualized self-experimentation. RL offers potential benefits that make it a promising candidate for providing automated, adaptive recommendations for individualized self-management in an illness context with a lot of complexity and uncertainty. Since it uses the results of each iteration to update the policy and inform future algorithmic decisions, it is particularly well-suited to help users self-experiment with self-management strategies, when it is not known beforehand what will be helpful, when, or for whom. In short, RL engages algorithmically in a trial-and-error process similar to the goal-directed, sequential learning that individuals follow when working to develop their own individualized self-management regimens [230]. RL can handle the dynamics of sequential interactions between the user and a system by utilizing feedback from the environment to adjust its actions; RL can account for long-term user engagement with a system; and RL can optimize a policy by sequentially interacting with the environment without requiring explicit user input [231].

In the context of endometriosis, RL-based recommendations offer several advantages. The goal of an RL agent is not only to optimize interactions with the environment, but also to learn insights that allow individualized interventions. Since RL allows for individual-level optimization, rather than focusing on population-level estimates, given that endometriosis shows strong person-to-person variation in symptoms and treatment responses, person-level personalized self-management recommendation is needed. RL is a sequential decision-making algorithm that can leverage multiple correlated time points during learning, making it suitable for learning with correlated longitudinal data. Since RL is able to provide sequential, adaptable, and individualized recommendations, it could be set up to conduct n-of-1 trials to facilitate self-experimentation at the individual-level — and helpful for other chronic diseases beyond endometriosis.

RL has shown exceptional performance in multiple scenarios [232, 233]. However, despite the previous successes of this approach and numerous benefits, there are also critical unresolved impracticalities in the context of this study: (*i*) successful examples have been limited to highly structured, controlled, and well-defined environments, i.e., they do not translate easily to practical, yet more complex situations; (*ii*) the state-of-the-art, deep-learning based RL techniques are data-hungry, i.e., they require inordinately large, repeated interactions with the environment to learn effective policies; and (*iii*) they are black-box models, i.e., it is inherently difficult to interpret the complex features learned by these models, and why the recommendations are made.

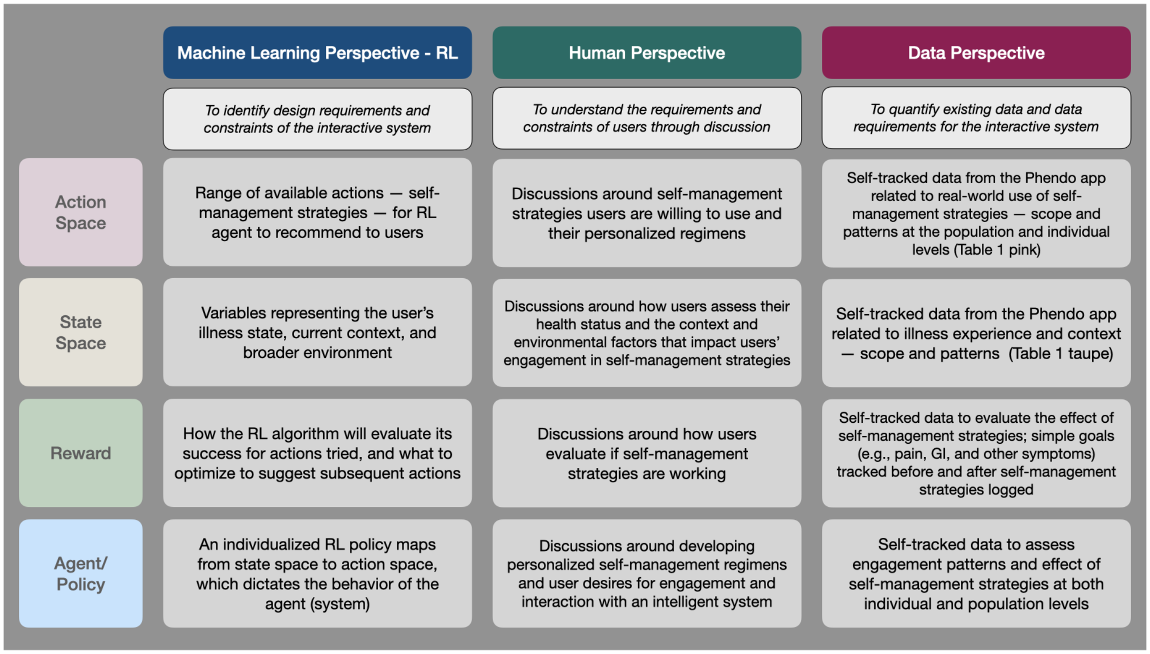
Problem formulation in the RL framework. The real-world task of figuring out what self-management strategies might work for an individual with an illness like endometriosis is an inherently interactive process that calls for a sequential decision-making process. Thus, it is possible to formulate the problem as a Markov decision process (MDP) and solve it with RL [234, 235]. The fundamental trade-off between exploitation (optimizing recommendations) and exploration (learning insights about the environment it is interacting with) is critical for efficiently updating the RL policies that decide which action to recommend next. To support individuals in developing a personalized management routine, an RL agent must learn a policy that can prescribe, based on observed context (i.e., user state), appropriate self-management strategies (i.e., actions at each interaction) that best improve an individual's symptoms/health status (i.e., the reward signal).

In real-world situations, especially when the RL directly interacts with and learns from humans' actions, there are multiple challenges to consider in designing an RL agent [236]. On one hand, the operational success of an RL policy is constrained by the size of the data it can learn from, the type of information it has access to, the number of interactions it can leverage, and the type of actions it can recommend. On the other, users of the system have their own preferences (e.g., the framing and tone of a recommendation) and values (e.g., preserving agency in their self-management activities).

Misalignment between human and machine can create unhelpful recommendations (e.g., a 20-minute jump rope session for an individual with chronic pelvic pain) and as a consequence, at best break trust and at worst produce harm. Hence, understanding user-centered requirements for

RL and its different components is critical to designing human-centered RL systems. In addition, an intelligent system that uses RL may ease the cognitive burden put on individuals to find their optimal self-management regimen. At the same time, this type of system requires delivery as a mobile intervention, which must incorporate human feedback. Delivery via a mobile self-tracking app requires the RL system to be able to determine when and how to best deliver the intervention components at different decision points. An interactive solution that incorporates human feedback in the learning process is necessary [237].

5.2 Methods: A novel human-centered AI framework — Multi-Perspective Directed Analysis (MPDA)



	Machine Learning Perspective - RL	Human Perspective	Data Perspective
	To identify design requirements and constraints of the interactive system	*To understand the requirements and constraints of users through discussion*	*To quantify existing data and data requirements for the interactive system*
Action Space	Range of available actions — self-management strategies — for RL agent to recommend to users	Discussions around self-management strategies users are willing to use and their personalized regimens	Self-tracked data from the Phendo app related to real-world use of self-management strategies — scope and patterns at the population and individual levels (Table 1 pink)
State Space	Variables representing the user's illness state, current context, and broader environment	Discussions around how users assess their health status and the context and environmental factors that impact users' engagement in self-management strategies	Self-tracked data from the Phendo app related to illness experience and context — scope and patterns (Table 1 taupe)
Reward	How the RL algorithm will evaluate its success for actions tried, and what to optimize to suggest subsequent actions	Discussions around how users evaluate if self-management strategies are working	Self-tracked data to evaluate the effect of self-management strategies; simple goals (e.g., pain, GI, and other symptoms) tracked before and after self-management strategies logged
Agent/ Policy	An individualized RL policy maps from state space to action space, which dictates the behavior of the agent (system)	Discussions around developing personalized self-management regimens and user desires for engagement and interaction with an intelligent system	Self-tracked data to assess engagement patterns and effect of self-management strategies at both individual and population levels

Figure 5.1: Overview of analytical approach with organizing principles.

In order to align design requirements of an intelligent system for supporting experimentation with self-management across human, data, and machine learning perspectives, we propose and implement a novel HAI framework to conduct a mixed-methods study. From the **ML perspective**, we use high-level concepts from RL to conduct directed coding. From the **human perspective**, we rely on the input of individuals with endometriosis to understand their needs and values when conducting self-management and receiving recommendations. The original high-level RL concepts form the basis for the directed content analysis of the qualitative data. From the **data perspective**, we conduct analysis of “in the wild” self-tracking data from current users of the Phendo app, who log their experiences in their day-to-day lives (i.e., in the absence of any ML algorithm) to understand patterns of illness experiences and self-management behavior from real-world data. The organizing principles from RL, which are revised according to the empirical data from the qualitative analysis, provide structure to the quantitative analysis of usage logs. An overview of the organizing principles and data sources is presented in Figure 5.1.

Through iterative and ongoing discussions amongst the researchers, we synthesize the findings across methods and data sources, guided and organized by the final categories of the directed content analysis. We use each perspective to guide the others and to organize the evidence, and we consider how the findings align with the theoretical and practical requirements of an intelligent system for self-management. Our novel methodology enables us to identify and understand the tensions and trade-offs of designing a tool in alignment with human needs and technological constraints. We expect that this approach will help to integrate human needs, challenges, and aspirations with ML and computational feasibility, so that we can identify design requirements that are not only desirable but also achievable.

5.2.1 Data and analysis: Machine learning perspective — RL

We select high-level RL concepts (i.e., *action space, state space, reward, agent/policy* — identified by Sutton and Barto [228]) as the basis for this analysis. These concepts are presented in the first column of Figure 5.1. In the context of this study, the **action space** represents the items available for the intelligent interactive system to recommend — self-management strategies. The **state space** represents the illness, contextual, and broader environmental variables that the agent will take into account for its recommendation. The **reward** is the function that will be used to evaluate the success of each recommended strategy. The **agent/policy** refers to the actual individualized self-management recommendations and behavior of the intelligent interactive system. A

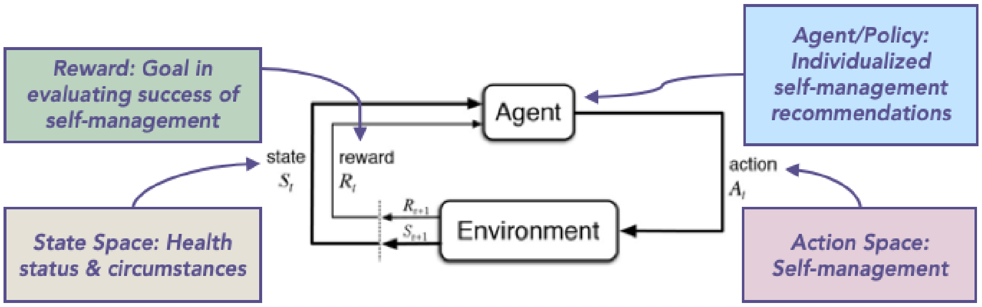

Reward: Goal in evaluating success of self-management
Agent/Policy: Individualized self-management recommendations
Agent
state
S_t
reward
R_t
R_{t+1}
S_{t+1}
Environment
action
A_t
State Space: Health status & circumstances
Action Space: Self-management

Figure 5.2: RL concepts mapped to the corresponding components of an intelligent interactive system to support individualized self-management.

mapping between the RL concepts and the corresponding components of an intelligent system for self-management is presented in Figure 5.2.

Since the aim in the current study is to support individuals in developing a personalized self-management routine that might evolve over time, we focus our attention on learning individualized sequences of strategies that maximize users' well-being as they interact with them (i.e., online learning). At the same time, we can leverage available historical data (i.e., offline data) for the design and modeling of the algorithm, inform its value and policy functions, and to avoid the cold-start problem.

An intelligent system that will be used and useful to individuals with endometriosis will have to meet their needs, operate acceptably and within expectations, and fit into their lives and self-management routines. To ensure these requirements are met, a key step undertaken in this study is to map the RL concepts and intervention components to the real-world aspects of human end-users. **The design items to consider are:** the scope of the action space; how to represent the state space and context; how to model the reward signal; how to balance exploitation vs exploration; what are the decision points (e.g., every hour vs daily opportunities) and decision rules; how much time can users dedicate for each recommendation; how often (dose schedule) and for how long (study duration) are users willing to experiment with the automated recommendation system. We answer these questions based on input from Phendo users and already available Phendo data. Below, we illustrate how our framework can use these RL-specific constructs to structure both qualitative analysis of user interviews and quantitative analysis of Phendo usage logs.

5.2.2 Data and analysis: Human perspective

For the qualitative analysis, we analyze transcripts from discussions with people with endometriosis. We use directed content analysis (as described by Hsieh [238]) based around the high-level RL concepts described in the previous section to code, understand, and organize the data into meaningful findings. The way that the RL concepts are applied in this part of the study is depicted in the second column of Figure 5.1. Note that data from the quantitative analysis also represent the real-world experiences of users, but are discussed separately within the data perspective in the next section.

Data. We re-analyze focus group transcripts from prior studies and conduct additional interviews with Phendo users with follow-up questions specifically relating to an interactive system for self-management. In total, we re-analyze 10 transcripts from focus groups with 48 participants (all women with a formal diagnosis of endometriosis) from prior work and data from three follow-up interviews. Focus groups from 2019 (5 groups, n = 21) were initially used to elicit design needs for tools to support care and self-management of endometriosis, detailed in Chapter 3 of this thesis (the focus group guide is available in Appendix A.2). Focus groups from 2016 (5 groups, n = 27) originally informed the design of the Phendo app (the focus group guide is available in Appendix A.4). This secondary analysis of focus group transcripts provides a valuable source of information about Phendo users' management practices, how people with endometriosis select and evaluate strategies, and where participants have unmet care needs that technology could support. But while these focus group transcripts do directly address self-management practices and technology needs, the discussions do not specifically address an interactive system for experimenting with self-management strategies. To fill this gap and strengthen the evidence collected from the secondary analysis, we conduct in-depth follow-up interviews (n = 3) to ask potential users directly about an interactive system for developing individualized regimens. We consider these interviews a sort of “member check” to ensure that the secondary analysis aligns with user perspectives and is not missing important perspectives related to the current research question. Interviews probe about participants’ experiences of self-management, the process by which participants have developed their own regimens, and what opportunities and constraints participants foresee in an automated system that provides intelligent recommendations that could support their management. The interview guide can be found in Appendix A.3. From these data sources, we are able to directly and indirectly elicit constraints and boundaries of an interactive system for self-management. Additional details on these methods, including a description of the study sample and topics covered in the interviews and focus groups, can be found in Appendix C.

Analysis. The qualitative analysis focuses on predetermined high-level RL concepts in order to guide the analysis towards evidence that answers our specific research question. Our inquiry seeks to identify technical requirements and parameters for an intelligent interactive system for providing adaptive self-management recommendations, to figure out a range of what users want and are willing to do when using the system, to outline constraints, and to assess potential feasibility of such a system. We use directed content analysis to code the data, starting from the four high-level RL concepts as predetermined codes (i.e., **action space, state space, reward, agent/policy**),

Table 5.1: Overview of themes across qualitative and quantitative analyses.

Initial coding categories	**Final coding categories**
(1) Action space — Self-management strategies available for the intelligent system to recommend	(1a) Broad action space
	(1b) Explore, with user autonomy
(2) State space — Health status and circumstances considered by the system for recommendations	(2a) Illness, everyday, and context variables
	(2b) Context for relevant recommendations
(3) Reward — Goal used to evaluate the success of strategies	(3) Short-term indicators
(4) Agent/policy — Behavior of the system when determining individualized self-management recommendations	(4a) Heterogeneity, so individualized policies
	(4b) Explainable recommendations
	(4c) Engagement patterns sufficient for RL

then refine the codes during analysis based on empirical data — the initial codes and final coding categories used as the framework for the study are presented in Table 5.1. During the analysis, the authors worked together to review, refine, and name the categories and extract key examples to present. Given the reflexive and collaborative nature of the analysis methods, we focused more on exploration of themes and consensus building rather than consistency in coding [239]. The codes generated through the analysis are used solely for conceptual purposes, serving to establish foundations for an RL model rather than for direct use in RL training. Future work developing the RL agent may employ an MPDA-based approach, where coding consistency will be critical.

5.2.3 Data and analysis: Data perspective

For the quantitative analysis, we leverage data collected from the Phendo app to capture the illness experiences and self-management behaviors of users "in the wild," i.e., without any intervention or recommendation, how individuals track their illness and what strategies they rely on for self-management. The Phendo data and analyses used for each of the RL concepts can be found in the final column of Figure 5.1.

Data. Phendo users self-track a variety of details about their illness (see Tables 5.2 and 5.3). We identify two broad domains of questions from the Phendo app that are relevant for this analysis: those related to the state space, or the personal experience of illness (Table 5.3), and those related to the action space, or self-management (Table 5.2). For each of the customizable self-management strategies (e.g., exercise, foods), we manually map and normalize all responses to a smaller, coherent set of pre-determined strategies in order to reduce the number of strategies in the dataset and to facilitate a population- vs individual-level analysis. E.g., the user-entered exercises 'walk,' 'hiking,' and 'walking dog' are all harmonized to a 'walking' entry, and the curated 'strength' strategy includes responses from customized answers like 'crunches,' 'sit ups,' 'core exercises,' 'squats,' 'planks,' 'push ups,' and 'weightlifting.'

Since this study focuses on self-management, we extract all data for users who have self-tracked at least one of the Phendo self-management questions marked with an asterisk (*) in Table 5.2. For each participant, we collect their longitudinal self-tracking data as entered in the Phendo app, so that we can quantify both their self-management habits and information on their health status through time. The dataset contains information from 10,463 users, with n = 399,786 self-tracked responses, of which n = 290,503 are self-management instances.

Analysis. The quantitative analysis was conducted in alignment with the pre-selected coding categories, and iteratively revised alongside the qualitative analysis, as the categories were revised inductively based on empirical data. For the **action space**, we analyze the answers to the self-management Phendo questions detailed in Table 5.2 to characterize the *breadth of strategies* users experiment with, as well as for *how long and how often do they engage with each strategy*. Similarly, for the **state space**, we analyze the answers to the questions related to the personal experience of illness, detailed in Table 5.3. For the analysis related to the **reward**, we define simple goals related to pain, GI symptoms, and other symptoms. We quantify how frequently users track this information before and after engaging in self-management activities. An exam-

Phendo Question	**Answer Type**	**Examples**
What did you do to self-manage? *	Pre-set: 14 multiple choice items	heat pack, massage, talk therapy
Did you do any of these exercises that hurt? *	User-specified multiple choice	running, situps, lunges, kickboxing
Did you do any of these exercises that help? *	User-specified multiple choice	yoga, pilates, swimming
Did you eat any foods that worsen symptoms? *	User-specified multiple choice	sugar, gluten, white flour, beer
Did you eat any foods that improve symptoms? *	User-specified multiple choice	fresh veggies, lean meat, nuts
Take any supplements?	User-specified multiple choice	CBD oil (15 mg), magnesium (500mg)
Take any hormones?	User-specified multiple choice	progestin(implant), microgestin (1.5 mg)
Take any medication?	User-specified multiple choice	Percocet (10mg), Oxycodone (7mg)

Table 5.2: *Questions related to self-management strategies —* ***action space***. Description of relevant questions in the Phendo app, the vocabulary type (pre-set, customized, or free-text), and the available answers. The analysis for this study included all users who had tracked at least one instance of self-management marked with an asterisk (*).

Phendo Question	Answer Type	Examples
How was your day?	Pre-set: 5 single-choice items	good, manageable, bad, unbearable
Do you have your period?	Pre-set: 2 single choice items	yes, no
Are you in pain now? (body location, severity) †	Pre-set: 39 multiple choice items	ovaries; cramping; moderate
Any GI/Urine issues? (description, severity) †	Pre-set: 15 multiple choice items	endo belly, vomiting, constipation; severe
Experiencing something else, other symptoms? (description, severity) †	Pre-set: 21 multiple choice items	fatigue, headache, swelling; mild
How is your mood?	Pre-set: 30 multiple choice items	calm, happy, angry, anxious
Are you bleeding?	Pre-set: 3 multiple choice items	clots, spotting, breakthrough bleeding
Which activities were hard to do?	Pre-set: 20 multiple choice items	sleep, shower, work, sit down, walk
How was sex?	Pre-set: 5 multiple choice items	painful during, painful after, avoided
Daily journal	Free-text	

Table 5.3: *Questions related to the personal experience of illness — **state space***. Description of relevant questions in the Phendo app, the vocabulary type (pre-set, customized, or free-text), and the available answers.

RL Concept	Description	Example
*Data related to evaluating the success of self-management strategies — **reward***	Difference in symptom severity scores, with a focus on: pain, GI issues, and other common endometriosis symptoms — the relevant Phendo questions are shown with †. We quantify how frequently this reward information is tracked before and after self-management activities are logged.	Reduction in pain severity tracked before and after engaging in self-management
*Data related to self-management trials — **agent/policy***	Self-management trials consist of an event (self-management strategy, highlighted above in pink) tracked, along with pre-event and post-event goal data (i.e., self-management strategy, with reward data tracked before and after). Effect estimates are calculated for each trial (the difference between pre- and post- strategy goal response), at both the population-level and for individuals.	*Walking* to *pain* trial with a negative effect estimate indicates a reduction in pain between before and after implementing the strategy

Table 5.4: Description of data and examples for the ***reward*** and ***agent/policy*** categories of analysis.

ple is highlighted in Table 5.4. For the analysis related to the **agent/policy** concepts of RL, we quantify the effects of strategies through an analysis of self-management trials, and assess engagement based on how many trials users log "in the wild." We define a *self-management trial* as an event where a user tracks a self-management strategy, as well as goal information a day before and after the tracked strategy. That is, a trial consists of a triad of pre-management goal information, a self-management instance, and post-management goal information, all contained within a day of the self-tracked strategy. We then quantify *effect estimates* of each strategy on symptom severity: i.e., we compute the difference between the post- and pre-strategy goal response for each trial. E.g., a self-management pre-post negative pain-severity effect estimate indicates the strategy reduces pain, while a positive pain-severity effect showcases worsening of the experienced outcomes of the self-management strategy. For a given set of per-strategy trial effects, we compute effect estimates both at the population level (by averaging per-strategy effects) and for individuals (by averaging per-individual longitudinal per-strategy effect). An example is shown in Table 5.4. In investigating if an RL policy is feasible for this task, this component of the analysis can help us understand if data tracking patterns "in the wild" will align with the ML and human requirements and constraints identified in the qualitative analysis.

5.3 Results: Human and technical specifications of an intelligent system for individualized self-management

We integrate results from the qualitative and quantitative analysis, with the RL concepts as the guiding framework. In synthesizing these findings, we identify promising circumstances where an intelligent system for self-experimentation with self-management could be successfully deployed; but we also find barriers and challenges that will need to be resolved. We present eight findings across the four RL categories (Figure 5.3): **1. Action Space — (1a)** The space of actions/self-management activities individuals experiment with is large; **(1b)** Individuals are willing to explore the action space as long as they can set some boundaries; **2. State Space — (2a)** The state space, that determines which actions to recommend, is complex and can vary across individuals; **(2b)** Individuals favor recommendations that are responsive to the user's current context; **3. Reward — (3)** Individuals assess the success of a strategy in the short term; **4. Agent/policy — (4a)** Success of specific strategies varies across the population, rendering the need for tailored recommendations from individualized policies; **(4b)** Individuals want to understand why they received recommendations and why they should try recommended strategies; and **(4c)** "In the wild" user engagement with self-management and self-tracking is frequent enough to satisfy RL data requirements. Illustrative quotes are included with each of the results (where the interviews are labeled "Int" and the focus groups are labeled "FG"), and Recommendations for design are included in the tables in Section 5.3.9.

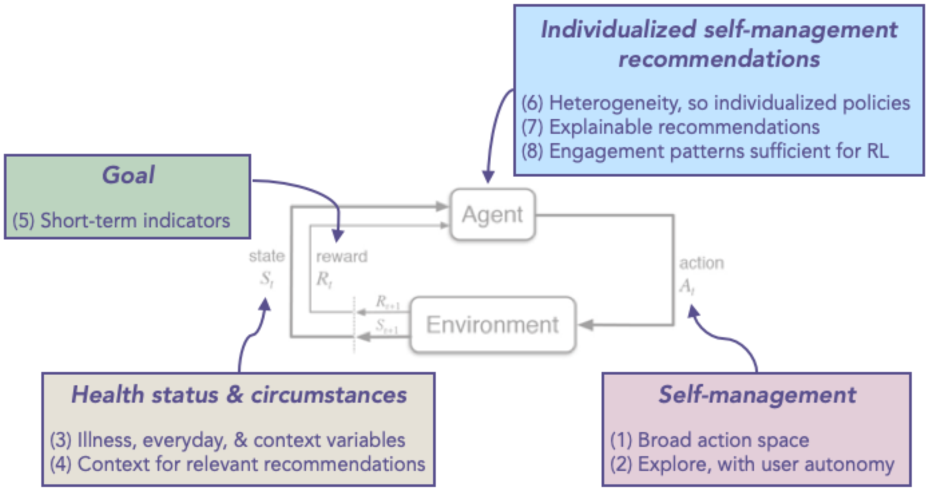

Individualized self-management recommendations
(6) Heterogeneity, so individualized policies
(7) Explainable recommendations
(8) Engagement patterns sufficient for RL
Goal
(5) Short-term indicators
Agent
state
S_t
reward
R_t
R_{t+1}
S_{t+1}
Environment
action
A_t
Health status & circumstances
(3) Illness, everyday, & context variables
(4) Context for relevant recommendations
Self-management
(1) Broad action space
(2) Explore, with user autonomy

Figure 5.3: Overview of study findings, organized around the high-level RL concepts used as initial directed coding categories, and sub-themes derived empirically through data analysis.

5.3.1 (Action Space - 1a) The scope of the action space is broad across the population of users, and individuals often combine multiple strategies.

In the context of chronic disease self-management, the action space corresponds to the space of possible self-management actions that an individual can take. As such, we code all participant descriptions of self-management approaches under this category and analyze self-tracking data related to self-management. Participants detail a multitude of different self-management strategies and self-care approaches for coping with symptoms and for promoting health more broadly. Their regimens incorporate strategies that are included in the Phendo self-tracking vocabulary (see Table 5.2), both pre-set (e.g., *"In the middle of the pain, just give me my heating pad and my drugs."* (FG 2)) and user-customized, such as exercise and diet (e.g., *"For me — no dairy, no gluten, very low sugar. Minimize raw food intake — mostly cooked vegetables, dried fruits. Fried foods. Preservatives. Processed foods. So, an endo-friendly diet."* (Int 3)). Participants also describe strategies and self-care tactics beyond what is included in Phendo, for example: *"I've found hobbies that I can do with two hands help me feel better. I crochet, I am kind of a plant lady, I have a garden."* (FG 7). Strategies like these could be added into Phendo as a user-customized vocabulary. Participants report using various strategies in combination and as needed, for example: *"I like to take a holistic approach — I put marijuana in tea and I drink it, sometimes I use the CBD pen. I have ginger and turmeric tea on a regular basis. And I try to exercise at least three times a week."* (FG 7).

From existing "in the wild" data, we corroborate these findings: current Phendo users track a wide range of strategies across the population and individual users often track several in combination. First, we showcase in Figure 5.4 the count of users that have self-tracked at least 10

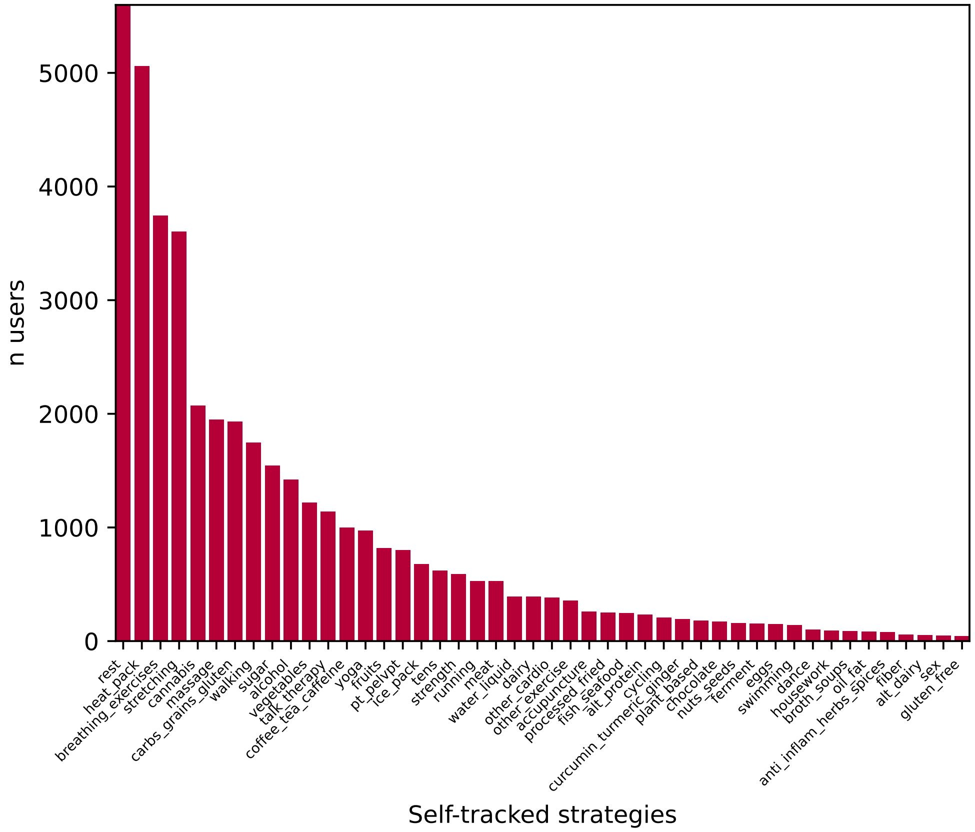


Figure 5.4: Histogram of Phendo users with self-tracking data (at least 10 instances) for each self-management strategy. We observe that users across the population engage with a broad range of strategies.

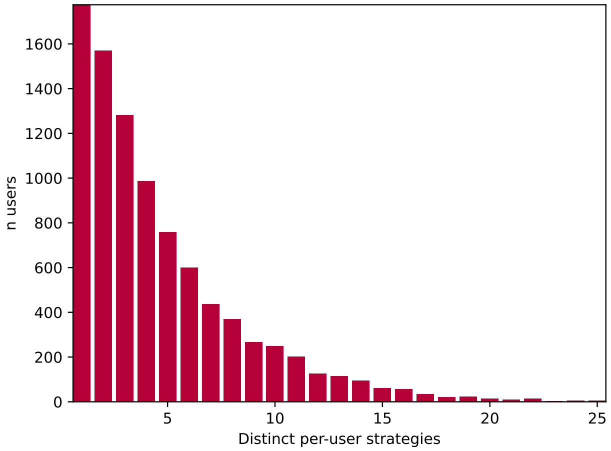


Figure 5.5: Histogram of Phendo users that track several strategies in combination. We observe that users often engage with a combination of several self-management strategies within their timeline.

instances of each strategy within the self-management Phendo questions. These results align with the participant reports of the wide variety of strategies they use (and are willing to use). Second, we illustrate in Figure 5.5 the count of users that self-track strategies in combination. We notice that, although a lot of Phendo users focus on one or two distinct self-management strategies, there are many Phendo users already tracking regimens in the app that combine several strategies together. From the ML perspective, it is beneficial that users engage with a range of strategies, yet an RL algorithm will benefit from a constrained action space to effectively learn meaningful policies. Design decisions will need to be made in order to satisfy both human user requirements and the RL requirements. Taken together, the data from users and self-tracking data suggests that there is a broad range of strategies used at the population level, but individuals are likely to only experiment with a few strategies at a time.

5.3.2 (Action Space - 1b) Users are willing to explore the action space, but control of the action space is important to participants.

The second code within the category of the action space relates to users' willingness to explore different strategy recommendations. Here, we code discussions about what participants are willing to try and limits and barriers that constrain what recommendations they are open to. Participants frequently discuss turning to exploration and trial-and-error self-experiments, since individuals managing endometriosis face a lot of uncertainty in care and are often left without effective treatments. Participants tell us they are interested in trying a broad range of self-management options and are willing to try strategies that don't necessarily make sense to them. When one interview participant was asked about trying strategies the intelligent system might recommend, she reports: *"I'll always give the new thing a try and put the old thing on hold."* (Int 3). Further, while users are willing to explore a wide variety of strategies and try suggestions that they may not otherwise, they do want an intelligent system that becomes more tailored to them and personalized over time (shifting from explore to exploit in the RL framework). As this same participant explains: *"In the beginning I'd be open to explore all options and see what there is to offer. And then, throughout usage it becomes more fine-tailored to me, I think that would be ideal."* (Int 3).

Participants warn that adding too many strategies to someone's care routine may become overwhelming, so generally limit experiments to a single strategy. At the same time, they describe benefits to engaging in multiple strategies that target different symptoms (e.g., one food strategy, one physical activity strategy, and one pain management strategy for flares). As one participant explains:

> *One thing, because that's all I can handle. When I'm trying a new thing, I try to just 'suss' that out for a while just to see if it's working. But at the same time, not like two things to cut out or add to my diet, but maybe something with diet and something with activity or movement — a combination of things.* (Int 3)

While participants are open to exploring different self-management strategies, which is beneficial to the learning process of RL, they are not willing to attempt very exhausting or taxing activities (e.g., *"I do think that there's a healthy limit to pushing yourself a little bit. But anything that causes me more pain is never going to be helpful."* (Int 1)). While participants describe a range of barriers that limit the options that are feasible for them, cost, logistics, and emotional challenges are the most often mentioned. Participants see value in setting hard limits ahead of time about what strategies are recommended. At the same time, they also imagine advantages to receiving recommendations with a variety of novel strategies that they otherwise might not try and to give feedback to the system in real-time as recommendations are given. For example, one interview participant explains:

> *Based on my previous experience with that type of activity, I already know it doesn't work for me. [...] I would want to tell the app, 'Please don't ever recommend that to me again.' But, I would say not setting all of the limits at the beginning, because there is something about 'pushing yourself' a little bit and trying something a little bit new. [...] So I actually think it would be helpful to get the things, but there are going to be some things for certain women, that are just going to be a no-go.* (Int 1)

Participants imagine giving input or feedback to the system about what they want to try (or not try), for example, interview participants asked about the intelligent system propose features that could provide options alongside recommendations to see if users are willing to try a strategy, such as: "yes," "not now," "not ever." They also suggest offering different options of strategies across difficulty level (from gentle to rigorous) to choose from when presented with self-management recommendations. These suggestions from participants could enhance user control and integrate human intuition into the decision-making process, while enabling efficient RL learning with constrained action spaces.

5.3.3 (State Space - 2a) Experiences of illness and everyday variables and context are important for characterizing the state space.

In the case of chronic disease self-management, the state space corresponds to the ways that an individual's illness state, current context, and the broader environment are represented. In this category, we code all descriptions of how individuals assess their health status and the context and environmental factors that impact their engagement in self-management and analyze self-tracking data related to illness experience and context. Participants describe rich, nuanced accounts of illness and discuss highly personalized and often embodied ways that they characterize and assess changes in their health status.

In detailing the ways that they understand their own disease, participants emphasize the importance of holistic and individualized representations of their illness experiences and talk about factors that are not always deemed clinically relevant. Pain and fatigue are the most frequently discussed indicators, but GI symptoms (e.g., cramping, diarrhea, nausea, bloating), breakthrough bleeding or spotting, skin issues (pimples/acne, rashes, hives), migraines, emotions (e.g., "mood swings," depression, anxiety), manifestations of pain (pain-somnia, pain-aggression), and other personal signs of inflammation (e.g., in the face or belly) are seen as necessary to provide a holistic and individualized picture of a person's health status. Participants also distinguish between dealing with typical, day-to-day endometriosis symptoms and the flare-ups that some individuals experience, as one focus group participant explains: *"When it comes to endo, you have the pain aspect but then you also have the flare-up aspect and it kind of goes hand in hand and sometimes it doesn't go hand in hand."* (FG 2).

Some participants report on indicators that suggest they are in a flare or a flare is coming. Participants recount their unique manifestations of "bad days" when they "feel terrible" or "horrible" (even though they often "don't look sick"). They talk about being "incapacitated" in bed because of pain (sometimes tied to their menstrual cycle, ovulation, or menses) and feeling so ill they can't move, are exhausted, or "collapse" when they get home. They talk about knowing their bodies, so they can know when something is not right (*"You just have to know your body, and you'd be like, 'Okay, well, something's off. How can I help it?' "* (FG 5)).

The analysis of Phendo data corroborates, as shown in Figure 5.6, that current users self-track the experience of illness via a multitude of dimensions that are available in the app — the list of available Phendo questions directly related to the individual's disease experience is found in Table 5.3. These variables align with the description of what users consider important to represent their illness experiences to facilitate holistic state space representations and that can be used to modulate the recommendations from an intelligent system.

5.3.4 (State Space - 2b) Users need recommendations that are suited to their current context.

The second code within the category of the state space relates to the relevant context features where users are open to receiving different recommendations. As such, in this category, we code participant discussions about what details they consider in deciding on self-management strategies to use and analyze contextual data tracked immediately before self-management strategies are logged. Users describe wanting intelligent systems to consider contextual details related to their illness, their environment, and their day-to-day lives when tailoring strategies to recommend — many of which may be subjective, esoteric, and/or highly personalized. Users request for recommendations to be responsive to their current context, as one interview participant explains:

> *I track emotional moments when things arise for me. So if the Phendo app could know, 'OK she's been logging these things, she's been anxious, or there's fatigue.' Then now is not a good time for the app to recommend training for the 5k. But if I log those things and the app suggests meditating today for 30 minutes, or 'Why don't you journal?' — it would be useful if the recommendations coincide with whatever symptoms I'm tracking.* (Int 3)

On top of this request for context-awareness, users also want strategies to tailor for particular

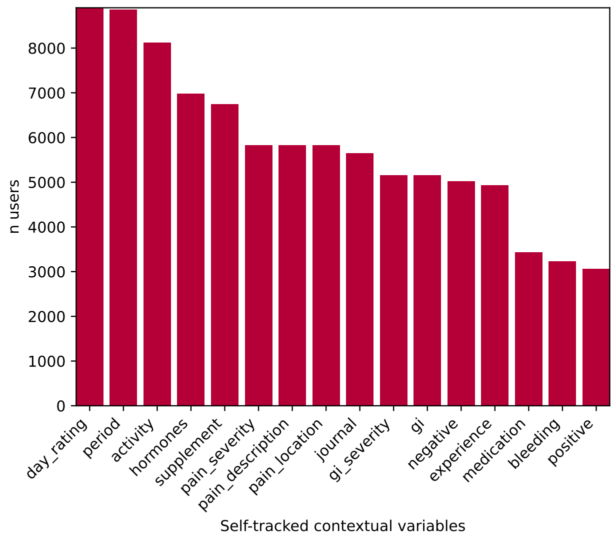


Figure 5.6: Histogram of Phendo users tracking questions related to their illness experience.

symptoms and for daily management compared to a debilitating flare.

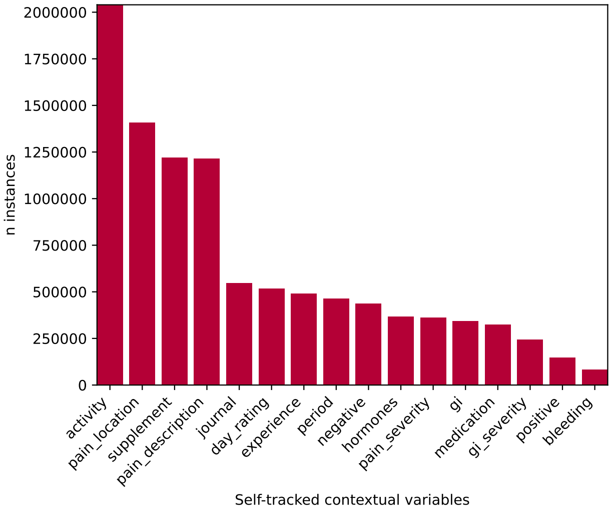


Figure 5.7: Histogram of instances for questions related to Phendo users' illness experience, tracked a day before a self-management strategy.

Participants discuss fitting strategies into their daily lives, and mention wanting a system to consider their schedules and daily routines and fit strategies into them in a way that makes sense. Participants also see value in a system that could push recommendations for strategies that would be helpful in the moment. Some users are open to pre-scheduled time set aside to do the self-management strategies offered by the system, but users desire a system that can tailor strategies to the particular context of situations as they arise. As one interview participant describes:

> *There could be benefits to both. I like the push of, 'Hey stretch right now.' However I typically implement those types of things when my threshold for pain has gotten so high to the point where I can't think, and that's where I would step away and do that, as opposed to just getting home later tonight and stretching.* (Int 1)

Participants imagine personalized recommendations would be responsive to a user's state, including health status and more broadly how are they feeling. But context about the day-to-day activities of users and their environments is currently infeasible, difficult, or impossible to capture (e.g., *"What headspace am I in when that is suggested?"* (Int 3)). Nonetheless, potential users want a system to take these aspects of their day-to-day lives into account.

We corroborate in Figure 5.7 that there are many existing instances of the self-tracked illness experiences discussed as important in Section 5.3.3, which are tracked by users within a day before they track a self-management strategy, suggesting that the contextual data currently tracked by users in the app can be useful for tailoring recommended actions. We observe that Phendo users are very likely to report information about their experiences with activities of daily living, as well as with information on symptoms they are experiencing the day before and the day after they self-manage. Therefore, there are already existing Phendo answers that can be used to compute a meaningful state space for an intelligent system and facilitate contextually-tailored recommendations, although we would still miss things that are not currently tracked or cannot be tracked (e.g., "current headspace," "timing is not working out today," or "pain so bad can't think"). In addition, there are sparsity and heterogeneity challenges that are raised by the "in the wild" data analysis with current Phendo data: i.e., not all users track the same set of experiences, and they do not track them consistently for every self-management strategy they try.

5.3.5 (Reward - 3) Users are looking for short-term self-management, but even then there are many ways to compute the reward function.

In an intelligent system for chronic disease self-management, the reward corresponds to the signal that will be optimized when suggesting actions and how the success of actions tried will be evaluated. As such, we code participant descriptions of how individuals evaluate if self-management strategies are working under this category and use simple self-tracked data to evaluate the effect of strategies tracked "in the wild." In endometriosis, symptoms often persist and sometimes progress, since there is no cure, which means that getting rid of symptoms altogether is often impossible. So, people with endometriosis report having to rely on assessing how they are feeling in the moment and trends or deviations from their individualized baseline to figure out if strategies are effective or not.

When asked about assessing the success of self-management strategies and evaluating the progress of their health goals, individuals largely describe short-term symptom or health-related quality of life indicators. While patients often have longer-term goals for their care and management, these are not what participants report relying on for developing their care regimens. Participants talk about having difficulty figuring out if strategies are working and tell us that they generally rely on checking in with how they feel in the moment (*"How do I know something is working? I'm not great at figuring that out all the time. [...] For me, if the pain goes away in the moment, then I will stick with that."* (FG 6)). They report relying on mind and body pain and sensations, immediately or within a short time-frame. Even still, users are not looking to eliminate pain and symptoms (which is often not possible), and instead focus on evaluating if strategies are helping to improve their symptoms (*"That's the key word, 'Not as bad.' So, better."* (FG 1)).

When evaluating if self-management strategies are working, participants all talk about assessing their pain symptoms; GI symptoms are also frequently used as an indicator. Participants also describe clues or indicators that are specific to them, which they use to help them determine if they are feeling better or worse (e.g., skin rashes, acne breakouts, and the ability to sit in a chair without high pain).

In order to measure the success of self-management strategies in a data-driven fashion, we need to determine the effect signal for each individual: i.e., we need to define the reward function that an intelligent system will use as feedback. When looking at existing Phendo data, we observe that pain, GI, and other symptomatic experiences are frequently self-tracked by current Phendo users, both before and after self-tracked self-management strategies. More specifically, in Figure 5.8 we showcase the abundance of Phendo self-management trials for which there is pain-related pre-post information. These results suggest that Phendo users track sufficient information related to their self-management goals (pain results are shown here, but similar stories hold for GI and other symptoms) for the design of a meaningful reward function. We have used simple, human-driven functions in this analysis, illustrating that an RL algorithm could have sufficient reward signal to detect an effect, based only on "in the wild" data. The reward function is critical for any RL agent, as it provides the signal from which to learn how to find the right exploration-exploitation tradeoff. Although the simple functions suggest sufficient data and effects exist to learn RL policies, the

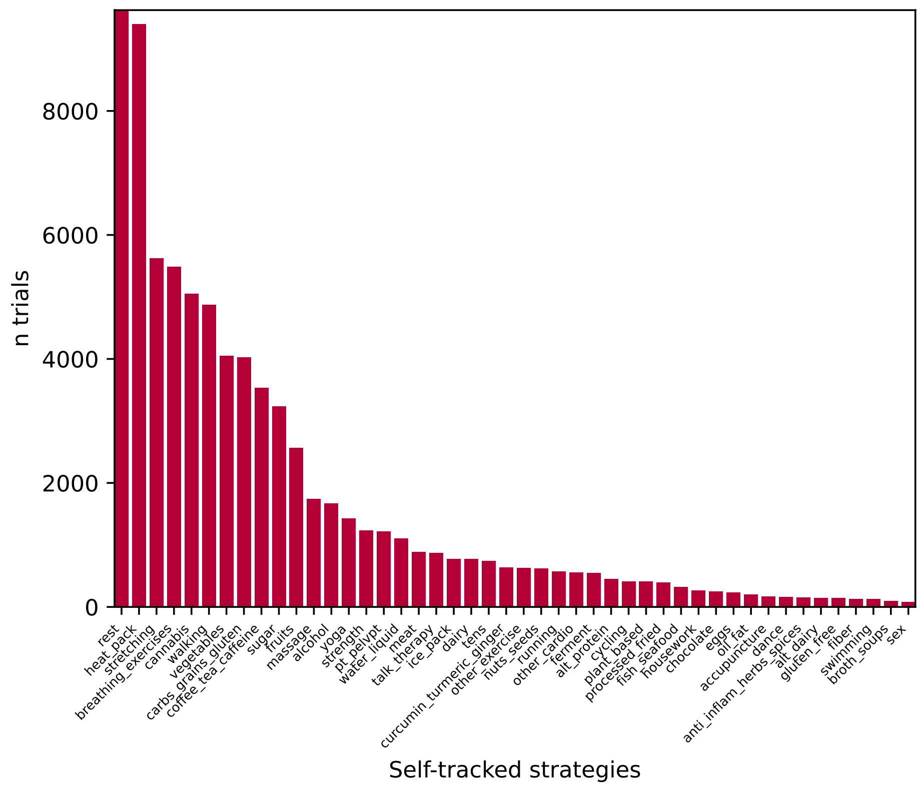


Figure 5.8: Histogram of Phendo self-management to pain trials. Phendo users engage with a wide variety of self-management strategies towards their pain management goal.

user perspective emphasizes that we will need to expand on these reward functions to represent individualized metrics.

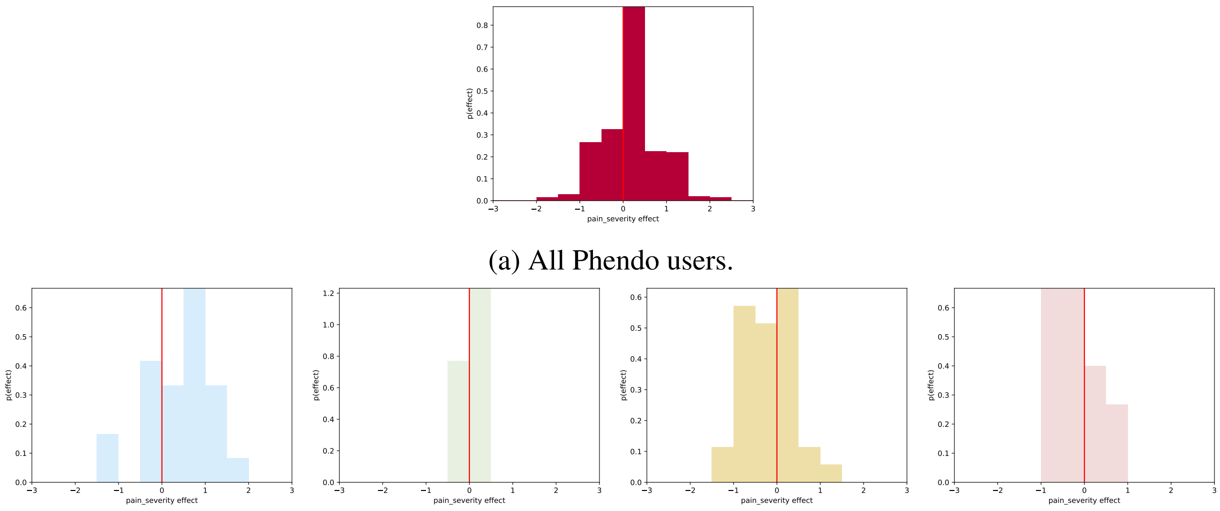
5.3.6 (Agent/Policy - 4a) Heterogeneity in self-management responses calls for tailored recommendations.

In the context of chronic disease self-management, the agent/policy dictates the behavior of the intelligent system. As such, we code participant discussions about developing their own self-management regimens under this category and analyze engagement and existing self-management trials with self-tracking data. This category has been separated into three codes during analysis; conversations discussing the personalized nature of self-management regimens are coded here. Users are enthusiastic about an intelligent system that could customize personalized recommendations based on their prior experiences, rather than suggesting strategies because it worked for someone else who is “similar” to them. They describe their desire for recommendations to be tailored to them based on their own data (*“Based on my stuff I logged — the experiences that I’ve had so far.”* (FG 2), since each person is different.

Phendo users emphasize that a personalized approach to self-management will be required to meet their needs, particularly since they are heterogeneous in which strategies individuals use for different symptoms and health states. Across the wide array of self-management strategies, individuals turn to different activities to address particular symptoms. Some examples of strategies users enact based on various symptoms include:

> *You can pretty much always do deep breathing. I don’t care how much pain you are in, as a default.* (Int 1)

> *I’ll read and write a lot when I’m having a bad day, because it’s a way to escape and be in a space that’s more enjoyable. When I’m having a good day, I’m more of a physical person. I’ll be like, ‘Oh, I’d love to go kayaking,’ or ‘Oh, let me go work in my garden.’* (FG 7)



(a) All Phendo users.

(b) Phendo individual A. (c) Phendo individual B. (d) Phendo individual C. (e) Phendo individual D.

Figure 5.9: Probability distribution of the pre-post effect of *walking to pain* in the Phendo cohort (a) and different Phendo individuals (b-e). The heterogeneity is evident, as the effect ranges from very hurtful to helpful for different individuals within the Phendo cohort: e.g., mostly positive pain effects in (b), null pain effects in (c), and mostly negative pain effects in (e).

> *I went to acupuncture when I was having really severe pain before my surgery.* (FG 6)

> *Cooking has been helpful to me. I have a task in front of me, I can get into a flow, and it is also good because it makes me feel like I'm in control of what I'm eating, what I'm creating. I can just close my eyes, deep breath and be distracted by something else.* (FG 6)

> *When I'm ovulating and menstruating, soups and broths help for that.* (Int 3)

Aligned with participant perspectives, we find that current Phendo users' responses to the same management strategy is heterogeneous, both at the population and individual levels. We showcase in Figure 5.9 *walking* to *pain severity* self-management pre-post effects for the population of Phendo users and specific individuals: i.e., a negatively skewed pre-post pain severity effect histogram implies that walking mostly reduces pain, while a positively skewed histogram showcases worsening of the experienced pain after walking. In Figure 5.9a, we observe how, although the effect of walking is null for a vast majority of Phendo users (notice the spike near the histogram origin), many users report both positive and negative effects in their pain severity within a day window. When looking at the individual (i.e., n-of-1) effects, we observe that the effect of the same self-management strategy (e.g., walking) is wildly heterogeneous for different users in the Phendo cohort. We illustrate the wide range of pain severity reported effects within a day of walking in Figure 5.9b-Figure 5.9e, where we observe how *walking* clearly helps reduce pain for the individual depicted in Figure 5.9e, but hurts the individual in Figure 5.9b; the same strategy does not have any impact for the individual in Figure 5.9c, while it is unclear on the effect for the individual in Figure 5.9d. We also corroborate the heterogeneity in population and individual effects for different strategies, as well as different goals (pain, GI, other symptoms) at the population level. This finding justifies the need for fully personalized self-management recommendation policies: i.e., one strategy does not suit all individuals with endometriosis.

5.3.7 (Agent/Policy - 4b) Users want an intelligent system with explainable recommendations to help understand why they were given particular recommendations and why they should try to implement the strategy.

The second code within the agent/policy category relates to the desire for explainability of recommendations. We code discussions about information participants want alongside recommendations in this category. Participants are interested in the reasons for trying a particular strategy and want to set their expectations about how long it may take to see an improvement in symptoms. All three participants interviewed about the intelligent system specifically tell us that they want explanations for recommendations provided by an intelligent system. They explain that they want to know why a strategy is being recommended, especially when the agent suggests strategies that have not yet been effective for an individual. Explanations would ideally include information about why the recommendation was chosen or tailored for the symptom or context. Participants suggest that explanations for what is offered and why would be helpful to users, especially when exploring options that may not make sense to the user or if the suggestions conflict with existing strategies that work. One person suggests that explanations might help motivate them to try the suggested strategies: *"I would be more prone to try it if there was something behind it."* (Int 2). These participant perspectives indicate that explanations could also help users build trust with the system.

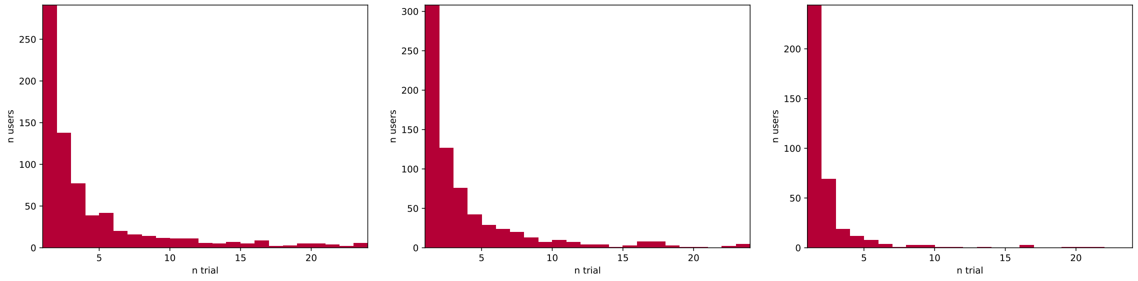
Participants are interested in implementing strategies that are responsive to their contextual environment, and report that insight from an intelligent system could help their understanding and exploration. One interview participant suggests that she be able to inspect and explore self-management options that are personally tailored for her, so that she may use the system's insights along with her own to decide on an appropriate action that is *"connected to how [I'm] actually feeling"* (Int 1).

5.3.8 (Agent/Policy - 4c) Engagement patterns suggest user interactions will be numerous and frequent enough for RL requirements.

The final code within the category of the agent/policy relates to users' willingness to engage with self-management and an intelligent system to support them. We code conversations about engaging in self-management and interacting with an intelligent system in this category. Users tell us that they are willing to extensively self-track their illness experiences and understand that an intelligent system for self-management would require active engagement. Participants are not concerned about the burden of using such a system (e.g., *"Logging observations is not cumbersome to me."* (Int 1)).

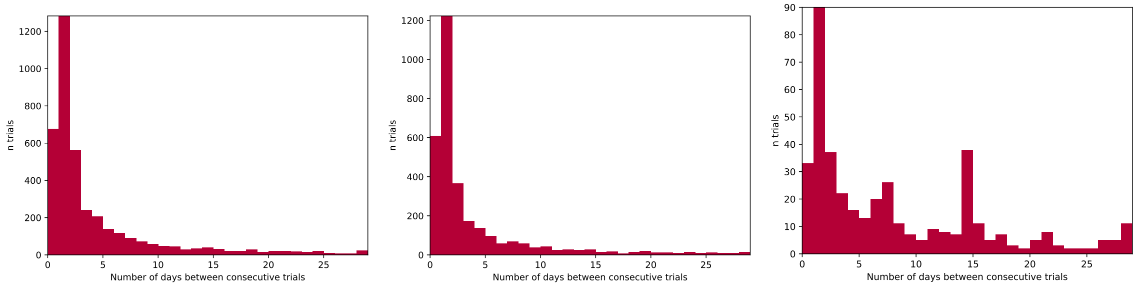
The analysis of current Phendo users' self-tracking engagement confirms that the frequency at which they track different self-management strategies is high — we find that many users track self-management strategies with associated goals many times within their timeline. In Figure 5.10, we observe that there are many Phendo users who have tracked up to 10 trials of *pain* to various strategies: *walking* (Figure 5.10a), *carbs, grains or gluten based foods* (Figure 5.10b), or *talk-therapy* (Figure 5.10c). In Figure 5.11, we also observe regular engagement: self-management trials occur as often as every day or every other day (see Figure 5.11a and Figure 5.11b), but also periodically, as in weekly or bi-weekly for *talk-therapy* to *pain* trials (observe the spikes at 8 and 15 days in Figure 5.11c). The observed "in the wild" number of interactions and frequency provides evidence for the feasibility of an RL-enabled system and the acceptability to users, but the specifics for programming an RL algorithm in practice will need to be determined.

Participants tell us that self-management is already part of their day-to-day routines and that they already dedicate significant time to these tasks, so committing to the self-experiments recommended by the system is seen to fit into their existing care regimens. Interview participants tell us they could easily incorporate 10-30 minutes into their daily routines, but also mention they would be open to recommendations that take one, two, or even three hours if those strategies might help their symptoms. However, participants explain that some strategies could be carried out more frequently than others.



(a) *Walking to pain* trials.
(b) *Carbs, grains or gluten foods to pain* trials.
(c) *Talk-therapy to pain* trials.

Figure 5.10: Histogram of the number of users per number of trials.



(a) *Walking to pain* trials.
(b) *Carbs, grains or gluten foods to pain* trials.
(c) *Talk-therapy to pain* trials.

Figure 5.11: Number of days between consecutive trials.

Participants generally agree that they try self-management strategies for about a month before deciding if they work or not, although sometimes they can tell sooner, even immediately. At the longer end of the scale, participants report willingness to experiment with strategies for several months or even up to a year to develop a regimen that works (e.g., *"Sometimes it was 60 days, sometimes it was a whole year."* (FG 9)). Further, participants explain that self-management will likely remain part of their ongoing care plan. While some users may use an intelligent system for a short time or on-and-off (e.g., during a flare-up) and then set it aside until the next episode where they need support (e.g., *"I would probably just move on until I have another acute pain episode, then I would be likely to use it again. I think, you try a couple of things. Maybe you learned something helpful."* (Int 1)), others envision using a system to support individualized self-management long-term (e.g., *"That's something I'm going to do the rest of my life, I've committed to that. [...] So, ongoing for sure. I don't think I'll ever reach my peak where I don't need this* *anymore, no.*" (Int 3)). These levels of engagement, while not guaranteed to result in a feasible RL algorithm, are promising from both the human and data perspectives.

5.3.9 Outline of findings and recommendations

Here, we lay out guidance for designing an intelligent interactive system that aligns with the human-data-machine insights and requirements elicited in this study. These recommendations address tradeoffs with the capabilities and constraints elaborated in this study and provide a concrete mechanism to convert the ideals articulated in the findings into real-world design decisions. The recommendations presented here provide a workable starting point for design. They also help to identify places where the sociotechnical gap [220] has been exacerbated by the use of ML or by the complex illness context, which provide guidance for future work and need for innovation. These recommendations are presented in the tables below.

Finding	**Human-Data-Machine Value Recommendations**
(1a) The scope of the action space is broad across the population of users, and individuals often combine multiple strategies.	***Balance diverse and personalized action space, with the need to constrain the set of available actions.*** Provide a diverse action space for the user population, to enhance user control and autonomy, as well as personalization of their experimentation. Leverage both pre-set Phendo vocabulary, and allow users to input their own user-customized strategies into the Phendo vocabulary. The action space should be discrete, but the cardinality (i.e., the number of possible actions to choose from) will need to be determined and agreed upon for each individual. A smaller set of actions to choose from means the system will be faster at finding an optimal self-management policy.
(1b) Users are willing to explore the action space, but control of the action space is important to participants.	***Support both experimentation with strategies already identified by user, and discovery of new strategies that may be effective. Facilitate exploration of the action space to support learning, shifting to exploitation over time as the RL learns an effective regimen.*** Provide strategy recommendations to experiment with one thing at a time for a particular health state, which can prevent users from getting overwhelmed and also help constrain the action space. ***Leverage user input to enable user control and autonomy, while meaningfuly constraining the action space.*** Facilitate user input, both globally and with each interaction. First, enable user input up front to elicit what users want to experiment with and what they are not willing to do based on individual constraints (e.g., remove strategies based on cost, schedule, or health status). Second, enable real-time input as recommendations are given, e.g., to decline a recommendation at that time or remove it from available options. Incorporating human-in-the-loop aspects of interactive RL could offset computational challenges by improving the efficiency of the model, while also resulting in a more personalized system that enhances user control and autonomy.

Table 5.5: Results and recommendations for (1) Action Space — Self-management strategies

Finding	**Human-Data-Machine Value Recommendations**
(2a) Experiences of illness and everyday variables and context are important for characterizing the state space.	***Leverage a diversity of user data about their health and day-to-day lives, to create tractable representations of users and their context that can inform the RL's learning process.*** Define personalized user states through careful human-centered machine learning: enable open user modeling questions (e.g., What is a good user-state? How to learn from both data and user input?) as well as data-driven learning of context (state space representation learning). Determine the appropriate dimensionality of the state space, since the rate for learning optimal policies depends sublinearly in the cardinality of the state space.
	Use existing computational approaches to enable representation of personalized, complex illness experiences in low-dimensional spaces, while innovating novel methods to representing the very human aspects of illness. Innovate ways to capture, compute, and represent embodied assessments and abstract conceptualizations for characterizing how a user is feeling. Digital phenotyping methods are one potential solution to capturing rich representations while constraining the dimensionality of the state space. Also devise mechanisms to handle these personalized user models as trajectories over time.
(2b) Users need recommendations that are suited to their current context.	***Use historical data and augment with sensors to provide real-time context.*** Address data, human and algorithmic challenges related to capturing and computing meaningful, real-world, real-time context from individuals. Rather than requiring that every user inputs granular data, which is burdensome and unrealistic, expand the data capturing capabilities of Phendo, e.g., with passive sensing, and combine them with state representation learning from historic data.
	Use existing self-tracked data about the user's current context and augment this information with more embodied sources of data, while at the same time exploring more creative methods to capture embodied experiences of illness. Innovate mechanisms to compute tractable representations of illness that can translate unique, multi-faceted, complex, and often subjective experiences into objective measures that could be quantified, measured, and compared across time in order to tailor recommendations based on these contextual experiences. For example, data could be captured via voice recording or artistic expression.

Table 5.6: Results and recommendations for (2) State Space — Health status and circumstances

Finding	Human-Data-Machine Value Recommendations
(3) Users are looking for short-term self-management, but even then there are many ways to compute the reward function.	***Start with simple, short-term, symptom-based reward functions as the baseline.*** Focus on short-term metrics for the design of RL reward functions. Conduct experiments to determine whether to use discrete (e.g., binary increase/decrease in pain reports) or continuous (e.g., difference in pain scores) reward functions. ***Expand to more complex, domain-derived, data-driven, individualized functions.*** Innovate methods to translate individualized metrics, which may be multi-faceted, complex, and subjective, into metrics that could be measured and used by an intelligent system for evaluating if a recommended strategy worked or not. Expand beyond simple reward functions, towards domain-derived but data-driven individualized scoring functions (e.g., composite scores combining pain, GI, mood, and symptom self-reports). A data driven approach would automatically learn the best state/reward function for a human-defined goal. However, it is critical to incorporate domain expertise and user input as well, to align each individual's notion of success with the reward function. e.g., if the individual's goal is short-term pain reduction, the RL agent can focus on the change in self-tracked pain within the desired pre-post management time window.

Table 5.7: Results and recommendations for (3) Reward — Goal in evaluating the success of self-management

Finding	Human-Data-Machine Value Recommendations
(4a) Heterogeneity in self-management responses calls for tailored recommendations.	***Balance individualized with population-based policies — start by augmenting with similar users, then transition to more fully individualized policies over time.*** Consider the trade-off between providing individualized recommendations (critical to an intelligent system's success) and the data requirements (long sequence of interactions) to learn individualized policies. Explore and leverage modeling and statistical tradeoffs between fully individualized techniques and hierarchical or pooling models to learn from similar (state, action, or effect) evidence. Bayesian mixed model effects and hierarchical models can pull statistical power across population evidence; and clustering approaches can help with dimensionality reduction.
(4b) Users want an intelligent system with explainable recommendations to help understand why they were given particular recommendations and why they should try to implement the strategy.	***Focus on explainable models to provide users with information about decision-making and insights into their illness and management.*** Avoid complex and black-box system based RL solutions (e.g., Deep-RL), and instead resort to explainable options: e.g., model-based, Bayesian sequential decision policies that allow for statistical and explainable modeling of the state to reward functions, facilitating the discovery of individualized insights.
(4c) Engagement patterns suggest user interactions will be numerous and frequent enough for RL requirements.	***Leverage the alignment between system requirements and users' documented behaviors in self-tracking and self-management, and operationalize the specific system requirements with experiments.*** Conduct simulated and/or trial experiments to determine the minimum number and periodicity of interactions required to learn an RL policy in the context of endometriosis self-management.

Table 5.8: Results and recommendations for (4) Agent/ Policy — Individualized self-management recommendations

5.4 Discussion

Intelligent systems — powered by AI and large volumes of patient-contributed data — have the potential to support humans in managing chronic illness [240]. However, their real potential has not yet been fully realized [241, 129]. Arguably, one barrier is limitations in existing design approaches: user-centered design, while well-established in the context of more traditional software applications, does not account for the unique constrains and inherent structures of AI models and, thus, may lead to requirements not possible to meet with existing AI algorithms [208, 209]. On the other hand, technology-driven approaches to the development of AI may lead to interactive systems inconsistent with user needs and result in limited adoption and unintended harms [127]. To move towards integrating these systems into the management of illness, we propose and implement an HAI framework — which we have termed Multi-Perspective Directed Analysis. We use MPDA to conduct a mixed-methods study to map and synthesize human, data, and ML requirements and constraints to generate design recommendations for an AI-enabled solution in the context of uncertainty in chronic illness. This approach also enables us to identify and elaborate on several sociotechnical gaps [220], where we document a mismatch between the complex, nuanced demands of real-world self-management and the rigid limitations of existing technologies. Here, we discuss implications from synthesizing needs identified by users, "in the wild" self-tracked data, and constraints of an RL approach for management of a complex, poorly understood disease.

5.4.1 Reflection on the human-centered AI framework MPDA — Accounting for unique affordances of ML/AI

Developing AI-enabled systems presents unique challenges compared to traditional software development. While a conventional design process allows for user needs to be identified, refined, and integrated, AI models have inherent requirements and constraints that limit adaptability. ML systems are less flexible since they operate within predefined conceptual spaces, with each model having its own limitations. This makes it difficult to simply identify and refine user needs to adjust an ML model, as we would with traditional software development. And unlike traditional approaches, we cannot easily create new ML methods to fit specific user needs. When working with AI, each ML method has its own configuration that we must work within to design intelligent systems. For example, if we were interested in neural networks, the design process would require careful consideration of input and output features. This difference highlights the need for a specific approach when designing AI-enabled systems, requiring attention to both user needs and technical capabilities of ML.

While sociotechnical approaches, like participatory design [242, 243, 244], have long emphasized the importance of aligning and optimizing both human and technical aspects in system design [60], the emergence of AI calls for new approaches to address how this automation reshapes the relationship between humans and technology and creates more complex socio-technical environments [245]. While some work has been done in this area [246, 247, 248, 249], there is value in a framework that facilitates designing intelligent systems that can account for human, machine, and data perspectives explicitly.

We have proposed an HAI framework to synthesize human, data, and ML perspectives – we use high-level concepts of a particular ML approach, RL, as guiding principles to synthesize and triangulate findings across quantitative and qualitative data sources. We use each perspective (ML, human, data) to guide the others and to organize the evidence. MPDA is unique in how we structure inquiry around specific ML concepts, and each "perspective" serves the analysis in important ways. First, the ML perspective allows us to account for the specific requirements and constraints of a particular ML approach, which is a novel aspect of this framework. Second, the human perspective ensures that the system is grounded in and designed in accordance with real-world needs, perspectives, and practices; the qualitative data also guides the development of the final results categories. Third, the data perspective allows us to triangulate the perspectives reported by users, validate the findings from the qualitative analysis, and augment those results with a large sample. Since these data are self-tracked "in the wild," they represent actual user behavior and practices of what humans are willing to do, in the absence of an AI system, and can inform what type of data may need to be supplemented. This integrated approach provides a comprehensive understanding of both technical and human factors to inform the design of a real-world system.

We demonstrate that the MPDA framework enables structuring the design process. Unlike traditional mixed-methods studies that rely on triangulation through comparing quantitative and qualitative data, our approach compares information from three distinct perspectives, providing a more comprehensive understanding of the specific task and illness context. This approach has helped manage and reduce the complexity of the problem space, yielding valuable insights that capture the nuances of human experience and illustrates how to translate these recommendations into actionable next steps. While MPDA cannot guarantee the success of an RL algorithm, it offers a structured way to explore its potential by addressing key questions around system requirements, user acceptance, and data feasibility.

In addition to helping identify specific requirements for RL-based intelligent systems for endometriosis, the framework also has implications for human-centered AI and both re-affirms some of its established principles and also suggests less explored ones. Below we describe how our proposed framework connects to HAI.

5.4.2 Upholding human-centered AI

The framework facilitates human-centeredness in two key ways. First, the framework itself explicitly accounts for the perspectives of people, data, and a specific algorithm — relying on human users as experts helps to avoid dehumanizing them or creating harmful technologies [250]. Second, using this framework has enabled us to generate recommendations that foreground the human-centerdness required for a successful real-world intelligent system. Particularly in the realm of AI, technology development frequently centers the algorithmic perspective. In HAI, it is crucial to consider both human and technological perspectives systematically. Our framework introduces a principled way to incorporate a human-centered perspective into the design of intelligent systems, synthesizing it with algorithmic constraints. Crossing the ML, human, and data perspectives addresses this critical need in the development of intelligent interactive systems and at the same time enables us to foreground human-centered principles in design recommendations.

- **Designing for control and user autonomy.** Facilitating human intuition integrated with computational support is a fundamental goal of HAI [58]. Thus, enabling user control and autonomy are key features for designing human-centered AI systems. Using the MPDA framework, we were able to identify several areas where incorporating user input and human-in-the-loop aspects of RL may facilitate control and autonomy while at the same time could offset computational challenges. In particular, in defining and constraining the action space, users wish to input preferences and limits upfront, as well as provide feedback to the system as recommendations are given (e.g., if they want to engage in a strategy or how they might want to modify it so that it fits their current needs). Existing applications provide examples for how features of interactive RL could accomplish these goals [237], which can work to simultaneously meet the needs of users and the requirements of RL, while advancing principles of HAI.

- **Enabling explainability.** Explainability can enable personalization [251] and is itself central to HAI [252, 128, 50]. Explainability in this context is particularly important, given the uncertainty around the illness and the high demands of user involvement with an RL, so users will constantly be deciding when to carry out recommendations given by the system. For an explainable RL-based system, designers should avoid black-box algorithms [253, 232, 233], opting instead for more inherently transparent and explainable approaches [254, 255, 256]. A small but quickly growing body of literature highlights explainable RL approaches specifically [257, 258, 259, 260, 261, 262] that could be adapted to this personal health context.

- **The importance of safety, privacy, and trust.** Although safety, privacy, and trust are critical principles of HAI systems [56], participants in our study did not identify these as important features. However, this does not mean that they are not important to users. We suspect that participants did not mention these requirements because they assume they are already being addressed, which is dangerous in the design of AI. The fact that these important design requirements did not come up in our study represents a limitation to our approach. Other researchers adopting this approach will need to be mindful to explicitly consider critical HAI aspects that may not be identified by users.

In addition, we highlight several more key insights related to less commonly emphasized areas of HAI that arose in the context of endometriosis. Taken together, the user needs and design recommendations generated in this study are consequences of the complexity and unpredictability of managing endometriosis. This is not a given, and in fact there are other aspects of the disease that emerge as most prominent in different settings. When seeking a diagnosis, the invisible nature of the disease is prominent [263]. In the context of shared decision making, research has documented that needs are more closely related to the stigma of the disease, privacy around disclosures, and managing emotions associated with illness [132]. This dynamic around the complexity and uncertainty of the illness resulted in specific requirements and recommendations directly relevant to this study context. We identify opportunities and challenges of using ML in a design space with nuanced, holistic representations of patient data.

- **Personalization for individualized, internal context.** We document wide variations in illness experiences and identify diverse approaches to self-management. While the need for personalization in chronic disease self-management is well understood [264, 265], it also presents a challenge for ML and AI that tend to require large volumes of data and are often less amenable to personalized approaches. This study helped to highlight needs and opportunities for personalization from the perspective of RL, suggesting the need to individualize the *action space*, *state space*, *reward*, and the *agent/policy*. While there are practical and computational challenges in translating human experiences into informative and meaningful representations that can be used by RL for the learning process, in particular that require reconciling human and algorithmic inputs, personalized user modeling has been addressed from various perspectives across the literature [266, 267]. There are also new ML approaches to guide the development of personalized reward functions [268]. Furthermore, considering the prospect of designing a personalized ML-enabled system from the data perspective suggests that individuals may be willing to engage in self-monitoring that can pave the foundation for the successful application of RL. However, the study also showed the need to constrain the action space, outlining several opportunities.

- **Embodied health experiences.** Individuals' perceptions of their illness experiences are multi-faceted, complex, unique, and sometimes subjective. They often relate to messy, embodied sensations. At the same time, computational tools require tractable representations of health status to evaluate if self-management strategies offer improvements. In an RL-enabled system, it will be essential for the *state space* to be able to account for and represent complex and subjective health states, and also for the *reward* to be able to evaluate if strategies are effective using a signal that is responsive to how individuals uniquely experience and evaluate their health. This means that an interactive system will need some way to translate these embodied health experiences into something that could be quantified, measured, and compared across time. Feminist scholars [269] and information theorists [270] have discussed how these properties of data require simplifying reality and imposing classifications that may not align with an individual's experience. Pichon et al. has also documented the gap between the messy, embodied tracking needs of menstruators and existing technologies [271]. Responding to this requirement calls for creative, human-centered technical innovations. Solutions might involve capturing multi-modal data (e.g., voice, video, or artistic creations) and applying digital phenotyping methods to create rich representations of health statuses that users are involved in categorizing and labeling.

- **Real-time, real-world context.** An intelligent interactive system for self-management will need to adapt to individuals' needs and preferences, respond to life circumstances and uncertainty, and provide dynamic recommendations that account for real-time context. Thus, representations that account for real-time, real-word context are important for characterizing the *state space* and for learning individualized *policies* — i.e., tailoring recommendations, providing relevant and dynamic recommendations, recognizing triggers or patterns, and enabling explainability for why a recommendation was made [272]. Disregarding context prevents a holistic representation of users [273], and instead paints a flat, shallow picture of an individual that is detached from their non-clinical experiences and broader environment. These fragmented and decontextualized representations can negatively impact patients [274]. The need for incorporating context is not new [275, 276, 226, 276, 277, 278]. Nevertheless, in the context of RL for a complex chronic illness, it will be challenging to capture, represent, and use complex, real-world, real-time context in ways that are not at odds with an RL's requirement for a constrained state and action space.

Designing intelligent systems to support management in the context of complex chronic illness represents an example of Ackerman's sociotechnical gap [220] — i.e., there is a known discrepancy between the nuanced, flexible, and contextual real-world task of self-management, and the rigid and brittle capabilities of technology. These key challenges represent examples of this gap, where the shortcomings of technical systems cannot fully support the complexities of human experiences and activities required in the task of self-management that we documented in this study. In particular, people with endometriosis describe complex and unpredictable experiences with their illness, which are challenging to translate into solutions that align with current technical capabilities. In this work, we not only articulate where these gaps arise but also present recommendations for solutions that partially solve these problems with known tradeoffs and provide work-arounds, in particular that enhance human-centered elements of control and autonomy.

5.4.3 Implications for the state of care for endometriosis patients

Unlike conditions with clearly established and well-defined treatment guidelines, people with endometriosis lack reliable, evidence-based treatments, leaving them to navigate their care through self-management [102, 103, 100, 33]. Yet despite this uncertainty, self-management helps empower individuals with endometriosis in their own care, mitigate their symptoms, and cope with the burden of their illness [132]. Even though recommendations provided by intelligent systems may be limited or ineffective, the trial-and-error process of finding strategies that work for an individual remains an essential aspect of care. As the standard of care advances, patients would benefit from systems that support this exploratory process. Given the documented complexity and uncertainty surrounding endometriosis, technology that can help patients identify patterns and discover new management strategies has significant implications for improving the state of care for endometriosis.

We document that individuals already engage in a complex, personal trial-and-error process and are interested in computational tools to support their self-management. Even in the absence of an AI, individuals already log sufficient self-tracked data to suggest that an RL-enabled system could be effective in assisting endometriosis care. We advocate for solutions that offer computational support for both self-experimentation with strategies already identified (i.e., evaluating if strategies are effective) and for discovery (i.e., recommending new strategies). Such tools could empower individuals in their care, support them in developing effective individualized management regimens, and improve quality of life. And while HAI scholars warn against falling into the "solutionism trap" that frames ML as a simple solution to a difficult problem [279, 280], our research affirms that ML may be an appropriate approach to this complex, unresolved problem.

We also situate our work within a feminist research perspective [281, 282, 283], particularly recognizing that individuals coping with a burdensome women's health disease like endometriosis must contend with issues of bodily autonomy, health equity, and society's disregard for women's health concerns [100, 284, 285, 286]. The deeply personalized and embodied experiences of endometriosis are often minimized and neglected, underscoring the need for solutions that center the highly contextualized and varied symptomatic manifestations of the illness. We call for human-centered technologies that leverage self-tracking data to elevate the perspective of users, enabling complex, contextualized representations of illness that can facilitate solutions to support users in their care tasks. We also emphasize the importance of innovating methods to translate individual, nuanced experiences of illness to tractable representations that can be leveraged in intelligent systems to advance personal health goals.

Chapter 6: Conclusions and Future Work

6.1 Conclusion

In this thesis, we set out to understand and support the work of patients and providers in complex chronic illness, with endometriosis as the use case. In **Chapter 2**, we laid the groundwork for the rest of the thesis through a review of the relevant literature. In **Chapter 3**, we addressed Aim 1, which was to elicit patient and provider needs and gaps in technology. We engaged endometriosis patients and specialists to identify their health and technology needs. We found that patients and providers engage in significant work to care for and manage endometriosis, which is complicated by the complexity of the enigmatic illness. We documented a range of needs, and focus on helping individuals to characterize and make sense of their health status and also in engaging with the trial-and-error process of self-management in the rest of the thesis. In **Chapter 4**, we addressed Aim 2, which was to develop and evaluate interpretable, temporal health status phenotypes. We developed and evaluated digital phenotypes of health statuses to support individuals in making sense of their own health and also that might be used computationally in intelligent systems. In **Chapter 5**, we addressed Aim 3, which was to map the requirements and constraints for adaptive self-management recommendations. We developed and implemented a human-centered AI framework — Multi-Perspective Directed Analysis — to map the human, data, and ML requirements of an intelligent system. We identified promising directions for an RL-based intelligent interactive system for management of complex chronic illness, and elaborated on several sociotechnical gaps. The health status phenotypes from Chapter 4 could serve as the computational state space for this proposed solution. In **this chapter**, we synthesize the contributions of this thesis, enumerate limitations of the research, and elaborate on future work to carry this research forward.

6.2 Contributions

This dissertation makes several important contributions to advance the management of complex chronic illness, the development of personal informatics technologies, and the field of human-centered AI. This research can advance biomedical informatics, human-computer interactions, and endometriosis research. There are also implications for developing tools to support endometriosis and other complex chronic illnesses. Further, the HAI approach taken here can provide a resource to others seeking to address complex health contexts in an ethical and human-centered manner. The key contributions of this dissertation include:

- **Contributions for human-centered AI.** *First*, this work both leverages human-centered AI approaches and advances the available approaches for carrying out HAI research. AI-enabled personal informatics tools have the potential to offer real-world support, but their benefits have not yet been fully realized due to barriers in translating technologies into pragmatic systems. We rely on patient-generated data and studies with human end-users, which centers the patient perspective in the design of new technologies. We also innovate new approaches to identify and align both human needs and technological requirements within the frame of HAI.

  Our novel approach, Multi-Perspective Directed Analysis, provides a framework and an example how to map human needs to technical requirements, bridging social science and data science perspectives within HAI. The MPDA framework highlights the potential of HAI in translating patient needs into actionable computational design requirements. We also provide a reproducible approach for tackling open questions in health and other domains through human-centered AI. This approach could be adopted by other researchers to address open questions that require HAI solutions.

  We also identified other key HAI insights for the development of personal informatics tools, like the need to design mechanisms to incorporate embodied health experiences into data representations and intelligent systems, as well as the importance of attending to both internal and external context in the development of such systems.

- **Contributions for informatics and technology.** *Second*, this dissertation contributes to advancing personal informatics technologies. We document a range of technology gaps and opportunities to innovate solutions to address these gaps, in the context of complex chronic illness. This thesis has produced interpretable health status representations (temporal phenotypes) and recommendations for an AI-enabled personal informatics tool that offers individualized recommendations. These technologies and findings expand the literature on chronic illness support, and make contributions to the fields of biomedical informatics and human-computer interactions. The research throughout this dissertation leverages human-centered AI approaches to explore the potential of AI in real-world, complex health scenarios, a benefit not yet fully realized for chronic disease management. We demonstrated the benefits and potential value of these technologies. At the same time, we also document and expand on key sociotechnical gaps in developing solutions to meet user needs in this space, where current technology is unable to meet the complex human needs. Despite the barriers in translating these technologies into practical systems, this work explores how AI might enhance chronic disease management through pragmatic, user-centered solutions.

- **Contributions for endometriosis.** *Finally*, we advance illness-specific endometriosis research. The research in this dissertation addresses significant gaps in supporting endometriosis, a prevalent but under-researched and under-supported condition affecting a sizable subset of the population — about 10% of women. This is an illness context that substantially burdens individuals with the disease and the healthcare system more broadly, and further, significant gaps in research and technologies to support care and management persist.

  We have documented and described how the complexities of the illness impact and complicate the work of patients and providers in caring for this complex chronic illness. Along with this, we have identified a number of needs and technology gaps for individuals and their care teams. This information serves as a foundation for designing tools and systems tailored to this context. While we have addressed several of the identified needs and gaps (i.e., difficulties in making sense of data and health status, challenges with individualized self-management), we also document sundry other opportunities to develop computational supports and interactive systems to meet the needs of individuals and their care teams. By articulating the needs of patients and their care teams in this illness context and opportunities for technologies to support these needs, we make an important contribution to the literature. While we only address the needs and opportunities to support the care of a single complex chronic illness, some of the insights that we have generated may provide useful guidance to others seeking to support individuals in the care and management of other conditions, particularly when there are knowledge gaps and significant heterogeneity across the population.

6.3 Limitations

We acknowledge various limitations in the research described in this dissertation. Some are due to the study designs and resources selected for the research. Various limitations might be resolved with future research, or by others undertaking similar research. These limitations include the following:

- **Reliance on the Phendo app.** The Phendo app provides an established self-tracking tool from which to draw real-world user data for quantitative analysis and a pool of end-users to recruit research participants for qualitative analysis and participatory design. And while the tool also gives us a platform upon which to build the proposed RL-enabled intelligent system, it also limits the design of the proposed system.

  Further, the data for phenotyping represent the user perspective, but are already set and cannot be changed, thus some important aspects could be missing (e.g., data related to real-time context and environmental factors). The self-tracked data domains, options, and granularities have been pre-specified, through prior work with end-users; but since they have been established and programmed into the Phendo app, it is not trivial to change.

The focus on the Phendo app and its users could also impede the generalizability of this research, since participants are likely to be highly engaged in both their own care and technology to support their illness.

- **Missingness.** For the quantitative research included in this thesis, some of the patterns of missingness represent a limitation to the research. For one, missing data impacts the coverage for the Baseline phenotypes for Aim 2, where a significant portion of user-week instances are missing a Baseline phenotype assignment, due to not having data tracked for the "How was your day?" field in that week. It may be useful to explore what other data the patterns of missingness are associated with.

  Further, while some people only track their illness experiences in Phendo for a short duration, others track intermittently over a longer period of time but may have periods of sparse or no data. It is difficult to infer what might be happening during these periods of missing data. It would be useful in the future to explore various patterns around engagement and missingness to characterize someone's health status when they have no data tracked for that time period.

  One solution may be to impute missing data or to leverage other data sources to infer the missing data, e.g., passively tracked data.

- **Data saturation.** For the qualitative research included in this thesis, there are some lingering limitations around data saturation, particularly in the context of the secondary analysis conducted for Aim 3. While secondary data analysis can maximize available resources to provide valuable insights, it inherently carries the limitation of not being directly shaped by the specific research questions. In Aim 3, the analysis was grounded in existing data, which may not fully capture the diversity of viewpoints that would emerge from direct, primary data collection. To combat this, we conducted three primary interviews that served as member checks. Still, it is difficult to ascertain if there may be other perspectives relevant to the research questions that we did not capture. The small sample size of primary interviews limits the breadth of viewpoints gathered, and while the interviews were valuable for validating findings, they may not have provided the depth needed to fully capture the range of experiences that could contribute to the richness of the research.

- **Representativeness of samples.** For both quantitative and qualitative aspects of the research in this thesis, there are limitations to the representativeness of the samples. The samples in these studies might not be fully representative of all people with endometriosis, particularly those who are not engaged in tracking their symptoms or using health apps. While participants in this research represent an array of illness experiences, they are not representative of the broader population of endometriosis patients in other ways. There are limitations in terms of race and ethnicity. We have also engaged people who are engaged in their care, either formally or on their own (e.g., through self-management). Importantly, these individuals are also invested in and willing to engage in research, and especially citizen science research. They may be more health and technology literate than average. While some of these characteristics may be representative of individuals who would use intelligent systems to support them in the care of their complex chronic illness, the data used in this thesis also may not capture the full range of experiences of those who have endometriosis.

- **Phenotyping approach.** The approach to constructing meaningful and interpretable state space representations may have some limitations. With so much unknown about health status in this context, it is difficult to know what defines meaningful clusters or representations for this purpose. We also relied on a previously validated phenotyping approach, the mixed-membership model. While we have evidence that this approach is a valuable one and have justifications for relying on this method, we did not investigate other phenotyping approaches that could have been used.

- **Limited evaluations.** We conducted only small-scale evaluations. While this provided valuable insights into the performance of our methods and results, we are also limited in generalizability. Future work would benefit from broadening the scope of evaluations. This should include further user studies, in particular prospective studies.

- **Focus on endometriosis.** Our focus on endometriosis is important since it is such an understudied condition that significantly burdens individuals and the healthcare system. Nevertheless, the generalizability of this research is further limited, because we focused narrowly on a particular condition. Endometriosis is poorly understood and burdens patients and care teams in particular ways, thus the insights garnered here may not extend to other contexts. The generalizability of our findings to other conditions and patient populations was not assessed in this thesis.

- **Limited scope.** This thesis does not directly address the needs of providers, although they have been consulted. Additionally, this work sets up future research that could design intelligent systems to support the needs of providers alongside patients.

  This thesis also does not address system-level concerns, nor does it directly address concerns over tech equity/justice, data and health literacy, or access to technology and healthcare — all important considerations.

6.4 Future work

This work can be expanded upon in various directions in the future. Some of these opportunities arise out of the aforementioned limitations of this research. Other future work has been identified from some of the findings of this work, and thus this dissertation opens up opportunities to continue advancing the aims of this research. Some of the potential future work includes:

- **Address other identified needs.** Aim 1 of this thesis identified a number of needs of patients and their care teams, and a wide variety of technological gaps that intelligent systems could help address. We only addressed a small number of these needs, thus there is an opportunity to develop additional solutions to meet the needs of patients and providers with future work.

  For example, patients identified a need to support their biographical work in crafting their illness narratives through data, which may be somewhat supported via the phenotypes. Future research could directly incorporate patients own narratives into their data representations.

Patients and providers also identified a significant need and technology gap in being able to identify and correct misalignments in support of the patient-provider partnership. While the technologies addressed in this thesis do have some implications to support the clinical encounter, they are predominantly in support of patients in their independent work. Thus, future research should work to address the collaborative work of patients and providers.

- **Leverage validated phenotypes.** In Aim 2, we developed interpretable, temporal phenotypes of health status that could be useful to individuals and also used by AI-enabled intelligent systems. However, we only explored the use of the phenotypes in a single RL-enabled system for supporting personalized self-management. Future work could implement and evaluate the use of these phenotypes in the proposed RL-enabled solution, as well as many other types of intelligent systems for supporting health and management. While the learned phenotypes performed only marginally well in the user study and computational task, potential applications for summarization could be developed now (e.g., to help individuals aggregate their data by health status to look for trends in their data and prepare for clinical visits), and others could be developed after further work on the phenotypes.

- **Incorporate other data types.** The phenotypes presented in Aim 2 could be improved by incorporating other data types, including passively tracked data, sensor data, and biometrics (e.g., voice recordings). These data sources could mitigate the burden of tracking while providing the model with additional useful information about a person's health status and broader context. These data sources could also mitigate issues with missingness, and could be used to further triangulate information already incorporated into the model (e.g., could give insight as to whether someone's lack of tracking may be due to a good day vs a bad day, based on activity, location, or other relevant information).

  We could also incorporate data that are already available in Phendo to improve the phenotyping model. We could leverage additional ML methods, for example Natural Language Processing (NLP) to incorporate the open-ended journal text that Phendo users already track, for example by using topic modeling to extract topics that someone discusses or using sentiment analysis to classify the text based on positive, negative, or neutral emotional tone. Using NLP with the journal entries could be one way to incorporate the person's broader life and circumstances into the model.

There is also an opportunity to expand the way that existing user-entered fields are included in the phenotypes, for example extracting more useful details about medication. Future work should go beyond including medications as a binary feature and include a high-level mapping to dis-aggregate, for example, pain medication from birth control.

Finally, it would be useful to incorporate meaningful information around patterns of engagement and missingness. For example, is someone missing data within a user-week because they have had a particularly bad week or because it has been an unusually good week. Future studies could work towards phenotyping the underlying health states behind different engagement and missingness patterns.

- **Develop the RL-enabled intelligent system.** In Aim 3, we identified human and technological requirements and constraints for an intelligent system to support the trial-and-error process of developing personalized self-management regimens. Future work could take this research forward by developing the RL-enabled system explored here. This will require more iterative, human-centered AI work that engages users while at the same time conducting experiments to train and implement an RL agent for this task.

- **Iterative improvements and broader scale evaluation.** Future work should consider scaling the implementation and evaluation of the technologies presented in this thesis, especially after iterative improvement integrating feedback from end-users elicited in this research. To further develop the phenotypes, we could rely on various existing techniques to understand and improve performance. For example, we could leverage the critical incident technique [287] to better understand participants' perspectives on events or circumstances that they perceive as having a major impact on an outcome. Analysis of these incidents could help to identify patterns and insights, which could be translated to inform phenotyping methods. We could also use the multi-trait, multi-method technique [288] to understand the underlying dynamics of what is going on with health status and to improve the model, especially in expanding to incorporate different data sources, in calibrating the models, and in evaluating the generalizability of the health status phenotypes.

Both the phenotypes, from Aim 2, and future intelligent systems created using these phenotypes, such as the one proposed in Aim 3, could benefit from a larger-scale evaluation. In particular, a prospective study would be useful in evaluation of the phenotypes. The intelligent system proposed in Aim 3 could benefit from real-world user studies, as the technology is developed through participatory design.

References

[1] R. Milani and C. Lavie, “Health care 2020: Reengineering health care delivery to combat chronic disease,” *Am J Med*, 2015.

[2] C. Hajat and E. Stein, “The global burden of multiple chronic conditions: A narrative review,” *Preventive Medicine Reports*, 2018.

[3] C. G. N. Mascie-Taylor and E. Karim, “The burden of chronic disease,” *Science*, 2003.

[4] B. Klijs, W. J. Nusselder, C. W. Looman, and J. P. Mackenbach, “Contribution of chronic disease to the burden of disability,” *PLOS ONE*, 2011.

[5] D. Yach, C. Hawkes, C. L. Gould, and K. J. Hofman, “The global burden of chronic DiseasesOvercoming impediments to prevention and control,” *JAMA*, 2004.

[6] M. Zanin, J. M. Tuñas, and E. Menasalvas, “Understanding diseases as increased heterogeneity: A complex network computational framework,” *Journal of The Royal Society Interface*, 2018.

[7] A. Ostropolets, R. Chen, L. Zhang, and G. Hripcsak, “Characterizing physicians’ information needs related to a gap in knowledge unmet by current evidence,” *JAMIA Open*, 2020.

[8] J Corbin and A Strauss, “Managing chronic illness at home: Three lines of work,” *Qual Sociol*, 1985.

[9] A Strauss, S Fagerhaugh, B Suczek, and C Wiener, “Sentimental work in the technologized hospital,” *Sociol Health Illn*, 1982.

[10] Institute of Medicine Committee on Quality of Health Care in America, *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington (DC): National Academies Press (US), 2001.

[11] D. A. Epstein *et al.*, “Mapping and taking stock of the personal informatics literature,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2020.

[12] G Demiris and L Kneale, “Informatics systems and tools to facilitate patient-centered care coordination,” *Yearb Med Inform*, 2015.

[13] F. Rajabiyazdi *et al.*, “Involving patients in their care plan: Patients’ and care providers’ perspectives,” in *Proceedings of the CHI Workshop on Interactive Systems in Healthcare (WISH’16)*, 2016.

[14] M. C. Figueiredo and Y. Chen, “Patient-generated health data: Dimensions, challenges, and open questions,” *Foundations and Trends® in Human–Computer Interaction*, 2020.

[15] F. C. Udegbe, O. R. Ebulue, C. C. Ebulue, and C. S. Ekesiobi, “The role of artificial intelligence in healthcare: A systematic review of applications and challenges,” *International Medical Science Research Journal*, 2024.

[16] R Karkar *et al.*, “TummyTrials: A feasibility study of using self-experimentation to detect individualized food triggers,” in *Proc ACM CHI Conf*, Denver, Colorado, USA: ACM Press, 2017.

[17] J Schroeder, J Hoffswell, C. Chung, J Fogarty, S Munson, and J Zia, “Supporting patient-provider collaboration to identify individual triggers using food and symptom journals,” in *Proc ACM CSCW Conf*, ser. CSCW ’17, Portland, Oregon, USA: ACM, 2017.

[18] D Buls and J Rooksby, “Technology for self-management of rosacea: A survey and field trial,” in *Proc ACM CHI Conf*, ser. CHI EA ’17, New York, NY, USA: ACM, 2017.

[19] J Schroeder *et al.*, “Examining self-tracking by people with migraine: Goals, needs, and opportunities in a chronic health condition,” in *Proceedings of the 2018 Designing Interactive Systems Conference*, ser. DIS ’18, New York, NY, USA: ACM, 2018.

[20] M. I. Ahmed, B. Spooner, J. Isherwood, M. Lane, E. Orrock, and A. Dennison, “A systematic review of the barriers to the implementation of artificial intelligence in healthcare,” *Cureus*, 2023.

[21] M. Hassan, A. Kushniruk, and E. Borycki, “Barriers to and facilitators of artificial intelligence adoption in health care: Scoping review,” *JMIR Human Factors*, 2024.

[22] S. Grimme, S. M. Spoerl, S. Boll, and M. Koelle, “My data, my choice, my insights: Women’s requirements when collecting, interpreting and sharing their personal health data,” in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ser. CHI ’24, New York, NY, USA: Association for Computing Machinery, 2024.

[23] K. Zondervan, C. Becker, K Koga, S. Missmer, R. Taylor, and P Viganò, “Endometriosis (primer),” *Nature Reviews: Disease Primers*, 2018.

[24] S. Agarwal *et al.*, “Clinical diagnosis of endometriosis: A call to action,” *Am J Obstet Gynecol*, 2019.

[25] P Acién and I Velasco, "Endometriosis: A disease that remains enigmatic," *Int Sch Res Notices*, 2013.

[26] J Brown and C Farquhar, "Endometriosis: An overview of cochrane reviews," *Cochrane Database Syst Rev*, 2014.

[27] I Urteaga, M McKillop, S Lipsky-Gorman, and N Elhadad, "Phenotyping endometriosis through mixed membership models of self-tracking data," in *Proc Machine Learning for Health Care*, ser. MLHC '18, 2018.

[28] E. Dancet, S Apers, J. Kremer, W. Nelen, W Sermeus, and T. D'Hooghe, "The patient-centeredness of endometriosis care and targets for improvement: A systematic review," *Gynecol Obstet Invest*, 2014.

[29] E. I. Geukens, S. Apers, C. Meuleman, T. M. D'Hooghe, and E. A. F. Dancet, "Patient-centeredness and endometriosis: Definition, measurement, and current status," *Best Practice & Research Clinical Obstetrics & Gynaecology*, Endometriosis: Impact and pathogenesis, 2018.

[30] C. J. Peek, M. A. Baird, and E. Coleman, "Primary care for patient complexity, not only disease," *Families, Systems, & Health*, 2009.

[31] J. D. Lee and A. Hohler, "Communication challenges in complex medical environments," *Continuum (Minneapolis, Minn.)*, 2014.

[32] E. H. Wagner, B. T. Austin, and M. Von Korff, "Organizing care for patients with chronic illness," *The Milbank Quarterly*, 1996.

[33] K Seear, "The third shift: Health, work and expertise among women with endometriosis," *Health Sociol Rev*, 2009.

[34] C. S. Sayegh *et al.*, "Designing an mHealth roadmap for the journey to self-management: A qualitative study with adolescents and young adults living with chronic illness," *Chronic Illness*, 2023.

[35] S. M. Mohsenizadeh, Z. S. Manzari, H. Vossoughinia, and H. Ebrahimipour, "Reconstruction of individual, social, and professional life: Self-management experience of patients with inflammatory bowel disease," *Journal of Education and Health Promotion*, 2021.

[36] H MacLeod, A Tang, and S Carpendale, "Personal informatics in chronic illness management," in *Proceedings of Graphics Interface 2013*, ser. GI '13, Toronto, Ont., Canada: Canadian Information Processing Society, 2013.

[37] I. Scott, P Scuffham, D Gupta, T. Harch, J Borchi, and B Richards, “Going digital: A narrative overview of the effects, quality and utility of mobile apps in chronic disease self-management,” *Aust Health Rev*, 2018.

[38] L. Crosby, N. Joffe, J Peugh, R. Ware, and M. Britto, “Pilot of the chronic disease self-management program for adolescents and young adults with sickle cell disease,” *J Adolesc Health*, 2017.

[39] W Choi, S Wang, Y Lee, H Oh, and Z Zheng, “A systematic review of mobile health technologies to support self-management of concurrent diabetes and hypertension,” *JAMIA*, 2020.

[40] C. Hui, R Walton, B McKinstry, T Jackson, R Parker, and H Pinnock, “The use of mobile applications to support self-management for people with asthma: A systematic review of controlled studies to identify features associated with clinical effectiveness and adherence,” *JAMIA*, 2017.

[41] A Gupta, X Tong, C Shaw, L Li, and L Feehan, “FitViz: A personal informatics tool for self-management of rheumatoid arthritis,” in *HCI International 2017 – Posters’ Extended Abstracts*, C. Stephanidis, Ed., ser. Communications in Computer and Information Science, Cham: Springer International Publishing, 2017.

[42] P. Desai, E. Mitchell, M. Hwang, M. Levine, D. Albers, and L Mamykina, “Personal health oracle: Explorations of personalized predictions in diabetes self-management,” in *Proc ACM CHI Conf*, New York, NY, USA: ACM, 2019.

[43] A. Pollack *et al.*, “Closing the gap: Supporting patients’ transition to self-management after hospitalization,” in *Proc ACM CHI Conf*, ser. CHI ’16, New York, NY, USA: ACM, 2016.

[44] N Daskalova, K Desingh, A Papoutsaki, D Schulze, H Sha, and J Huang, “Lessons learned from two cohorts of personal informatics self-experiments,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2017.

[45] N Daskalova *et al.*, “Self-e: Smartphone-supported guidance for customizable self-experimentation,” in *Proc ACM CHI Conf*, Yokohama Japan: ACM, 2021.

[46] R Karkar *et al.*, “A framework for self-experimentation in personalized health,” *JAMIA*, 2016.

[47] J Lee, E Walker, W Burleson, M Kay, M Buman, and E. Hekler, “Self-experimentation for behavior change: Design and formative evaluation of two approaches,” in *Proc ACM CHI Conf*, ser. CHI ’17, New York, NY, USA, 2017.

[48] A Ayobi, P Marshall, A. Cox, and Y Chen, “Quantifying the body and caring for the mind: Self-tracking in multiple sclerosis,” in *Proc ACM CHI Conf*, ser. CHI ’17, Denver, Colorado, USA: ACM Press, 2017.

[49] I Nahum-Shani *et al.*, “Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support,” *Ann Behav Med*, 2017.

[50] W Xu, “Toward human-centered AI: A perspective from human-computer interaction,” *Interactions*, 2019.

[51] S. Schmager, I. Pappas, and P. Vassilakopoulou, “Defining human-centered AI: A comprehensive review of HCAI literature,” *MCIS 2023 Proceedings*, 2023.

[52] U. A. Usmani, A. Happonen, and J. Watada, “Human-centered artificial intelligence: Designing for user empowerment and ethical considerations,” in *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2023.

[53] S Chancellor, “Toward practices for human-centered machine learning,” *Commun ACM*, 2023.

[54] T. Capel and M. Brereton, “What is human-centered about human-centered AI? a map of the research landscape,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg Germany: ACM, 2023.

[55] B. Shneiderman, *Human-Centered AI*. Oxford University Press, 2022.

[56] O Ozmen Garibay *et al.*, “Six human-centered artificial intelligence grand challenges,” *Int J Hum Comput Interact*, 2023.

[57] B. Shneiderman, “Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems,” *ACM Transactions on Interactive Intelligent Systems*, 2020.

[58] B Shneiderman, “Human-centered artificial intelligence: Reliable, safe & trustworthy,” *Int J Hum Comput Interact*, 2020.

[59] B. Shneiderman, “Human-centered artificial intelligence: Three fresh ideas,” *AIS Transactions on Human-Computer Interaction*, 2020.

[60] E Trist, “The evolution of socio-technical systems: A conceptual framework and an action research program,” *Ontario Quality of Working Life Centre*, 1981.

[61] B. Shneiderman, “Human-centered artificial intelligence: Reliable, safe & trustworthy,” *International Journal of Human–Computer Interaction*, 2020.

[62] U. B. Kirk *et al.*, "Understanding endometriosis underfunding and its detrimental impact on awareness and research," *npj Women's Health*, 2024.

[63] G. Márki, D. Vásárhelyi, A. Rigó, Z. Kaló, N. Ács, and A. Bokor, "Challenges of and possible solutions for living with endometriosis: A qualitative study," *BMC Women's Health*, 2022.

[64] L. Buggio, G. Barbara, F. Facchin, M. P. Frattaruolo, G. Aimi, and N. Berlanda, "Self-management and psychological-sexological interventions in patients with endometriosis: Strategies, outcomes, and integration into clinical care," *International Journal of Women's Health*, 2017.

[65] L. Giudice, "Endometriosis," *N Engl J Med*, 2010.

[66] M. Barry and S Edgman-Levitan, "Shared decision making – the pinnacle of patient-centered care," *N Engl J Med*, 2012.

[67] B. Paterson, C Russell, and S Thorne, "Critical analysis of everyday self-care decision making in chronic illness," *J Adv Nurs*, 2001.

[68] E. Wagner, B. Austin, C Davis, M Hindmarsh, J Schaefer, and A Bonomi, "Improving chronic illness care: Translating evidence into action," *Health aff*, 2001.

[69] T Bodenheimer, E. Wagner, and K Grumbach, "Improving primary care for patients with chronic illness," *Jama*, 2002.

[70] T Bodenheimer, K Lorig, H Holman, and K Grumbach, "Patient self-management of chronic disease in primary care," *Jama*, 2002.

[71] J. Corbin and A Strauss, *Unending work and care: Managing chronic illness at home* (Unending work and care: Managing chronic illness at home). San Francisco, CA, US: Jossey-Bass, 1988.

[72] H Holman and K Lorig, "Patient self-management: A key to effectiveness and efficiency in care of chronic disease," *Public Health Rep*, 2004.

[73] P. Carayon *et al.*, "Work system design for patient safety: The seips model," *BMJ Quality & Safety*, 2006.

[74] E. Shortliffe and M. Sepĺveda, "Clinical decision support in the era of artificial intelligence," *JAMA*, 2018.

[75] P. Beeler, D. Bates, and B. Hug, "Clinical decision support systems," *Swiss Med Wkly*, 2014.

[76] N Huba and Y Zhang, "Designing patient-centered personal health records (PHRs): Health care professionals' perspective on patient-generated data," *J Med Syst*, 2012.

[77] E. Jung, J Kim, K. Chung, and D. Park, "Mobile healthcare application with EMR interoperability for diabetes patients," *Cluster Comput*, 2014.

[78] M Rabbi, M. Aung, M Zhang, and T Choudhury, "MyBehavior: Automatic personalized health feedback from user behaviors and preferences using smartphones," in *Proc of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*, ser. UbiComp, Osaka, Japan: ACM, 2015.

[79] S. Mishra *et al.*, "Supporting coping with parkinson's disease through self tracking," in *Proc ACM CHI Conf*, New York, NY, USA: ACM, 2019.

[80] R Schnall, H Cho, A Mangone, A Pichon, and H Jia, "Mobile health technology for improving symptom management in low income persons living with HIV," *AIDS Behav*, 2018.

[81] T Davies, S. Jones, and R. Kelly, "Patient perspectives on self-management technologies for chronic fatigue syndrome," in *Proc ACM CHI Conf*, ser. CHI '19, New York, NY, USA: ACM, 2019.

[82] C. Chung, J Cook, E Bales, J Zia, and S. Munson, "More than telemonitoring: Health provider use and nonuse of life-log data in irritable bowel syndrome and weight management," *J Med Internet Res*, 2015.

[83] Y Kim *et al.*, "Prescribing 10,000 steps like aspirin: Designing a novel interface for data-driven medical consultations," in *Proc ACM CHI Conf*, ser. CHI 2017, Denver, Colorado, USA: ACM, 2017.

[84] P West, R Giordano, M Van Kleek, and N Shadbolt, "The quantified patient in the doctor's office: Challenges & opportunities," in *Proc ACM CHI Conf*, ser. CHI '16, Santa Clara, California, USA: ACM Press, 2016.

[85] C. Ruland, "Improving patient safety through informatics tools for shared decision making and risk communication," *Int J Med Inform*, Improving Patient Safety with Technology, 2004.

[86] A. Fiks, "Designing computerized decision support that works for clinicians and families," *Curr Probl Pediatr Adolesc Health Care*, 2011.

[87] H. Mentis *et al.*, "Crafting a view of self-tracking data in the clinical visit," in *Proc ACM CHI Conf*, ser. CHI '17, Denver, Colorado, USA: ACM Press, 2017.

[88] L Graham, A Tang, and C Neustaedter, “Help me help you: Shared reflection for personal data,” in *Proceedings of the 19th International Conference on Supporting Group Work*, ser. GROUP ’16, New York, NY, USA: ACM, 2016.

[89] C. Lim *et al.*, “Facilitating self-reflection about values and self-care among individuals with chronic conditions,” in *Proc ACM CHI Conf*, ser. CHI ’19, New York, NY, USA, 2019.

[90] A. Berry *et al.*, “Supporting communication about values between people with multiple chronic conditions and their providers,” in *Proc ACM CHI Conf*, ser. CHI ’19, New York, NY, USA, 2019.

[91] T Owen, J Pearson, H Thimbleby, and G Buchanan, “ConCap: Designing to empower individual reflection on chronic conditions using mobile apps,” in *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, ser. MobileHCI ’15, New York, NY, USA: ACM, 2015.

[92] A Bussone, S Stumpf, and S Wilson, “Designing for reflection on shared HIV health information,” in *Proceedings of the 13th Biannual Conference of the Italian SIGCHI Chapter: Designing the Next Interaction*, ser. CHItaly ’19, New York, NY, USA: ACM, 2019.

[93] M. Hong, U Lakshmi, T. Olson, and L Wilcox, “Visual ODLs: Co-designing patient-generated observations of daily living to support data-driven conversations in pediatric care,” in *Proc ACM CHI Conf*, ser. CHI ’18, New York, NY, USA: ACM, 2018.

[94] A. Ayobi, P. Marshall, and A. L. Cox, “Trackly: A customisable and pictorial self-tracking app to support agency in multiple sclerosis self-care,” in *Proc ACM CHI Conf*, ser. CHI ’20, Honolulu, HI, USA: ACM, 2020.

[95] S. Mishra *et al.*, “Supporting collaborative health tracking in the hospital: Patients’ perspectives,” in *Proc ACM CHI Conf*, ser. CHI ’18, New York, NY, USA: ACM, 2018.

[96] A. Piper and J. Hollan, “Supporting medical conversations between deaf and hearing individuals with tabletop displays,” in *Proc ACM CSCW Conf*, ACM, 2008.

[97] G Marcu, A. Dey, and S. Kiesler, “Designing for collaborative reflection,” in *PervasiveHealth*, 2014.

[98] A. Hartzler and W. Pratt, “Managing the personal side of health: How patient expertise differs from the expertise of clinicians,” *Journal of Medical Internet Research*, 2011.

[99] K Young, J Fisher, and M Kirkman, “Partners instead of patients: Women negotiating power and knowledge within medical encounters for endometriosis,” *Fem Psychol*, 2019.

[100] K Seear, "'nobody really knows what it is or how to treat it': Why women with endometriosis do not comply with healthcare advice," *Health Risk Soc*, 2009.

[101] E. Fillion, "How is medical decision-making shared? the case of haemophilia patients and doctors: The aftermath of the infected blood affair in france," *Health Expectations*, 2003.

[102] A. Young and A. Miller, ""this girl is on fire": Sensemaking in an online health community for vulvodynia," in *Proc ACM CHI Conf*, ser. CHI '19, New York, NY, USA: ACM, 2019.

[103] K Young, J Fisher, and M Kirkman, "Women's experiences of endometriosis: A systematic review and synthesis of qualitative research," *J Fam Plann Reprod Health Care*, 2015.

[104] N Daskalova *et al.*, "SleepCoacher: A personalized automated self-experimentation system for sleep recommendations," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, ser. UIST '16, New York, NY, USA: ACM, 2016.

[105] S Taylor, A Sano, C Ferguson, A Mohan, and R. Picard, "QuantifyMe: An open-source automated single-case experimental design platform," *Sensors (Basel, Switzerland)*, 2018.

[106] I Li, A Dey, J Forlizzi, K Höök, and Y Medynskiy, "Personal informatics and HCI: Design, theory, and social implications," in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11*, Vancouver, BC, Canada: ACM Press, 2011.

[107] F. Bentley *et al.*, "Health mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change," *ACM Transactions on Computer-Human Interaction (TOCHI)*, 2013.

[108] M. Rabbi, M. H. Aung, M. Zhang, and T. Choudhury, "Mybehavior: Automatic personalized health feedback from user behaviors and preferences using smartphones," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015.

[109] F. Cordeiro, E. Bales, E. Cherry, and J. Fogarty, "Rethinking the mobile food journal: Exploring opportunities for lightweight photo-based capture," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015.

[110] P. Karaturhan, E. Arıkan, P. Durak, A. E. Yantac, and K. Kuscu, "Combining momentary and retrospective self-reflection in a mobile photo-based journaling application," in *Nordic Human-Computer Interaction Conference*, ser. NordiCHI '22, New York, NY, USA: Association for Computing Machinery, 2022.

[111] C. Chung *et al.*, "Identifying and planning for individualized change: Patient-provider collaboration using lightweight food diaries in healthy eating and irritable bowel syndrome," *Proc ACM Interact Mob Wearable Ubiquitous Technol*, 2019.

[112] S. Kim *et al.*, "Toward becoming a better self: Understanding self-tracking experiences of adolescents with autism spectrum disorder using custom trackers," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, ser. PervasiveHealth'19, New York, NY, USA: ACM, 2019.

[113] L Mamykina *et al.*, "Grand challenges for personal informatics and ai," in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '22, New Orleans, LA, USA: ACM, 2022.

[114] E. G. Mitchell *et al.*, "From reflection to action: Combining machine learning with expert knowledge for nutrition goal recommendations," in *Proc ACM CHI Conf*, ser. CHI '21, New York, NY, USA: ACM, 2021.

[115] F Nunes, N Verdezoto, G Fitzpatrick, M Kyng, E Grönvall, and C Storni, "Self-care technologies in HCI: Trends, tensions, and opportunities," *ACM Trans Comput-Hum Interact*, 2015.

[116] N Elhadad, *Phendo app available at Apple's App store*, https://itunes.apple.com/us/app/phendo/id1145512423, 2021.

[117] N Elhadad, *Phendo app available at Google Play*, https://play.google.com/store/apps/ details?id=com.appliedinformaticsinc.phendo, 2021.

[118] H. Schmitz, C. L. Howe, D. G. Armstrong, and V. Subbian, "Leveraging mobile health applications for biomedical research and citizen science: A scoping review," *Journal of the American Medical Informatics Association*, 2018.

[119] M McKillop, N Voigt, R Schnall, and N Elhadad, "Exploring self-tracking as a participatory research activity among women with endometriosis," *J Particip Med*, e17 2016.

[120] M McKillop, L Mamykina, and N Elhadad, "Designing in the dark: Eliciting self-tracking dimensions for understanding enigmatic disease," in *Proc ACM CHI Conf*, 2018.

[121] M. Currie, B. S. Paris, I. Pasquetto, and J. Pierre, "The conundrum of police officer-involved homicides: Counter-data in los angeles county," *Big Data & Society*, 2016.

[122] I Ensari, A Pichon, S Lipsky-Gorman, S Bakken, and N Elhadad, "Augmenting the clinical data sources for enigmatic diseases: A cross-sectional study of self-tracking data and clinical documentation in endometriosis," *Appl Clin Inform*, 2020.

[123] N Elhadad, I Urteaga, S Lipsky-Gorman, and M McKillop, "User engagement metrics and patterns in phendo, an endometriosis research mobile app," *preprint*, 2022.

[124] A. F. Vitonis *et al.*, “World endometriosis research foundation endometriosis phenome and biobanking harmonization project: Ii. clinical and covariate phenotype data collection in endometriosis research,” *Fertility and sterility*, 2014.

[125] I Urteaga, M McKillop, and N Elhadad, “Learning endometriosis phenotypes from patient-generated data,” *NPJ digital medicine*, 2020.

[126] I Ensari, S Lipsky-Gorman, E. Horan, S Bakken, and N Elhadad, “Associations between physical exercise patterns and pain symptoms in individuals with endometriosis: A cross-sectional mhealth-based investigation,” *BMJ open*, 2022.

[127] E. Baumer, “Toward human-centered algorithm design,” *Big Data Soc*, 2017.

[128] R. Bond, M Mulvenna, and H Wang, “Human centered artificial intelligence: Weaving UX into algorithmic decision making,” 2019.

[129] T. Andersen, F Nunes, L Wilcox, E Coiera, and Y Rogers, “Introduction to the special issue on human-centred AI in healthcare: Challenges appearing in the wild,” *ACM Trans Comput-Hum Interact*, 2023.

[130] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, “Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI,” *Berkman Klein Center Research Publication*, 2020.

[131] R. Chatila and J. C. Havens, “The IEEE global initiative on ethics of autonomous and intelligent systems,” in *Robotics and Well-Being*, M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, and E. E. Kadar, Eds., Cham: Springer International Publishing, 2019.

[132] A Pichon *et al.*, “Divided we stand: The collaborative work of patients and providers in an enigmatic chronic disease,” *Proc ACM CSCW Conf*, 2020.

[133] C Charles, A Gafni, and T Whelan, “Shared decision-making in the medical encounter: What does it mean? (or it takes, at least two to tango),” *Soc Sci Med*, 1997.

[134] A. Fossa, S. Bell, and C DesRoches, “OpenNotes and shared decision making: A growing practice in clinical transparency and how it can support patient-centered care,” *J Am Med Inform Assoc*, 2018.

[135] S. Haldar, S. R. Mishra, M. Khelifi, A. H. Pollack, and W. Pratt, “Beyond the patient portal: Supporting needs of hospitalized patients,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’19, New York, NY, USA: ACM, 2019.

[136] D Kralik, T Koch, K Price, and N Howard, “Chronic illness self-management: Taking action to create order,” *J Clin Nurs*, 2004.

[137] C Storni, "Design challenges for ubiquitous and personal computing in chronic disease care and patient empowerment: A case study rethinking diabetes self-monitoring," *Pers Ubiquitous Comput*, 2014.

[138] F Hill-Briggs, "Problem solving in diabetes self-management: A model of chronic illness self-management behavior," *Ann Behav Med*, 2003.

[139] C Storni, "Patients' lay expertise in chronic self-care: A case study in type 1 diabetes," *Health Expect*, 2015.

[140] H. Clark and S. Brennan, "Grounding in communication," in *Perspectives on socially shared cognition*, Washington, DC, US: APA, 1991.

[141] E Coiera, "When conversation is better than computation," *J Am Med Inform Assoc*, 2000.

[142] S. Collins *et al.*, "In search of common ground in handoff documentation in an intensive care unit," *J Biomed Inform*, 2012.

[143] S. Collins, L Mamykina, D. Jordan, and D. Kaufman, "Clinical artifacts as a treasure map to navigate handoff complexity," in *Cognitive Informatics in Health and Biomedicine: Case Studies on Critical Care, Complexity and Errors*, ser. Health Informatics, V. Patel, D. Kaufman, and T Cohen, Eds., London: Springer London, 2014.

[144] A Mol, *The Logic of Care: Health and the Problem of Patient Choice*. Routledge, 2008.

[145] A Mol, *The Body Multiple: Ontology in Medical Practice*. Durham: Duke University Press, 2003, 216 pp.

[146] R. Deber, N Kraetschmer, and J Irvine, "What role do patients wish to play in treatment decision making?" *Arch Intern Med*, 1996.

[147] R. Deber, N Kraetschmer, S Urowitz, and N Sharpe, "Do people want to be autonomous patients? preferred roles in treatment decision-making in several patient populations," *Health Expect*, 2007.

[148] B Chewning, C. Bylund, B Shah, N. Arora, J. Gueguen, and G Makoul, "Patient preferences for shared decisions: A systematic review," *Patient Educ Couns*, 2012.

[149] A. Kon, "The shared decision-making continuum," *JAMA*, 2010.

[150] N Kraetschmer, N Sharpe, S Urowitz, and R. Deber, "How does trust affect patient preferences for participation in decision-making?" *Health Expect*, 2004.

[151] C. P. Lee, “Boundary negotiating artifacts: Unbinding the routine of boundary objects and embracing chaos in collaborative work,” *Computer Supported Cooperative Work (CSCW)*, 2007.

[152] C.-F. Chung *et al.*, “Boundary negotiating artifacts in personal informatics: Patient-provider collaboration with patient-generated data,” in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, ser. CSCW ’16, New York, NY, USA: ACM, 2016.

[153] E. Piras and F Miele, “Clinical self-tracking and monitoring technologies: Negotiations in the ICT-mediated patient-provider relationship,” *Health Sociol Rev*, 2017.

[154] I. Li, A. Dey, and J. Forlizzi, “A stage-based model of personal informatics systems,” *CHI*, 2010.

[155] D. A. Epstein, A. Ping, J. Fogarty, and S. A. Munson, “A lived informatics model of personal informatics,” in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, New York, NY, USA: ACM, 2015.

[156] R. S. Valdez, R. J. Holden, L. L. Novak, and T. C. Veinot, “Transforming consumer health informatics through a patient work framework: Connecting patients to context,” *JAMIA*, 2015.

[157] K Yin *et al.*, “Context-aware systems for chronic disease patients: Scoping review,” *J Med Internet Res*, 2019.

[158] H. Mentis, M Reddy, and M. Rosson, “Invisible emotion: Information and interaction in an emergency room,” in *Proc ACM CSCW Conf*, ser. CSCW ’10, New York, NY, USA: ACM, 2010.

[159] A Strauss, “Work and the division of labor,” *Sociol Q*, 1985.

[160] A Strauss, “The articulation of project work: An organizational process,” *Sociol Q*, 1988.

[161] I Hampson and A Junor, “Invisible work, invisible skills: Interactive customer service as articulation work,” *New Technol Work Employ*, 2005.

[162] M Berg, “Accumulating and coordinating: Occasions for information technologies in medical work,” *CSCW*, 1999.

[163] M Berg and G Bowker, “The multiple bodies of the medical record: Toward a sociology of an artifact,” *Sociol Q*, 1997.

[164] A. Büyüktür and M. Ackerman, "Information work in bone marrow transplant: Reducing misalignment of perspectives," in *Proc ACM CSCW Conf*, ser. CSCW '17, New York, NY, USA: ACM, 2017.

[165] A Mathieu-Fritz and C Guillot, "Diabetes self-monitoring devices and transformations in "patient work"," *Revue d'anthropologie des connaissances*, 2017.

[166] V Braun and V Clarke, "Using thematic analysis in psychology," *Qual Res Psychol*, 2006.

[167] J Cohen, "A coefficient of agreement for nominal scales," *Educ Psychol Meas*, 1960.

[168] C. Miaskowski *et al.*, "Advancing symptom science through symptom cluster research: Expert panel proceedings and recommendations," *JNCI: Journal of the National Cancer Institute*, 2017.

[169] J.-P. Onnela and S. L. Rauch, "Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health," *Neuropsychopharmacology*, 2016.

[170] J. Torous, M. V. Kiang, J. Lorme, J.-P. Onnela, *et al.*, "New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research," *JMIR mental health*, 2016.

[171] A. Vaidyam, J. Halamka, and J. Torous, "Enabling research and clinical use of patient-generated health data (the mindLAMP platform): Digital phenotyping study," *JMIR mHealth and uHealth*, 2022.

[172] K. Radhakrishnan *et al.*, "The potential of digital phenotyping to advance the contributions of mobile health to self-management science," *Nursing Outlook*, 2020.

[173] A. M. Bernardos, M. Pires, D. Ollé, and J. R. Casar, "Digital phenotyping as a tool for personalized mental healthcare," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, New York, NY, USA: ACM, 2019.

[174] X. Wang *et al.*, "HOPES: An integrative digital phenotyping platform for data collection, monitoring, and machine learning," *JMIR*, 2021.

[175] J.-P. Onnela, "Opportunities and challenges in the collection and analysis of digital phenotyping data," *Neuropsychopharmacology*, 2021.

[176] I. Moura, A. Teles, L. Coutinho, and F. Silva, "Towards identifying context-enriched multimodal behavioral patterns for digital phenotyping of human behaviors," *Future Generation Computer Systems*, 2022.

[177] G. Guo *et al.*, "MSLife: Digital behavioral phenotyping of multiple sclerosis symptoms in the wild using wearables and graph-based statistical analysis," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2022.

[178] T. R. Insel, "Digital phenotyping: Technology for a new science of behavior," *JAMA*, 2017.

[179] C. Molina, B. Prados-Suarez, and n, "Digital phenotypes for personalized medicine," *pHealth 2021*, 2021.

[180] C. Musto, G. Semeraro, C. Lovascio, M. de Gemmis, and P. Lops, "A framework for holistic user modeling merging heterogeneous digital footprints," in *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, New York, NY, USA: ACM, 2018.

[181] C. Wu *et al.*, "Multi-modal data collection for measuring health, behavior, and living environment of large-scale participant cohorts," *GigaScience*, 2021.

[182] N. Nag, V. Pandey, P. J. Putzel, H. Bhimaraju, S. Krishnan, and R. Jain, "Cross-modal health state estimation," in *Proceedings of the 26th ACM international conference on Multimedia*, New York, NY, USA: ACM, 2018-10-15.

[183] G. Hripcsak and D. J. Albers, "Next-generation phenotyping of electronic health records," *Journal of the American Medical Informatics Association*, 2013.

[184] J. Onnela, "Opportunities and challenges in the collection and analysis of digital phenotyping data," *Neuropsychopharmacol*, 2021.

[185] C. Allen, J. Hu, V. Kumar, M. A. Ahmad, and A. Teredesai, "Interpretable phenotyping for electronic health records," in *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, 2021.

[186] T. Aledavood, S. Lehmann, and J. Saramäki, "Digital daily cycles of individuals," *Frontiers in Physics*, 2015.

[187] M. Hodgman, C. Minoccheri, M. Mathis, E. Wittrup, and K. Najarian, "A comparison of interpretable machine learning approaches to identify outpatient clinical phenotypes predictive of first acute myocardial infarction," *Diagnostics*, 2024.

[188] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, 2012.

[189] R. Pivovarov, A. J. Perotte, E. Grave, J. Angiolillo, C. H. Wiggins, and N. Elhadad, "Learning probabilistic phenotypes from heterogeneous EHR data," *Journal of Biomedical Informatics*, 2015.

[190] I. Urteaga, M. McKillop, and N. Elhadad, “Learning endometriosis phenotypes from patient-generated data,” *npj Digital Medicine*, 2020.

[191] A Pichon, J Blumberg, L Mamykina, and N Elhadad, “The voice of endo: Leveraging speech for an intelligent system that can forecast illness flare-ups,” in *Accepted for publication in: Proc ACM CHI Conf*, 2025.

[192] A Pichon, I Urteaga, L Mamykina, and N Elhadad, “Informing the design of individualized self-management regimens from the human, data, and machine learning perspectives,” *Accepted for publication in: Transactions on Computer-Human Interaction (TOCHI)*, 2025.

[193] D Wang, Q Yang, A Abdul, and B. Lim, “Designing theory-driven user-centric explainable AI,” in *Proc ACM CHI Conf*, New York, NY, USA, 2019.

[194] C Morrison *et al.*, “Visualizing ubiquitously sensed measures of motor ability in multiple sclerosis: Reflections on communicating machine learning in practice,” *ACM Transactions on Interactive Intelligent Systems*, 2018.

[195] J Heer, “Agency plus automation: Designing artificial intelligence into interactive systems,” *Proc Natl Acad Sci*, 2019.

[196] J Auernhammer, “Human-centered AI: The role of human-centered design research in the development of AI,” *DRS Biennial Conf Series*, 2020.

[197] S Chandran, A Al-Sa’di, and E Ahmad, “Exploring user centered design in healthcare: A literature review,” in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2020.

[198] R De Croon, T De Buyser, J Klerkx, and E Duval, “Applying a user-centered, rapid-prototyping methodology with quantified self: A case study with triathletes,” in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2014.

[199] Y. Korpershoek, S Hermsen, L Schoonhoven, M. Schuurmans, and J. Trappenburg, “User-centered design of a mobile health intervention to enhance exacerbation-related self-management in patients with chronic obstructive pulmonary disease (copilot): Mixed methods study,” *JMIR*, 2020.

[200] E. Eaves *et al.*, “Applying user-centered design in the development of a supportive mHealth app for women in substance use recovery,” *Am J Health Promot*, 2023.

[201] N Tuzcu, A White, B Leonard, and S Geofrey, “Unraveling the complexity: A user-centered design process for narrative visualization,” in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA ’23, New York, NY, USA: ACM, 2023.

[202] S Chancellor, E. P. Baumer, and M De Choudhury, "Who is the "human" in human-centered machine learning: The case of predicting mental health from social media," *Proc ACM CSCW Conf*, 2019.

[203] J. Vaughn and H Wallach, "A human-centered agenda for intelligible machine learning," in *Machines We Trust: Perspectives on Dependable AI*, 2021.

[204] T Capel and M Brereton, "What is human-centered about human-centered ai? a map of the research landscape," in *Proc ACM CHI Conf*, 2023.

[205] J Fjeld, N Achten, H Hilligoss, A Nagy, and M Srikumar, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai," *Berkman Klein Center Research Publication*, 2020.

[206] H Zhu, B Yu, A Halfaker, and L Terveen, "Value-sensitive algorithm design: Method, case study, and lessons," *Proc ACM CSCW Conf*, 2018.

[207] A Zhang, A Boltz, J Lynn, C. Wang, and M. Lee, "Stakeholder-centered AI design: Co-designing worker tools with gig workers through data probes," in *Proc ACM CHI Conf*, ser. CHI '23, New York, NY, USA: ACM, 2023.

[208] T Bratteteig and G Verne, "Does AI make PD obsolete? exploring challenges from artificial intelligence to participatory design," in *Proc Participatory Design Conference*, ser. PDC '18, New York, NY, USA: ACM, 2018.

[209] D Schiff, B Rakova, A Ayesh, A Fanti, and M Lennon, *Principles to practices for responsible AI: Closing the gap*, 2020.

[210] C. Miller, J. Kristeller, A Headings, and H Nagaraja, "Comparison of a mindful eating intervention to a diabetes self-management intervention among adults with type 2 diabetes: A randomized controlled trial," *Health Educ Behav*, 2014.

[211] N New, "Teaching so they hear: Using a co-created diabetes self-management education approach," *J Am Acad Nurse Pract*, 2010.

[212] H. Cameron-Tucker, R Wood-Baker, C Owen, L Joseph, and E. Walters, "Chronic disease self-management and exercise in COPD as pulmonary rehabilitation: A randomized controlled trial," *Int J Chron Obstruct Pulmon Dis*, 2014.

[213] T Effing, G Zielhuis, H Kerstjens, P van der Valk, and J van der Palen, "Community based physiotherapeutic exercise in COPD self-management: A randomised controlled trial," *Respir Med*, 2011.

[214] K. Mitchell *et al.*, "A self-management programme for COPD: A randomised controlled trial," *Eur Respir J*, 2014.

[215] S Zhang, T Kang, L Qiu, W Zhang, Y Yu, and N Elhadad, "Cataloguing treatments discussed and used in online autism communities," in *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[216] S Chopra, R Zehrung, T. Shanmugam, and E. Choe, "Living with uncertainty and stigma: Self-experimentation and support-seeking around polycystic ovary syndrome," in *Proc ACM CHI Conf*, New York, NY, USA: ACM, 2021.

[217] L Mamykina, A. M. Smaldone, and S. R. Bakken, "Adopting the sensemaking perspective for chronic disease self-management," *J Biomed Inform*, 2015.

[218] L Mamykina *et al.*, "Structured scaffolding for reflection and problem solving in diabetes self-management: Qualitative study of mobile diabetes detective," *JAMIA*, 2016.

[219] R Brown, B Ploderer, L. Da Seng, P Lazzarini, and J van Netten, "MyFootCare: A mobile self-tracking tool to promote self-care amongst people with diabetic foot ulcers," in *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, ser. OZCHI '17, New York, NY, USA: ACM, 2017.

[220] M. Ackerman, "The intellectual challenge of cscw: The gap between social requirements and technical feasibility," *Hum-Comput Interact*, 2000.

[221] M Ghassemi, T Naumann, P Schulam, A. Beam, I. Chen, and R Ranganath, *A review of challenges and opportunities in machine learning for health*, 2019. arXiv: `1806.00388`.

[222] J Schroeder, R Karkar, J Fogarty, J. Kientz, S. Munson, and M Kay, "A patient-centered proposal for bayesian analysis of self-experiments for health," *J Healthc Inform Res*, 2019.

[223] N Daskalova *et al.*, "SleepBandits: Guided flexible self-experiments for sleep," in *Proc ACM CHI Conf*, ser. CHI '20, New York, NY, USA: ACM, 2020.

[224] D Almirall, S. Compton, M Gunlicks-Stoessel, N Duan, and S. Murphy, "Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy," *Stat Med*, 2012.

[225] E. Mitchell *et al.*, "From reflection to action: Combining machine learning with expert knowledge for nutrition goal recommendations," in *Proc ACM CHI Conf*, 2021.

[226] P Klasnja *et al.*, "Efficacy of contextually tailored suggestions for physical activity: A micro-randomized optimization trial of HeartSteps," *Ann Behav Med*, 2019.

[227] L. Coughlin *et al.*, "Toward a just-in-time adaptive intervention to reduce emerging adult alcohol use: Testing approaches for identifying when to intervene," *Substance Use & Misuse*, 2021.

[228] R. Sutton and A. Barto, *Reinforcement learning: an introduction* (Adaptive computation and machine learning series), Second edition. Cambridge, Massachusetts: The MIT Press, 2018, 526 pp.

[229] C Szepesvári, "Algorithms for reinforcement learning," *Synthesis lectures on artificial intelligence and machine learning*, 2010.

[230] Y Lin *et al.*, "A survey on reinforcement learning for recommender systems," *arXiv preprint arXiv:2109.10665*, 2022.

[231] M. M. Afsar, T. Crump, and B. Far, "Reinforcement learning based recommender systems: A survey," *ACM Computing Surveys*, 2022.

[232] V Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nat*, 2015.

[233] D Silver *et al.*, "Mastering the game of go without human knowledge," *Nat*, 2017.

[234] G Shani, D Heckerman, and R. Brafman, "An MDP-based recommender system," *Journal of Machine Learning Research*, 2005.

[235] A Zimdars, D. Chickering, and C Meek, "Using temporal data for making recommendations," *UAI*, 2001.

[236] O Gottesman *et al.*, "Guidelines for reinforcement learning in healthcare," *Nat Med*, 2019.

[237] C Arzate Cruz and T Igarashi, "A survey on interactive reinforcement learning: Design principles and open challenges," in *Proc Designing Interactive Systems Conf*, New York, NY, USA: ACM, 2020.

[238] H. Hsieh and S. Shannon, "Three approaches to qualitative content analysis," *Qual Health Res*, 2005.

[239] D Byrne, "A worked example of braun and clarke's approach to reflexive thematic analysis," *Qual Quant*, 2022.

[240] O Asan, E Choi, and X Wang, "Artificial intelligence–based consumer health informatics application: Scoping review," *JMIR*, 2023.

[241] E Coiera, "The last mile: Where artificial intelligence meets reality," *JMIR*, 2019.

[242] M. Muller, "Participatory design: The third space in HCI," in *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*, 2002.

[243] D Schuler and A Namioka, *Participatory Design: Principles and Practices*. CRC Press, 1993.

[244] P Ehn, “Participation in design things,” in *Proceedings of the Tenth Anniversary Conference on Participatory Design*, ACM Digital Library, 2008.

[245] F Delgado, S Yang, M Madaio, and Q Yang, “The participatory turn in AI design: Theoretical foundations and the current state of practice,” in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, ser. EAAMO ’23, New York, NY, USA: ACM, 2023.

[246] S Amershi *et al.*, “Guidelines for human-AI interaction,” in *Proc ACM CHI Conf*, ser. CHI ’19, New York, NY, USA, 2019.

[247] D. Olsen, “Evaluating user interface systems research,” in *Proceedings of the 20th annual ACM symposium on User interface software and technology*, ser. UIST ’07, New York, NY, USA: ACM, 2007.

[248] K Inkpen, S Chancellor, M De Choudhury, M Veale, and E. P. Baumer, “Where is the human? bridging the gap between AI and HCI,” in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA ’19, New York, NY, USA: ACM, 2019.

[249] D Norman, *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books, 2013.

[250] S. Chancellor, E. P. Baumer, and M. De Choudhury, “Who is the" human" in human-centered machine learning: The case of predicting mental health from social media,” *Proceedings of the ACM on Human-Computer Interaction*, 2019.

[251] F Sørmo, J Cassens, and A Aamodt, “Explanation in case-based reasoning–perspectives and goals,” *Artif Intell Rev*, 2005.

[252] U Ehsan, Q. Liao, M Muller, M. Riedl, and J. Weisz, “Expanding explainability: Towards social transparency in AI systems,” in *Proc ACM CHI Conf*, ser. CHI ’21, New York, NY: ACM, 2021.

[253] K Arulkumaran, M. Deisenroth, M Brundage, and A. Bharath, “Deep reinforcement learning: A brief survey,” *IEEE Signal Process Mag*, 2017.

[254] S Agrawal and N Goyal, “Thompson Sampling for Contextual Bandits with Linear Payoffs,” in *ICML*, 2013.

[255] I Urteaga and C Wiggins, “Variational inference for the multi-armed contextual bandit,” in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, A. Storkey and F. Perez-Cruz, Eds., ser. Proceedings of Machine Learning Research, Playa Blanca, Lanzarote, Canary Islands: PMLR, 2018.

[256] S Tomkins, P Liao, P Klasnja, and S. Murphy, “IntelligentPooling: Practical thompson sampling for mHealth,” *Mach Learn*, 2021.

[257] S Milani, N Topin, M Veloso, and F Fang, “A survey of explainable reinforcement learning,” *arXiv*, 2022.

[258] R Dazeley, P Vamplew, and F Cruz, *Explainable reinforcement learning for broad-XAI: A conceptual framework and survey*, 2021.

[259] A Krajna, M Brcic, T Lipic, and J Doncevic, *Explainability in reinforcement learning: Perspective and position*, 2022.

[260] C Glanois *et al.*, *A survey on interpretable reinforcement learning*, 2022.

[261] E Puiutta and E. Veith, *Explainable reinforcement learning: A survey*, 2020.

[262] L Wells and T Bednarz, “Explainable AI and reinforcement learning—a systematic review of current approaches and trends,” *Frontiers in Artificial Intelligence*, 2021.

[263] N Hudson, “The missed disease? endometriosis as an example of ‘undone science’,” *Reproductive biomedicine & society online*, 2022.

[264] S Dineen-Griffin, V Garcia-Cardenas, K Williams, and S. Benrimoj, “Helping patients help themselves: A systematic review of self-management support strategies in primary health care practice,” *PloS one*, 2019.

[265] S. Taylor *et al.*, “A rapid synthesis of the evidence on interventions supporting self-management for people with long-term conditions.(prisms practical systematic review of self-management support for long-term conditions),” *Health services and delivery research*, 2014.

[266] G Fischer, “User modeling in human–computer interaction,” *User Modeling and User-Adapted Interaction*, 2001.

[267] G. Webb, M. Pazzani, and D Billsus, “Machine learning for user modeling,” *User Modeling and User-Adapted Interaction*, 2001.

[268] A Mandyam, M Jorke, W Denton, B. Engelhardt, and E Brunskill, “Adaptive interventions with user-defined goals for health behavior change,” *Proc Mach Learn Res*, 2024.

[269] M Posner and L. Klein, “Editor’s IntroductionData as media,” *Feminist Media Histories*, 2017.

[270] G. Bowker and S. Star, *Sorting Things Out: Classification and Its Consequences*. MIT Press, 2000.

[271] A Pichon, K. Jackman, I. Winkler, C Bobel, and N Elhadad, “The messiness of the menstruator: Assessing personas and functionalities of menstrual tracking apps,” *JAMIA*, 2022.

[272] G Adomavicius and A Tuzhilin, “Context-aware recommender systems,” in *Recommender Systems Handbook*, F Ricci, L Rokach, B Shapira, and P. Kantor, Eds., Boston, MA: Springer US, 2011.

[273] K. Stange, “The problem of fragmentation and the need for integrative solutions,” *Ann Fam Med*, 2009.

[274] E. Bayliss *et al.*, “Understanding the context of health for persons with multiple chronic conditions: Moving from what is the matter to what matters,” *Ann Fam Med*, 2014.

[275] A. Dey, “Understanding and using context,” *Personal Ubiquitous Comput*, 2001.

[276] N. Villegas, C Sánchez, J Díaz-Cely, and G Tamura, “Characterizing context-aware recommender systems: A systematic literature review,” *Knowledge-Based Systems*, 2018.

[277] G Adomavicius and A Tuzhilin, “Context-aware recommender systems,” in *Recommender systems handbook*, Springer, 2010.

[278] S Raza and C Ding, “Progress in context-aware recommender systems—an overview,” *Comput Sci Rev*, 2019.

[279] E Morozov, *To Save Everything, Click Here: The Folly of Technological Solutionism*. PublicAffairs, 2013.

[280] A. Selbst, D Boyd, S Friedler, S Venkatasubramanian, and J Vertesi, “Fairness and abstraction in sociotechnical systems,” Social Science Research Network, Rochester, NY, 2018.

[281] C D’ignazio and L. Klein, *Data feminism*. MIT press, 2023.

[282] S Bardzell, “Feminist HCI: Taking stock and outlining an agenda for design,” in *Proc ACM CHI Conf*, New York, NY, USA, 2010.

[283] S Bardzell and J Bardzell, “Towards a feminist HCI methodology: Social science, feminism, and HCI,” in *Proc ACM CHI Conf*, ser. CHI ’11, New York, NY, USA: ACM, 2011.

[284] S Westwood *et al.*, “Disparities in women with endometriosis regarding access to care, diagnosis, treatment, and management in the united states: A scoping review,” *Cureus*, 2023.

[285] A Pettersson and C. Berterö, “How women with endometriosis experience health care encounters,” *Women’s Health Reports*, 2020.

[286] S Cunnington, A Cunnington, and A Hirose, “Disregarded, devalued and lacking diversity: An exploration into women’s experiences with endometriosis. a systematic review and narrative synthesis of qualitative data,” *Journal of Endometriosis and Uterine Disorders*, 2024.

[287] J. C. Flanagan, “The critical incident technique.,” *Psychological bulletin*, 1954.

[288] D. T. Campbell and D. W. Fiske, “Convergent and discriminant validation by the multitrait-multimethod matrix.,” *Psychological bulletin*, 1959.

Appendix A: Focus Group and Interview Guides

A.1 Interview Guide — 2019 Interviews with Providers

INTRO & INSTRUCTIONS

We have developed the Phendo app, an instrument to help endometriosis patients self-track a range of variables related to their condition. Now, we want to determine what would be useful for providers when meeting with these patients—both to get an overview update of the patient and as a basis to ground/anchor the clinical visit.

Endo patients see various types of providers. As we go through the scenarios, think about someone in your specialty. We will present a range of situations and are interested to hear how you or someone with your clinical expertise would approach them. The purpose of today's session is to help us develop tools that will work best for you and your patients.

We will be focusing on typical interactions with endo patients, using the patient scenarios below to ground our discussion (but please use your own patient experiences, too; cases are only guiding examples). Remember that you should think about patients that you already have a relationship/history with, not new patients or first-time visits. I am going to ask you to tell me about your typical interactions and will also show you several tools that may or may not help to get your feedback and improve the tools.

Before we start, could you tell me a bit about your background?

- Gender, Latest Degree, Type of Provider & Specialty, Endo Experience & Years
- What type of management do you provide for endometriosis? (e.g., surgical, hormonal treatments)

PATIENT SCENARIO #1 (Alice)

Alice is a 23yo G0P0 with endometriosis. She began to experience dysmenorrhea without menorrhagia as a 16yo resulting in missed school days and social outings. Her mom told her that her heavy menstrual flow was normal, and similar to hers at her age. She also suffers from stomach aches, asthma, headaches, and fatigue. When she was 18yo her doctor recommended starting OCP's as her history was concerning for endometriosis. Unfortunately, she never got any relief from OCP's and developed dyspareunia. She underwent a diagnostic laparoscopy 6mo ago with ablation of endometriotic implants. She is not currently on medication post-procedure.

She is casually dating and not currently trying to get pregnant; sex is still painful. She exercises 3x/wk, eats healthy, avoids gluten but does not have a diagnosed sensitivity. She has a close social support system. She is worried her doctors don't understand her and are ignoring her symptoms.

PATIENT SCENARIO #2 (Sophia)

Sophia is a 38yo G0P0 with endometriosis. She has had endometriosis as long as she can remember, having undergone 3x laparoscopic procedures in the past. Her symptoms are debilitating and have negatively impacted her personal (overwhelming symptoms prevented childbearing) and work life (unable to travel). She suffers from anxiety and fatigue. She is a "mature" endo patient with a holistic understanding of her disease, regularly trying new treatments in mini-self-experiments (tracking her progress!). She also reads the scientific literature to learn about endo.

She owns her home in Manhattan with her wife Olivia. She's an ambitious professional but slightly anti-social and prefers to spend weekends at home, sometimes inviting friends over for brunch. She feels frustrated with the healthcare system, feeling ignored and misunderstood and has largely disengaged from care.

I. TRADITIONAL VISIT (no tools)

1. How long is your new patient visit? Routine return visit?

2. Thinking about the two patient scenarios, what might a typical visit look like for each patient?

(a) How much time do you spend with patients for each of those activities?

(b) What is your goal for a patient visit? What is your overall goal across multiple visits?

- How are goals and priorities set? What does this depend on? Who participates (patient/provider) and how (roles)?
- Can you describe your typical patient-provider relationship?

(c) What is the process for determining treatment options and making decisions?

- How do/would visual aids or patient-materials fit into decision-making process?

(d) What do you think about "shared decision-making"? What would it look like to have patients and providers working together? Is that happening? Why/how or why/how not? (Positive/Negative)

- How often do you encounter more engaged patients, or those with strong opinions? How do you consider their preferences?
- Have you used shared decision making in other clinical scenarios? Have you & your patients found it to be helpful?

3. How would you assess the progress of your endo patient? Over what period of time?

(a) Are there specific symptoms you ask about? What are the specific symptoms?

- Pain only? Pelvic pain? Other symptoms, like fatigue? Other groups of symptoms (GI, touch sensitivity, other frequent symptoms)?

(b) What patterns do you look for? Are you interested in trends, changes? Absolutes?

- Over what period of time?
- Time of day?

(c) Does relationship to menses matter?

- How do your patients share/report this info to you?

4. For a patient under your care for some time, what type of info beyond symptoms are you interested in?

    (a) Emotional status? Satisfaction? Sex?

    (b) Personal factors in the patient's life? Routines? Stressors? Major life events or transitions?

5. How do you assess if a treatment or management plan was "successful"?

    (a) How might you assess efficacy of past or current medication history, treatments, or self-management? What do you ask about?

    (b) What's the best next step? Anything to be done for better management?

    (c) How do you help patients not feel discouraged?

    (d) To see improvements or positive outcomes, especially if patient experiences severe symptoms?

    (e) How do you manage failures in treatment or management plan?

Patient-Generated, Self-Tracking Data

6. Could you tell me about an experience when you worked with patient-collected data, specifically with Endometriosis?

    (a) Have you asked patients to track their experiences? What apps did they use/you recommend?

    (b) What symptoms were focused on?

    (c) What type of data?

    (d) Did you ask them to bring it, or did they bring it on their own?

    (e) Did you put this into the patient's medical record? If so, how?

    (f) How much time could be spent reviewing data? Before or during appointment?

7. Can you think about potential benefits of having patients record self-tracking data? Potential downfalls?

   (a) To the patient? Provider?

   (b) Do you think these data could be valuable to patients even if not shared with providers? How?

8. In an ideal world, how would you like to see patient-generated data and self-tracking used/not used in your clinic population?

   (a) Can you think about how utilizing self-tracking data might benefit or harm the patient-provider relationship? Can you think of benefits and/or downfalls in reviewing data with patients?

   (b) How may it impact outcomes?

   (c) How do you think your patients would feel about utilizing patient-generated data?

II. DASHBOARD TOOL

9. Imagine each of the two patients brought you this dashboard tool (see Fig A.1). Thinking of the tasks you just described with our patient scenarios, how would this tool change the visit?

   (a) Helpful? Not helpful? Problematic?

   (b) Would you rather see it before the visit or during? How long would you review?

   (c) How might the tool facilitate communication between you and the patient?

   (d) One complaint about self-tracked data is that there's often too much information tracked; what information is important/useful for you to see and what is unnecessary or too much?

10. How would you use the tool to understand the success of a new treatment (diet, exercise, hormones, surgery)?

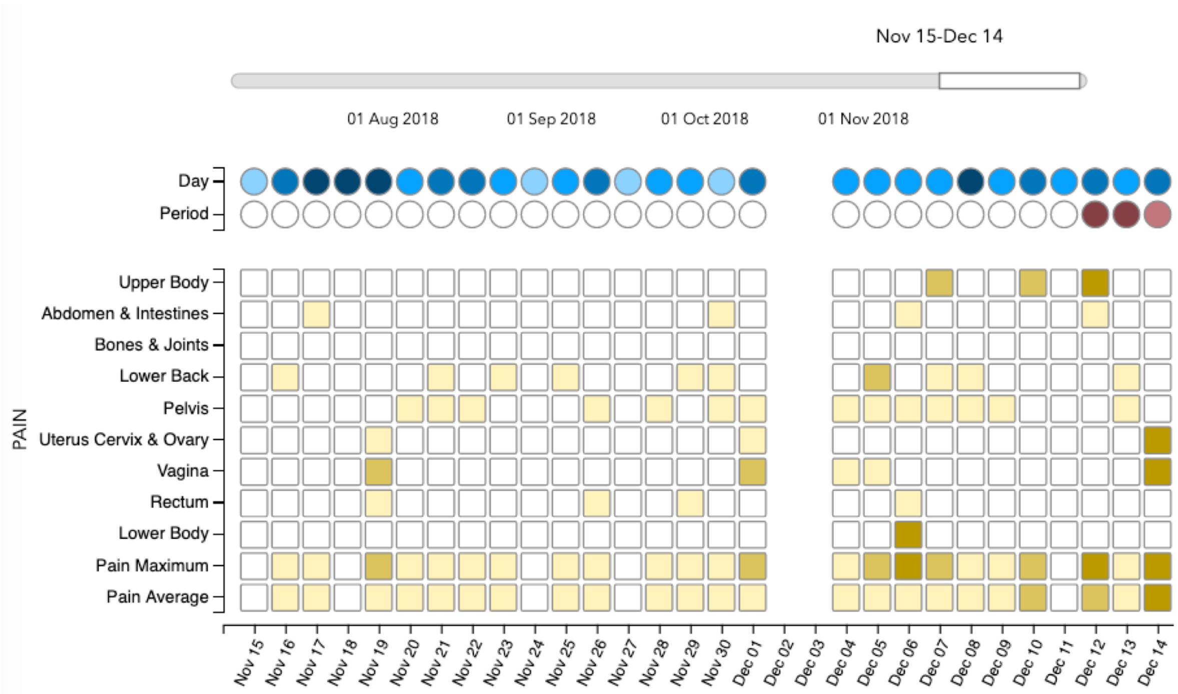


Figure A.1: Heatmap visualization shown to providers. They were able to view two simulated cases, and the visualization was interactive (they could scroll left-right through the timeline and up-down through the domains).

(a) How long until you expect to see results? What kind of results are you looking for?

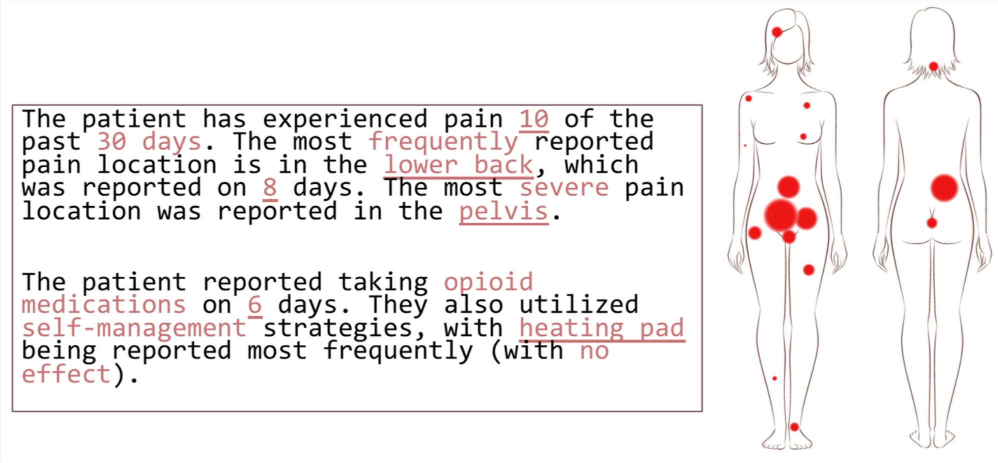
Additional questions

11. When a patient has a specific concern aside from surgery, who generally provides this treatment or consult? Do endo surgeons refer to gynecologist for hormonal treatments, or do the surgeons act as the endo specialist?

12. When you evaluate an endo patient, do you compare to a cohort? A typical patient? Would this be useful to you? Do you compare the individual to themselves at previous timepoint?

13. Before we wrap up, is there anything else you think we should know? Anything else we should have asked?

III. SUMMARY TOOL



The patient has experienced pain 10 of the past 30 days. The most frequently reported pain location is in the lower back, which was reported on 8 days. The most severe pain location was reported in the pelvis.

The patient reported taking opioid medications on 6 days. They also utilized self-management strategies, with heating pad being reported most frequently (with no effect).

Figure A.2: Summary visualization shown to providers.

14. Imagine each of the two patients brought you this summary tool (see Fig A.2). Thinking of the tasks you just described with our patient scenarios, how would this tool change the visit?

    (a) Helpful? Not helpful? Problematic?

    (b) Would you rather see it before the visit or during? How long would you review?

    (c) How might the tool facilitate communication between you and the patient?

A.2 Focus Group Guide — 2019 Focus Groups with Patients

Introductions

You are all here because you have endometriosis and have had symptoms for at least the past three months. We are going to talk about them probably for most of our time here! For now, just as a quick icebreaker, let's go around the room and share our names, can you tell us how long you've had symptoms for, how long you've been diagnosed. If you have one fun fact about endo, please feel free to share!

Self-Assessment of Health Status

Our first set of questions to you is about your symptoms and your general health status, and in particular, how you assess how well you are doing beyond day-to-day experiences. Here, and for the rest of the session, we are looking into health status as a whole, not only aspects that are endometriosis-specific. There are two reasons for this. First, we know that many patients, and maybe some of you here, have other chronic conditions in addition to endometriosis, and our goal is to build tools that support you as a whole person. The second reason is that it is hard to know what symptoms or signs are due to endometriosis or something else. So, let's assume we want to know about your general health status as a patient of endometriosis.

1. I can imagine a provider asking you this question of "how have you been feeling in the past three months," how often do you ask yourself this question?
    - Daily? Weekly? Some specific event triggers the question? Why or why not?
    - How do you interpret this question of how have you been doing? (overall functionality, emotions, specific symptoms that you know are particularly important for you to monitor)?
    - Typically how far back in time would you go to assess your health status and reflect upon it? (e.g., Would it make more sense for you to ask "in the past week, how have I been feeling?" or rather on the other extreme "in the past year, how have I been feeling?")
2. What part of that assessment and answering the question of "how have you been feeling" is easy for you? What is hard?
    - Are there specific symptoms that are easier than others for you to assess status? Any that are harder? (e.g., Maybe because of time over which they occur or because they change a lot or some other patterns?)

3. Do you use any technology to help you recall or reflect back on your health status through time?

    - If so, what tools do you use and how do you find them useful? Why? Why not?
    - If not or if you are dissatisfied with current tools, what are the functionalities you'd like to see in technology that could support you in reflecting back on your health and assessing your status through time?

Patient-Provider Communication

For the rest of this session, we are going to focus on you and your care team. By care team, we mean healthcare providers who help manage your endometriosis symptoms. Examples of care team members could be a surgeon or a physical therapist.

1. Who does your endometriosis care team consist of?

    - Do they communicate with each other?
    - What does a typical visit look like?

Now, we are interested in how you and your care team communicate. Because we want to build tools that help with shared decision making, that is coming to a treatment decision that make sense for you as a patient and your provider as well, we are very interested in the kind of role you play in your care and the role your provider has. For instance, some patients see themselves as their own advocates, or the primary communicator between the different providers on your care team, if you have one.

2. What is the process for decision-making around your care?

    - What role do you play in your care? Why?
    - How satisfied are you with your role? Why? Why not?

3. How do you convey your goals to your providers? What about your preferences? An example of goal would be “I want to be pain free for a week.” An example of preference is “I’d rather not go on hormonal therapy because I am afraid of the side effects.”

    - Do you have a specific time when you meet with your provider to go through your goals and preferences? Or do you let it be more organic during encounters? (e.g., “if it comes up” or “if my doctor asks me”)
    - How do you tell if your provider has heard you? Specifically, when it comes to treatments, how do you tell if your provider has taken your preferences and goals into account?
    - Are you satisfied with how you convey your goals and preferences? If not, what would you want to see differently? (e.g., have enough time/opportunity other than during encounter to convey them)
    - How would technology help you with conveying your goals and preferences with your care team? (e.g., check list for patients to review during encounters?)

4. What works and what is missing with your communication with your provider?

    - Does it differ among the different members of your care team?
    - How do you ensure your concerns are heard?

Self-Management of Health

Now, we want to move the discussion away from assessing your health status, and thinking about self-management. By self-management, we mean all the different actions that you take to help you deal on a day to day basis with your symptoms, whether you know they work for you or you are experimenting with.

1. Would you say you self-manage your endometriosis?

    - Why or why not?

- What factors make you decide to try it or not? Other endometriosis patients' experiences? Trusted source? Feasibility with your daily life? (i.e., cost, time, effort)

2. How many different self-management strategies would you say you experimented with in the past year?

    - What were they?
    - How do you evaluate if a specific strategy works for you?
    - How long do you typically give a specific strategy a go before deciding whether it is helpful to you or not?
    - How do you stick with them? What are the challenges with doing so? What would help you stick with them better?
    - Thinking about technology, how would you ideally use it to help you evaluate a self-management strategy? How would technology help with this? Think of Phendo or any other self-tracking app, what is missing from it that would help you determine whether a specific self-management strategy work for you?

3. How do you find out about different self-management strategies?

    - Reading? Online? Social media, blogs? Your doctors? Your social network (friends, family)? Other patients?
    - Would you say there are a lot of self-management strategies for endometriosis? And if so, you have no problem identifying new ones, or on the contrary do not know where to find them?

Before we wrap up, is there anything else you think we should know? Anything else we should have asked?

A.3 Interview Guide — 2020 Interviews with Patients

We have developed the Phendo app, an instrument to help endometriosis patients self-track a range of variables related to their condition. Now, we want to determine what would be useful for patients to use to support self-management.

The purpose of today's session is to help us develop tools that will work best for you. We will ask you about how you currently self-manage your condition, and what kinds of tools you imagine would be helpful. Remember there are no wrong answers (we just want to hear your experiences and opinion) and you won't hurt our feelings! Oh, and please make yourself comfortable (stand or sit, snacks, bathroom, etc)...

First we are interested in your current approach to self-management and experimentation... By self-management, we mean all the different actions that you take to help you deal on a day to day basis with your symptoms, whether you know they work for you or you are experimenting with.

1. So, let me ask you: given the definition we just gave, would you say you self-manage your endometriosis?

    - If not, why not? Are you interested in trying different options?

2. Can you think about the last time you tried to self-manage your endometriosis, to improve your symptoms? (examples: diet, exercise, hormones, rest, etc)

    - Can you tell me about the process of deciding to use the strategy, incorporating it in your life, and how you decided if it was working or not?
    - Probes: How did you choose this self-management technique? How many did you experiment with at a time? How many were you considering in your head?
    - How long did you try for?
    - How did you keep track?

    - How did you determine if it works? When did you expect to see any effects? Did you expect to see effects more in the short-term or long-term?
    - When you try a new strategy, are you looking for holistic improvements in your health or specific improvements?

3. What was your approach to developing your own self-management regimen?
    - Has your approach to self-management and/or experimentation changed over the course of your illness? Depending on other factors in your life? How have you adjusted according to changes in your life?

OK now let's pretend that we've developed this cool tool and we've asked you to use it.... The app might ask you to log your experiences from a few days to a few weeks or maybe months to learn about you, and then the app would enter into an experimental phase where it gives you strategies to try for a few days, weeks, or months, and then finally the app would know enough about you to give you real suggestions but not until after it gets a chance to know you—so we want to figure out what this app might look like, what you would be willing to do and what would be helpful to you...

1. Imagine a tool for experimenting with self-management has been created like described above (observations -> trying things -> learns recommendations).... We are going to spend the rest of the interview talking about this, but do you have any first impression? How might this be useful to you?

2. *(If described approach to self-management above)*: Thinking about the self-management strategy you described before... How might this tool change the way you self-manage your condition? *(If no strategy described above)*: How might a tool like this support you in figuring out your own self-management routine?

Logging & Experimenting

3. How would you feel about logging your data in the app for a while before it gave you useful recommendations? (for example, a meal planning app that uses similar methods requires 40 meals logged before providing first recommendation)

4. How much would you be open to experimenting?

    - One strategy at a time or multiple ones? Why? How would it help you? What types of strategies would you be interested in and why would they make sense together or on their own? Do you want to know about all of the options or just whichever are suggested as they come up?

    - During the exploration phase and for the app to learn about you and how you respond to different self-management strategies, the app might suggest a wide range of strategies that might not make sense to you. By a wide range, I mean for instance when learning about your exercise patterns, would provide recommendations that might seem like they won't help (2 minutes walking per day) to they are just might not be feasible for you (sprint back and forth for an hour). I'd like to understand from you what types of recommendations you would be open to or feel like you would not carry out at all: go for very fast run after work for 1 hour (why? Is it time, is effort required based on your health status, is it just not motivating?; acupuncture 4 times per week (why? Is it too much time? Too far? Don't know where to find an acupuncturist? Too expensive?); Marijuana? (why? Legal?)

    - You might have already a routine for your self-management, where you know some specific strategy works for you (for instance, stretching every morning), how would you react to recommendations from the app that disrupt or cause you to deviate from that routine? Why?

5. How much of a commitment are you willing to make for self-management? Every day? A few times a week? 10 mins? 1 hr? (WHY?) How many weeks or months are you willing to try a new strategy? How long would you want to keep trying strategies?

   Recommendations (when, where)

6. Mechanisms underlying adherence and retentions

   - Would it help if you made a commitment with the app to set aside time for self-management, without knowing what the specific strategies will be in advance? Why? How much heads up in time would you need when getting these recommendations (eg, the day before, or a prescription for the week)? Why?
   - What kind of feedback or other messaging would encourage you to stick to this? (positive encouragement when log success, when log failure; social support)... Content, frequency, ways of communicating of message?
   - Would knowing that something worked for other patients similar to you be useful?
   - What would turn you off or make you stop using the app?
   - Mechanisms underlying health conditions
   - How does the way you are feeling would change your response to a recommendation? For instance, if a specific day where you are supposed to try a recommendation you have severe bloating, how would you respond to the recommendation? Why?

   Recommendations (what)

7. What kind of information about the recommendations would you expect to see?

8. How do you imagine this information could be presented? Would you like it to be more personal and encouraging, or more clinical and neutral?

9. Do you want an explanation of why such a message is sent out? Would you feel that be more engaging, encouraging?

Outcomes/rewards

10. When would you like to provide feedback for a recommendation you receive?
11. How long until you expect to see results? What kind of results are you looking for?
12. What kind of messages would be helpful to keep you on track of a self-management tool?

Additional questions: Before we wrap up, anything else you think we should know? Do you have any worries about using a tool like this?

A.4 Focus Group Guide — 2016 Focus Groups with Patients

Symptom tracking – Our first set of questions to you is about your symptoms of endometriosis and whether you have any method to track them. By tracking we mean, keep recording them say in a notebook, or in your calendar, or using an app, it could even be in your diary. So let's start. Again, remember, no right or wrong answer!

1. How have you tracked your symptoms in the past? What symptoms have you tracked? Anything other than symptoms per se that you think was relevant to track with respect to endometriosis?
2. If you tracked, what were the reasons you decided to track, what did you like about tracking, what did you not like? How did you track and how often?
3. If you didn't track, why not? What makes tracking your symptoms difficult?

Endo app. We are now moving to a set of questions about what this mobile app we are working on would look like and how useful it could be for you.

1. Do you use any health apps? If so, which ones? What do you like or not like about these apps?

- Do you use any apps to help you with tracking your menstrual cycle?
- What do you like or not like about these apps?

2. Imagine there is an app that can track your symptoms throughout your cycles. How would a tool for tracking your symptoms of endometriosis be useful to you? *(Write categories on flip chart and write down people's suggestions)*

    - What would want out of such a tool?
    - What would you definitely NOT want in the tool we are proposing?

3. Thinking about endometriosis as a disease that is not well understood by the medical world yet, what do you think would be important aspects of your disease that need to be tracked in the app? There is a wide range of ways in which endometriosis manifests itself in women, so please tell us what makes sense for you.

    - Pain? Intensity? nature of pain? Location? (could be abdominal pain, could be pain during sex)
    - Digestive issues? Bloating? And where?
    - Sleep issues?
    - How you feel (physically, mentally, emotionally)?
    - Any "weird" symptoms you don't usually tell your doctor about but think might be endometriosis related? If so, what are they?

4. Thinking about how you manage your endometriosis symptoms, treatments, and how they affect your life... what features would you like in an application? *(Write categories on flip chart and write down people's suggestions)*

    - Educational or information resources
        - Advice? From clinicians, peers?

    - Treatment options

  - Reminders for medications and/or treatments?
  - Methods to self-reflection? Thoughts diary
  - Methods for changing your mood?
  - Experimenting with medications

5. Off all the features we discussed today *(make sure list is visible)* what are the one or two most important ones an endometriosis app must have?

Contributing to research, citizen science

1. What is your current medical understanding of the disease?
2. Do you feel like you understand what is going on with your disease and why it affects you the way it does?
3. What would you like to know about the disease? Either for yourself, for science.
   - Who is affected and why?
   - What are the different types of endometriosis, and which type am I?
   - What happens to patients like me?
   - What works and what doesn't work for me and patients like me?
4. What do you think that patients with endometriosis (you) could contribute that is novel information to the scientific community about endometriosis?
5. Some people participate in citizen science, where they partner with researchers to answer challenging questions. Would you be interested in contributing to research on endometriosis, for instance, through working out this app together?

Do you feel there is anything else you would like to discuss?

Appendix B: Phenotyping Appendix

B.1 Full Phendo vocabulary and mapping for phenotyping model

Domain	Feature	Self-tracked variables mapped to this feature
How was your day?	Great	great
	Good	good
	Manageable	manageable
	Bad	bad
	Unbearable	unbearable
Symptoms	Pain symptom	pelvis (left pelvis, right pelvis, pelvis, left ovary, right ovary, uterus, cervix), vagina (vagina entrance, deep vagina), rectum, abdomen (left side abdomen, right side abdomen, whole abdomen, upper abdomen, intestines), lower back (left lower back, right lower back, lower back), lower body (left outer hip, right outer hip, inner thighs, right left, left leg, legs), upper body (right shoulder, left shoulder, left ribs, right ribs, left arm, right arm, right breast, left breast, lower chest, upper chest, diaphragm), head, neck
	GI/UI symptom	urination problems (can't urinate, painful urination, frequent urination), nausea, vomiting, stomach problems (uncomfortably full, stomach upset, stomach upset, heartburn, mouth sores), endo belly, intestinal problems (blood in stool, painful bowel movement, gas, constipation, diarrhea)
	Other symptom	skin symptoms (eczema, itchy, rash, hives), temperature regulation (ever, hot flash, sweaty), respiratory and breathing (asthma, chest pressure, sinus congestion), auditory symptoms (noise sensitivity, ringing in ears), dizziness, blurry vision), allergies, fatigue, headache, mentally foggy, numbness, swelling, touch sensitivity
Pain severity	Mild	mild
	Moderate	moderate
	Severe	severe
GI/UI severity	Mild	mild
	Moderate	moderate
	Severe	severe
Other symptom severity	Mild	mild
	Moderate	moderate
	Severe	severe

Domain	Feature	Self-tracked variables mapped to this feature
Period	Yes period No period Light flow Medium flow Heavy flow	yes period no period light flow medium flow heavy flow
Bleeding	Breakthrough bleeding Clots Spotting No bleeding	breakthrough bleeding clots spotting no bleeding
Sex experience	No sex	no sex
	Sex felt good	sex felt good
	Painful sex or bleeding	painful after sex, painful during sex, bleeding from sex
	Sex ADL	sex ADL
	Avoided sex	avoided sex
	No penetration	no penetration
	Any penetration	vaginal penetration, anal penetration
Difficult activities of daily living (ADLs)	No hard activities	none
	Physical (active)	running, climbing stairs, walking, lifting, jumping, shopping, housework, stretching
	Physical (stationary)	sitting, lying down, getting out of bed, standing, kneeling, bathing, dressing, preparing food
	Mental	sleeping, working, socializing
	GI-related	eating, toilet
Self-management	Acupuncture	acupuncture
	Alcohol	alcohol
	Breathing	breathing
	Cannabis	medical marijuana, cbd oil
	Heat pack	heat pack
	Ice pack	ice pack
	Massage	massage
	PT or pelvic PT	physical therapy, pelvic pt
	Rest	rest
	Stretching	stretching
	Talk therapy	talk therapy
	TENS	tens
	None	none
Self-management Effect	Helped Didn't help No effect	helped didn't help no effect
Positive mood	(Not included in the analysis)	enthusiastic, excited, affectionate, social, relaxed, calm, happy, motivated, productive, optimistic
Negative mood	(Not included in the analysis)	antisocial, lonely, isolated, angry, contemptuous, belligerent, anxious, worried, defensive, disgusted, erratic, frustrated, guilty, indifferent, irritable, overwhelmed, sad, scared, whiny

Domain	**Feature**	**Self-tracked variables mapped to this feature**
Exercise	No exercise	no exercise
	Strength	e.g., sit ups, weight training, squats
	Walking	e.g., hikes, treadmill, walk dogs
	Cardio	e.g., elliptical, bicycle, rowing
	Yoga	e.g., basic yoga stretching, Bikram yoga, chair yoga
	Other exercise	e.g., tai chi, bowling, fencing
Food	No food	no food
	Carbs/grains/gluten	e.g., wheat, rice, bagels
	Gluten free	e.g., wheat-free, gf bagels, gf cookie
	Dairy	e.g., milk, ice cream, cheese
	Alternative dairy	e.g., almond milk, coconut milk, dairy-free milk
	Produce	e.g., berries, pineapples, greens
	Protein	e.g., poultry, bacon, fish
	Liquid/broth	e.g., lots of water, bone broth soup, hot drinks
	Fermented foods	e.g., apple cider vinegar, kimchi, kombucha
	Spices/herbs	e.g., turmeric, ginger, curcumin
	Spicy food	e.g., spicy foods, sriracha hot chili sauce, spicy salsas
	Alcohol/caffeine/chocolate	e.g., wine, coffee, dark chocolate
	Salt/sugar/processed foods	e.g., high-sugar foods, junk foods, fried foods
Medications and supplements	Medications	e.g., ibuprofen, paracetamol, Orlissa
	Supplements	e.g., magnesium, probiotics, milk thistle
Hormones	(Not included in the analysis)	e.g., progestin, mirena IUD, GnRH

Note: Phenotype mapping of user-customized options. For food and exercises, users are able to add entries that "hurt" and entries that "help" their symptoms. For these user-customized options, we manually map and normalize all responses to a smaller, coherent set of pre-determined strategies in order to reduce the number of strategies in the dataset.

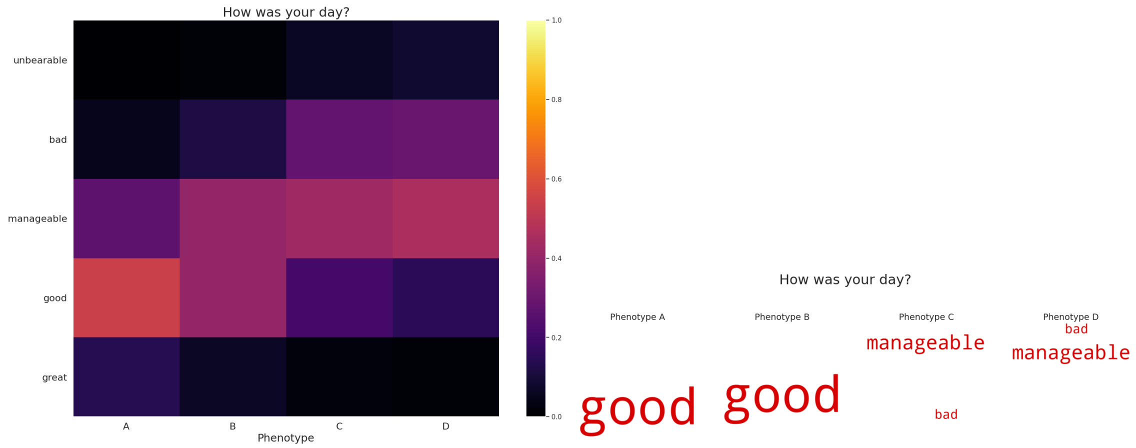
Domain	**Feature**	**Self-tracked variables mapped to this feature**
Exercise: Bad Effect	No effect	no effect
	Mild	mild
	Moderate	moderate
	Severe	severe
Exercise: Good Effect	No effect	no effect
	Mild	mild
	Moderate	moderate
	Severe	severe
Food: Bad Effect	No effect	no effect
	Mild	mild
	Moderate	moderate
	Severe	severe
Food: Good Effect	No effect	no effect
	Mild	mild
	Moderate	moderate
	Severe	severe

Note: Users log these data alongside the user-customized Food and Exercise entries from the table above.

Other Phendo data that is not included in the phenotyping analysis:

- Users are able to enter an **open-text journal entry** field each day. Including features from these unstructured entries would require the use of NLP, and is beyond the scope of this thesis.
- **Profile information** is also available, which is information about a person that does not change ever or very often (e.g., demographics).
- **Health history data** is also available, in the form of the WERF survey. This survey information is also unlikely to change rapidly over time.

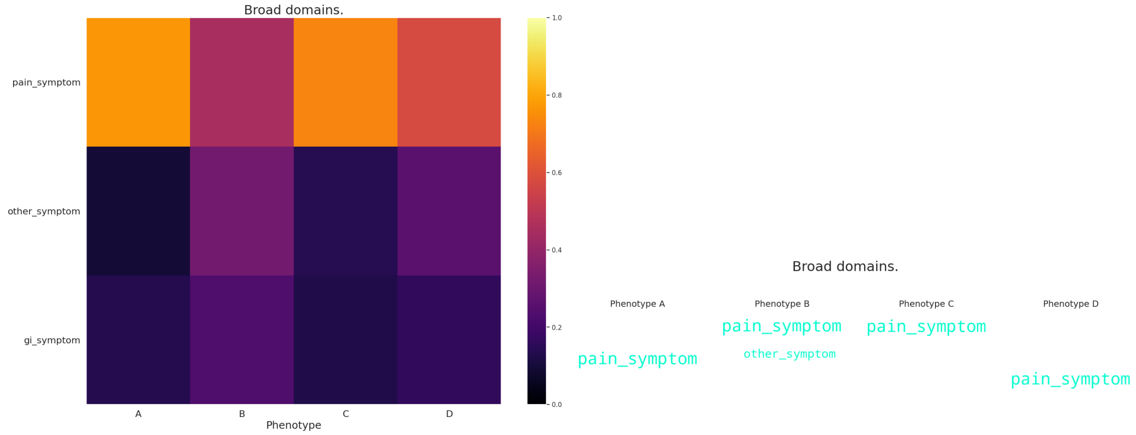
B.2 Final learned phenotyping model



(a) Heatmap — 'How was your day?'

(b) Word Cloud — 'How was your day?'

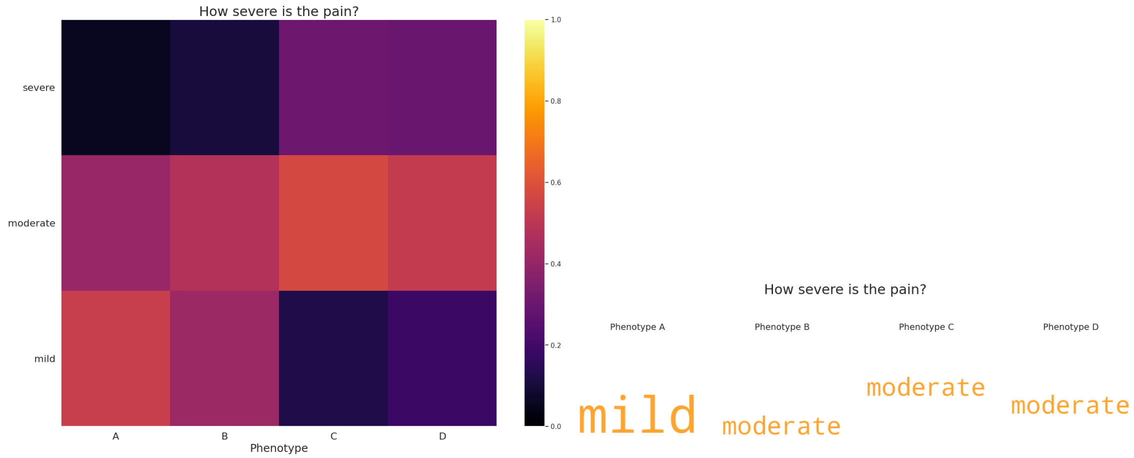
Figure B.1: Learned Phenotype Model: 'How was your day?'



(a) Heatmap — Symptoms

(b) Word Cloud — Symptoms

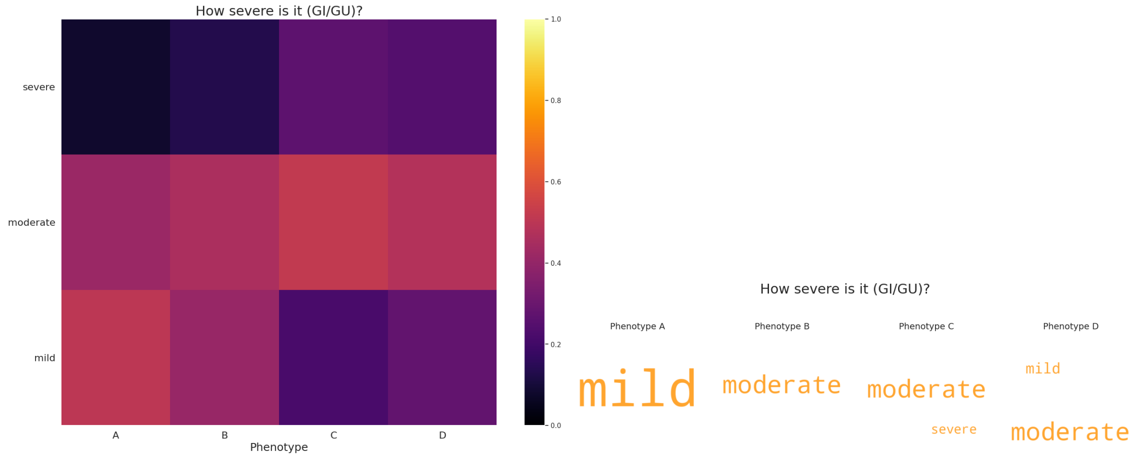
Figure B.2: Learned Phenotype Model: 'What symptoms are you experiencing?'



(a) Heatmap — Pain Severity

(b) Word Cloud — Pain Severity

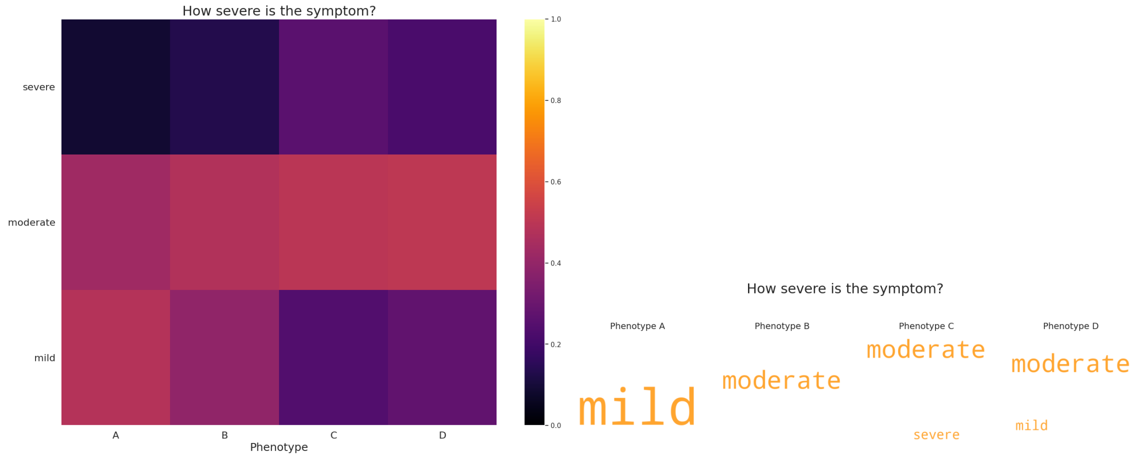
Figure B.3: Learned Phenotype Model: ‘How severe is the pain?’



(a) Heatmap — GI Severity

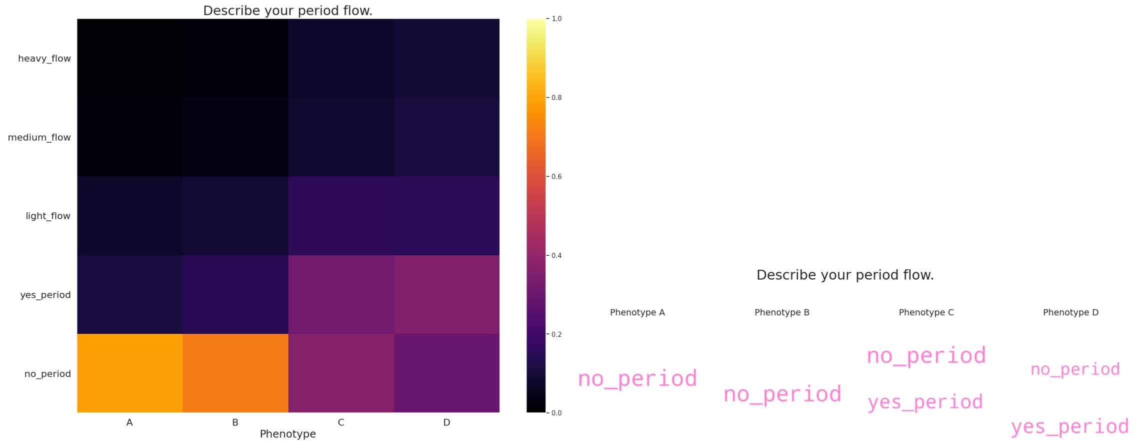
(b) Word Cloud — GI Severity

Figure B.4: Learned Phenotype Model: ‘How severe is the GI symptom?’



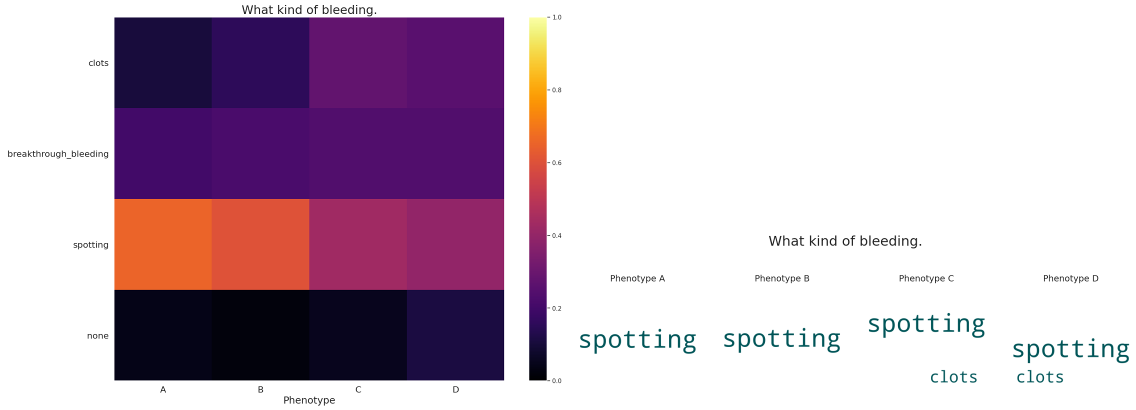
(a) Heatmap — Other Symptom Severity (b) Word Cloud — Other Symptom Severity

Figure B.5: Learned Phenotype Model: ‘How severe is the other symptom?’



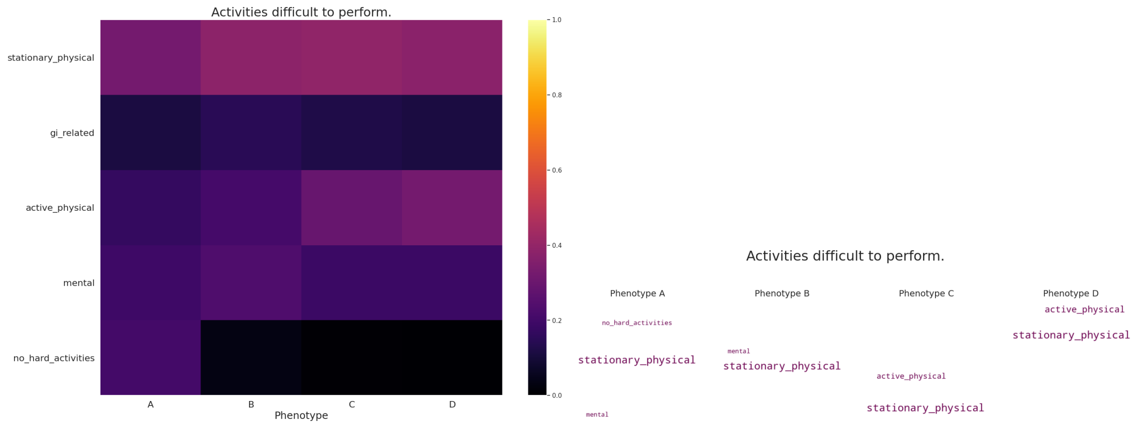
(a) Heatmap — Period (b) Word Cloud — Period

Figure B.6: Learned Phenotype Model: ‘What is your period flow?’



(a) Heatmap — Bleeding (b) Word Cloud — Bleeding

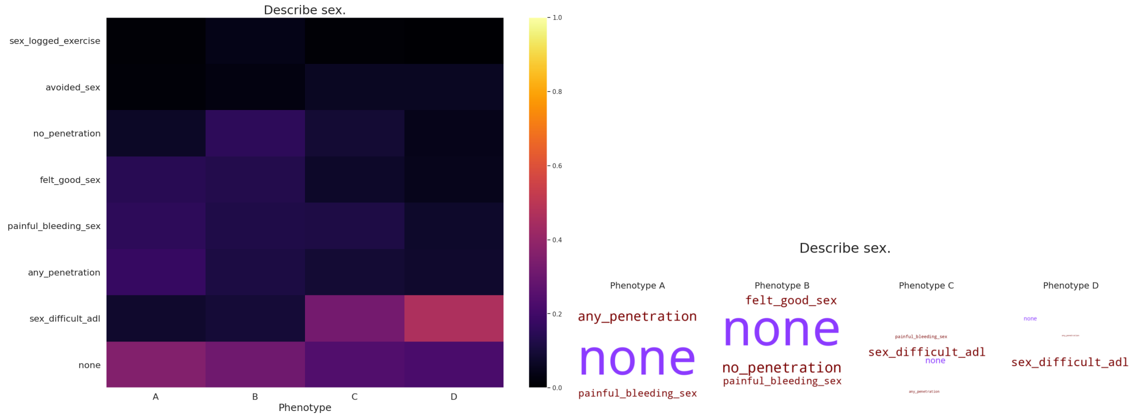
Figure B.7: Learned Phenotype Model: ‘What kind of bleeding?’



(a) Heatmap — ADL

(b) Word Cloud — ADL

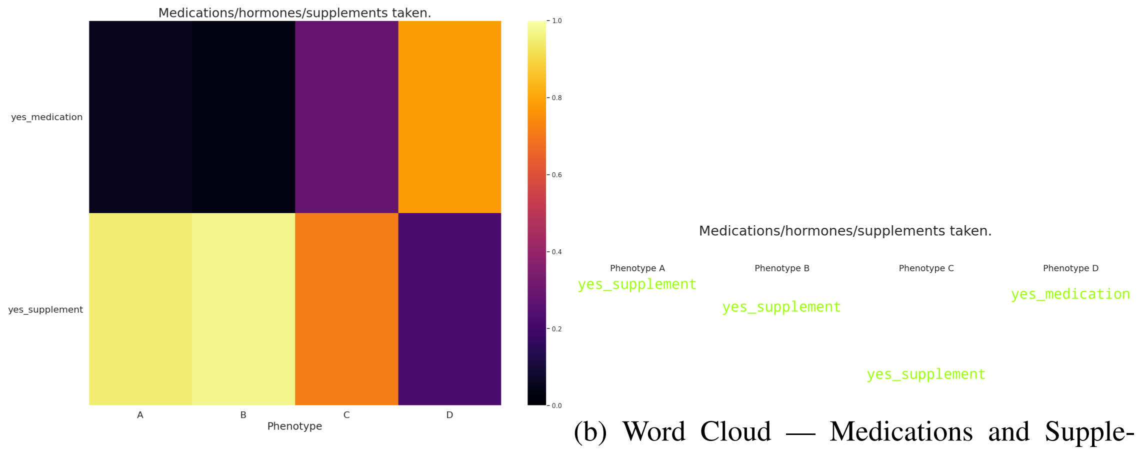
Figure B.8: Learned Phenotype Model: 'What activities were hard to do?'



(a) Heatmap — Sex Experiences

(b) Word Cloud — Sex Experiences

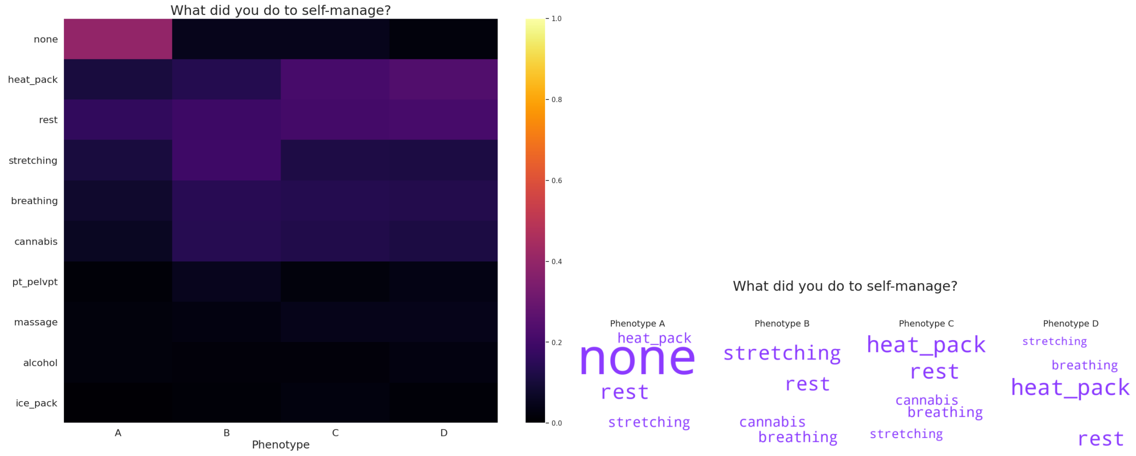
Figure B.9: Learned Phenotype Model: 'How was sex?'



(a) Heatmap — Medications and Supplements

(b) Word Cloud — Medications and Supplements

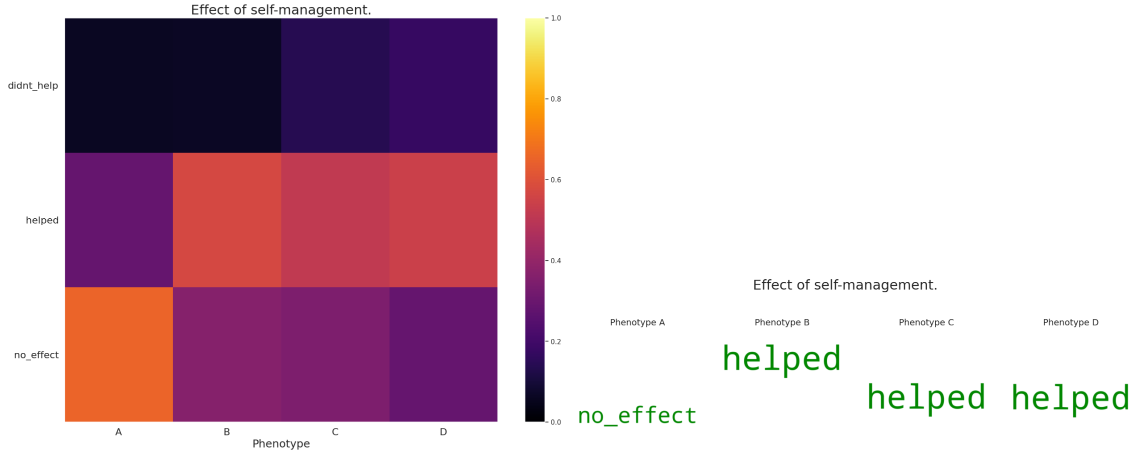
Figure B.10: Learned Phenotype Model: 'Did you take any medication or supplements?'



(a) Heatmap — Self-Management

(b) Word Cloud — Self-Management

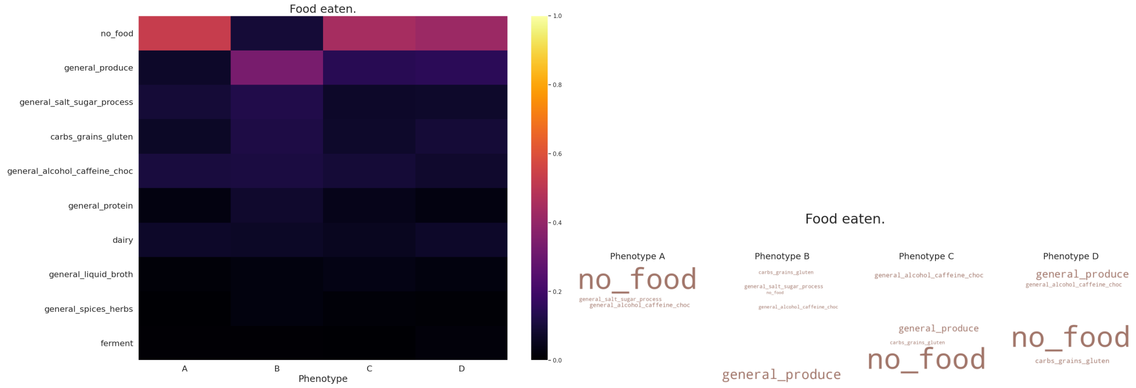
Figure B.11: Learned Phenotype Model: ‘What did you do to self-manage?’



(a) Heatmap — Self-Management Effect

(b) Word Cloud — Self-Management Effect

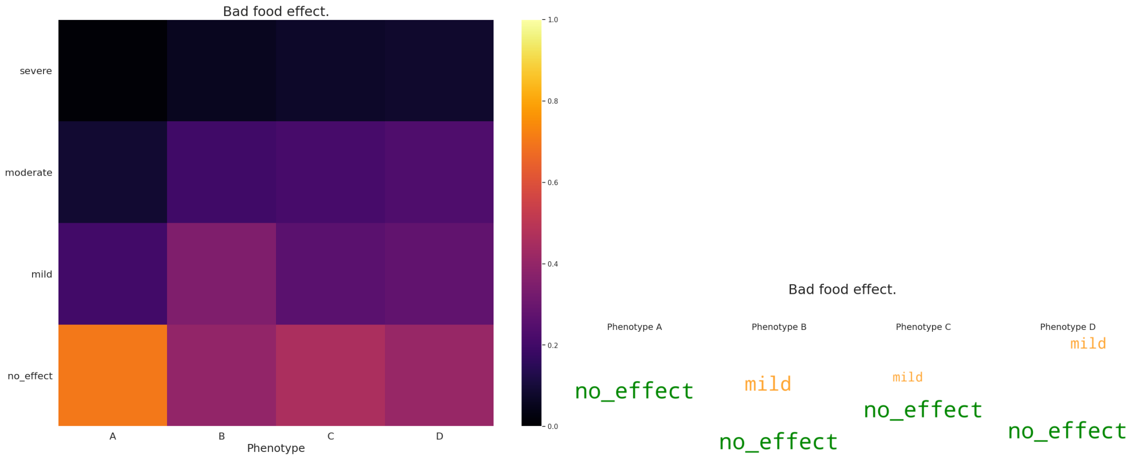
Figure B.12: Learned Phenotype Model: ‘Was self-management effective?’



(a) Heatmap — Food

(b) Word Cloud — Food

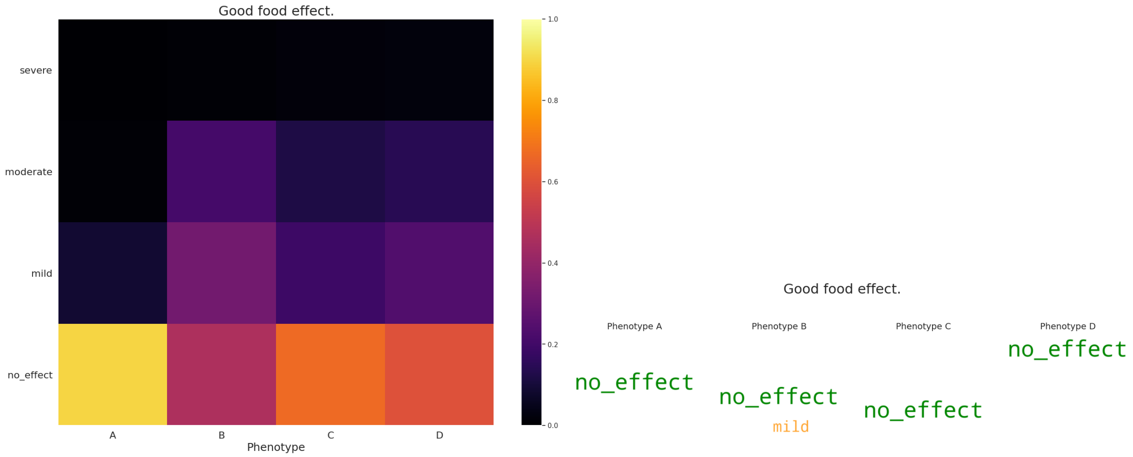
Figure B.13: Learned Phenotype Model: ‘What food did you eat?’



(a) Heatmap — Food, Bad Effect

(b) Word Cloud — Food, Bad Effect

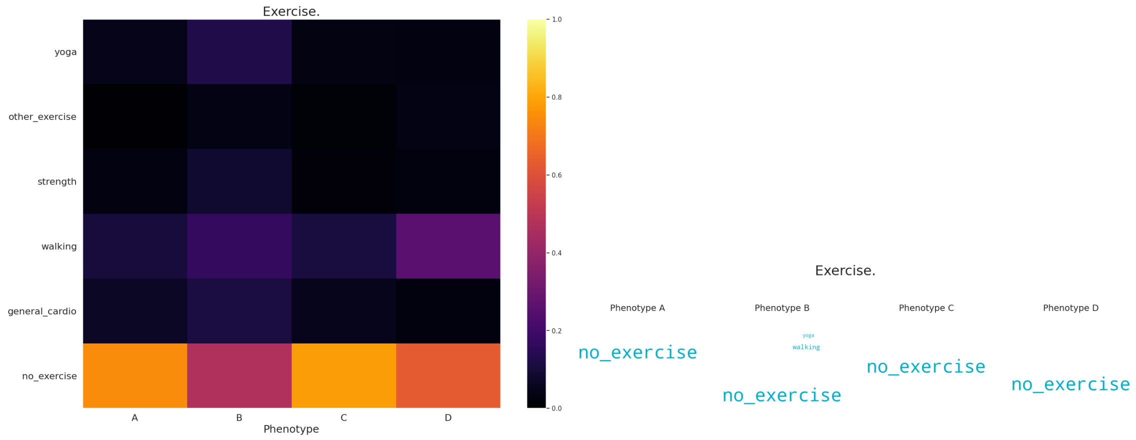
Figure B.14: Learned Phenotype Model: ‘Did you experience negative effects from the food?’



(a) Heatmap — Food, Good Effect

(b) Word Cloud — Food, Good Effect

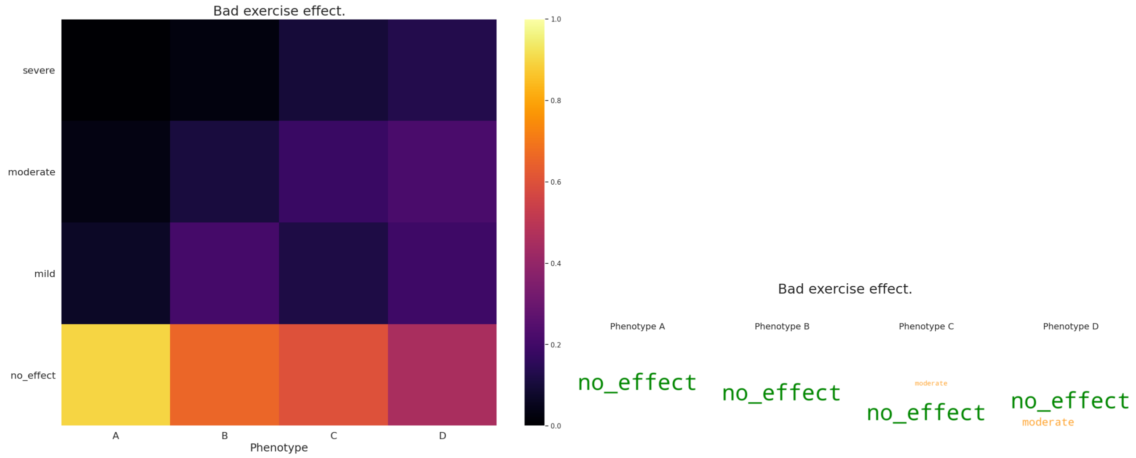
Figure B.15: Learned Phenotype Model: ‘Did you experience positive effects from the food?’



(a) Heatmap — Exercise

(b) Word Cloud — Exercise

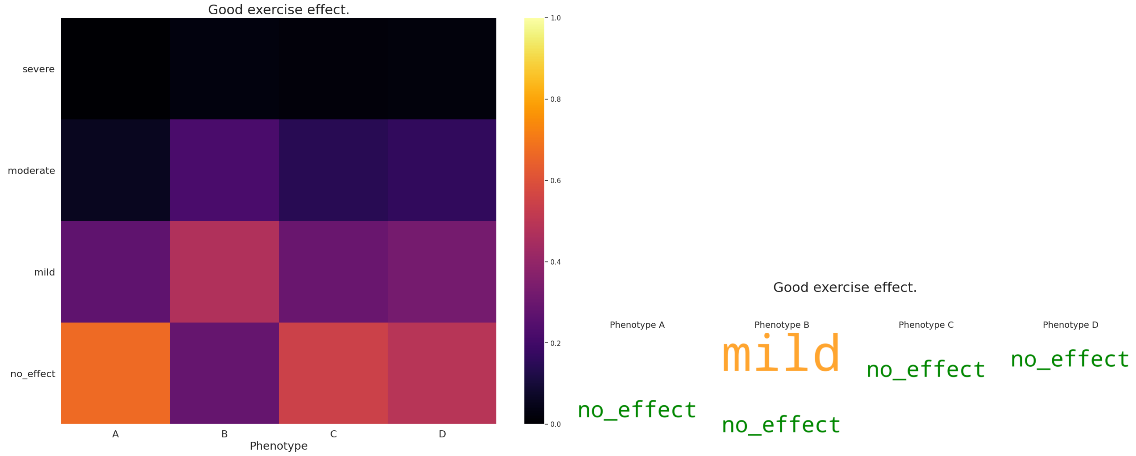
Figure B.16: Learned Phenotype Model: 'What exercise did you do?'



(a) Heatmap — Exercise, Bad Effect

(b) Word Cloud — Exercise, Bad Effect

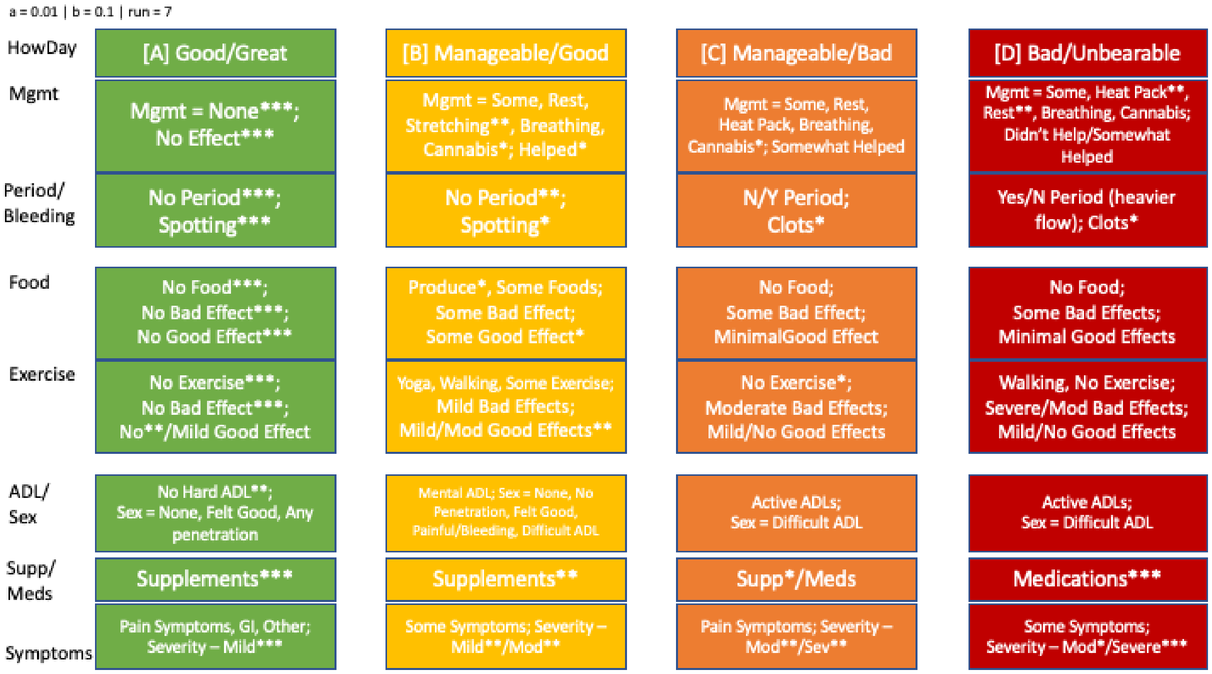
Figure B.17: Learned Phenotype Model: 'Did you experience negative effects?'



(a) Heatmap — Exercise, Good Effect

(b) Word Cloud — Exercise, Good Effect

Figure B.18: Learned Phenotype Model: ‘Did you experience positive effects?’



a = 0.01 | b = 0.1 | run = 7

HowDay	[A] Good/Great	[B] Manageable/Good	[C] Manageable/Bad	[D] Bad/Unbearable
Mgmt	Mgmt = None***; No Effect***	Mgmt = Some, Rest, Stretching**, Breathing, Cannabis*; Helped*	Mgmt = Some, Rest, Heat Pack, Breathing, Cannabis*; Somewhat Helped	Mgmt = Some, Heat Pack**, Rest**, Breathing, Cannabis; Didn’t Help/Somewhat Helped
Period/ Bleeding	No Period***; Spotting***	No Period**; Spotting*	N/Y Period; Clots*	Yes/N Period (heavier flow); Clots*
Food	No Food***; No Bad Effect***; No Good Effect***	Produce*, Some Foods; Some Bad Effect; Some Good Effect*	No Food; Some Bad Effect; MinimalGood Effect	No Food; Some Bad Effects; Minimal Good Effects
Exercise	No Exercise***; No Bad Effect***; No**/Mild Good Effect	Yoga, Walking, Some Exercise; Mild Bad Effects; Mild/Mod Good Effects**	No Exercise*; Moderate Bad Effects; Mild/No Good Effects	Walking, No Exercise; Severe/Mod Bad Effects; Mild/No Good Effects
ADL/ Sex	No Hard ADL**; Sex = None, Felt Good, Any penetration	Mental ADL; Sex = None, No Penetration, Felt Good, Painful/Bleeding, Difficult ADL	Active ADLs; Sex = Difficult ADL	Active ADLs; Sex = Difficult ADL
Supp/ Meds	Supplements***	Supplements**	Supp*/Meds	Medications***
Symptoms	Pain Symptoms, GI, Other; Severity – Mild***	Some Symptoms; Severity – Mild**/Mod**	Pain Symptoms; Severity – Mod**/Sev**	Some Symptoms; Severity – Mod*/Severe***

Figure B.19: Learned Phenotype Model: Overview of Phenotypes

B.3 User study task survey

WEEK X

What health status do you think fits this data best?

- Health Status A (Good)
- Health Status B (Manageable/Good)
- Health Status C (Manageable/Bad)
- Health Status D (Bad)

How certain are you about that assignment?

- Extremely Certain
- Very Certain
- Moderately Certain
- Slightly Certain
- Not at all Certain

How hard or easy was it to make the assignment?

- Very Hard
- Hard
- Neither Hard Nor Easy
- Easy
- Very Easy

[Tell participant the AI-generated health status]

WEEK X — Now that you've seen the AI-generated health status...

Do you agree/disagree with the AI's assignment?

- Strongly Agree
- Agree
- Undecided
- Disagree
- Strongly Disagree

What health status do you think fits this data best, after seeing the AI-generated health status?

- Health Status A (Good)
- Health Status B (Manageable/Good)
- Health Status C (Manageable/Bad)
- Health Status D (Bad)

How certain are you about that assignment?

- Extremely Certain
- Very Certain
- Moderately Certain
- Slightly Certain
- Not at all Certain

How hard or easy was it to make the assignment?

- Very Hard
- Hard
- Neither Hard Nor Easy
- Easy
- Very Easy

Appendix C: Additional Details on the Qualitative Methods for Chapter 5

In this study, we conducted interviews with three key informants who were active Phendo users, asking questions specifically relating to the use of an interactive system for self-management. We also re-analyzed focus group transcripts from prior studies. In total, we analyze transcripts from three in-depth interviews and re-analyze 10 transcripts from focus groups with 48 participants from prior work. A summary of the methods for each of these qualitative components is included here, but additional details for the focus groups can be found in the relevant publications. The demographics for participants across the interviews and focus groups are presented in Table C.1. The interview and focus group guides for all interviews and focus groups analyzed for this study are included above in Appendix A.

Table C.1: Demographics of all participants

	Interviews n = 3	**2019 Focus Groups** n = 21	**2016 Focus Groups** n = 27
Age *mean (range)*	34 (31-41)	32 (21-41)	38 (27-60)
Race or Ethnicity			
White	1 (33%)	14 (67%)	21 (78%)
Black	1 (33%)	6 (29%)	2 (7%)
Latina	1 (33%)	2 (10%)	2 (7%)
Asian		1 (5%)	2 (7%)
Years Diagnosed			
Less than 5	1 (33%)	12 (57%)	
5 to 10	1 (33%)	6 (29%)	
10 or more	1 (33%)	3 (14%)	
Age at Diagnosis *mean (range)*	26 (15-37)		29 (18-40)

C.1 Primary Interviews with Patients

In-depth interviews with key informants were conducted in 2020 to help further illuminate the needs of patients actively managing endometriosis symptoms in their day-to-day lives. Active Phendo users were recruited from social media. Eligibility criteria included adults with recent endometriosis symptoms and use of the Phendo app. The sessions lasted sixty minutes and individuals were compensated with a $25 pre-paid card for participating. In-depth discussions centered around the needs and desires of users for a tool to support self-management of endometriosis, and sought to elicit the constraints and acceptability of an AI-enabled tool for experimenting with and identifying self-management regimens that are personalized and effective for individuals. In total, three participants were interviewed. Participants were all women ranging in age from 31 to 41 (34 years mean). Time since diagnosis ranged from 3 to 16 years (8 years mean).

C.2 2019 Focus Groups with Patients

Focus groups that were conducted in 2019 were initially used to elicit design needs for tools to support care and self-management of endometriosis. Current endometriosis patients were recruited using social media and flyers hung near clinics. Eligibility for participation were English-speaking adults with a diagnosis of endometriosis, having experienced endometriosis symptoms in the past three months, and having engaged in care for endometriosis in the past year. The sessions lasted ninety minutes and individuals were compensated with a $25 pre-paid card for participating. Semi-structured focus group discussions centered around the ways that patients assess their own health status, practices around self-managing their condition outside of the clinical context, how they communicate with their care teams, and the ways that they evaluate if they are making progress towards their goals. In total, five groups were conducted with a total of 21 participants. Participants were all women ranging in age from 21 to 41 years old (32 years mean). Time since diagnosis ranged from less than one year to 21 years (5 years mean). Additional details on the primary study, including findings, can be found in the original publication [132].

C.3 2016 Focus Groups with Patients

Focus groups that were conducted in 2016 originally informed the design of the Phendo app. Individuals with endometriosis were recruited through convenience sampling through the Endometriosis Foundation of America email listserv, through social media, and through flyers hung near gynecological clinics. Eligibility for participation included adults with an endometriosis diagnosis through laparoscopic surgery. The sessions lasted ninety minutes and individuals were compensated with $25 in cash for participating. In these, semi-structured focus group discussions prompted participants to discuss the domains of health relevant to their illness experience, how they care for their illness and manage their symptoms, and what self-tracking would help them with. In total, five groups were conducted with a total of 27 participants. Participants were all women ranging in age from 27 to 60 years old (38 years mean), and age at diagnosis ranged from 18 to 40 years old (29 years mean). Additional details on the primary study, including findings, can be found in the original publication [119, 120].