UNIVERSITY OF CALIFORNIA,
IRVINE

Advancing Collaborative Health Monitoring Amidst Infrastructural Constraints

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Informatics

by

Eunkyung Jo

Dissertation Committee:
Associate Professor Daniel A. Epstein, Chair
Professor Yunan Chen
Professor Stephen M. Schueller
Research Scientist Young-Ho Kim

2025

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, I am deeply grateful to my advisor, Dr. Daniel Epstein. Thank you so much for believing in me, guiding me, and supporting me through all the ups and downs. I feel incredibly fortunate to have built this lifelong relationship with someone I can turn to for advice and continue to look up to as a role model.

I extend my sincere thanks to my committee members, Dr. Yunan Chen, Dr. Stephen Schueller, and Dr. Young-Ho Kim, for generously sharing their remarkable expertise and insights. Yunan, thanks so much for always asking thoughtful questions that helped me strengthen my work across all my PhD program milestones. Stephen, while it did not end up in my dissertation, I truly enjoyed collaborating with you on the teletherapy app review project and learned so much from your expertise in digital mental health. Young-Ho, I am grateful for the opportunity to intern with you during your first year at NAVER AI Lab. That first experience led to a second internship and eventually to three collaborative studies, two of which became core parts of my dissertation. Working with you across these projects broadened my research perspective and equipped me with valuable new skill sets.

This work could not have been done without the help of my collaborators: Dr. Alexandra Papoutsaki, Dr. Bryan Shapiro, Rachael Zehrung, Katherine E. Genuario, Whitney-Jocelyn Kouaho, and Myeonghan Ryu. I also thank my collaborators at NAVER—Dr. Young-Ho Kim, Dr. SoHyun Park, Yuin Jeong, Sang-Houn Ok, and Hyunhoon Jung—for their invaluable contributions as industry partners in our joint work.

I am also appreciative of my mentors. I am grateful to Dr. Hwajung Hong and Dr. Austin Toombs for inspiring my early interest in HCI when I was just beginning to explore research as an undergrad and a Master's student. I am deeply thankful to Dr. Yun Jung Kim and Dr. Jennifer Kim for their most kind support and thoughtful advice during difficult times.

I am thankful for the friendships I have formed with incredible people during my PhD. I feel very lucky to have you as lab mates—Lucas M. Silva, Xi Lu, Dennis Wang, Whitney-Jocelyn Kouaho, Matthew Dressa, Weijun Li, and Ziqi Yang from PIE Lab. I also want to thank my other friends at UCI—Ariel Han, Je Seok Lee, Seungmin Jeong, Aehong Min, Kyuha Jung, Rachael Zehrung, Sohyeon Park, and Hannah Hertenstein. I am so grateful that we were able to share both joys and challenges together in grad school—You all made the hard days easier and the good days even better. I will always cherish my time in Irvine because of you. A special thanks goes to my oldest friends back in Korea who have been by my side through so many stages of life: Hyemin Yoon, Jungeun Seo, and Minji Kim. Thanks so much for staying close no matter the distance or time. Having you in my life made everything feel a little lighter and a lot more meaningful.

I cannot thank my parents enough for their unconditional love and support. Dad, when I first started college, I never imagined I would end up doing AI research like you—but I guess your influence shaped me more than I realized. Thank you for showing me the value of integrity and dedication through the way you live your life. Mom, it turns out I did not just take

after you in appearance—I believe I inherited your curiosity, passion, and resilience as well. Thank you for always being there for me and being my greatest source of strength during difficult times. I also thank my sister and brother for their constant support. Knowing you are always there has given me strength throughout this journey.

There is someone I held close who is no longer with us, but who I know would be incredibly proud of me. I am grateful to him for always believing in me, even when I doubted myself, and for carrying me through countless challenges.

This work was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health (grant UL1 TR001414), the National Science Foundation (awards IIS-1850389, IIS-2237389), research internships at NAVER AI Lab, and the Google PhD Fellowship.

Chapter 3 is an adaptation of the material as it appears in [145], used with permission from the Association for Computing Machinery (ACM). The co-authors listed in this publication are Myeonghan Ryu, Georgia Kenderova, Samuel So, Bryan Shapiro, Alexandra Papoutsaki, and Daniel A. Epstein. Daniel A. Epstein directed and supervised research which forms the basis for the dissertation.

Chapter 4 is an adaptation of the material as it appears in [146], used with permission from ACM. The co-authors listed in this publication are Rachael Zehrung, Katherine E. Genuario, Alexandra Papoutsaki, and Daniel A. Epstein. Daniel A. Epstein directed and supervised research which forms the basis for the dissertation.

Chapter 5 is an adaptation of the material as it appears in [141], used with permission from ACM. The co-authors listed in this publication are Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. Young-Ho Kim directed and supervised research which forms the basis for the dissertation.

Chapter 6 is an adaptation of the material as it appears in [143], used with permission from ACM. The co-authors listed in this publication are Young-Ho Kim, Sang-Houn Ok, and Daniel A. Epstein. Daniel A. Epstein directed and supervised research which forms the basis for the dissertation.

VITA

Eunkyung Jo

EDUCATION

Doctor of Philosophy in Informatics **2025**
University of California, Irvine *Irvine, California*

Master of Science in Computer Graphics Technology **2019**
Purdue University *West Lafayette, Indiana*

Bachelor of Arts in Information Science and Culture **2017**
Bachelor of Arts in Child and Consumer Studies
Seoul National University *Seoul, South Korea*

RESEARCH EXPERIENCE

Graduate Research Assistant **2020–2025**
University of California, Irvine *Irvine, California*

Research Intern **2022-2023**
NAVER AI Lab *Seongnam, South Korea*

Researcher **2019-2020**
National Center for Mental Health *Seoul, South Korea*

Graduate Research Assistant **2017-2019**
Purdue University *West Lafayette, Indiana*

TEACHING EXPERIENCE

Teaching Assistant **2020–2021**
University of California, Irvine *Irvine, California*

Teaching Assistant **2017**
Purdue University *West Lafayette, Indiana*

REFEREED JOURNAL PUBLICATIONS

Exploring Patient-Generated Annotations to Digital Clinical Symptom Measures for Patient-Centered Communication **2024**

Eunkyung Jo, Rachael Zehrung, Katherine E. Genuario, Alexandra Papooutsaki, Daniel A. Epstein

Proceedings of the ACM on Human-Computer Interaction 8, CSCW2

Exploring User Perspectives of and Ethical Experiences with Teletherapy Apps: Qualitative Analysis of User Reviews **2023**

Eunkyung Jo, Whitney-Jocelyn Kouaho, Stephen M. Schueller, Daniel A. Epstein

JMIR Mental Health. Volume 10, 2023

Understanding Cultural Influence on Perspectives Around Contact Tracing Strategies **2022**

Xi Lu, Eunkyung Jo, Seora Park, Hwajung Hong, Yunan Chen, Daniel A. Epstein

Proceedings of the ACM on Human-Computer Interaction 6, CSCW2

GeniAuti: Toward Data-Driven Interventions to Challenging Behaviors of Autistic Children through Caregivers' Tracking **2022**

Eunkyung Jo, Seora Park, Hyeonseok Bang, Youngeun Hong, Yeni Kim, Jungwon Choi, Bung Nyun Kim, Daniel A. Epstein, Hwajung Hong

Proceedings of the ACM on Human-Computer Interaction 6, CSCW1

MAMAS: Supporting Parent-Child Mealtime Interactions Using Automated Tracking and Speech Recognition **2020**

Eunkyung Jo, Hyeonseok Bang, Myeonghan Ryu, Eun Jee Sung, Sungmook Leem, Hwajung Hong

Proceedings of the ACM on Human-Computer Interaction 4, CSCW1

The Social Infrastructure of Co-spaces: Home, Work, and Sociable Places for Digital Nomads **2019**

Ahreum Lee, Austin L. Toombs, Ingrid Erickson, David Nemer, Eunkyung Jo, Yushen Ho, Zhaung Guo

Proceedings of the ACM on Human-Computer Interaction 3, CSCW1

REFEREED CONFERENCE PUBLICATIONS

Understanding Public Agencies' Expectations and Realities of AI-Driven Chatbots for Public Health Monitoring **2025**

Eunkyung Jo, Young-Ho Kim, Sang-Houn Ok, Daniel A. Epstein

Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems

Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention **2024**

Eunkyung Jo, Yuin Jeong, SoHyun Park, Daniel A. Epstein, Young-Ho Kim

Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems

Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention **2023**

Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, Young-Ho Kim

Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems

Designing Flexible Longitudinal Regimens: Supporting Clinician Planning for Discontinuation of Psychiatric Drugs **2022**

Eunkyung Jo, Myeonghan Ryu, Georgia Kenderova, Samuel So, Bryan Shapiro, Alexandra Papoutsaki, Daniel A. Epstein

Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems

Comparing Perspectives around Human and Technology Support for Contact Tracing **2021**

Xi Lu, Tera L. Reynolds, Eunkyung Jo, Hwajung Hong, Xinru Page, Yunan Chen, Daniel A. Epstein

Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems

Understanding Parenting Stress through Co-Designed Self-Trackers **2020**

Eunkyung Jo, Austin L. Toombs, Colin M. Gray, Hwajung Hong

Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems

Toward Becoming a Better Self: Understanding Self-Tracking Experiences of Adolescents with Autism Spectrum Disorder Using Custom Trackers **2019**

Sung-In Kim, Eunkyung Jo, Myeonghan Ryu, Inha Cha, Young-Ho Kim, Heejeong Yoo, Hwajung Hong

PervasiveHealth 2019

LIGHTLY-REVIEWED PUBLICATIONS

Incorporating Multi-Stakeholder Perspectives in Evaluating and Auditing of Health Chatbots Driven by Large Language Models **2024**

Eunkyung Jo, Young-Ho Kim, Yuin Jeong, SoHyun Park, Daniel A. Epstein

CHI 2024 Workshop – Human-Centered Evaluation and Auditing of Language Models

GeniAuti: Towards Data-Driven Interventions to Challenging Behaviors of Autistic Children through Caregivers' Tracking **2023**

Eunkyung Jo, Seora Park, Hyeonseok Bang, Youngeun Hong, Yeni Kim, Jungwon Choi, Bung Nyun Kim, Daniel A. Epstein, Hwajung Hong

CHI 2023 Workshop – Workshop in Interactive Systems for Healthcare

Examining the Role of Conversational AI in Personal Informatics Systems for Collaborative Health Work and Care **2022**

Eunkyung Jo, Young-Ho Kim, Yuin Jeong, Hyeri Kim, Hyun Jung Park, Daniel A. Epstein

CHI 2022 Workshop – Grand Challenges in Personal Informatics and AI

Development of a Mobile Application that Tracks Challenging Behaviors of Children with Autism Spectrum Disorders for Supporting Data-Driven Interventions **2020**

Hwajung Hong, Eunkyung Jo, Yeni Kim, Youngeun Hong

The International Society for Autism Research (INSAR) 2020 Annual Meeting

Investigating the Self-Tracking Use of Managing Stress of New Parents **2019**

Eunkyung Jo

Doctoral Consortium of PervasiveHealth 2019

A Study for Designing a Cooperative Self-Management Program on a Smartphone for Adolescents with Autism Spectrum Disorder **2018**

Eunkyung Jo, Sung-In Kim, Myeonghan Ryu, Hwajung Hong, Heejung Yoo

Grace Hopper Celebration 2018

COSMA: Cooperative Self-Management Tool for Adolescents with Autism **2017**

Myeonghan Ryu, Eunkyung Jo, Sung-In Kim

ASSETS 2017: ACM SIGACCESS Conference on Computers and Accessibility

AWARDS AND HONOR

Google PhD Fellowship	**2023-2025**
Best Paper Award, CHI 2023	**2023**
University of California, Irvine Dean's Recruitment Fellowship	**2020**
Gold Medal, ASSETS 2017 Student Research Competition (Undergraduate Section)	**2017**
Seoul National University Student Directed Education Program Research Excellence Award	**2017**

ABSTRACT OF THE DISSERTATION

Advancing Collaborative Health Monitoring Amidst Infrastructural Constraints

By

Eunkyung Jo

Doctor of Philosophy in Informatics

University of California, Irvine, 2025

Associate Professor Daniel A. Epstein, Chair

Health monitoring technologies—such as mobile apps, wearables, and recent AI-powered chatbots—are increasingly embedded in clinical and public health systems. While these tools have gained popularity for supporting individual health management, they also hold potential to facilitate collaboration in clinical care and public health interventions. However, in practice, such technologies often fall short when their design fails to align with the complex human infrastructures that shape real-world healthcare delivery. My dissertation adopts the lens of human infrastructures to examine how health monitoring technologies can better support collaboration across clinical and public health settings by attending to the sociotechnical realities that influence their use.

Through my dissertation, I demonstrate that health monitoring technology that accounts for infrastructural complexity can improve collaboration in clinical care and public health interventions. In clinical care, I first present the design and evaluation of AT Planner—a clinical decision-support tool to help providers develop antidepressant taper plans—to show how flexible tools can help navigate infrastructural constraints in care planning. I then explore how technology can support patients in conveying their health experiences amidst the logistical constraints of clinical visits. Through the design and evaluation of AT Annotator—an annotation tool that allows patients to enrich clinical self-report measures with various

forms of input, I contribute design opportunities for digital symptom measures that better address patients' communication needs.

In public health interventions, I examine how technology—particularly AI chatbots—might support stakeholder practices within public health infrastructures. Through multi-stakeholder interviews with users, frontline workers, and developers involved in deploying an LLM-driven chatbot named CareCall as part of public health interventions, I identify key design considerations for managing tensions around system expectations among stakeholders. Lastly, I investigate how AI chatbots like CareCall might support public agencies in scaling up care within public health infrastructures. Through interviews with public agency workers across various roles, I highlight the need to consider the impact of AI implementation on labor demands in public health infrastructures.

Across these projects, I demonstrate that health monitoring technology can more effectively support collaboration when it accounts for infrastructural complexity, specifically by incorporating flexibility to balance stakeholder constraints and amplifying existing stakeholder practices. Together, insights from my work advance our understanding of how technologies should be designed to better reflect the realities of clinical and public health infrastructures. My dissertation contributes conceptual foundations for designing collaborative health technologies that are responsive to the infrastructural realities. By framing flexibility and amplification as key strategies for navigating infrastructural complexity and by highlighting the importance of managing system expectations across stakeholders, my dissertation offers a lens for rethinking how health technologies can support—rather than override—the human systems that sustain care.

Chapter 1

Introduction

Digital technologies have become deeply embedded in healthcare systems around the world. From electronic health records and telemedicine platforms to algorithmic decision support tools and mobile health apps, technologies now mediate many aspects of clinical and public health work. While such innovations are often framed as solutions for making healthcare more efficient, patient-centered, and accessible, their real-world implementation often falls short when their design does not align with the human systems that they are intended to support. For example, these technologies often encounter friction when they fail to account for some important stakeholders—such as pharmacy workflows and insurance company policies—that shape what care is actually deliverable. In the United States, a recent report found that more than one-third of Americans were told that their insurance would not cover a medication prescribed by their doctor in the past year [239], which often leads patients to abandon their prescriptions because of high prices [237]. As physician order entry systems often fail to account for such insurance-related constraints, providers are frequently forced into redundant communication processes, such as navigating prior authorization protocols, creating delays and administrative burdens [241, 239]. Over time, such inefficiencies in clinical systems can reduce trust in clinical technologies and contribute to their eventual

abandonment. Similarly, in Saudi Arabia, the introduction of the national electronic prescribing system encountered significant friction due to poor interoperability with existing pharmacy workflows, leading to delays in medication delivery and increased communication burden on both providers and pharmacists [9]. This example highlights how digital tools can inadvertently complicate rather than streamline care delivery when they fail to reflect the human systems that shape healthcare delivery.

Further, public health contexts introduce greater challenges of alignment between health technologies and existing infrastructures. The COVID-19 pandemic revealed critical gaps in many countries' public health infrastructures, particularly in how digital technologies were adopted at scale. In the United States, many state governments and major technology companies developed contact tracing apps to track down positive cases and identify their contacts early in the pandemic. However, the adoption rates remained below 10% across most states, limiting their intended benefits [316]. Public health experts attributed the low uptake in part to the absence of a centralized healthcare system capable of providing coordinated guidance and support [316, 238] and a substantial lack of public health workforce to operate such technology [240]. Similarly, in New Zealand, the initial rollout of digital contact tracing tools was fragmented without a centralized approach [52]. As private entities developed contact tracing applications with varying functionalities and data management practices, public health officials struggled to navigate and integrate data across these fragmented systems, hindering their ability to scale contact tracing efforts in response to emerging outbreaks [52]. Together, these examples highlight the importance of designing health technology that aligns with the broader human infrastructure to ensure its effective real-world implementation.

Although clinical and public health infrastructures differ in scale and structure, both heavily rely on human efforts to implement and maintain technologies within their complex healthcare environments. In the HCI and CSCW communities, this human work among various stakeholders has been characterized as *human infrastructures* [184]. While infrastructures

are traditionally understood as the physical and technological foundations of human activities (e.g., electric grids, telecommunication networks), human infrastructure is another critical aspect of understanding modern healthcare services, which requires collaboration among diverse stakeholders [304]. In this thesis, I adopt the lens of human infrastructures to examine the complexity of collaboration surrounding the use of health monitoring technologies in both clinical and public health settings, and how health monitoring technologies can support—or strain—these collaborative efforts.

Self-tracking technologies—such as mobile apps and wearables—represent one type of technology that has gained popularity for supporting individual health management, particularly through monitoring physical activity, sleep, nutrition, mood, and other aspects of health. While research on self-tracking has focused on individuals' practices of monitoring various aspects of their health and wellbeing, tracking practices are often collaborative, involving different stakeholders [90]. A significant body of literature has suggested the potential benefits of incorporating personal health tracking into clinical or public health decision-making [90, 148, 38]. For example, personal health tracking can empower patients to more effectively convey their illness experiences during clinical visits [62, 135, 63, 289]. Beyond research, personal health tracking technologies are increasingly being integrated into routine clinical care. For instance, the Apple Watch—FDA-approved for detecting irregular heart rhythms such as atrial fibrillation—is now commonly recommended by doctors to help patients monitor potential episodes [313]. Similarly, the Mayo Clinic has equipped heart surgery patients with Fitbit activity trackers to encourage physical activity and support their post-operative recovery [315]. Personal health tracking can also help extend public reach while reducing the burden of public health authorities in monitoring people at scale by automating aspects of data collection [131, 245, 130]. Personal health tracking technologies have been used across the world to scale up the efforts for contact tracing during the pandemic [292]. Technologies like CommCare [85] and Community Health Toolkit [217] have also been widely used in the Global South to help community health workers collect health data

from populations in low-resource settings. Recently, AI-powered chatbots have emerged as a promising tool to support large-scale health monitoring, particularly by automating data collection and engaging in empathetic conversations [125].

Integrating personal health tracking into clinical care or public health interventions affects a wide range of stakeholders who collectively make up human infrastructures that shape real-world healthcare settings—including, but not limited to, users, healthcare providers, insurers, pharmacies, tech companies, and government agencies. Unfortunately, the role of human infrastructure is often overlooked in the design and adoption of health monitoring technologies. This oversight can introduce significant risks in clinical care, as providers often struggle to review patient-tracked data during clinical visits due to time constraints [332, 333, 61, 351, 284], potentially compromising the quality of care. Similarly, public health workers may face logistical challenges in managing the vast amounts of data collected through technical monitoring solutions [261, 130, 151, 245, 322]. Despite growing interest in integrating personal health monitoring into formal care systems, we lack an understanding of how these technologies can better account for the complexity of multi-stakeholder needs around collaborative health monitoring.

My dissertation offers insights for designing and implementing collaborative health monitoring technology in ways that account for infrastructural complexity. Through my dissertation, I provide an empirical understanding of how technology helps or hinders collaboration in clinical care and public health interventions. Through engaging with multi-stakeholder perspectives around collaborative health monitoring technologies within healthcare infrastructures, I contribute a comprehensive understanding of how technology could help stakeholders navigate infrastructural constraints or introduce new tensions, along with design insights for better aligning these technologies with the infrastructural complexity of real-world healthcare settings.

1.1 Thesis Statement

My thesis claim is as follows:

Health monitoring technology that accounts for infrastructural complexity can improve collaboration in (T1) clinical care and (T2) public health interventions.

A range of terms are used across disciplines, such as Computer Science, Health Informatics, and Medicine, to describe the collaboration around personal health data in healthcare settings [71]. In medical research, *remote patient monitoring* is typically used to refer to the collection of health-related data by patients between their visits, which is then transmitted directly to providers' databases. These systems often limit patients' access to or control over the data [57, 71]. Another commonly used term is *self-monitoring,* which has its origins in behavioral psychology, where individuals record their own thoughts, feelings, and behaviors as part of therapeutic processes [58, 59]. Self-monitoring is also often used interchangeably with self-tracking or personal informatics [59, 71] in HCI research. The term *personal informatics* (used as a synonym of self-tracking) [196, 197]—the most commonly used concept in HCI—describes systems that enable individuals to collect personally relevant data for the purpose of self-reflection and self-knowledge [196]. Self-tracking encompasses not only health-related data but also personal data from other areas of daily lives, such as financial spending or productivity [18, 277]. In my thesis, I adopt the term *health monitoring* to describe the collection and use of personal health data not only for individual use but also to support clinical care or public health interventions. I chose *health monitoring* over other terms (e.g., self-monitoring, self-tracking) to highlight the collaborative and infrastructural dimensions of these practices when they are integrated into formal care systems. While *remote patient monitoring* in medical research typically refers to systems where patients passively generate data with limited access or agency, I use *health monitoring* to encompass contexts where individuals play an active role in data collection and communication with

Table 1.1: Thesis Organization

Research Question	**Addressed in**
RQ1: How might technology help **providers develop care plans** within clinical infrastructures?	**Chapter 3**, through the design and evaluation of AT Planner, a clinical decision-support tool that project tentative prescription plans that providers can flexibly adjust to fit their clinical regimens
RQ2: How might technology help **patients convey their health experiences** within clinical infrastructures?	**Chapter 4**, through the design and evaluation of AT Annotator, a tool that allows patients to enhance clinical self-report measures with various annotation types
RQ3: How might technology support **stakeholder practices** within public health infrastructures?	**Chapter 5**, through generating design guidelines for AI chatbots for public health monitoring to consider tensions among stakeholders around system expectations
RQ4: How might technology support **public agencies in scaling up care** within public health infrastructures?	**Chapter 6**, through developing guidelines to account for the changes in human labor with the introduction of AI chatbots for public health monitoring

other stakeholders.

Drawing on Lee's characterization of *human infrastructures* as an organization of human labor required for collaborative work [184], I focus on how health monitoring technologies are shaped and embedded within these infrastructures. I define *"collaboration"* in clinical care or public health interventions as the joint effort of individuals and various professionals towards shared health goals—including disease management, disease surveillance and control, health promotion, and emergency response—which involves activities such as deciding whether and what technology to adopt, coordinating efforts among different types of professionals, collecting personal health data, analyzing the data to gain insights, developing care plans, and offering the necessary support.

Table 1.2: Stakeholder Involvement Across Chapters

		Patients / Care Recipients	Healthcare Professionals	Insurers	Pharmacies	Tech Companies	Government Agencies
Clinical Infrastructures	Chapter 3 on clinician planning amidst infrastructural constraints	✓	✓	✓	✓		
	Chapter 4 on patient-centered communication amidst logistical constraints	✓	✓				
Public Health Infrastructures	Chapter 5 on multi-stakeholder perspectives on AI chatbots for public health monitoring	✓	✓			✓	✓
	Chapter 6 on public agency perspectives on AI chatbots for public health monitoring		✓				✓

1.2 Thesis Overview

I report on findings from four studies that investigate how health monitoring technology can better support collaboration in (T1) clinical care and (T2) public health interventions by accounting for the infrastructural complexity. Table 1.1 outlines the overall structure of my dissertation, aligning each chapter with the corresponding research question that contributes to my thesis claim.

Chapter 2 reviews prior work on human infrastructures in healthcare, focusing on the theoretical concepts that inform my work. I then review studies on personal health monitoring in clinical care, focusing on patient-provider collaboration around patient-tracked data. Finally, I review research on personal health monitoring for public health, focusing on how the introduction of technology has supported both frontline workers and care recipients.

Chapter 3 examines how technology could assist clinician planning amidst infrastructural constraints through the design and evaluation of AT Planner, a clinician decision-support tool that projects tentative prescription plans that providers can flexibly adjust to fit their clinical regimens for tapering antidepressants. Through a formative study and a feedback study on AT Planner with providers, I identify the need to support provider needs in balancing infrastructural constraints when developing plans for the discontinuation of antidepressants, accounting for various stakeholder needs such as patients, insurers, and pharmacies (Table 1.2). These findings contribute design insights into how clinical health monitoring technology can accommodate the infrastructural constraints clinicians face in care planning.

Chapter 4 investigates how technology could assist patients in conveying their health experiences within the logistical constraints of clinical visits through the design and evaluation of a research prototype, AT Annotator, an annotation tool that allows patients to enhance clinical self-report measures with various annotation types while discontinuing their antidepressants. Through an interview study with patients with AT Annotator, I explore the idea of patient annotations to clinical self-report measures as one way to promote patient-centered communication amidst providers' logistical constraints (Table 1.2). Based on the interview findings, I propose design opportunities for annotation tools to cater to patients' communication needs and preferences. Studies presented in Chapters 3 and 4 were conducted in the United States, where clinical infrastructures are characterized by a multi-payer healthcare system that involves fragmented insurance coverage and requires patients and providers to navigate access to medications individually.

Chapter 5 explores how health monitoring technologies, particularly AI chatbots, might support stakeholder practices within public health infrastructures. Through the case of CareCall, a large language model (LLM)-driven voice chatbot for monitoring the health and wellbeing of socially isolated individuals, I elicit multi-stakeholder perspectives on its deployment for public health monitoring. Through interviews with users, frontline health

workers, and developers who were involved in the deployment of CareCall (Table 1.2), I identify key design considerations for designing AI chatbots for public health monitoring, focusing on tensions around system expectations among stakeholders that make up public health infrastructures.

Chapter 6 delves into public agency perspectives on leveraging AI chatbots to scale up care within public health infrastructures. Through interviews with public agency workers across various roles, including decision-making, administration, and frontline monitoring (Table 1.2), I elicit public agencies' expectations and realities of adopting AI chatbots for public health monitoring. Findings from the study highlight the need to consider the impact of AI implementation on labor demands within public health infrastructures. Studies presented in Chapters 5 and 6 were conducted in South Korea, where centralized public health governance and national coordination shaped how AI chatbots like CareCall were integrated into existing infrastructures through local public agencies.

Chapter 7 revisits my thesis statement and discusses how findings from previous chapters support my overall argument for the need to account for infrastructural complexity in health monitoring technologies. In clinical contexts, I argue that this can be achieved by implementing flexibility in care planning and patient input, though tensions may arise due to providers' varying levels of experience and their perceptions of flexible data forms. In public health contexts, I argue that health monitoring technologies can better address infrastructural constraints by amplifying stakeholder practices to enhance and scale their existing efforts. However, tensions around managing stakeholder expectations should be carefully considered, particularly regarding the types of tasks that such technologies can perform, the labor they require, and how users might engage with them. I conclude by briefly describing my plans for future work.

All projects described in my dissertation were done in collaboration with others. In each chapter, I specify my own contributions and those of my collaborators.

Chapter 2

Related Work

In this chapter, I first review prior work on human infrastructures in healthcare, focusing on the theoretical concepts that inform my work. I then review prior research on personal health monitoring in clinical care, focusing on patient-provider collaboration around patient-tracked data. I finally review prior work on personal health monitoring for public health, focusing on how the introduction of technical monitoring solutions has supported both frontline workers and care recipients.

2.1 Human Infrastructures in Healthcare

While infrastructures are traditionally understood as the physical and technological foundations of human activities (e.g., electric grids, telecommunication networks), *human infrastructure* is another critical aspect of understanding modern healthcare services, which necessitates collaboration among diverse professionals [304]. A growing body of HCI and CSCW research has thus focused on the humans who make technology work in complex healthcare settings through the lens of human infrastructures, or organizations of human

labor required for collaborative work [184].

In clinical infrastructures, the planning and delivery of healthcare is shaped by stakeholders beyond healthcare providers, such as pharmacies, insurance companies, and pharmaceutical companies. Prior work in Medical Informatics has thus highlighted that ignoring the collaborative aspects of healthcare deliveries can lead to interruptions in the clinical workflow [1, 235]. For example, Pontefract et al. found that the introduction of physician order entry systems led to an increased communication load between providers and pharmacists, as these systems overlooked pharmacists' routine practices in paper-based prescribing environments, such as annotating prescription sheets to make low-risk amendments [264]. Similarly, Patterson et al. [257] showed that relative to paper-based free-text notes, the introduction of a physician order entry system made it inefficient for pharmacists to process non-standard prescriptions such as tapering medication doses. Studies have also highlighted the influence of insurance policies in the planning and delivery of healthcare. Wang et al. [326] described how Chinese insurance policies, such as rejecting reimbursements for cases that were deemed as inappropriate (e.g., overuse of antibiotics), impacted providers' prescription practices. Groot et al. [113] illustrated another example in which many health insurers in the Netherlands refused to reimburse tapering antidepressants to patients with severe withdrawal symptoms whose providers wanted to administer gradual tapering schedules over longer periods of time. Recent studies further criticized the lack of dose pill options provided by pharmaceutical companies, as it can lead to frequent prescriptions of tablet-splitting despite the potential risk of dose inaccuracy [136, 294]. These examples highlight how the introduction of new technologies can disrupt clinical routines when they fail to account for the broader healthcare infrastructures.

Introducing new technology into complex public health infrastructures also requires continuous human efforts to adapt to unanticipated real-world scenarios by adjusting plans, reallocating resources, and coordinating efforts [107, 308, 287, 29]. This type of work is

characterized as *articulation work*, or "a set of activities required to manage the distributed nature of cooperative work" [308]. Studies have also examined *maintenance work*, as a form of articulation work [287], that frontline health workers perform to anticipate, repair, and reconfigure infrastructural arrangements, thereby maintaining the community health infrastructures [283, 322, 310]. For example, Verdezoto et al. highlighted maintenance work that community health workers performed to repair the lack of social and material arrangements (e.g., limited availability of data entry personnel, a shortage of computers with internet connection) for managing survey data from populations [322].

Berg et al. highlighted that the process of introducing new systems to healthcare settings involves continuous negotiations among various professionals with differing perspectives [29]. Since technologies can embed values and assumptions not shared by everyone, stakeholders may perceive them as tools to achieve their potentially conflicting goals, especially by reshaping roles and distribution of responsibilities of others within the care infrastructures [29, 31]. Studies have shown that tools designed to support community health work often prioritized meeting top-down data requirements, such as task completion and data entry counts, which conflicted with some of the key values that frontline workers held [130, 244].

In summary, components of healthcare infrastructures involve not only providers but also pharmacies, insurers, and pharmaceutical companies, whose interactions shape healthcare delivery. When new technologies overlook these collaborative dynamics, they often disrupt established workflows and increase burdens on health workers. Therefore, a successful introduction of technologies into healthcare infrastructures requires ongoing efforts to adapt systems to real-world complexities and negotiate competing stakeholder interests. In my thesis, I adopt the lens of human infrastructures to examine the complexity of collaboration surrounding the use of health monitoring technologies in healthcare settings.

2.2 Personal Health Monitoring in Clinical Care

Research in HCI, CSCW, and Health Informatics has explored ways to promote patient-provider collaboration around patient-tracked data. Patient-provider collaboration around patient-tracked data is often described as combining patients' experiential knowledge of their everyday health, routines, and lifestyle with providers' medical expertise based on disease-specific knowledge and clinical experiences [62, 269, 289, 72]. Literature has demonstrated many benefits of patient data practices for patient-provider collaboration. From patient perspectives, patient-tracked data can serve as a reference to help them effectively communicate their illness experiences, empowering them to assert their voice in clinical consultations. Patients perceive that their self-monitoring data supports constructing and sharing their own narratives [62, 262]. Pichon et al. suggested that patient-tracked data could act as objective evidence that is potentially more acceptable for providers than a verbal narrative [262]. Patient-tracked data can also assist them in making sense of their conditions through providers' input. When patients collect data on their own, they often experience difficulties interpreting their data and figuring out what actions to take based on the analysis [61, 62]. By engaging providers in their personal data practices, patients can seek feedback from their providers during their visits [122]. Patients' data practices can be particularly valuable to help formulate treatment plans when conditions lack scientific knowledge around appropriate care strategies [262, 72]. When managing these conditions, patients and providers can effectively structure the process of their trial-and-error approach to treatment through self-monitoring [262].

Prior work has also demonstrated the benefits of patients' data practices from providers' perspectives. Patient-tracked data can help providers gain a better understanding of individual patients, such as patient preferences, routines, goals, and values, which could lead to personalized care [61, 135, 62]. Oftentimes, patient-tracked data providing such contexts is subjective in nature. For example, while validated instruments such as PHQ-9 [177] often

ask patients to encode their perceived wellbeing as numbers between 0 and 3, different patients could mean different things by each number. Providers perceive that the subjective meanings of patient-tracked data are critical for understanding how individual patients perceive their quality of life, general wellbeing, and burden of symptoms [333, 122], which is essential for making decisions related to diagnosis or treatment [62]. Patient-tracked data can also help providers organize clinical consultations more efficiently and effectively by allowing them to focus on specific areas, compared to when patients verbally report how they have been doing [135]. Having the ability to access patient-tracked data in advance can also allow providers to plan the agenda for upcoming appointments [61, 255, 243].

Despite the aforementioned benefits, researchers have pointed to several challenges of leveraging patient-tracked data in clinical settings. Patients and providers often have different perspectives on provider involvement in patients' data practices. Studies showed that patients often feel dissatisfied with how much providers engaged with their data [62] and providers are concerned about patients' unrealistic expectations of their engagement in patients' data practices [284, 120, 289, 290, 299]. For example, while patients wish that their providers reviewed their data thoroughly to make sense of it, providers might only intend to give high-level comments [62]. Or, patients wish their providers routinely review their data regardless of having a particular concern, whereas providers might prefer to engage in regular reviews only for high-risk patients [284]. Providers often experience information overload when patients bring large amounts of data, given the time constraints of typical clinic visits [351, 167, 333]. Sharing data with providers outside visits presents another challenge. When patient-tracked data are continuously shared with providers, providers are concerned if patients unrealistically expect providers to monitor patient data continuously and intervene when the data changes for the worse [120]. Real-time access to patient-tracked data could also expose providers to liability issues for unreviewed data [284, 120], particularly for serious consequences such as suicide [299].

In addition, patients and providers may have different opinions on what data should be shared. For example, while providers might be interested in monitoring a broader range of health factors, patients often hold back some data to maintain communication boundaries [135, 198, 332, 333]. Patients often struggle to communicate personal aspects of illness, withholding what matters most to them (i.e., family obligations, personal values) due to a perceived communication boundary, even when they are closely related to their care decisions and quality of life [199, 198]. Oh et al. showed that patients often considered their personal data clinically irrelevant and thus held back on sharing their data with their providers [243]. West et al. similarly demonstrated that patients might not share their data with their providers due to privacy concerns [332]. Individuals managing mental health conditions may have even stronger concerns over their ability to control and selectively disclose their data [230, 120]. Hoefer et al. highlighted the tension between patients' desire not to share some of their self-monitoring data and providers' needs in gaining complete data [120]. While providers often have valid needs for complete data, granting providers all data permissions likely impacts patients' data collection practices, potentially harming the authenticity of data or adherence to self-monitoring [120].

Studies commonly mention that the lack of standardization in measures, formats, and representation of patient-tracked data could hinder their use in clinical settings. When providers initiate patients' self-monitoring, they often give standardized guidance on how and what data should be collected [332, 263]. However, patients tend to desire flexible monitoring tools that account for their individual experiences rather than relying on pre-defined options based on medical terms [212, 284, 122, 223, 17, 121]. When patients initiate self-monitoring, therefore, their data tend to take different forms that might not be perceived as clinically relevant to providers [332, 333]. Providers are less receptive to reviewing such free-form or unstandardized patient-tracked data because the heterogeneity makes it more difficult for providers to interpret the data [212, 61, 351, 284, 333]. In addition, providers often question the validity of patient-tracked data [61, 10, 351, 284]. Providers are often concerned about

patients' data reporting errors due to the lack of knowledge necessary for rigorous data collection methods [351]. Providers also prefer to have contextual information, such as what the patient was doing during the time of data collection, but such information is often missing in patient-tracked data [333]. The incompleteness of data is another significant barrier to using patient-tracked data because missing data creates ambiguity around the patient's condition during those times [126, 333, 202].

As previously mentioned, providers often experience information overload when patients bring large amounts of data, given the time constraints of typical clinic visits [351, 167, 333]. To mitigate this concern, prior work has proposed different types of technology features that support the curation of patient-tracked data to make it easier for providers to review meaningful data points within the time constraints of clinical visits. Filtering features can support patients and providers in focusing on self-monitoring data that is most relevant to their health goals [63, 333]. Features for flagging or sorting based on a range of factors, such as the level of patient concerns or severity of symptoms, can also aid in communicating patients' priorities during visits [284, 122, 262, 12, 13]. Supporting annotations, or allowing patients to augment standardized self-report measures with their own ideas, could further enable patients to mark exceptional events and add contextual information to set an agenda before visits [62, 289, 12, 13]. However, we have a limited understanding of how patients and providers perceive these adaptations to better convey their illness experiences.

To summarize, prior work has shown that patient-tracked data can support patient-provider collaboration by combining patients' experiential knowledge with providers' clinical expertise. These data can help patients articulate their illness experiences and enable more personalized care from providers. However, the subjective nature of patient-tracked data, misaligned expectations between patients and providers, concerns about provider burden, relevance and validity, and privacy issues further complicate collecting and sharing patients' personal health data. While studies have proposed various mechanisms to curate patient-

tracked data to promote patient-provider collaboration, there is a lack of understanding of how patients might adapt clinical measures with their own ideas to better convey their illness experiences. In Chapters 3 and 4, I focus on how personal health monitoring technology could assist patient-provider collaboration, particularly by helping providers develop care plans and helping patients convey their lived experiences amidst the logistical constraints of clinical settings.

2.3 Personal Health Monitoring for Public Health

The HCI community has offered insights into the use of technology by different stakeholders involved in public health work, including government officers, community health workers, and care recipients. One major line of research on technology interventions in public health settings has focused on automating aspects of care that public health workers typically have to provide manually, such as answering common questions [340] and identifying public resources [23, 298]. For example, Pendse et al. [260] highlighted that institutional limitations often interfere with providing support through helpline systems, suggesting that automating some aspects of these systems could help care recipients better navigate the barriers. Relevant to our work, technology is often used to automate the collection of personal health information from care recipients, to reduce the burden of public health authorities in monitoring people at scale. For example, Ismail and Kumar found that health workers often perceive collecting such data to be mundane and redundant, and technology offloading that burden could enable workers to focus on more care-driven tasks [131]. A range of systems, including chatbots [131, 340] and mobile apps [205, 204], have been proposed and examined to support care recipients in self-reporting aspects of their health and well-being to public and community health infrastructures.

Beyond logistical advantages, a benefit of these automated approaches is that care recipients

may feel more comfortable disclosing sensitive information to a digital system rather than a human, such as a positive test result [205, 204, 340]. However, a core concern is that these systems may not be as empathetic or unable to provide emotional support to people going through difficult health experiences in the same way direct communication with a human would [205, 204]. Researchers reiterate that these systems should thus not fully replace public health workers in collection roles but aim to be complementary support [340, 270].

Although the introduction of technology could reduce the burdens of public health work, those experiences may be uneven across stakeholders. For example, in reflecting on years of deploying FeedFinder, Simpson et al. highlighted the uncompensated maintenance and communication labor the service required, despite it being beneficial for care recipients [298]. Further, research often does not capture the attitudes of the people on the front lines of using these technologies, such as community health workers, to understand the technology's benefits and tradeoffs [129]. In my thesis, I gathered perspectives from as many stakeholders as possible to offer a holistic understanding of the use and perspectives of health monitoring technologies.

Through Chapters 5 and 6, my work adds to prior work by unpacking multi-stakeholder perspectives around emerging technologies like AI chatbots to assist public health monitoring.

Chapter 3

Assisting Clinician Planning Amidst Infrastructural Constraints

Many medical conditions require longitudinal care planning, which involves continuous patient health monitoring and frequent treatment adaptations. Patient health monitoring can be particularly challenging for providers when dealing with subjective symptoms, as they need to rely on patient self-reports. In these cases, providers typically follow the guide of patients and make adjustments to personalize their care plans. Adjusting treatment plans is further complicated by the need for buy-in from stakeholders within the clinical infrastructure, such as pharmacies and insurance companies.

In this chapter, I investigate how technology might help provider needs in longitudinal care planning while accounting for infrastructure constraints. I specifically examine clinical decision support technology for longitudinal planning of discontinuation of antidepressants. Since the discontinuation of antidepressants often involves debilitating withdrawal symptoms that can last for months [76, 271], clinical organizations are increasingly recommending gradual discontinuation over months (a *taper*) rather than abrupt cessation, while carefully monitor-

ing symptoms [123]. Tapering antidepressants often requires iterative revision of the schedule to adapt to the patient's reactions and may involve non-standard prescriptions [216, 136, 294], which require buy-in from stakeholders such as pharmacies and insurance companies [291]. Discontinuation of antidepressants is, therefore, a useful case study for understanding how technology can support providers in developing longitudinal care plans amidst infrastructural challenges.

Through an iterative design and implementation of a research prototype, **AT Planner**, a clinical decision support tool that projects tentative prescription plans for discontinuing antidepressants, and multi-round interviews with 12 providers—including psychiatrists, general practitioners, and nurse practitioners, I sought to answer the following research question:

RQ1: How might technology help providers develop care plans within clinical infrastructures?

Toward my thesis claim T1, this study demonstrates that providers' taper planning practices are under interpersonal and infrastructural constraints, facing barriers from pharmacies and insurance companies in creating the complex prescriptions required for longitudinal plans. I also found that providers desired different types of technology support based on their varying levels of experience. Providers with more experience in tapering antidepressants, typically psychiatrists, preferred that technology supports greater flexibility in planning to allow them to adapt taper schedules to their current practice and react to patients' experiences. Conversely, providers with less experience in tapering antidepressants, such as general practitioners, often wished technology could automate the process of creating the taper plans, such as suggesting and generating standard taper schedules.

This chapter makes several contributions to the design of health monitoring technology to support clinical care. First, it contributes design guidelines for clinical decision support tools that assist longitudinal planning, emphasizing the need for flexibly supporting

providers' various regimens, scaffolding longitudinal decision-making through iterative planning, and seamlessly integrating into clinical workflows. Second, it contributes the design and implementation of AT Planner, which scaffolds longitudinal taper planning by projecting schedules, allowing flexible adjustment, and generating notes to connect to the EMRs. Lastly, it contributes design implications for research on clinical decision support tools, particularly around providing greater flexibility, even allowing some loopholes to help providers balance interpersonal and infrastructural constraints.

This work was published in CHI 2022 with co-authors Myeonghan Ryu, Georgia Kenderova, Samuel So, Bryan Shapiro, Alexandra Papoutsaki, and Daniel A. Epstein [145]. Development of AT Planner was done in collaboration with Myeonghan Ryu and Daniel A. Epstein, and participant recruitment was done in collaboration with Bryan Shapiro, a licensed psychiatrist at UCI Medical Center. Georgia Kenderova and Samuel So assisted with data collection, and Alexandra Papoutsaki and Daniel A. Epstein served in supervisory roles, providing guidance and feedback throughout the research process. I led the design of the system, interviews, data analysis, and paper writing.

3.1 Background and Related Work

3.1.1 Prescribing and Discontinuing Antidepressants

With the rise in diagnoses of mental health disorders in the United States, the prescription of psychiatric drugs has rapidly increased [3]. One in six adults in the United States receives a prescription for one or more psychiatric drugs per year, with antidepressants being the most commonly prescribed class of psychiatric medication [8, 258]. Recently, clinical guidelines have recommended discontinuing antidepressants when patients achieve complete symptom remission over an extended period of time. According to the American Psychiatric Associa-

tion, antidepressants can be stopped in stable patients, although the precise timing has not been specified [15]. The United Kingdom's National Institute for Health and Care Excellence suggests stopping an antidepressant six months after symptom remission is achieved [233]. Discontinuation of antidepressants is also considered in other circumstances, such as when patients report intolerable side effects [272], when the medication is ineffective in treating a target condition [47], or when special conditions (e.g., pregnancy or breastfeeding) exist that may adversely affect ongoing antidepressant treatment [328].

Discontinuing antidepressants is a challenging task. A systematic review article reported that more than half of the people who attempt to come off antidepressants experience withdrawal symptoms which may include flu-like symptoms, insomnia, nausea, or sensory disturbances [272, 271, 328, 76]. Withdrawal symptoms have been frequently reported with SSRI (Selective-Serotonin Reuptake Inhibitor) and SNRI (Serotonin-Norepinephrine Reuptake Inhibitor) antidepressants [272]. Antidepressants with shorter half-lives (i.e., the length of time required for a drug to decrease to half of its starting dose in the body [116]) are more likely to evoke withdrawal symptoms with greater severity compared to antidepressants with longer half-lives [123, 271]. These withdrawal symptoms can be debilitating and last for months and even years [76, 271], and can lead to serious psychiatric problems such as suicidal ideation [115].

Research has provided theoretical evidence that exponential taper of antidepressants (i.e., making dose changes based on fixed percentage reductions in dose such that tapers become more gradual towards the end) over months, as opposed to linear taper (i.e., making fixed numerical dosage amount reductions), might help prevent withdrawal symptoms [123]. Clinical recommendations are increasingly advising gradual discontinuation (a taper) rather than abrupt cessation [233, 267, 15].

Past work suggests that tapering plans should be tailored to individual patients along with careful symptom monitoring because tolerance for dose reductions could vary by individ-

ual [272, 123, 113, 99]. For example, providers may adjust the reduction rate, switch to another drug form (e.g., from tablet to liquid), or use a different kind of drug. Switching to a different drug is called *"cross-taper"* [158], which typically means switching from an antidepressant with a shorter half-life to another antidepressant with a longer half-life to mitigate withdrawal symptoms.

Unfortunately, there is no standard approach on **how** to plan for the gradual discontinuation of antidepressants [242, 123]. Instead, recommendations at the time of this study are vague, suggesting tapers *"over the course of at least several weeks"* [15] or *"at a rate proportional to the duration of treatment."* [233] Therefore, providers must independently devise taper regimens for individual patients based on their training and experiences. In the United States, antidepressants are widely prescribed by primary care providers, including general practitioners and nurse practitioners, as well as psychiatric providers [335]. While primary care has benefits in terms of consistency, continuity, and accessibility [139, 301], the training that primary care practitioners receive on antidepressant treatment can be highly variable [335]. As a result, general practitioners may feel less confident about administering tapering regimens compared to psychiatrists [242, 158]

In addition, tapering antidepressants involves complex pharmacological considerations and may require incorporating different dosage formulations to achieve a sufficiently gradual taper. Antidepressants can be available for prescription as tablets, capsules, or liquid formulations. A survey in 2020 with general practitioners and psychiatrists demonstrated that both providers predominantly used tablets or capsules (93-96%) whereas using liquid form was relatively uncommon (19-21%) [216]. Tablets are preferred by providers for both general-purpose and tapering prescriptions as they are more economical for patients and help facilitate flexible dose changes, especially when scored (i.e., embossed with a line to facilitate splitting in half) [136]. However, the available tablet dosage strengths of antidepressants are generally too high to allow for a significant gradual taper and several antidepressants are

available only in the form of capsules or unscored tablets [294]. Liquid formulations allow even greater flexibility for creating smaller or intermediate doses to facilitate a gradual taper [159, 123] and can be an alternative for people who have difficulty swallowing pills [2]. However, they tend to be costly [294] and measuring accurate dosages can be challenging [2, 294]. In addition, insurance companies may reject or require prior authorization for non-standard formulations of antidepressants such as liquid [119, 291].

3.1.2 Technology Support for Clinical Decision-Making

Clinical decision support tools have increasingly been proposed as promising ways to assist providers with computational support in various medical domains such as detection or diagnosis of a disease [50, 51, 27, 114], patient assessment [191, 190], and prognosis prediction [345, 344]. Prior clinical decision support tools have leveraged automation to reduce human errors and mitigate the cognitive burden on providers [104, 195, 220]. Although previous studies have shown the potential benefits of technology in assisting clinical decision-making, those systems have rarely been adopted in clinical practice [89, 138, 156, 235]. A frequently-mentioned barrier is that decision support tools are often a poor contextual fit in clinical workflows [329, 156, 220]. For better adoption of clinical decision support tools into clinical practice, studies have highlighted that such tools should be integrated into the existing healthcare systems such as EMRs [345, 344, 147, 134].

Additionally, providers are not likely to adopt clinical decision support technology if they feel it undermines or infringes their expertise [345, 160, 326]. Therefore, researchers increasingly highlight the need for reconsidering the relationship between the agency of providers and automation in the design of clinical decision support tools [345, 344, 50, 326, 173, 195]. Wang et al. [326] proposed framing clinical decision support tools as *"doctor assistants"* rather than replacements or replications of doctors, emphasizing the need for a clear division

between what tasks can be automated and what tasks should be administered by providers. Studies have described different types of tasks that providers and machines can perform well, respectively. While computational support can be of great help with numeric-based tasks that generate objective output, human attention is required for tasks that need initiative and creativity [195]. In addition, while data-driven technology can help clarify and monitor patient conditions, clinical intuition is necessary when the decision requires balancing various clinical evidence and complex social evaluations [345]. Recent systems have thus leveraged AI for decision-making and interpretation, showing that giving providers agency to collaborate with technology can improve acceptance and effectiveness of clinical decision support tools [190, 50].

3.1.3 Technology Support for Planning and Scheduling for Health

Prior clinical decision support tools have predominantly focused on supporting a single decision-making at a time [345]. However, care for many medical conditions, such as irritable bowel syndrome [289, 63, 149], infertility [70, 71], and depression [230, 134], often requires longitudinal planning and multiple decisions over time. HCI researchers increasingly investigate the need for technology to support longitudinal clinical decisions [345, 147, 114]. For example, Yang et al. [345] found that a heart implant decision requires many smaller clinical decisions that clarify, adapt, and optimize based on evolving patient conditions. Similarly, through their field study with clinical tools for assisting volume therapy decisions in intensive care units, Kaltenhauser et al. [147] suggested that clinical support tools should support decisions over time rather than providing a conclusive diagnosis or prognosis. Previous studies have pointed to the challenges of longitudinal care planning, as it requires continuous monitoring and assessment of patient conditions and adapting treatment plans accordingly. Technology has an opportunity to support longitudinal care planning by providing informational and logistical support to help providers adapt care plans.

Outside clinical settings, research has investigated technology's role in assisting planning for managing and improving diverse domains in health, such as exercise [7, 6], diet [7], sleep [187], and mental health [275, 186]. Technology can provide substantial informational support to help people generate effective plans. For example, by providing expert guidelines and a database of physical activities, Agapie et al. [6] helped crowdworkers plan out the amounts and what kind of health activities to do. Planning support systems such as MUBS [275] have also provided personalized recommendations for activities to do. In addition, technology can help people develop plans which provide answers to specific questions that people might have, such as TummyTrials [149] helping people plan a scientific process to determine what causes their symptoms and SleepCoacher [75] helping determine what causes their symptoms or health outcomes.

Creating effective long-term plans for health requires articulating and scheduling small steps to reach the goal. For example, when creating exercise plans, one needs to identify and articulate what activities to perform, how much activity to perform, and when to perform activities each time [6]. Technology can provide logistical support to help long-term planning by offloading the burden of articulating and scheduling small steps to reach individuals' health goals. For example, MindForecaster [186] and MUBS [275] supported scheduling time to work on health activities, such as intervention plans to cope with anticipated stressful situations or activities to promote mental health. Studies have suggested that humans' insights are vital for balancing complex individual preferences, constraints, and expert guidelines for personalizing plans [6]. Research further highlights the importance of making adaptations in light of individuals' evolving health status, knowledge, and contexts. For example, CrowdFit [6] let planners adjust exercise plans based on the clients' feedback, and Lee et al.'s system [187] allowed individuals to modify what behavior change technique their plan leveraged.

3.2 Methods

To understand providers' perspectives on the role of clinician tools for antidepressant taper planning, our research team conducted two studies: a formative study and a feedback study. The formative study aimed to understand the current practices of providers when tapering antidepressants and how technology can support their needs. The formative study consisted of two rounds of interviews with eight providers who have prior experience supporting patients in tapering antidepressants but come from different clinical backgrounds. I conducted the first round of interviews for need-finding, developing a low-fidelity prototype based on the insights, and the second round of interviews for verification of providers' design needs, with seven out of eight providers returning. Based on the study findings, I developed design guidelines for a clinical support tool for planning antidepressant tapers, which guided the implementation of AT Planner. I then conducted a feedback study with AT Planner, consisting of interviews with eight providers. The feedback study examined the potential impact of technology that supports flexible planning for discontinuing antidepressants in clinical practice, using AT Planner to elicit providers' thoughts. I interviewed 12 different providers in total, with four participating in both studies.

3.2.1 Study Procedures

Formative Study

The formative study consisted of two interviews: the first round of interviews for need-finding and the second round of interviews to verify my interpretations of providers' design needs. Due to the COVID-19 outbreak, all research activities were conducted remotely using video conferencing via Zoom. Two members of the research team conducted each interview, with one asking protocol questions and the other asking probing questions.

In the first round of interviews, I sought to understand providers' perspectives on tapering antidepressants and how technology can and should support their practices. I first understood their current practices of tapering a patient off of an SSRI/SNRI antidepressant by asking them to imagine developing a taper for a patient based on their dosage and mental health condition. I then sought to better understand their approaches to managing tapering antidepressants, including how they calculate taper dosages and adjust plans when patients experience withdrawal symptoms or relapses of depressive symptoms. Finally, I asked them how technology could improve the tapering management process, such as what kind of features they would like a taper planning tool to have and in what contexts they would like to use such a tool. Based on findings from these interviews, I developed five preliminary design guidelines for supporting planning for tapering antidepressants. Following these guidelines, I developed a low-fidelity prototype of a tool to support tapering using Mockflow.

In the second round of interviews, I sought to verify my interpretation of providers' design needs with the low-fidelity prototype. I walked participants through the prototype and asked questions to elicit feedback, such as whether the features would be helpful for configuring tapers in their practice and if there would be any additional features that might be useful. After the interviews, I revised my design guidelines into four guidelines. Section 3.3 describes the refined guidelines.

System Design and Development

Our research team then developed a high-fidelity prototype, AT Planner, which is a realized version of my design guidelines derived from formative study findings, which I describe in detail in Section 3.4. Our research team implemented AT Planner using React in TypeScript. We chose to develop a web application to enable providers to participate in the feedback study from home or in their office by running it on their machines without having to install dedicated software. Because we only implemented the client-side, participants' input and

output data (e.g., participant-generated projected schedules) only persisted in the browser session and was not stored. The application is publicly accessible at: `https://pielab-uci.github.io/antidepressant-tapering/`.

Feedback Study

I conducted a feedback study with eight providers, using AT Planner to scaffold broader conversations around the utility of the tool's concepts for supporting providers in tapering antidepressants and the benefits and challenges of integrating those concepts into clinical practice. The feedback study involved a 60-minute Zoom meeting with each participant. During the study, I sent providers the link to AT Planner and asked them to share their screen and think aloud as they interacted with the tool. My study used AT Planner as a backdrop for understanding the role of technology in the space of planning tapers, rather than evaluating providers' ability to use AT Planner to complete predefined tasks or the usability of the specific interfaces. Therefore, I provided guidance to the providers when they found an aspect of the interface confusing or unintuitive.

I first explained the overall concept of AT Planner, focusing on the iterative planning aspect. Next, I asked participants to come up with an example patient who is planning to be entirely off of an antidepressant and use AT Planner to develop the tapering plan. I then asked providers to imagine that they found the patient struggling with withdrawal symptoms in their follow-up visit, and to make adjustments to the tapering plan accordingly. Once participants had a sense of the capabilities and structure of AT Planner, the interview broadened to ask providers about the potential impact of technology like AT Planner on their practice. The interview questions asked how they felt about using technology to create and adjust a tapering plan compared to their current approach and which of the typical care plans they administer would be well- or poorly supported through AT Planner or technology more broadly.

During the feedback study, our research team regularly met and reflected on the participants' feedback to better fit into their clinical practice. We iteratively refined some aspects of the tool design after each interview and fixed small usability and performance bugs. For example, the default projection mode of AT Planner was exponential, but we added a feature allowing providers to switch to linear per participants' feedback. We also made iterative wording and format changes to patient instructions and notes for pharmacies based on participants' feedback.

Data Analysis

All interviews were video-recorded, automatically transcribed through Zoom, and manually revised to correct errors afterward. I used thematic analysis [40] to qualitatively analyze both interview studies. I open-coded the transcripts to identify patterns in the dataset. The full research team discussed and identified themes. The final codebook contained nine parent codes and 32 child codes for the formative study and seven parent codes and 22 child codes for the feedback study. One of our co-authors, a licensed psychiatrist, regularly reviewed our findings to verify if the conclusions from our analysis aligned or conflicted with medical expertise.

3.2.2 Participants

Our research team recruited providers through mailing lists associated with the Psychiatry Department at our University's Medical Center, other affiliated psychiatric care sites, and through direct recommendations of a psychiatric provider in our research team. We required providers to have prior experience supporting patients in tapering SSRI or SNRI antidepressants. Both formative and feedback studies were classified as exempt by our University's Institutional Review Board because the interview methodology did not involve more than

Table 3.1: Participant demographics and study participation. PS# denotes psychiatrist or psychiatric resident. GP# denotes general practitioners in Family Medicine, and NP# denotes nurse practitioners.

Participant ID	Years post-residency	Study participation	
		Formative	Feedback
PS1 (M, 33)	2	✓	✓
PS2 (M, 50)	18	✓	✓
PS3 (M, 59)	30	✓	✓
PS4 (F, 37)	5		✓
PS5 (F, 29)	4th-year resident		✓
NP1 (F, 36)	1	✓	
NP2 (M, 37)	2	✓	
NP3 (M, 41)	9	✓	
GP1 (M, 33)	3	✓	✓
GP2 (F, 35)	11	✓	
GP3 (M, 37)	7		✓
GP4 (M, 59)	26		✓

minimal risk to participants, and any disclosure of participants' responses would not place participants at the risk of damaging their employability or reputation.

Three psychiatrists, three nurse practitioners, and two general practitioners participated in the formative study (Table 6.2). All eight providers participated in the first interview, with all returning for the second interview except GP2. Our research team compensated each participant $25 cash or a gift card for two 20-30 minute individual interview sessions. Four practicing psychiatrists, one psychiatric resident, and three general practitioners participated in the feedback study (Table 6.2). We compensated each participant $50 cash or a gift card for a one-hour individual interview session. Most participants' primary affiliation was either the Family Medicine or Psychiatry Departments of our University Medical Center. All nurse practitioners were affiliated with private practices, and GP4 was affiliated with a Community Mental Health Center. Participants varied in clinical experience, from last year of residency or a few years post-residency to a decade or more of practice.

3.2.3 Limitations

I sought to understand provider perspectives on the main design components of AT Planner, such as scaffolding longitudinal planning and enabling flexible adjustments. I expected that provider burdens of using a tool operating outside of EMR settings in clinical environments would make it challenging to get feedback on the design components, which I discuss in detail in Section 3.6.3. I thus decided to focus on gaining feedback on prototypes through interview studies without imposing that burden on providers rather than to conduct a field deployment of AT Planner. Further longitudinal evaluation of AT Planner in clinical environments, particularly around integrating into EMR systems, is likely to surface additional challenges in designing clinical tools and further contribute to our understanding of how to support longitudinal planning.

Past work has highlighted that patients are often self-motivated to discontinue psychiatric drugs, with or without the support of a provider [99, 113, 252]. I focused on investigating clinical support for taper planning because the clinical recommendation to support tapering is increasing, but guidance is currently low, which indicates a need to understand how to support providers in developing tapers. Engaging patients in the design process is likely to reveal additional needs for clinician taper planning to account for their perspectives. For example, documenting patients' withdrawal symptoms can benefit both patients and providers as it allows providers to adjust their prescriptions based on patients' felt experiences, particularly when structured for easy review [252]. Having a deeper understanding and incorporating patient needs into the design of taper support tools is an important area of future research.

All of our participants except GP4 were working for either a University Medical Center or in private practice in a relatively wealthy city in the United States. Our participants described their practices in prescribing and discontinuing antidepressants as being influenced by the socioeconomic status of the patients. For example, patients in Community Mental Health

Centers tend to have less frequent consultation times with their providers compared to University Medical Centers or private clinics. Providers' experience in this specific domain, providers' relationship with pharmacies and insurance companies, what EMR systems are being used in providers' medical systems, and how healthcare systems are designed also impact their clinical practices around discontinuation of antidepressants. Therefore, my findings might not generalize as well to different medical settings in different geographical locations in the United States or other countries with healthcare systems designed in a different way.

3.3 Formative Study Findings: Design Guidelines

Based on the insights that I gained through the formative study with eight providers, I developed the following design guidelines that clinical decision support tools for tapering antidepressants should follow.

3.3.1 Flexibly Supporting Providers' Various Regimens

Participant perspectives suggest that taper planning tools should allow for a range of different tapering regimens. Previous work has pointed out that there are no specific clinical guidelines for how providers should support tapering off antidepressants to minimize withdrawal symptoms [242, 123]. Our participants acknowledged the lack of clinical guidelines and described how they devised their own strategies for tapering off antidepressants using their clinical intuition. All of them agreed with the need for gradual discontinuation of antidepressants as abrupt discontinuation can provoke withdrawal symptoms or reemergence of depressive symptoms. For example, NP2 said, *"I am really on the conservative side, so we would usually taper over four-week intervals (i.e., the duration of time on a certain dosage*

strength of a drug before reducing to a lower dosage), between 25% to 50% rate." However, I noted significant variability in each provider's typical regimens. In terms of the reduction rate, two participants (GP2, PS3) said they would reduce antidepressants by 25% every interval, two participants (NP2, NP3) said they would reduce them by 25% to 50% every interval, and three participants (NP1, PS2, GP1) said they would reduce them by 50% every interval. Providers' regimens varied in interval length, as well, from frequent reductions every one-two weeks (GP1) and every two weeks (GP2, NP1) to longer or more varied intervals like four to eight weeks (PS3) or every couple of months (PS2). In addition, providers often desired gradually introducing another antidepressant while decreasing the other simultaneously (cross-taper), as well as tapering off a single antidepressant. They hoped that decision support tools could also assist cross-tapering.

Even though the providers usually applied their own rules in terms of certain reduction rates and intervals, they still noted tapers should be *"customized to the patient"* (PS1) considering various factors specific to each patient. The factors that the providers considered included symptom history (e.g., how long the patient has experienced depression), medication history (e.g., how long the patient has been on the medication), history of the patient's reaction to the medication (e.g., whether they have experienced withdrawal symptoms in their past tapering attempts), and current dosage (e.g., whether it is the maximum available one). For example, if a patient has experienced multiple episodes of depression and been on antidepressants for a few years, some providers described implementing a more gradual taper than if they were tapering a patient off of medication after a single episode that lasted a few months. Similarly, if patients had unsuccessful attempts to discontinue their antidepressants previously, providers wanted to take a more careful approach. Providers also aimed to respect patients' willingness and feelings towards the tapering speed and potential withdrawal symptoms. When patients felt confident and comfortable about tapering at a faster rate (e.g., 25% every two weeks rather than every four weeks), providers often supported them in doing so. PS3 noted, *"I will probably follow the guide of the patient. If*

they feel they can tolerate this, sometimes I may give them a short-term one."

In addition, when providers wanted to incorporate smaller and intermediary doses that they could not directly order from a pharmacy into the taper, they used different strategies. Providers mentioned using compound pharmacies and liquid formulations. However, there were some challenges associated with these methods, such as insurance coverage and pharmacy availability. Therefore, the most common and standard strategies to obtain smaller dosages include using pill cutters to cut scored tablets in half. NP3 preferred using a pill cutter to get smaller dosages as it is a simpler method than others: *"Compound pharmacy or liquid is not always available or covered by insurance, so they are kind of down the list. So I would either use the next dosage available or cut the pills in half to decrease the dosage."* Providers (GP1, PS1, PS3) also wanted to make use of patients' leftover pills for getting smaller dosages by using a pill cutter. GP1 said, *"If you have extra 100 (mg pills), and the next step I wanted to do was to go down to 50 (mg), you can just cut it in half, instead of having to buy more."*

Considering the complex needs in the tapering process, taper planning tools need to support flexibility in terms of reduction rate, intervals, cross-taper, and different methods for micro-dosing.

3.3.2 Seamlessly Integrating into Clinical Workflows

Providers repeatedly emphasized the difficulties of integrating taper planning into their workflows, suggesting that planning needed to be quick and easy as their typical follow-up appointments are short (e.g., from three minutes to twenty minutes). GP1 said, *"being able to do it fast is really important. If you have a 20-minute appointment and that (taper planning) really ends up being like ten minutes, then you don't have a lot of time to do face-to-face time with patients."* I found that providers often had to enter the same information when

prescribing medications to multiple platforms such as EMRs and after-visit notes for patients. Therefore, they were concerned that introducing a new tool to clinical settings might create even greater redundancy. PS1 noted, *"Doing a taper, residents would write it down on the prescription sheet and then put it in the after-visit summary which should be given to the patient. And then if they are going to program it into an app, that's a lot of the exact same data, so that might be annoying."* To mitigate the burden from double-charting, they hoped that the tool would generate prescription information that could be easily transferred to their existing platforms. PS2 imagined, *"If it calculates the taper for you and you can just copy and paste it into the after-visit summary, it would be nice and easy."* Similarly, PS1 stated, *"If it generates the [prescription] text on its own and it's easy enough to copy-paste into the order for prescription, that'd be really convenient."*

3.3.3 Articulating Longer-Term Plans

Participants wanted taper tools to describe the full taper plan, preferring to plan out all dosage changes at once. GP1 preferred planning the taper out on the first day as his typical patients did not come back frequently. Therefore, he usually gave patients instructions involving a couple of steps of dosage reductions: *"I would tell them to cut it in half for two weeks, then cut it in half again for two weeks. If you have any problems, you can always come back. If not, you'll be back in a month, and hopefully, by then, you'll be completely off of it."* Since the tapering schedules might involve multiple steps of dosage change, providers highlighted the need for considering ways to better communicate the complexity of the tapering schedule. NP2 said he would often experience challenges in communicating the taper schedules with patients. He described, *"We try to write it [taper schedule], and they get a medication label from the pharmacy, but still, miscommunication does happen. Anything that would make it easier for patients to comply with the taper schedule would be beneficial."* NP1 empathized with this need and hoped to have a clearer way to communicate

the multiple steps of prescriptions to patients during the taper as well: *"Our notes are clear to us; we know what we're thinking. But that can sometimes get misconstrued by patients. I do find tapering sometimes messed up because patients are just getting confused about when to increase or when to decrease."*

3.3.4 Configurable through an Iterative Process

Providers suggested that the taper process needed to be adjustable. While some of the participants preferred to plan out the whole taper, including multiple dose changes all at once, they often iteratively revised the taper plans in light of patients' reactions to dose changes. For instance, GP2 explained her tapering strategy as to *"move to the next dose down in the follow-up every two weeks."* This practice ensures that they can prevent and manage withdrawal symptoms and relapse of depressive symptoms. A follow-up visit was an important space for providers to determine whether they should stick with or adjust tapering schedules based on patients' tolerance of the previous dosage change. If patients reported at their follow-up visits that they had experienced a relapse of depressive or withdrawal symptoms, providers adjusted the tapering schedules to mitigate the symptoms. NP3 described the tapering process as *"reducing the dosage and then having follow-up visits with patients to determine the response."* Their standard practice for adjusting the taper was to *"go back to the previous dose that symptoms were controlled"* (PS1, NP2, GP1) and *"slower the taper down"* (GP1, PS3, GP2) by increasing the taper rate and/or increasing the length of intervals. PS3 would thus emphasize to his patients that the tapering schedule of antidepressants is tentative and subject to change: *"It doesn't have to be hard and fast. It's not like an antibiotic that you really have to finish the course even if they're feeling better. We convey that there may need to be potential adjustments."*

3.4 The AT Planner Design

Informed by the formative interviews, our research team designed and developed AT Planner, a web application to help providers iteratively develop plans for tapering antidepressants while accommodating their care regimens.

3.4.1 System Overview

AT Planner scaffolds the process of planning antidepressant tapers by allowing providers to select from different options of available dosages for each drug and formulation. Based on the selected projection mode and the difference between the current and next dosage, AT Planner populates a tentative taper schedule projecting future prescriptions until the patient reaches the goal dosage. Then, AT Planner allows providers to flexibly adjust the taper plans to customize to their regimens. I envision that providers could use AT Planner to revisit the projected schedule in the patient's follow-up consultations, making iterative adjustments in light of patients' reactions to the taper. Once providers select what intervals they would like to prescribe, the system automatically creates notes for pharmacies and patients in plain text to be easily shared through existing healthcare systems.

3.4.2 System Features

Informed by my design guidelines, AT Planner has four main features: scaffolding taper planning, projecting tentative schedules, providing flexibility to adjust taper plans, and generating notes to communicate with pharmacies and patients.

Scaffolding Taper Planning

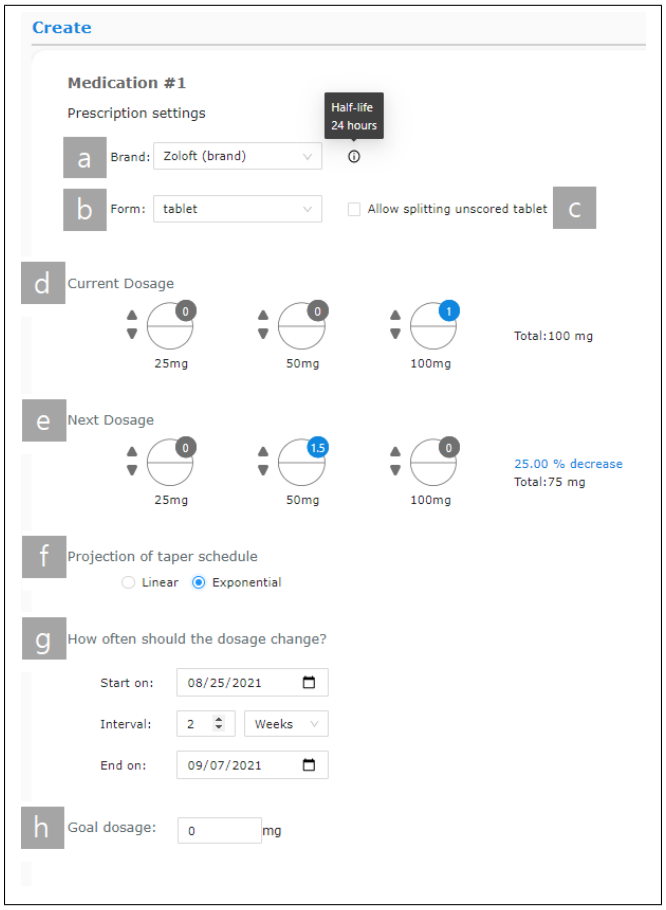

Figure 3.1: To configure a taper plan in AT Planner, providers choose (a) the medication brand, (b) an available drug form, (c) and whether splitting unscored tablets is allowed. Providers then indicate (d) the dose a patient is currently on, (e) the dose to be prescribed next, (f) what mode future projections should take, (g) the duration of each interval, and (h) the goal dosage for the end of the taper.

To scaffold taper planning, AT Planner provides information relevant to prescribing SSRIs and SNRIs. Providers are first given medication options of five of the most frequently prescribed SSRIs and SNRIs [234] in both brand-name and generic version: Prozac / Fluoxetine, Citalopram / Celexa, Sertraline / Zoloft, Paroxetine / Paxil, and Escitalopram / Lexapro. These medications were selected with the guidance of our psychiatrist co-author. AT Planner also provides information about the half-life of a chosen medication when hovering next to the medication options input (Figure 3.1 (a)). After selecting a type of medication

(Figure 3.1 (a)), providers can choose from different formulations, either tablet, capsule, or liquid, depending on what options are available on the market for a prescription for each drug (Figure 3.1 (b)). Providers frequently look up information about medication's available formulations, doses, and half-lives when developing tapers, so we retrieved them from GoodRx [108] and individual medications' package inserts and incorporated them into AT Planner.

AT Planner also scaffolds taper planning by providing a visual interface to configure prescriptions. Once the brand and form of the medication are chosen, AT Planner shows registered dosage options for the corresponding formulation in the current dosage and the next dosage (Figure 3.1 (d, e)). The current dosage is the dose that a patient is currently on, and the next dosage is the dose that the provider intends to prescribe in the current consultation. For tablets and capsules, providers are given available dosage options to select how many of different strength pills to prescribe (e.g., 100 mg, 50 mg; Figure 3.1 (d, e)). The icons of pills change depending on whether they are capsules, scored tablets, or unscored tablets. By default, the count of capsules and unscored tablets change by one, but AT Planner allows prescribing half-doses for scored tablets. If providers feel their patients would be interested and capable of cutting unscored tablets, they can check *"Allow splitting unscored tablet"* (Figure 3.1 (c)). For liquid, when a dosage input is entered in mg, the input in ml is automatically calculated, and vice versa.

After setting the current and next dosage, AT Planner allows providers to choose from two options on how to project the taper schedule: linear or exponential (Figure 3.1 (f)). This input is used to populate the upcoming dosages of medication and create a tentative taper schedule. Depending on the chosen projection mode, projected dosages are calculated based on the difference between the current and the next dosage by rate (exponential) or amount (linear). The duration of each interval could be set by defining two of the start date, interval, and the end date (Figure 3.1 (g)). Lastly, providers could select a goal dosage to

determine when the projection of schedules should stop, with the default being 0 mg (full discontinuation) (Figure 3.1 (h)). Providers can also develop cross-tapering plans. Once they add a new medication and follow the same steps, AT Planner generates projected schedules for both medications in parallel.

Projecting Tentative Schedules

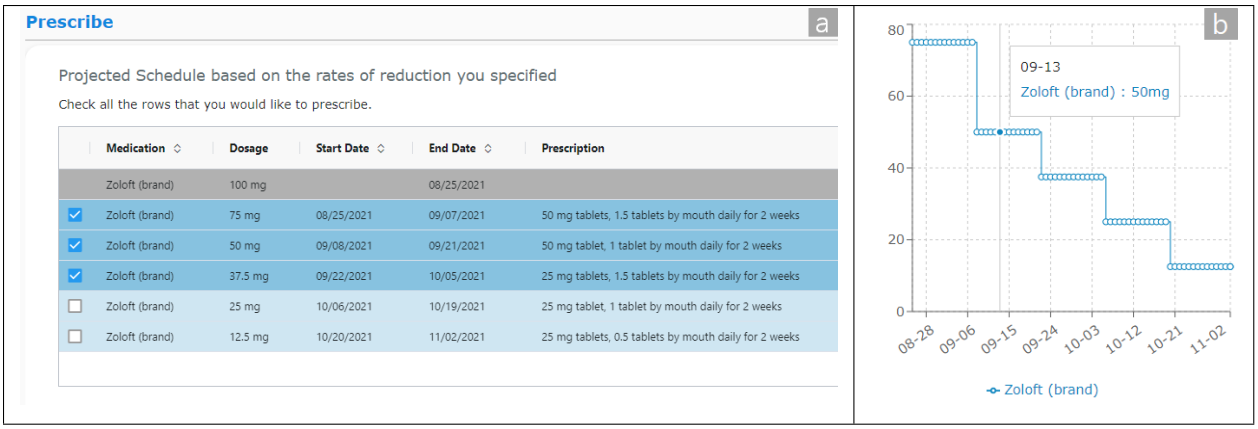

Figure 3.2: Based on the configured settings, AT Planner projects a potential taper schedule in a table and a line chart. (a) Each row in the table represents an interval, and selected are included in the notes for patient and pharmacy (see Figure 3.3 (b)). (b) The line chart highlights the dosages and reduction rate across the schedule.

AT Planner enables providers to plan the entire course of a taper by projecting tentative schedules. Based on the entered dosage reduction projection type, interval duration, and goal, AT Planner generates a tentative taper schedule until reaching the goal dosage (Figure 3.2). AT Planner will approximate the reduction rate, opting for the closest lower available doses of tablets or capsules registered at pharmacies when exact dosages are not available. For example, if a provider developing a taper for Zoloft selected 100 mg for the current dosage and 75 mg for the next dosage, an exponential projection will project future dosages estimating a 25% reduction for each interval. The exponential projection would therefore project 75 mg, 50 mg, 37.5 mg, 25 mg, and 12.5 mg to approximate a 25% reduction with available dosages.

Conversely, a linear projection would reduce by 25 mg each interval, projecting 75 mg, 50 mg, and 25 mg. For liquid, AT Planner will approximate the projected doses to the nearest whole number of milliliters (e.g., 5 ml). Each row in the table represents individual dose changes and corresponding intervals (Figure 3.2 (a)). In the prescription text, AT Planner suggests a combination of available dose options for the projected dose of each interval in a way that minimizes the total count of tablets or capsules. Let's revisit the example above. Considering that Zoloft is available in 25 mg, 50 mg, and 100 mg scored tablets, AT Planner will suggest the combinations to obtain projected dosages as one and a half of 50 mg tablets (75 mg), one 50 mg tablet (50 mg), one and a half of 25 mg tablets (37.5 mg), one 25 mg tablet (25 mg), and a half of 25 mg tablets (12.5 mg). For liquid, AT Planner suggests how many and what size of bottles should be prescribed based on the total milliliters.

The projected dosages are also visualized in a line chart (Figure 3.2 (b)), allowing providers to see the overall plans at a glance. When providers select intervals that they would like to prescribe at a time, the selected rows are highlighted.

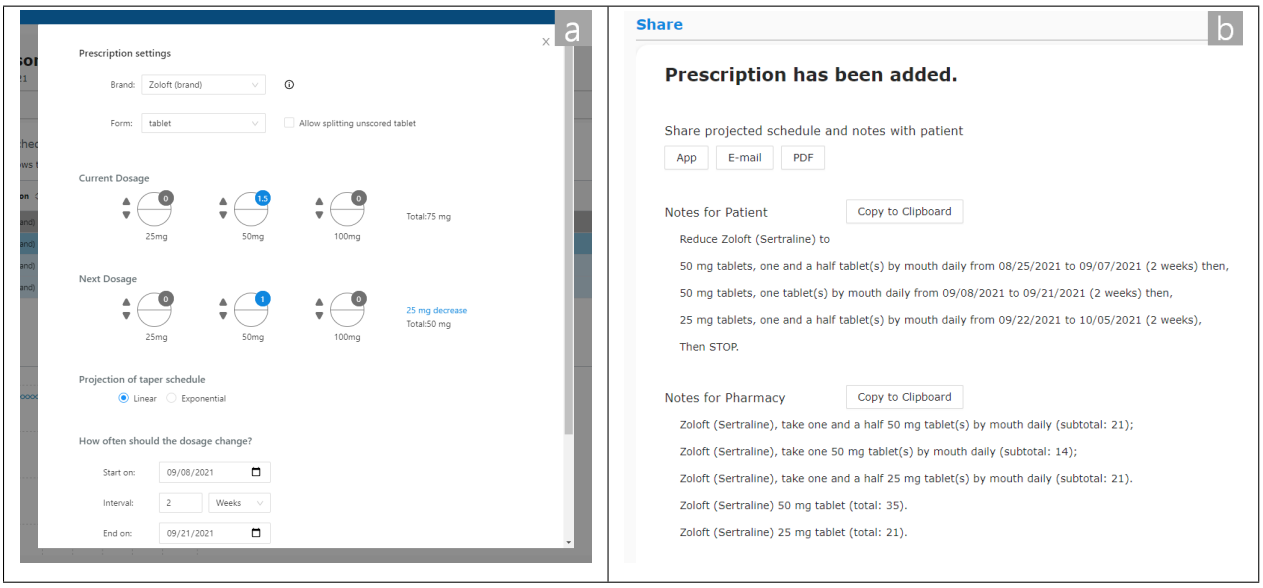

Figure 3.3: (a) Providers can edit the drug prescribed, reduction rate, or duration of projected intervals in AT Planner (see Figure 3.2 (a)), which updates the projection further. (b) To aid in medication-taking and prescription, AT Planner automatically generates notes for patient and pharmacy for the selected intervals (see Figure 3.2 (a)).

Providing Flexibility to Adjust Taper Plans

AT Planner allows providers to flexibly adjust taper schedules to fit their clinical practice. Providers can modify the automatically-generated schedule to fit their regimens. When any row is clicked in the table, a modal window pops up and lets providers change the type of medication, next dosage, projection mode (linear or exponential), length of intervals, and goal dosage (see Figure 3.3 (a)). When the changes are submitted, AT Planner repopulates the selected row and all following rows with updated dosages based on the changed condition. For example, in the previous example, if a provider changed the third interval's dosage from 37.5 mg to 25 mg and instead increased the duration of intervals from 2 weeks to 4 weeks, those changes are projected to the rest of the schedule to lengthen all subsequent intervals and adjust the subsequent reduction rate. AT Planner then saves the projected schedule to a patient profile in the tool (Note that the system did not save the projected schedules in practice, and schedules would disappear upon reloading AT Planner). We envision that when a patient returns for a follow-up consultation, the provider could revisit the projected schedule to make iterative adjustments in light of patients' reactions to the taper.

Generating Notes for Communication

To allow providers to communicate the prescriptions with pharmacies and patients, AT Planner automatically creates notes based on the projected schedule and selected intervals in plain text (Figure 3.3 (b)). The notes for patients provide instructions on how to take the prescribed medication in each interval, such as the combination of each strength of pills. The notes for pharmacy give a brief version of patient instructions with a subtotal and a total number of each strength of pills to be prescribed. I envisioned ways that AT Planner could help providers connect the plans to existing healthcare systems and communicate them to their patients if they were adopted in practice. Using the copy-to-clipboard feature, providers

could copy and paste the notes to their EMR systems or share them with patients. Providers could also send the information to patients via email, a patient-facing app, or print a PDF file (Figure 3.3 (b)). Note that we did not implement these sharing methods; instead, we used the buttons to invite conversations about their utility in the interviews.

3.5 Feedback Study Findings

I found that providers' tapering planning practices were influenced by interpersonal and infrastructural constraints, resulting in taper plans which balanced conflicting needs. Providers' feedback on AT Planner also pointed to desires for different types of support through technology based on their clinical experience in tapering antidepressants.

3.5.1 Impact of Interpersonal and Infrastructural Needs

Providers' taper planning processes often involved careful consideration of the needs of individual patients, pharmacies, and insurance companies. Therefore, the tapering plans that providers developed were often the result of balancing conflicting interpersonal and infrastructural needs.

Consideration of Patients' Health History, Financial Circumstances, and Health Literacy

Consistent with the findings from the formative study, providers indicated that they consider various factors about the patients they prescribe for when developing taper plans. As mentioned in Section 3.3.1, providers took patients' overall health and medication history into account, such as administering more careful tapering regimens if a patient had shown

sensitivity towards withdrawal symptoms previously or had been on an antidepressant for years. For example, PS3 mentioned that he would prescribe alternating doses rather than tapering down to a fixed dose right away for patients who have been on antidepressants for an extended period of time (e.g., ten years). If providers expected that patients might experience withdrawal symptoms, they would often prescribe extra pills in case they cannot tolerate the dose changes. PS1 noted: *"If I'm anticipating a rocky taper, they might have to back up and go to a higher dose again."* PS1 thus added eight pills to the total prescription that AT Planner calculated based on the doses prescribed and the duration of each interval.

Patients' financial circumstances were also an important consideration for providers. GP4 worked for a federally qualified health center focused on treating underserved populations, which impacted his taper regimens. Many of his patients did not have insurance or were on Medicaid, which is operated by the U.S. federal and state governments to provide health coverage for people living in poverty [218]. When tapering such patients, GP4 would opt for filling the fewest prescriptions versus using the best tapering strategies to accommodate their financial circumstances: *"They are really averse to having to spend extra money [getting another prescription]. What I tend to do for uninsured patients is to try to work with what they've got. I would get them to break it [tablet] in half for a couple of weeks and then get them to take one every other day or every third or fifth day. It's not so regimented, but I would rather do what will be helpful for them than have them not be interested in following directions."*

In addition, providers also considered patients' health literacy when developing tapering plans, particularly when using complex tapering strategies such as cutting a tablet in quarters (GP3, PS4), measuring liquid formulations with a syringe (PS2), alternating between different dosages (PS3), or taking different strengths of pills at a time (PS1, GP1, PS3). For instance, PS4 mentioned that she would instruct patients to cut pills only when she was confident they would be able to: *"You have to make sure the person is going to be able to cut*

the pills and be reliable about it. A lot of people just forget, won't know what pill cutter is, or won't cut them in half." GP3 similarly could give complex instructions such as cutting pills into quarters for some patients, but not others: *"We always want to be aware of their abilities for safety for this [cutting tablets in quarters]. It's very hard to cut a pill into quarters, but if they are very high-functioning, and have a very good health literacy, then I would think they would go do this."* Although GP3 understood that prescribing pills into quarters is not standard practice, he wanted AT Planner to support it. PS2 was also mindful of patients' health literacy when deciding what drug formulation and how much dose to prescribe. Since measuring liquid with a syringe could be a challenging task for patients, PS2 did not like the precision involved in liquid prescriptions suggested by AT Planner even though it mathematically fit his taper plan: *"I'm not going to tell them to measure all this, like 7.5 ml, 5.6 ml, over the next six months."* To minimize asking patients to precisely measure liquid, PS2 said he would start the taper with pills and change to liquid formulations once the patient gets down to lower doses, such as 25 mg.

Constraints by Pharmacies and Insurance Companies Influencing Clinical Decision-Making

Providers also often considered the constraints that pharmacies and insurance companies would put on what they were able to prescribe, including requiring authorization processes (PS2, PS5), altering prescriptions (PS3, GP1), or rejecting filling medications (PS1, PS2, PS4, PS5). They frequently mentioned they would face the constraints when sending non-standard prescriptions such as multiple strengths of pills at a time and felt that AT Planner's taper schedules which involved such prescriptions would not be feasible in their practice. PS3 said: *"The pharmacy and many health plans will reject that. They won't allow the patient to get 40 [mg] and 20 [mg]."* PS5 similarly thought, *"If I put multiple prescriptions for 100 mg, 50mg, and 25mg, they're like, 'what?' Then I usually get a kickback."* PS4 added that

some health plans also restrict prescribing high quantities of smaller sizes of pills: *"I tend to like prescribing the fewest amount of different pills. But sometimes, insurance companies will say there's a quantity limit. Some insurances may not let me prescribe just four 5 mg pills. They'll say you have to prescribe a higher dose, like one 20 mg pill. A lot of times, the cost per pill is pretty similar, so it is cheaper to get the highest dose possible."* PS4 thus valued that AT Planner suggested the minimum total count of pills in the projected schedules. Furthermore, providers said some health plans would not cover particular registered doses, and prescriptions generated by AT Planner might not be useful in those cases. PS5 explained: *"Some doses might not be covered [by patients' health plans]. I've run into problems prescribing Prozac 30 mg. 20 [mg] is covered. 10 [mg] is covered. But 30 [mg] is not. Then I'll have to jigsaw puzzle the next dose."*

Providers described that pharmacies and insurance companies also had particular volumes of prescriptions that they preferred prescribing. PS2 explained the reasons: *"What they [some pharmacies] are saying is if you're just prescribing the same drug and if it's not changing over month by month, it's most convenient for the patients, and they [insurance companies] save a little money to do a 90-day prescription."* PS3 similarly described: *"For most of the insurances, I would probably send 90 pills. What often happens is that I'll write a prescription for 30-day, for example, and I'll get a message back the next day asking if the patient can get a three-month supply because they prefer that."* Therefore, he perceived AT Planner using the taper plan to automatically calculate the number of pills to be prescribed would be ineffective: *"Here, it came up with 72 pills because of the next appointment. But if it goes to the technician who receives that prescription, they're going to think that's odd. I may get a message in the evening asking to confirm if I only wanted 72 pills. So I'm going to tell the patient that I want you to take 20 mg until our next appointment, but I would never prescribe 72 pills."*

Since different pharmacies and insurance companies had varied constraints and expectations,

providers would often try to prescribe and see what happens with individual cases. PS5 noted: *"I don't really have a good way of knowing it beforehand. I usually just submit it, and then something happens, like either the patient can't pick it [medications] up, or the pharmacy will try calling us."* To mitigate the providers' burden, PS4 mentioned that her EMR system recently introduced a new feature that provides information about the quantity limit of patients' insurances: *"They just brought out a new tool where we can do a payment estimate using their insurance. Just how much it is or if there's any quantity limit, like 90 tabs per month."* However, it did not provide information about other constraints that the health plans might have. Therefore, she said she would still have to *"wait until you get a rejection by something."*

Accommodating and Circumventing Constraints

Even though providers were generally aware of the benefits of administering a gradual taper and making prescriptions easier for patients to follow, pharmacies and insurance companies imposed various constraints that impede such strategies. Taper planning often involved non-standard prescriptions such as making frequent dose changes, requiring providers to communicate with pharmacists to clarify. GP1 explained: *"When you have the same medication with two different doses at the same time, they don't know if it was a mistake or was on purpose. They'll call you back and ask, 'Did you want to do 40 [mg] or 20 [mg]?"'* PS5 also noted the same issue: *"I've had to make a lot of calls to pharmacies, to clarify what the goal with all these dosages are."* Providers perceived that such practices were time-consuming and interfered with their workflow in developing tapering plans. GP1 said: *"Sometimes it's really frustrating. I have to call them, and it takes another 10 minutes."* Providers would sometimes take extra steps to prevent time-consuming communication with pharmacies. GP1 noted: *"So what I found helpful is to include in the note to the pharmacy that the patient needs to take '20 [mg] plus 10 [mg].' And then they know that I'm mean-*

ing to send those two at the same time. It's an extra step, but it'll definitely save you a phone call and a delay in getting the medication to the patient." GP4 would also call the pharmacy when sending prescriptions for tapering, expecting he would get phone calls from pharmacists otherwise: *"If it's going to be something more complicated, I will usually just call the pharmacy after I send the prescription over and talk with the pharmacist so that they understand what my rationale is. That way, they're not going to be calling me constantly."* He expected that sharing notes for pharmacies generated by AT Planner would allow him to save such phone calls: *"With this [notes generated by AT Planner], they'd understand what you're doing. So it saves the hassle of a phone call."*

Notably, providers often developed workarounds for the constraints of pharmacies and insurance companies. Providers sometimes sent pharmacies prescriptions that did not match with what they actually wanted patients to do. For example, PS3 illustrated how he had worked around the pharmacies' practices when he prescribed alternating doses as tapering strategies, explaining: *"What I prescribe sometimes doesn't translate to what's being filled. They [pharmacies] may make adjustments with the pill size based on my prescription. So I would say take 20 [mg] twice a day, one in the morning and the other at night, in the prescription [being sent to the pharmacy]. And I'll tell the patient that I still want them to take both [two 20mg pills] all at once every other day. I'm doing this for the purpose of the pharmacy giving them 20 [mg] for sure. That way, when I send that prescription, they [pharmacies] can't automatically give her 40[mg]. This is a way to ensure she could do 40 [mg], 20 [mg], 40 [mg], 20 [mg] without any interference."* Therefore, he disliked that AT Planner only enabled daily prescriptions, even though a daily prescription is standard practice for antidepressants [81]. PS1 would similarly work around constraints by prescribing a big supply of the lowest available dose if he knew that the patient's health plan and pharmacy would push for a 90-day supply: *"What I would probably do is to give them a big honking supply of 10 mg [pills]. And then just do [taper] it in increments of the 10 mg tablets. So we could give 'a 90-day supply' to do a month-long taper. I'll just write out 'Take three [10*

mg] tablets daily' in the patient instructions without that being the record. That's one way around it." PS1 therefore wished that AT Planner could suggest different combinations of pills to get a certain dose, so that he could select from them in light of other constraints: *"If this [AT Planner] had multiple options, I could say 'I know this insurance company. I know this patient,' then actually pick and choose from those options."*

Furthermore, providers often adjusted the complexity, speed, and combinations of pills of taper schedules to accommodate infrastructural constraints and patient factors. For example, PS2 would typically prescribe and reuse a single strength of pill to avoid a time-consuming authorization process with insurance companies and pharmacies. PS2 explained: *"If you started at 200 [mg] and you want to reduce to 50 [mg], I'm probably not gonna want to deal with these 25 mg tablets. I might just say take one and a half of 100 [mg] for the next month, 100 [mg] the following month, and then half of 100 [mg]. So that way, you don't have to deal with the pharmacy. I'm not necessarily prescribing separate 25 mg tablets, and having the insurance company give me a hassle about doing authorization. Ergh...forget about it."* As a result, PS2 wanted to be able to administer a faster taper than what AT Planner suggested, but would more gradually taper if the circumstances required it: *"If the patient has demonstrated that sensitivity to withdrawal symptoms and they want to use the smaller doses to have smaller steps, I would definitely do a longer taper."*

3.5.2 Desiring Flexibility versus Automation Based on Clinical Experience

Providers' training and level of experience influenced their desires for tool support. Psychiatrists' ample experience in tapering antidepressants led them to desire greater flexibility in the tool to support their current strategies. Conversely, general practitioners' relative lack of experience in tapering antidepressants led them to wish for a greater degree of automation

to guide their taper planning.

Appreciation of Flexibility over Automation

In general, psychiatrists perceived that AT Planner was a flexible tool which could support their complex tapering regimens. For example, psychiatric providers valued that they were able to flexibly change the reduction mode between linear and exponential taper. While PS2 thought *"most people will be fine with just linear taper,"* he saw value in administering exponential taper for patients who might be vulnerable to withdrawal symptoms: *"Doing this [exponential taper] would allow you to help the small percentage of people avoid very unpleasant withdrawal symptoms."* Psychiatrists also appreciated that AT Planner allowed them to flexibly adapt taper schedules, especially in later phases of the taper, where patients are more likely to experience withdrawal symptoms. PS1 said, *"Oftentimes, the end of the taper is the hardest. That's when we want to be able to break it down into smaller and smaller increments."* PS4 agreed with the need for adjusting taper schedules towards the end: *"There are so many individualized changes that occur that I don't know [in advance]. It's hard to plan for the entire taper because so much of it depends on how people react to parts of the taper."* Therefore, he perceived that AT Planner supporting adjustments would allow them to react to patients' experiences.

Some psychiatrists desired even greater flexibility to support their complex practices relative to what AT Planner provided. For example, PS3 often instructed his patients to alternate doses every other day: *"In the next two months, the patient is going to be on 40 mg one day and 20 mg the other day."* He would also prescribe a higher dose on a certain week for patients with other health conditions: *"There could be a premenstrual dysmorphic disorder where their depression is more pronounced one week before they have menstruation. I would say we want to keep that one week at 40 [mg], but the other weeks I'm comfortable with 20 [mg]."* PS4 wished that AT Planner supported prescribing tablets and capsules

together: *"We would want to have more options, like being able to combine capsules and tablets. Because there's not a lot of flexibility with capsules, sometimes for the end of the taper, I actually mix it up and add some of the instant release tablets in it."* In addition, PS1 hoped that he could flexibly change the combinations of pills to gain a certain dose: *"If this [AT Planner] had multiple options to select from one 30 [mg] tablet, or one 10 [mg] and one 20 [mg] tablet, or three 10 mg tablets, [I'd] then actually pick and choose from those options."*

Psychiatrists generally did not feel that the automatic generation of the projected schedule would be useful, since they were confident in their own ability to support tapering antidepressants. PS3 noted, *"As a long-term psychiatrist, this is easy stuff. It really is. Depression is our primary area of treatment. We know all the antidepressants; when to introduce them, how to decrease them."* PS2 similarly felt schedule generation would mostly not help him: *"The math isn't as complicated when you are just doing each step a month."* However, he thought automatic schedule generation could be useful when schedules become complex: *"The most complicated ones are where you make a change each week or every other week because then it starts to get very tricky with the number of pills you're getting in a monthly prescription. Then, I have to sit here and do some mental math to make sure the total number [of pills] is right. That's when it's most helpful to have something like this. (PS2)"* Similarly, PS4 similarly perceived that it would be helpful for more complicated tapering strategies such as cross-tapering: *"I don't know if I would need this [AT Planner] for just tapering off one medication because I can do it easily in our EMR. It's harder if I'm doing cross-tapering. In that case, it would be more helpful."*

Desiring Automation Relative to Flexibility

Compared to psychiatrists who generally desired flexibility in technology support for taper planning, general practitioners desired a greater degree of automation. GP3 appreciated that

AT Planner automatically generated aspects of the taper schedule that his EMR system did not provide: *"With the current EMR system, there's no automation. There's no way of just saying taper by 50% every two weeks until you're done. So having this functionality for a taper would be very useful."* GP1 also perceived that AT Planner would make his practice of planning tapers more efficient: *"I think it helps because you don't have to type it out. Otherwise, you have to write an individual prescription for each step by clicking a bunch of buttons in the EMR system."* Similarly, GP4 valued that using AT Planner would help him save time creating multiple prescriptions for planning for tapers: *"Honestly, it's a real hassle to have to do that because prescriptions are discrete data points in our EMR. So I just type everything out [when tapering], and that's time-consuming. This [AT Planner] probably even took 10 seconds. You have all this stuff in each prescription. I can just copy and paste. It's just so much easier."* Although most psychiatrists felt confident in their ability to schedule tapers on their own, a few agreed with the general practitioners and appreciated AT Planner's automation. PS1 perceived that the projection of tentative taper schedules could improve efficiency in his practice: *"I love that the calculations were behind the scene. It was really neat to see the prediction that I wanted over here. It saves me the time of having to calculate it and think of the dates. It's just handy enough."* Likewise, PS5 thought AT Planner's automatic generation of instructions would be easier for her to write prescriptions than her current EMR system which does not align well with tapering medications: *"The way our EMR system works, if I wanted to give them like a 100 mg pill, then a 50 [mg] pill, and then a 25 [mg] pill, those would be all kind of separate prescriptions that I would have to put in. It's kind of a pain. It [AT Planner] generates all those instructions that make my life easier, rather than typing it all out."*

Furthermore, general practitioners wished AT Planner incorporated an even greater degree of automation to guide their taper plans. For example, GP1 wished AT Planner indicated what a standard taper for a drug would be: *"I want it to be more automatic somehow. Instead of specifying a current and upcoming dose, you might want to have something automatically*

pops up saying this is the standard taper." Similarly, GP3 also hoped that a system could provide evidence-based recommendations and prognosis of tapering schedules: *"If there's some evidence base behind it to inform clinical decision making, that could be helpful. What does the evidence say about how quickly you should take somebody off of Paxil 20 mg, and then it would tell you to do two weeks or four weeks. Or it can be used for people to input what they are doing and then the system can spit out the likelihood of withdrawal symptoms, etc."* General practitioners were particularly less confident about dealing with complex tapering strategies, such as cross-tapers. GP1 noted: *"For cross-taper, as a primary doctor, we are not as comfortable doing this because we don't do this all the time. It would be helpful to get more guidance on what a normal taper is."* GP3 also resonated with the need for technological guidance on cross-tapering: *"I think the more complex ones where a tool like this can be more helpful are when a patient wants to go from Paxil to Prozac. What is the best way? Is there some sort of cross-taper where you just discontinue one and start the other?"*

The finding on primary care providers' needs for automatic guidance through a tapering support tool is consistent with the needs that other formative study participants mentioned. In the formative study, all general practitioners and nurse practitioners mentioned their desire to gain guidance through tapering support tools. For instance, NP2 stated: *"If you indicate tapering in certain conditions and there's maybe standard tapering doses, that would be helpful."* NP1 similarly expressed the desire to gain automated recommendations: *"Ideally, I would like a recommendation of tapering based on current dosage, length of time patient has been on medication, age, and average health of the patient. If I plug those in and be given recommendations of how to decrease it, such as decreasing by 50% every two weeks, that would be helpful."*

3.6 Discussion

My findings from designing AT Planner and getting feedback from providers on its design suggest opportunities for improving clinical tools to support planning practices. I found that providers played a role in balancing different interpersonal and infrastructural constraints, which might have been exacerbated by a lack of established clinical guidelines and the requirements of longitudinal planning. Based on the findings, I discuss how technology can better support providers in balancing the influencing constraints on developing care plans. I also found that providers with varying levels of experience in tapering antidepressants had different design needs for clinical decision support tools. Providers with more experience were likely to desire a flexible tool to support their current practice, whereas providers with less experience were likely to seek automated guidance from technology. I suggest that providers' varying levels of experience need to be carefully considered in the design of clinical decision support tools. Lastly, I consider the opportunities and challenges of clinical decision support tools operating outside the EMRs in research and in practice.

3.6.1 Developing Flexible Planning Tools to Assist Providers in Balancing Influencing Infrastructural Constraints

Through this study, I found that providers' planning practices for tapering antidepressants were influenced by various constraints arising from the human infrastructures of clinical care, including individual patients, pharmacies, and insurance companies. In my work, providers played a significant role in balancing such constraints. Providers are often engaged in complex evaluations on what combinations of pills they should prescribe to meet individual pharmacies' and health plans' constraints while ensuring the instructions are not too complex for patients to follow, nor too expensive for patients. Prior work has emphasized the importance of considering human infrastructures in the design of clinical support

tools [30, 31, 235, 264, 134]. Consistent with prior work, my findings suggest that clinical decision support tools should be mindful of other interpersonal and infrastructural factors in play [184]. I posit the challenges in balancing various infrastructural constraints are particularly likely to emerge in longitudinal care planning than in one-time clinical decision-making contexts. In my work, longitudinal taper planning required multiple dose changes over time and iterative adjustment of schedules. Adapting plans to provide care that was responsive and accommodating to patient needs often required providers to make non-standard prescriptions, such as combining different strengths of pills or many small pills that pharmacies or insurance companies may not allow.

Prior medical literature has frequently mentioned that providers often used workarounds to circumvent the requirements of EMR systems [30, 235, 174]. These workarounds mainly occur when the human infrastructures in which the systems are embedded have not been accounted for in the design [30, 174]. Our participants similarly indicated using workarounds to circumvent the interpersonal and infrastructural constraints. Specifically, I suspect that the lack of standardization in clinical guidelines around tapering antidepressants is likely to have led providers to seek out loopholes in EMR systems because the pharmacies and insurance companies they worked with lacked policies and protocols that align with the care regimens required for tapering as a result. Prior work on EMR systems has suggested allowing space for providers to indicate their needs to other groups [254]. Extending this, my study suggests the value of flexibility in clinical decision support tools to accommodate providers' needs, particularly around longitudinal planning and a lack of standardized guidelines. However, addressing the larger problem of getting care regimens recognized by other stakeholders requires creating and formalizing guidelines through clinical trials. Before clinical guidelines are established, technology can help support flexibility in developing care plans for patients.

3.6.2 Different Levels of Experience Impacting the Design Needs for Clinical Decision Support Tools

I found that providers with relatively less experience, such as primary care providers, often faced challenges in developing taper plans due to their unfamiliarity and lack of clinical guidance. Therefore, they desired different types of support from what providers with more experience, such as psychiatrists, desired. While the most experienced psychiatrists were often skeptical that technology-driven decision support would even be helpful and preferred to rely solely on their own experience, primary care providers desired more automated guidance to offload the decision-making burden to technology. Many health conditions, including diabetes, HIV, and depression, are treated in both primary and specialist care settings. While primary care has benefits in providing continuous and accessible care for patients throughout their life course [139, 301], prior work frequently mentioned primary care providers' concerns about knowledge and experience with regimens that can be complex or risky [139, 301, 158]. Although both groups might face challenges in developing care regimens with a lack of clinical guidance, specialists are more likely to develop their own regimens over time based on their empirical knowledge resulting from many trials and errors; on the other hand, primary care providers are less likely to feel confident about developing their own regimens due to their limited experience in specific domains relative to specialists [242, 158].

This finding indicates that providers' varying levels of experience should be carefully considered in the design of clinical decision support tools. Prior work on clinical decision support tools suggested the potential of AI-powered systems in providing data-driven recommendations [195, 191, 190, 27, 114, 50]. This approach could help providers with less experience navigate the challenges of developing care regimens with a lack of clinical guidance. On the other hand, because providers with more experience are likely to have developed their unique regimens with the lack of standard clinical guidelines, providing automatic recommendations without the ability to make adjustments might make them feel that their expertise is not

recognized [345, 160, 326]. Providers with more experience might instead benefit the most from a flexible tool to support their regimens involving complex strategies and iterative adjustments. Or alternatively, workflows could enable experienced providers not to engage with tool support at all. A challenge in this space is that it is difficult to design a single tool to provide both automated recommendations and flexibility to support various regimens of providers. When possible, understanding the experience levels and needs of the target providers before designing new clinical decision support tools can help tailor them to the audience. In cases where it is not possible to tailor such tools to a specific target audience, we should consider developing tools that offer guidance only when needed or requested but let providers leverage their own expertise.

3.6.3 Opportunities and Challenges of Clinical Decision Support Tools Operating in and outside the EMRs

Though I did not aim to evaluate the clinical efficacy of AT Planner, its development led me to reflect on the relative advantages of prototyping clinical decision support technology in and outside of EMRs. My motivations for designing AT Planner as an online tool outside of an EMR were predominantly practical. Doing so allowed me to implement the design without having to be tied to a specific EMR system for participant recruitment (e.g., recruiting only participants who work in medical systems using Epic), and avoiding regulatory needs around developing an add-on to an EMR tool and software installation on computers in hospitals and clinics. This decision greatly reduced prototyping time and participant recruitment, enabling me to efficiently evaluate my design ideas and contribute a broader understanding around clinical decision support tools. I thus see significant advantages for HCI researchers to prototype clinical decision support tools outside the EMRs for developing an understanding of design approaches.

Despite the benefits of implementing tools operating outside the EMRs, significant challenges emerge when considering the utility of the approach in field deployments or extending from a design idea to clinical adoption. A main challenge is that the barrier to provider adoption remains high. Previous studies primarily highlighted the need for clinical decision support tools to be integrated into the existing health care systems, such as EMRs, to minimize interruptions and fit into providers' workflow [345, 344, 147, 134]. Similarly in my study, even though AT Planner generated output for prescriptions that could theoretically be copied into their EMR, providers often expected using the tool would involve additional work for them to double-chart prescriptions or other patient information. They hoped that tools like AT Planner would be integrated into EMRs, and expected that such tools are unlikely to be used in clinical environments otherwise.

Another significant challenge mentioned by several HCI studies [351, 352, 62] is avoiding the use of any personal health data in the tool itself under the medical data regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the U.S. and the General Data Protection Regulation (GDPR) in the European Union. One model providers in my study frequently leveraged was using online resources to look up information about antidepressants (e.g., GoodRx [108]). Designing clinical decision support tools to similarly serve purely as resources, where inferences or plans can be developed and exported to EMRs, is one potential opportunity for standalone technology. However, this is often not possible for many kinds of tools, such as AI tools which aim to aid with the interpretation of patient health data.

In a domain like tapering antidepressants where medical advice is still evolving, operating outside the EMR enables tools to more quickly respond to the evolving medical literature, such as by integrating recommendations as clinical trials are published. Although tools operating outside EMRs can more easily integrate advice, they lack regulation important to ensuring the advice is clinically supported. Participants in my formative interviews,

particularly primary care providers, frequently mentioned desiring advice on effective taper plans for particular medications. However, our research team intentionally opted not to incorporate such advice for a few reasons. We sought to avoid going beyond our primary area of expertise, as the medical community is in the phase of conducting individual trials, and tapering recommendations have not yet been formalized into national guidelines. We were also concerned that the advice offered by AT Planner could be subject to regulation in future stages of research, reducing our ability to quickly evaluate the design ideas. However, opting not to include advice came at a cost to the needs of providers, particularly those with less experience with taper planning. Trading off the relative benefits of operating quickly outside of medical record systems with operating carefully inside those systems warrants further consideration in future projects.

3.7 Conclusion

Through designing and evaluating clinical decision support for tapering antidepressants, I found providers' planning practices were often influenced by interpersonal and infrastructural constraints and clinical experience influenced their design needs. Allowing some loopholes in clinical tools can be valuable for navigating infrastructural barriers, particularly in domains with a lack of standardized guidelines. Providers with more experience desire flexibility in decision support systems, while providers with less experience appreciate automated guidance from technology. Therefore, providers' varying levels of experience should be carefully considered in the design of clinical decision support tools. Lastly, I suggest that developing decision support tools which operate outside EMRs can allow HCI researchers to quickly implement the design and more quickly respond to evolving medical literature. However, the barrier to provider use of external tools remains high, which presents challenges when conducting field deployments or extending to clinical adoption.

Chapter 4

Assisting Patient-Centered Communication Amidst Logistical Constraints

Many clinical conditions involve subjective and idiosyncratic symptoms that affect individuals' daily lives in varying ways, including chronic pain [137, 5, 266, 4, 279], cancer [255, 122, 180, 121, 88], and mental illness [102, 24, 215, 229, 214]. In such cases, patients' self-reports—or any methods that rely on an individual's own descriptions of their symptoms, behaviors, feelings, and attitudes [97]—play a vital role in care management by allowing patients to communicate their illness experiences with their providers. Traditional methods to elicit patient self-reports include standardized paper-based measures, but they often fail to capture the richness of patients' lived experiences [137, 121] and may limit their agency in managing illness in everyday life [18, 17, 72].

To mitigate the limitations of the traditional approaches to eliciting patient self-reports, HCI, CSCW, and Health Informatics communities are increasingly developing and using

technology to support personalized and expressive self-reporting, such as pictorial trackers [17], digital storyboards [122, 121], and media probes [122, 63, 121]. However, making these formats practically usable in clinical settings remains challenging, particularly given the time constraints of typical clinical visits [61, 332, 333, 168]. One way to better center patient needs in clinical symptom measures routinely used in clinical care is through **annotations**, which allow patients to adapt and extend these self-report measures with their own ideas. Allowing patients to annotate standardized self-report measures with free-form data could potentially empower them to convey their lived experiences to their providers while preserving clinical relevance.

Through interviews with patients with AT Annotator, a research prototype that allows users to enhance standardized clinical self-report measures with five annotation types—free-text notes, emojis, animated GIFs, icons, and body parts, I sought to answer the following research question:

RQ2: How might technology help patients convey their health experiences within clinical infrastructures?

Extending Chapter 3, this chapter examines this research question in the context of discontinuing antidepressants. Since most withdrawal symptoms that patients experience when discontinuing antidepressants lack objective measures, providers typically rely on patients' self-reports using standardized tools. However, due to the subjective nature of withdrawal symptoms and variability of patient tolerance for dose reductions [123, 272], such measures might not fulfill providers' and patients' symptom monitoring needs. Discontinuation of antidepressants thus presents a useful case study for studying patient adaptations of clinical self-report measures.

Toward my thesis claim T1, this study shows that annotations could effectively support many patient communication goals amidst the logistical constraints of clinical settings by enriching

clinical measures with their individual illness experiences, such as symptom fluctuations and the impact of medication changes on daily life. Participants believed annotations could also empower them to highlight their primary concerns to their providers, aid in recalling symptom experiences, and alleviate the logging burden. However, participants were concerned that visual annotations, such as GIFs and emojis, might disrupt the professional relationship with their providers due to their casual or ambiguous nature. Participants further pointed out that annotations might not adequately address the sensitivity and complexity of mental health contexts, potentially imposing significant mental and emotional burdens while lacking the flexibility needed to convey mental health experiences.

Based on the findings, I discuss opportunities for annotations to support patients in conveying their personal experiences and influencing the direction of their communication with providers. I further propose incorporating customization support to cater to patients' diverse communication needs around data forms. Lastly, I suggest opportunities to enhance the clinical practicality of digital symptom measure annotations by guiding patients to focus on those more pertinent to their ongoing care when sharing them with providers.

This project was published in the Proceedings of the ACM on Human-Computer Interaction 8, CSCW2 in 2024 [146] with co-authors Rachael Zehrung, Katherine E. Genuario, Alexandra Papoutsaki, and Daniel A. Epstein. Rachael Zehrung and Katherine E. Genuario assisted with data collection, and Alexandra Papoutsaki and Daniel A. Epstein served in supervisory roles, providing guidance and feedback throughout the research process. I led the design of the system, interviews, data analysis, and paper writing.

I proposed the initial study ideas, developed the research prototype and interview protocols, recruited participants, and led the interviews, data analysis, and paper writing.

4.1 Background and Related Work

4.1.1 Symptom Monitoring for Antidepressant Discontinuation

In recent years, the prescription of psychiatric medications—especially antidepressants—has risen significantly in the United States [3, 8, 258], prompting growing clinical interest in safe discontinuation practices. While clinical guidelines recommend tapering after sustained remission or in other clinical circumstances [233, 15], providers face persistent challenges due to vague recommendations [242, 123] and the high prevalence of withdrawal symptoms [76, 272, 271, 328, 115]. For an in-depth discussion of prescribing patterns, clinical guidelines, and challenges associated with tapering, see Chapter 3.

Providers typically rely on patients' self-reports using clinician-rated instruments, such as the Discontinuation-Emergent Signs and Symptoms (DESS) checklist [278], to monitor whether patients have any new or worsening withdrawal symptoms or the Hamilton Depression Rating Scale (HDRS) [117], the Patient Health Questionnaire (PHQ-9) [177] and the Beck Depression Inventory (BDI) [26] to monitor potential relapse of depressive symptoms. If relapse occurs, providers often revert to a higher dose or consider alternative treatments such as different medications.

However, given the subjective nature of common withdrawal symptoms as well as depressive symptoms, this approach may be limited in capturing patients' lived experiences during the taper. Recent work shows that many patients feel that they do not receive adequate support for their safe discontinuation of psychiatric drugs from their providers and turn to online health communities [113, 99, 252]. These patients often monitor their own drug history and daily symptoms, with some following taper plans they found in online health communities. Acknowledging the importance of understanding patients' lived experiences and the need for provider engagement in the tapering process [113], Papoutsaki et al. argued that such

patient-generated logs could instead be leveraged to facilitate patient-provider collaboration for the safe discontinuation of psychiatric drugs [252]. For example, patient-generated logs could serve as evidence of patients' illness experiences and medication adherence, allowing providers to distinguish between withdrawal and relapse and adjust their prescriptions accordingly.

4.1.2 Self-Reports as Patient-Tracked Data

While section 2.2 examined the benefits and challenges of leveraging patient-tracked data for patient-provider collaboration, this subsection specifically narrows the focus to prior work on patient self-reports as a particular form of patient-tracked data. Among the various forms of patient-tracked data, patient self-reports is one of the most widely used and essential forms, often used for pain self-assessment [137, 5, 266, 4, 279], mood tracking [24, 215, 229, 214, 122, 121], and journals about daily routines [122, 180, 121]. Patients' self-reports play a vital role in care management, enabling them to share qualitative insights about their illness experiences and improving providers' understanding of their quality of life, well-being, and burden of symptoms [333, 122]. This information is particularly useful for understanding subjective and unique symptoms that impact daily life in conditions such as chronic pain [137, 5, 266, 4, 279], cancer [255, 122, 180, 121, 88], and mental illness [102, 24, 215, 229, 214].

Traditionally, self-reports are generated through clinical interviews conducted by trained professionals based on disease-specific measures [251, 61, 302, 266]. Although clinical interviews can ensure reliable patient assessment [201, 251], this approach has several drawbacks. These interviews rely on patients' retrospective recall, which might be biased by their symptom experience at the time of reporting or recent extremes [255]. In addition, time constraints of typical clinical visits may prevent providers from eliciting detailed accounts of patient experiences [201, 251, 266]. Thus, providers often find clinical interviews insufficient for

understanding what happens between visits [122].

Another common method to elicit patient self-reports is using self-report measures. Two terms closely related to self-reporting in clinical settings are patient-reported outcome (PRO) measures and observations of daily living (ODLs). PRO measures are validated tools for monitoring and assessing patients' symptoms, functional status, and health-related quality of life [37]. Numerous PRO measures exist, including disease-specific and generic ones, typically in the form of multiple-choice questionnaires [37]. Typical PRO measure questions include the frequency and severity of symptoms, their impact on daily life, and perceptions of conditions or treatments [80]. PROs are the gold standard for assessing symptoms in clinical trials [46, 285] and are also widely used in routine clinical care to monitor symptoms and evaluate treatment outcomes [80, 285]. As standardized, validated instruments, PROs reliably capture some aspects of patients' quality of life [80, 180]. However, relying on standard measures risks losing valuable nuances about patients' lived experiences [137, 121, 180] and diminishing their sense of agency in managing illness in everyday life [18, 19, 72]. Pichon et al. [262] suggested that patients could feel unheard or dismissed when standardized clinical measures used in their care fail to capture key aspects of their illness experiences.

In contrast to PROs, which are clinician-defined and patient-generated, ODLs are patient-defined and patient-generated data [42] with a greater focus on personal experiences of health. Brennan et al. highlighted that medical language often differs from patients' everyday experiences, suggesting ODLs as a new mechanism for patients to collect personally meaningful health data and express their unique experiences in their own words [42]. However, such patient-defined self-monitoring often produces free-form data that may not align with clinical standards and is often seen as clinically irrelevant to providers [332, 333]. The heterogeneity in data forms and representations further complicates provider interpretation of patient-tracked data [61, 332, 168, 333].

In this chapter, I explore how patient health monitoring technology might better center pa-

tient needs in communicating their illness experiences with their providers while maintaining clinical relevance.

4.1.3 Patient-Centered Communication and Care

While prior work has highlighted the role of patient-tracked data in patient-provider collaboration (see section 2.2), these data practices alone are often insufficient to fully bridge the gap between patients' lived experiences and providers' clinical priorities. Beyond patients' data practices, research has pointed to deeper communication mismatches in patient and provider perspectives and discussed how to address these gaps. This subsection reviews prior work on patient-centered communication and care beyond patient health monitoring.

Patients and providers tend to focus on different aspects of illness, leading to divergent concerns and priorities [48, 269, 11, 198, 118, 13]. While providers tend to base their care recommendations on biomedical data (e.g., lab results), patients experience illness as personal events that directly impact various aspects of their daily lives [118, 48, 12, 11, 13, 198]. For example, patients may prioritize certain activities (e.g., gardening) or relationships (e.g., newborn grandchildren) when planning treatments, but these personal needs are not routinely discussed in clinical settings [198, 199]. Providers may overlook non-medical aspects of illness that are important to patients due to their focus on medically oriented goals [32, 208]. Such communication gaps can lead to conflicts in care priorities, negatively influencing patient health outcomes [34].

Prior work suggests that aligning patient and provider perspectives requires a collaborative process to identify and translate meaningful patient concerns into actionable insights for providers [12, 11, 13]. Studies have proposed ways technology could help bridge the gap between patient and provider perspectives, particularly by empowering patients to communicate the unique, personal aspects of their illness experiences. Berry et al. [33, 34, 35] sug-

gested utilizing open-ended prompts [33, 36] or visual artifacts (e.g., photos and videos) [34, 35] to better equip patients to discuss their values during clinical visits. They argued that open-ended, exploratory tools could help overcome perceived communication boundaries, whereas tools explicitly designed to generate discussion topics with their providers might reinforce existing boundaries, highlighting the value of supporting exploratory reflection [36]. Chung et al. [62, 63] characterized personal tracking data in patient-provider collaboration as boundary-negotiating artifacts [183] to mediate different spheres of expertise. Consistent with Berry et al. [32, 34, 35], they suggested the potential benefits of incorporating visual artifacts to resolve disagreements between patients and providers [62, 63].

As a first step to address this gap, this work explores patient annotations to clinical measures as a means of facilitating patient-centered communication on subjective illness experiences.

4.1.4 Patient-Driven Tracking

Patients actively engage in data practices not only to collaborate with providers but also to support their own self-care and self-management. While section 2.2 focused on patients' tracking practices in the context of provider involvement, this subsection examines patients' independent tracking practices, without direct engagement from providers.

Prior work has consistently highlighted that patients' data practices allow them to gain awareness and insights even without provider involvement [102, 112, 229, 71]. While many health conditions involve a great variability in individuals' illness experiences, patients' data practices can help them identify patterns in their symptoms, routines, and daily activities, potentially leading to insights about how to cope better [102, 112, 229] and perceived agency [229, 121]. Furthermore, patients' data practices can facilitate their self-expression and mindfulness [229, 165, 5].

Studies show that patients often prefer configuring their own tracking regimens rather than relying on pre-defined options from providers [284, 122, 223, 17, 121, 88]. Past work has proposed ways to support customized tracking for patient needs. For example, Schroeder et al. [290] proposed goal-directed tracking, allowing patients to customize data types and frequency based on their health goals. Similarly, Karkar et al. [149] presented TummyTrials, a tool that generates customized study protocols to identify causal relationships between food triggers and symptoms of irritable bowel syndrome.

Further, HCI studies have proposed various ways to support free-form descriptions of patients' illness experiences. Maharjan et al. [211] and Lazar et al. [181] explored how a speech modality can be used to elicit patient self-reports and promote self-expression in the context of mental health and cognitive impairment, respectively. Yamashita et al. [342] and Luo et al. [206] showed that free-form texts enable individuals to describe their experiences in a flexible manner. Stawarz et al. [305] and Mishra et al. [224] and other studies in medical informatics [306, 300] demonstrated that emojis could be an effective way to convey emotions in clinical settings as it reduces cognitive burden by using symbols frequently used in everyday life [300]. Icons have been suggested as a way to allow for succinct and efficient communication on teletherapy platforms [341]. Adams et al. [5] proposed tailoring pain assessment measures (e.g., coloring, number picker, illustrative faces, slide bar) to consider patients' individuality in how they experience and process symptoms. Ayobi et al. [17] suggested pictorial symptom journaling to improve a sense of agency in patients with multiple sclerosis, surfacing the need to consider individual data collection preferences. Hong et al. [122, 121] similarly proposed digital storyboards for adolescent cancer patients to communicate their illness experiences in their preferred ways. Insights from prior work suggest that self-tracking can serve as a mechanism for patients to articulate their needs in personalized ways.

Despite the benefits of patient-driven tracking outside clinical settings, there are several challenges and risks in patients' tracking without provider involvement. Patients with com-

plex health needs often lack the necessary information or knowledge across their tracking processes and might find it challenging to decide how and what data needs to be monitored to achieve their health goals without providers' involvement [212, 288]. For example, women engaging in fertility tracking often struggle to understand how and what health indicators they need to track as fertility care comprises specialized, complex knowledge that laypeople do not commonly have [70]. Patients may also find it difficult to recognize and self-assess their symptoms by themselves, particularly for patients with mental illnesses [214], children and adolescents who are still developing their the necessary literacy and conversational skills to articulate their illness experiences, or patients who are going through treatment that could compromise their cognitive functions [122]. In addition, patients often find it difficult to make sense of their data [112]. Most self-monitoring technologies lack mechanisms to support patients' sensemaking process of their data [112, 149, 290]. Therefore, individuals struggle to answer specific questions about their health through tracking.

Further, patients often find it burdensome to engage in tracking. Patients experiencing debilitating symptoms likely have limited energy to engage in self-tracking practices [255, 24, 10, 135, 63, 5, 223]. Studies on individuals with bipolar disorder found that they might find it burdensome to engage in mood tracking because focus or energy is low during their manic or depressive episodes [229, 214]. Similarly, severe pain likely interferes with patients' ability to engage in their pain self-reporting [5]. In particular, patients' data practices could be emotionally draining [10, 5, 70, 69, 72]. Patients experience emotional consequences when they do not receive the desired results. Chen et al. demonstrated that diabetes patients often feel frustrated and disappointed even with slight glucose fluctuations, leading them to lose interest in disease management [54]. Similarly, in Ancker et al., patients with diabetes felt frustrated when seeing undesirable blood glucose levels [10]. Murnane et al. further noted that individuals with bipolar disorder could engage in rumination when realizing their undesirable states through tracking [229]. Matthews et al. also pointed to the general risk for individuals with bipolar disorder to engage in self-tracking, which could foster hyper

self-scrutiny and unrealistic normative expectations of health [214]. Eikey et al. showed that weight loss apps could exacerbate the obsessive logging behaviors of individuals with eating disorders through their heavy focus on numbers [87]. Figueiredo et al. suggested that fertility tracking data often have strong moral and emotional implications, contributing to anxiety and stress without achieving the desired result [69]. Figueiredo et al. further emphasized that patients' fertility tracking data are inherently emotion-laden because the fertility struggles disrupt their daily lives and are directly related to the patients' life plans [72]. Furthermore, tracking tools often provide data analysis and/or recommendations for action [229, 79], which involves risks for fueling or even triggering emotional instability. For example, Frost et al. showed that providers were opposed to the idea of presenting mood forecasts to patients with bipolar disorder because patients might end up feeling depressed due to a forecasted depression [102]. Matthews et al. noted that a streak feature, which displayed the current run of days in a row that the user achieved behavioral goals, could exacerbate the risk for individuals with bipolar disorder engaging in data practices [215].

Inspired by prior work, in this chapter, I explore how technology can better support patients in conveying their lived experiences while mitigating the potential risks of patient-driven tracking. As indicated in Chapter 1, I primarily use the term *monitoring* rather than *tracking* throughout the rest of the chapter to highlight the collaborative aspects of these practices in clinical care, extending beyond individual use.

4.2 Methods

To understand how annotations could support patient adaptation to clinical measures, I conducted interviews with 20 patients who were either interested in stopping or in the process of discontinuing their antidepressants. This study was approved by our institution's Institutional Review Board.

4.2.1 Study Procedures

I designed AT Annotator as a digital aid for understanding how annotations to clinical measures might help or hinder patients' communication needs when tapering antidepressants. The aim of designing AT Annotator as a low-fidelity prototype was to introduce the concept of annotations during the interviews.

Prototype Design

I designed a low-fidelity prototype, AT Annotator, to elicit conversations about how annotations to clinical measures might support their communication goals with their providers. The design of my prototype was influenced by findings from prior studies on clinicians' use of validated scales, both in the space of tapering antidepressants [278, 26, 117, 177] and outside [65, 67]. These studies suggested that providers regularly ask patients to complete clinical self-reports, review them prior to patient interactions, and use them to inform treatment plans. While AT Annotator was created for the specific case of antidepressant tapering, I expect the design principles could be applied to other domains where standardized clinical measures are used, such as chronic pain [65] or cancer [67].

My prototype illustrated the following features:

1. *Standardized clinical self-report measures.* AT Annotator includes self-report questions that are widely used in the context of tapering antidepressants as well as part of standardized clinical measures, such as the PHQ-9 [177] and the DESS checklist [278] (Figure 4.1). PHQ-9 [177] is a widely used self-report measure designed to assess the severity of depressive symptoms. This scale is relevant to tapering antidepressants because it is useful for monitoring the relapse of depressive symptoms with dose changes. The DESS checklist [278] is a self-report measure specifically designed to monitor with-

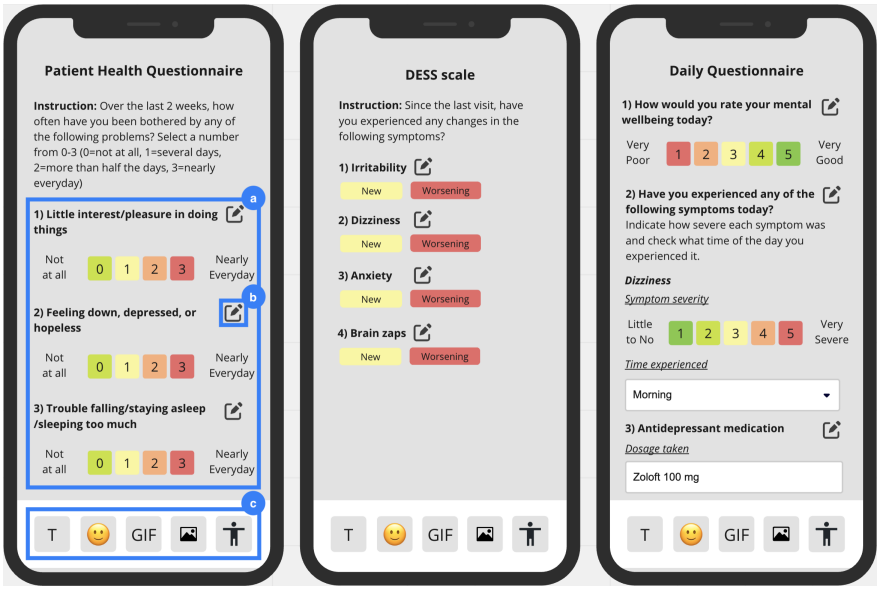

Figure 4.1: The AT Annotator low-fidelity prototype includes (a) Standardized self-report questions including the PHQ-9 [177] and the DESS checklist [278], (b) Features for adding annotations to responses for specific questions by clicking on the edit icon next to each question, (c) Different annotation methods in the bottom navigation bar, including free-text notes, emojis, animated GIFs, icons, and body parts.

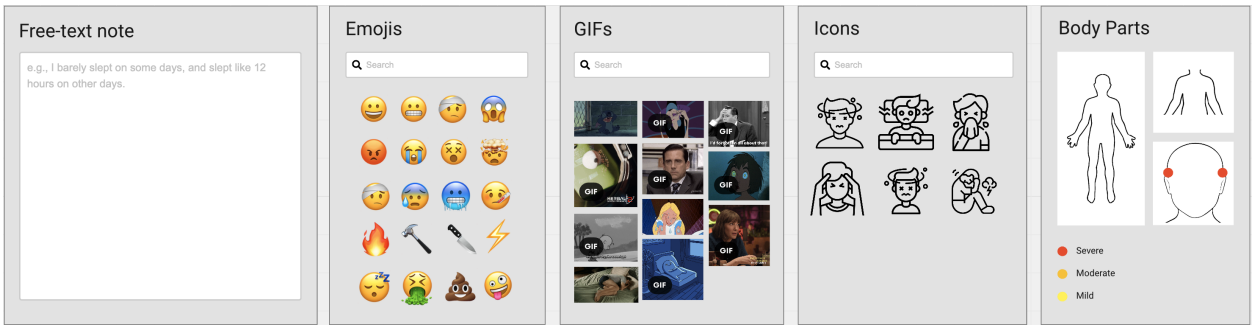

Figure 4.2: The AT Annotator low-fidelity prototype contains five types of annotation methods: free-text notes, emojis, animated GIFs, icons, and body parts. Users could reflect on the utility of adding these annotations to each question of their clinical survey responses. Icons and GIFs are from GIPHY.com and Flaticon.com.

drawal symptoms that patients might experience when discontinuing antidepressants. This scale is useful for monitoring what type of withdrawal symptoms, if any, patients are experiencing and how severe they are. I included these questionnaires as prior work suggested that providers often monitor withdrawal symptoms during the antidepressant tapering process [145]. To ensure their relevance, a board-certified psychiatrist validated these questions as representative of the most frequent and noteworthy symptoms in the context of tapering antidepressants. In addition, I added some questions as a daily questionnaire, including daily mental wellbeing, the severity of common withdrawal symptoms (e.g., dizziness), the time when the symptoms were experienced, and medication adherence, with the advice of the board-certified psychiatrist.

2. *Annotations to clinical self-report measures.* AT Annotator provides users with the ability to enhance their clinical survey responses using five annotation types, which include free-text notes, emojis, animated GIFs, icons, and body parts (Figure 4.2). Prior studies in HCI and Health Informatics have proposed that these forms could effectively support aspects of self-tracking and patient-provider communication. For example, free-form texts provide the flexibility for users to describe their thoughts and feelings [342, 206] and convey contextual information around standardized clinical measures [293]. Emojis and animated GIFs have been suggested as a way to better convey emotions within and outside clinical settings [305, 224, 306, 300, 22, 74, 140]. Similarly, icons have been suggested as a way to provide communication shortcuts on teletherapy platforms [341]. With free-text notes, users can type out anything that they want to convey to their doctor in free form. With emojis, GIFs, and icons, users can search for relevant images and add them to describe how they feel about particular symptoms or their overall health. With body parts, users can specify the affected area by clicking on a specific part of the body manikins (Figure 4.2). Users can add these annotations to their clinical survey responses by clicking on the edit icon next to each question (Figure 4.1 (b)) and selecting an annotation method in the bottom navigation

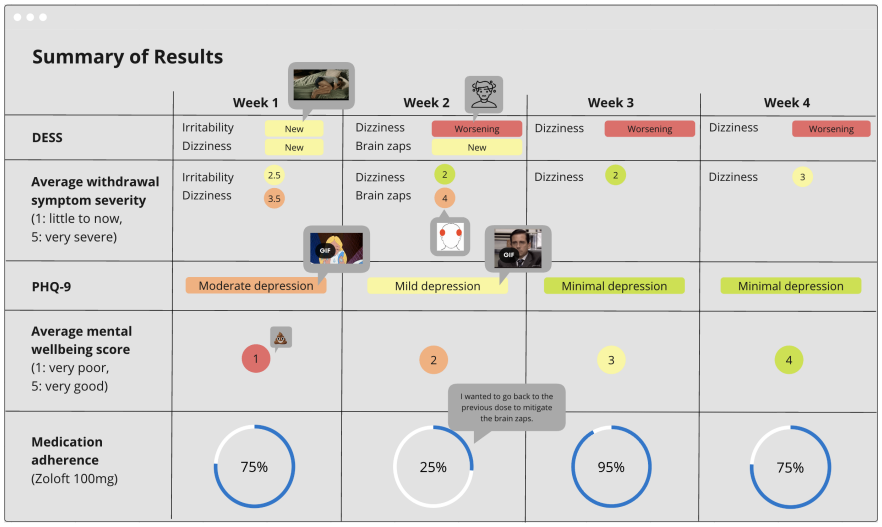

Figure 4.3: The AT Annotator prototype included a monthly symptom report that summarizes the annotations as well as responses to the clinical measures. The report provides (1) user responses to standardized questionnaires, such as DESS [278] and PHQ-9 [177], (2) average withdrawal symptom severity and mental well-being scores, (3) medication adherence, and (4) annotations added by users. The report utilizes color-coded indicators; green denotes positive trends in symptoms, while red indicates negative trends. Annotations appear in gray callouts attached to corresponding clinical survey responses. Icons and GIFs are from GIPHY.com and Flaticon.com.

bar (Figure 4.1 (c)). Given this study consisted of exploratory interviews with a low-fidelity prototype, the visual artifacts (e.g., emojis, GIFs, icons) in Figure 4.2 were mere examples that show the overall concept of annotations to clinical measures, rather than specific design outcomes to be evaluated. My goal for their representation was to aid participants in envisioning the use of annotation tools.

3. *Symptom summary report.* AT Annotator generates a monthly symptom report to provide a summary of the annotations as well as their responses to the clinical measures for facilitating patient-centered communication about their illness experiences during or between visits. These sorts of summary reports have been commonly proposed and created to present patient-generated data to clinicians [63, 289, 332, 333, 167, 262]. Some common features of these summary reports include numerical summaries and graphs to allow providers to get a quick overview of the data and decide whether they should take a close look at it if they find anything noteworthy, which were replicated

in AT Annotator. Figure 4.3 shows an example of a summary of annotations for an entire month. As prior work suggested that providers often decrease the medication dosage every month [145], I chose a month-long period to summarize patient-generated data through AT Annotator. The report provides (1) users' responses to standardized questionnaires, such as DESS [278] and PHQ-9 [177], (2) average withdrawal symptom severity and mental well-being scores, (3) medication adherence, and (4) annotations added by users. The report utilizes color-coded indicators; green denotes positive trends in symptoms, while red indicates negative trends. Annotations appear in gray callouts attached to corresponding clinical survey responses.

Interview Study

Our research team conducted all interviews remotely using video conferencing. I led each interview, lasting 50 to 60 minutes. This study used AT Annotator as a backdrop to understand the role of technology in communicating patient experiences to providers during the tapering process.

During the interviews, I first asked about participants' experiences of stopping their antidepressants, focusing on existing interactions with their providers and challenges with changes to their medication doses. Next, I showed an example of standardized clinical measures commonly used in the context of tapering antidepressants in the low-fidelity prototype, asking what participants felt about the topics covered by the survey and if there were other things that they may want to convey to their doctors. I then walked participants through different annotation methods (i.e., free-text notes, emojis, GIFs, icons, and body parts) in the prototype, asking whether and how different annotation methods might help patients convey their individual experiences to their providers. I emphasized to participants that all described annotations (e.g., the specific emojis and GIFs) shown in the prototype were mere examples of the overall types of annotations and encouraged them to think broadly about

how they might perceive the utility of that type of annotations. Lastly, I showed participants a prototype of a monthly report that provides a summary of the patient annotations as well as their responses to the clinical measures, asking their opinions about how such a report might impact communication with their providers.

Data Analysis

All interviews were video-recorded, automatically transcribed, and manually revised to correct errors afterward. I used inductive thematic analysis [40, 282] to qualitatively analyze interview transcripts without preconceived categories or theories. I open-coded the transcripts to identify patterns in the dataset. The full research team discussed and identified themes through multiple rounds of peer debriefing meetings. From this coding, we identified the main themes about patient goals being supported or left unmet by annotations, which I used to organize my findings. The final codebook contained six parent codes—including concerns about tapering, limitations of existing clinical communication practices, enriching clinical symptom measures, alleviating logging burden, establishing a professional communication environment, and considering the sensitivity and complexity of mental health contexts—and 20 child codes.

4.2.2 Participants

Our research team recruited participants through three channels: (1) ResearchMatch,[1] a non-profit program that helps connect people interested in clinical studies and researchers (17 participants), (2) recommendations of psychiatric providers at the medical center of our institution (two participants), and (3) word-of-mouth through our personal networks (one participant). Our eligibility criteria included individuals who are: (1) 18–65 years old, (2)

[1] https://www.researchmatch.org/

have been receiving treatment with an SSRI, SNRI, or other SRI antidepressants for at least three months, (3) are either interested in stopping their antidepressants or in the process of discontinuing them, and (4) do not have a comorbid DSM-5 diagnosis of schizophrenia, schizoaffective disorder, or a substance use disorder that is currently active. We settled on these criteria with the advice of the board-certified psychiatrist in order to (1) include patients who could consent and had relevant interests and experiences, (2) exclude patients who had a different serious mental health condition (e.g., schizophrenia) and were receiving psychiatric treatment that would benefit from monitoring additional clinical self-reports, as this could distract from our RQs, and (3) exclude patients who would require more active provider involvement than the typical, which would reduce the need for at-home clinical monitoring. We did not exclude patients who had other multiple chronic conditions. We compensated each participant with $30 cash or a gift card for a one-hour interview session.

Participants included 18 Caucasians, one Asian, and one Hispanic or Latino, consisting of four males and 16 females, ranging in age from 22 to 65 (median = 35.5) (Table 6.2). The median annual household income of our participants fell between $50,000 and $75,000. The majority of participants had been taking antidepressants for over four years (N=12). Four participants were in the process of discontinuing their antidepressants at the time of the study, twelve participants had prior experiences of attempting to discontinue their antidepressants, and four participants had no prior experience of discontinuing their antidepressants but were interested in doing so. Most participants (12) had taken antidepressants for over four years, while one participant had taken them for 3–4 years, five for 2–3 years, and two for 1–2 years. All our participants were based in the United States.

Table 4.1: Participant demographics, including gender, age, years on antidepressants, and tapering attempts

Alias	Years on antidepressants	Tapering attempts
P1 (F, 65)	≥ 4	Ongoing
P2 (M, 36)	3–4	3-4 years ago
P3 (F, 32)	≥ 4	In the last year
P4 (M, 57)	≥ 4	None
P5 (M, 25)	≥ 4	Ongoing
P6 (F, 53)	≥ 4	2-3 years ago
P7 (F, 22)	≥ 4	2-3 years ago
P8 (F, 23)	2–3	In the last year
P9 (M, 65)	≥ 4	None
P10 (F, 23)	2–3	Ongoing
P11 (F, 22)	≥ 4	More than 4 years ago
P12 (F, 26)	2–3	Ongoing
P13 (F, 38)	≥ 4	In the last year
P14 (F, 38)	≥ 4	2-3 years ago
P15 (F, 35)	≥ 4	More than 4 years ago
P16 (F, 48)	≥ 4	More than 4 years ago
P17 (F, 51)	1–2	None
P18 (F, 38)	2–3	In the last year
P19 (F, 29)	1–2	None
P20 (F, 27)	2–3	More than 4 years ago

4.2.3 Limitations

My study sample skews toward the experiences of Caucasian females in their 20s and 30s, potentially influencing the transferability of my findings [82]. According to the National Center for Health Statistics [43], the prevalence of antidepressant use was significantly higher among females (17.7%) than males (8.4%) during 2015-2018, indicating a nearly twofold difference. Thus, my study sample somewhat reflects the gender distribution of the overall population taking antidepressants. Males, older individuals, and people from diverse racial backgrounds may have distinct relationships with their healthcare providers and different perspectives on annotation methods than those identified in my study. For example, males may experience unique challenges in healthcare settings due to social expectations of masculinity, affecting their communication style and openness with healthcare providers [309]. Older individuals are more likely to have established long-term relationships with their healthcare providers and be more accustomed to traditional healthcare practices. Therefore, they might be less receptive to annotation methods that heavily rely on technology or less interested in adopting a novel tool that has not yet been integrated into traditional clinical practices.

In order to understand the utility of adapting provider-administered clinical scales, I narrowed my focus to the perspectives of patients and examined if such an approach could meet their needs. While this initial exploration is valuable, engaging with healthcare providers is likely to reveal potentially conflicting viewpoints. Given providers' scarce resources [332, 333, 61], I expect there are serious challenges to adopting this approach in clinical settings despite the opportunity for improved patient care. Future research on incorporating provider needs into the design of annotation tools would advance our understanding of how to present the annotations in ways providers could understand and accept, and whether providers would engage with them at all.

My study method was exploratory interviews with the aid of a low-fidelity prototype, AT

Annotator, to elicit patient perspectives on the potential use of annotation tools in their care processes. Therefore, the findings were grounded in participants' envisioned use of tools that incorporate such features instead of their real-world experiences. Although my low-fidelity prototype was useful for exploring how people envision using annotations for patient-centered communication, deploying a functional prototype in real-world settings would be necessary to examine how often patients would annotate, what sorts of annotations they would include, and how they would use the annotations to communicate with their providers in practice. Further, before deploying such a tool in real-world clinical settings, all contents and visual artifacts should be carefully designed to address specific patient needs and validated by providers who have domain expertise.

4.3 Findings

I present my findings on how annotations might support patient-centered communication in the context of tapering antidepressants, focusing on whether patients believed annotations to clinical self-reported measures could support their communication goals in clinical settings.

4.3.1 Goals Supported by Annotations

Overall, participants perceived that annotations to clinical measures could effectively support many of their goals in communicating with their providers, particularly by enriching clinical measures with individual symptom experiences and alleviating the cognitive and emotional burden of logging.

Enriching Clinical Measures with Individual Symptom Experiences

Participants perceived that annotations could enrich clinical measures with individual symptom experiences, including (1) symptoms not covered in standard surveys, (2) localized physical symptoms, (3) symptom severity, frequency, and duration, (4) personal circumstances impacting mental health, (5) complex mental health symptoms, and (6) the impact of medication changes on daily life.

Conveying symptoms that are not covered in standard clinical surveys: Several participants mentioned experiencing symptoms not covered in commonly used standardized clinical self-report measures during antidepressant tapering. For example, P2 had sexual side effects when taking antidepressants and desired to track his improvement as the medication dose decreased, but the clinical surveys he was asked to fill out during appointments lacked relevant questions. Participants also highlighted some aspects of psychiatric symptoms not covered in the clinical surveys. P13 was concerned about communicating her itching, which she had experienced during her previous episodes of medication nonadherence: *"My doctor says that itching can be a physical manifestation of depression. If I don't take it for a day or two by accident, I start getting itching and it's just horrible. I wish the survey covered those physical aspects."* P15 similarly wished that there was a way for her to communicate night terrors with her provider: *"There are things that are not part of the survey that I would talk to my doctor about. Sometimes I experience night terrors, and there's never any kind of question about that."* Therefore, she appreciated annotations, specifically free text, as a means to communicate symptoms not covered by clinical measures.

Describing localized physical symptoms: Participants thought annotations with body parts would help them describe localized physical symptoms. P2 believed that visual annotations that indicate specific body parts would help him convey withdrawal symptoms more

accurately: *"Sometimes you're having an issue in a certain part of the body, and you have a hard time explaining where an issue is on the body. For example, when I started tapering off my medication this time, I felt some pressure on my head. It would have been helpful to be able to say that these are the locations where I'm feeling this pressure and be able to annotate that with the graphic."* P20 further emphasized the value of visual annotations, particularly in telemedicine settings: *"You could annotate where the headache is and pinpoint exactly where that's happening, especially for telemedicine since you're not physically with your doctor to tell them where things are happening."*

Conveying symptom severity, frequency, and duration: Participants found clinical measures inadequate in accurately conveying details about the severity, frequency, and duration of symptoms during the tapering process, viewing annotations as an effective means to describe such details. P20 expressed concerns about solely relying on clinical measures to summarize her psychiatric symptoms between visits because it would not accurately reflect fluctuations in symptoms: *"I don't want them to just get a view of the week and be like, okay, she was three out of five on average. It's not a fair assessment for that week. I might have had three really great days and four terrible days. The only thing that they [providers] are seeing would be the average. That gives me anxiety."* Participants thus perceived the benefits of using free-text annotations in accurately conveying details about their psychiatric symptoms: *"If I felt like the survey options didn't fully encapsulate what my experience was, I would use the notes to clarify. Let's say I select, 'I felt depressed nearly every day,' but in the free text note, I could specify, 'Yes, I felt it every day, but it didn't last very long. Most days, it lasted a short period of time.' So I could use that to give some more context to my answer. (P12)"*

Describing personal circumstances impacting mental health: Participants further desired to convey personal circumstances that may contribute to their deteriorating mental

health during the taper. However, standard clinical surveys did not allow them to communicate recent major life events that likely impacted their mental health, suggesting the potential benefits of annotations in facilitating such communication. P1 said: *"One thing [that I would add] is any life changes. Found out my son is an addict, got divorced, mother died. Any life changes can play into this."* They thus envisioned that annotations could help convey significant life events as relevant contexts to their clinical survey responses: *"I think it [annotations] adds a lot more nuances to the results. I don't think that I could get my entire message across without being able to say like, yes, I had moderate depression during week one. But my cat just died. That's what brought down my mood. Maybe not the medication. (P11)"*

Participants further emphasized the need to convey daily events that may have influenced their mental health during the taper. P18 wished to convey personal circumstances surrounding her mental health flare-ups: *"Contexts of what was happening around me when the symptoms were coming up is the part that I would be trying to get across most. If both of my kids were sick and it was a particularly stressful day, I want to contextualize the symptoms more."* P14 expressed concerns about reporting her current mental health status solely through standard clinical surveys because it could misguide her provider in making care decisions: *"If I hit three [nearly every day] for little interest/pleasure in doing things, they [providers] will only see three. But there might have been some things behind it, like suffering from a UTI [urinary tract infection]. If I cannot explain why I picked three, I worry I'll be seen as my meds aren't working, I'm a horrible human being, and nothing helps. For me, not being able to explain why I chose three increases anxiety."* Participants thus highlighted that annotations to clinical surveys could help them convey potential factors influencing their mental health, aiding providers in making informed decisions about medication changes.

Articulating complex mental health symptoms: Participants perceived that annotations could help them articulate complex mental health symptoms during the taper. They often mentioned experiencing challenges in explaining complex psychiatric symptoms to their providers, wishing to have alternative ways to express their mental health status. P4 said: *"There are always complications in trying to explain to someone how you feel mentally, especially if that's someone you rely on for medical advice."* P20 similarly mentioned challenges in describing her mental health status: *"I guess the hardest part is to describe how you're feeling. When I think of how I'm feeling about the medication changes, 'good' doesn't necessarily cover everything. It's really difficult to describe."*

Some participants believed free-text annotations would add nuance to clinical survey responses. P12 said: *"When you're just looking at the survey answers without any of those annotations, it feels kind of robotic, like you're trying to squeeze your feelings into boxes. I think free-text notes could provide nuances during the [tapering] process."* Other participants perceived that visual annotations would better convey complex mental health states: *"I could see myself using GIFs to portray some feelings, just because it is sometimes easier to have visuals rather than put it into words. Using a GIF could be helpful when you don't have the words for certain things. (P20)"* Participants felt that conveying personal aspects of their mental health through annotations would help their providers understand them better. P14 noted: *"We want someone to understand how we're feeling. Emojis or GIFs could help my doctor be like, okay, that's how she's feeling right now. I get that. I physically see it."* P19 added that annotations might lead her provider to ask more personalized questions: *"It would be more specific to my personal needs than just the general questions that they typically ask. So my doctor will know what would be a better question when they see these versus when I talk to them."*

Conveying the impact of medication changes on daily life: Participants perceived annotations as a valuable tool for describing the impact of medication changes on their daily lives, assisting providers in making informed tapering decisions. P3 thought tolerability of her withdrawal symptoms would be important information for her provider: *"Is your anxiety tolerable with decreasing your dose? If you're dizzy, is it tolerable dizzy? Can you still function? Can you get yourself to the bathroom? Kind of wondering how my doctor would change my dose or plan based on this scale."* Since this information was not covered by clinical measures, P3 wanted to use icons for communicating the tolerability of withdrawal symptoms: *"If I were to take the survey and I'm experiencing worsening symptoms, I would definitely use the icons. This icon probably means that I can still function and things are tolerable in daily life."* P7 similarly thought free-text annotations would help convey tolerability of withdrawal symptoms during the taper: *"If I had this app while I was tapering off my meds, I probably would have put a free-text note in there like, 'I'm so dizzy that I'm unable to work."'*

Participants sought to highlight various dimensions of their daily lives influenced by medication changes, such as work, social activities, or self-care. P4 desired to figure out how medication changes might impact his psychiatric symptoms in different social settings, which can be included in free-text annotations: *"I wonder how I'm going to feel, doing things outside my home, at work, social activities. So being able to address some of that would be helpful for the provider and myself."* P8 wanted to specify what aspects of her daily life have been impacted by medication changes through free-text annotations: *"I feel like these [questions in the clinical survey] are just really broad. Some days, it was not an issue at all for me to take care of myself, but I couldn't take care of social aspects of my life, and then other days, I couldn't even get in the shower. So there needs to be an area to write specifically what that feeling was like."*

Alleviating Cognitive and Emotional Burden of Logging

Participants recognized the potential of annotations in helping them engage in more patient-centered communication by alleviating the burden of logging. They envisioned that annotations would reduce the cognitive burden by highlighting primary concerns and helping recall symptoms, as well as the emotional burden associated with logging challenging mental health experiences.

Empowering the expression of primary concerns: Annotations were seen as a valuable means for participants to offload the cognitive burden of conveying their primary concerns with their providers. P7 wished to use free-text annotations to ensure that her concerns about withdrawal symptoms receive attention from her provider: *"If I had this app while I was tapering off my meds, I would have put a free-text note saying, 'I'm so dizzy that I'm unable to work' to make it clear to my provider that this is something that I'm really struggling with, and I need some changes to continue on my daily tasks."* Visual annotations were also considered effective in conveying primary concerns. P7 believed these visual elements could highlight areas of significant impact: *"Amongst all of this data, it [an emoji] really highlights what I focused on most. I think the fact that I put an emoji there would show my psychiatrist that this issue is affecting me more heavily, whereas my average withdrawal symptom severity doesn't have anything like that."*

Enhancing recollection of symptom experiences: Participants found annotations valuable for reducing the cognitive burden of recalling their symptom experiences from the lengthy intervals between doctor's visits. Given that most participants saw their providers only once or twice a year, they often felt that their progress was not adequately monitored during the tapering process. P7 expressed a desire for more frequent and in-depth discussions about medications: *"I wish that I had more frequent visits with him [provider], just to*

go over the medication a lot more in-depth. But we only talk three to five minutes at a time, and it's basically just so he can refill my prescription. I didn't really feel like my progress was being monitored." Annotations were thus seen as a potential way to bridge this gap by providing helpful references. P8 mentioned that free-text annotations would help her remember specific situations that influenced her mood: *"I'm not going to remember a few days later what I was specifically irritated or upset about. So this is going to be helpful."* P5 perceived that free-text annotations would help him not leave out important things about his symptom experiences during the doctor's visits: *"I always kind of worry, 'What if I left something out? What if I forget to say something in a session?' Having a report like this leaves what-if questions on the table because I've already done the reporting. It gives me more confidence in my general review of my symptoms."* P14 emphasized that visual annotations could trigger recollection of mental health symptoms during discussions with providers: *"If I add that [emoji], my provider can be like, 'Do you remember feeling like this?' It might be like a trigger to recall how I was feeling at that moment."*

Alleviating the emotional burden of logging: Visual annotations were perceived as a potentially powerful tool for alleviating the emotional burden associated with logging, allowing them to have more control over how they monitor their symptoms. Participants envisioned that when experiencing poor mental health, visual annotations would provide a more accessible and expressive means of conveying emotions. Some participants noted they might experience a lack of mental energy to write down how they were feeling, but selecting visual annotations that match their feelings would be much easier. P14 explained: *"If I'm in a bad state, I might not have the energy to fully write down how I'm feeling. I feel like an emoji or a GIF is gonna be a lot easier to relay at that time."* P16 similarly emphasized the ease of expressing emotions with visuals: *"Sometimes you don't feel like you can express it in words. So I think those [GIFs and icons] would be useful. I could look at some of these GIFs and choose Alice in Wonderland to express that feeling of getting overwhelmed by*

emotion or bewildered." P19 also highlighted the value of visual annotations during times when verbal communication is difficult: *"Sometimes I don't feel like talking or even going to my appointment. If it's one of those days, I think that [visual annotations] would be a good way."*

Participants noted that visual annotations could help them document challenging mental health symptoms, which could be used later to convey those experiences to their providers. P18 explained that certain mental health experiences were overwhelming to verbalize or write about: *"Some of the stuff that is scarier to experience is also hard to verbalize. Writing it out makes it feel a lot more real and frightening."* The use of visual annotations alleviated some of this burden, allowing individuals to make a note for themselves without explicitly stating the details. P18 added: *"I do think that the icons are attractive. It is sort of a way that I can make a note for myself without having to deal with explicitly saying what it was, and then I can detail it later. If I'm verbalizing it at some time distance from the event, I think I would have an easier time describing it."*

4.3.2 Unmet Goals by Annotations

While participants perceived that annotations would help achieve the aforementioned communication goals in clinical settings, they felt annotations would fall short in (1) establishing a professional environment and (2) considering the sensitivity and complexity of mental health contexts.

Establishing a Professional Environment in Clinical Settings

Participants felt visual annotations might interrupt the professional relationship with providers by being too casual or ambiguous.

Too casual and light: Some participants felt that visual annotations often used in everyday interpersonal communication would not mesh with their professional communication environment in clinical settings because they might be too casual and light. P2 explained his perceptions about emojis and GIFs: *"I feel like those silly GIFs are for adding comedy and levity. But depression and anxiety are not something that should be taken lightly. I feel like it [annotations] adds a sense of 'It's not that big of a deal' and makes it cute and funny, so it just feels inappropriate."* Several participants viewed annotations with emojis and GIFs as *"not very sincere"* (P16), *"superficial and flippant"* (P17), and for *"lightening up"* (P18), which led them to think those would not be appropriate for mental health contexts. P8 felt that such visual annotations might be particularly inappropriate if she was in poor mental health: *"If I was in an okay mood, I would use them [emojis, GIFs, icons]. Sometimes, I'd like to joke about things, so I can see how it can be a funny, humorous way to cope. But if I'm in such a bad mood, an emoji or a little picture is not gonna sum that up."*

Participants thus thought that, unlike in casual contexts, visual annotations might not be appropriate for communication with their providers. P18 said: *"I send those [emojis and GIFs] to try to lighten up a text to a friend. But this [visits for antidepressant management] doesn't really feel like a lighten-it-up scenario."* P5 similarly highlighted the inappropriateness of visual annotations in clinical settings: *"Emojis feel like interpersonal communication. It doesn't generally mesh with the way that I communicate with a mental health professional."* P19 perceived that visual annotations were for casual interactions on social media: *"It's something that we'd use more on Facebook, and maybe that's how the doctor is going to see it, very casual."* They thus felt uncomfortable about using visual annotations for communication with their providers: *"Those [visual annotations] are more casual, and I would feel a little weird or uncomfortable about using a GIF in a note to a provider. (P15)"*

Too ambiguous and subjective: Some participants thought visual annotations might be too ambiguous and subjective, making it difficult for their providers to understand their intent. P11 was concerned about using visual annotations because the personal meanings may not be conveyed to her provider: *"I don't think that those [emojis, icons, GIFs] would be helpful for providers. I know what I meant when I used these, but for a provider, it might not mean anything. They might be like, why would they add the Alice in Wonderland thing?"* P12 was similarly worried about her providers' potential misunderstanding of her visual annotations: *"I would probably be less likely to use the GIFs with the doctor. Let's say I put a GIF. They might interpret it differently from how I do."*

Considering the Sensitivity and Complexity of Mental Health Contexts

Participants pointed out that annotations would fall short of considering the sensitivity and complexity of mental health contexts by requiring significant mental and emotional burdens and not providing the flexibility needed to convey complicated mental health experiences.

Mental and emotional burdens: Participants thought annotations might involve significant mental and emotional burdens for people when going through difficult mental health experiences. P1 stated she would not have the mental energy to add a free-text annotation to clinical surveys when in poor mental health: *"In the past when I was experiencing deep depression, I don't know if I could properly articulate it. I can point to something real quick to convey how I'm feeling, but I don't know whether I could add a note when things aren't going well. For me, when depressed, 'Go away. Don't talk to me.' I wanna put the covers over my head."* P3 noted that visual annotations might require her too much mental energy when depressed: *"I feel like for someone who's depressed, it [a GIF] is like brain overload. Like, 'God, damn it! I gotta go through all of that?' I'm exhausted just trying to choose which one."* Participants also mentioned that they would feel more emotional burden when

using personal language to log their mental health experiences through annotations. P18 said: *"I try to make things more objective and less emotional when I'm talking to my doctor, creating some distance between myself and my symptoms. When I started the medication, it was much easier to say, 'I was suicidal' than to talk about it in a more personal and emotional way. I am able to talk about more things if I'm able to put them in that [medical] language."*

Inflexibility to convey complicated mental health experiences: Participants perceived that visual annotations would not be flexible enough to support them in conveying complicated mental health status. P5 thought visual annotations would not support as flexible communication as written words would do: *"I feel like emojis and GIFs are set in their nature. They're not necessarily as flexible as the written word would be to me. They don't cover the breadth of symptoms or emotions."* P7 similarly felt: *"The emojis and GIFs can be used to convey emotion, but I think, with something so complicated, I would do better explaining how I really feel with written words. If I were to just pick one of the icons to represent how I'm feeling, I don't think that would be sufficient."*

4.4 Discussion

My findings suggest that patients see annotations as a valuable tool for communicating their illness experiences, offering resources to convey a more authentic representation, and influencing the direction of conversations during clinical visits. My findings further show that patients consider various factors when selecting an annotation form, particularly how their providers will receive and interpret it. I recommend providing customization support to cater to patients' communication needs, particularly considering provider perspectives. Lastly, it is crucial to consider the clinical practicality when designing annotation tools. I

highlight opportunities to enhance the utility of digital symptom measure annotations in clinical settings.

4.4.1 Annotations For Facilitating Patient-Centered Communication

Through this study, I learned that patients often entered clinical visits with a clear idea about what they wanted to convey to their providers yet lacked the means to guide conversations around their complex symptom experiences when the interactions are centered around standardized patient self-reports. Aligned with prior work [72, 62, 63, 122, 262, 12, 13], this finding highlights the gap between patients' desires to communicate their lived experiences and the limitations of current patient-provider communication practices. As the standardized self-report measures focus on the presence or absence of common symptoms, they often miss out on how the patient perceives and experiences those symptoms. My findings suggest that annotations to clinical measures could provide patients with additional resources to describe their experiences in greater detail, potentially bridging communication gaps with providers. Our participants valued the space that annotations provide for documenting and elaborating on their experiences, envisioning that annotations could convey a more authentic representation of their illness experiences, such as symptom fluctuations, beyond being distilled down to *"just a number"* on a clinical scale. My findings thus suggest that, unlike digital passive data collection methods, which are often held up as an alternative to self-report scales, annotations have the potential to offer patients a more authentic portrayal of patients' lived experiences between visits while remaining grounded in standardized clinical measures.

My findings further indicate the potential value of annotations in enhancing patient visibility in clinical settings. Our participants valued annotations for allowing them to convey *"why"*

behind their clinical survey responses, such as their personal circumstances impacting mental health, and *how* the medication changes impacted their daily lives. Prior work suggested that solely relying on standardized measures could hinder patients' sense of agency in managing their illness in everyday life, pointing to the need to provide mechanisms for patients to articulate and communicate personally meaningful aspects of their illness experiences [18, 17, 72, 262]. My findings further highlight that the act of annotating clinical scales could provide patients with a sense of validation within the existing medical infrastructure, as their lived experiences are contextualized within a form that providers already engage with. This technique can empower patients through concrete means to have their voices heard and influence the direction of their communication with providers.

4.4.2 Considerations of the Form of Patient-Generated Data in Clinical Settings

My study demonstrates that patients had particular forms of annotations that they felt comfortable presenting to their providers. I observed that patients considered various factors when selecting an annotation form, such as their mental health status, communication styles of themselves and their providers (e.g., level of formality), and the dynamics of patient-provider relationships (e.g., level of trust and bonding). For example, while some participants considered GIFs or emojis as an effective way of communicating their health experiences, others felt that such an approach lacked the necessary clarity for clinical settings. Further, while some participants felt GIFs and emojis could provide effective communication tools in clinical settings, others felt those forms were inappropriate for clinical contexts. These findings align with prior work, which highlighted patients' conflicting views of emojis as both lacking the seriousness needed for their challenging illness journey and as a valuable tool for reducing cognitive barriers in representing their personal health data by using symbols frequently used in everyday digital communication [300].

Factors such as patients' and providers' personalities, cultural or generational differences in experiences with different mediums, and perceived norms in patient-provider relationships may have influenced the contradictory viewpoints on annotation methods. In my study, participants who described themselves as more familiar with social media or messaging apps or were younger tended to prefer visual annotations, such as GIFs or emojis, over other forms like free-text notes, which aligns with previous research findings on the general use of GIFs outside health contexts [140, 203]. In contrast, patients who preferred more formal communication styles, were relatively less familiar with social media or messaging apps, or were older tended to be more reluctant to use GIFs or emojis and preferred data forms more common in clinical discussions, such as free-text notes and body parts. Preferences for different data forms are likely complex. For instance, some people might prefer not to use GIFs or emojis even if they are casual with their doctors because they do not regularly use these mediums in other communication in their daily lives.

I found it noteworthy that patients not only consider how helpful the data form will be for conveying their experiences but also how providers will receive and interpret it. While prior work on self-tracking has primarily focused on flexibility in data types to track more personally meaningful data [5, 17, 122, 121], my findings highlight that social settings warrant deeper consideration of the form and how other parties might interpret it. Studies on personal informatics in social media demonstrated that people often consider the form in which data are shared and how it would be perceived in social settings [92, 325]. My findings further show that such consideration could be particularly critical in clinical contexts where care decisions could be made based on the tracked data. When designing a tool for supporting patient-centered communication through patient-generated data, it is crucial to recognize that the interpretation of data forms by other parties, particularly healthcare providers, can influence the effectiveness and reception of different data forms.

Future annotation tools could help people formally indicate their communication preferences

and needs, particularly in clinical settings. Beyond presenting different styles of annotation forms, tools could allow people to explicitly specify communication preferences as part of aligning their annotations with their provider perspectives. Systems could then support customizing annotation strategies through the parameters indicating patients' communication needs. In particular, prior work highlighted that social contexts should be a primary consideration when designing for GIF and emoji use, suggesting opportunities to integrate GIF or emoji recommendations into keyboards personalized to each communication partner based on the chat log analysis [140, 163]. Using a similar approach, future health monitoring tools could provide recommendations for contents and data forms for their annotations based on patient input about their patient-provider dynamics. In addition, our study shows that patients often worry about the flexibility of GIFs and emojis to adequately describe their complicated mental health status due to the limited options available online. One possible approach to address this concern is leveraging generative AI features, such as DALL-E [248], that can generate images from natural language descriptions to allow patients to quickly create visual artifacts and iterate on the outcome until it meets their articulation needs. Such a feature could further cater to aspects of patients' relationships with their providers (e.g., the level of formality).

However, given the sensitivity of patient-tracked data, privacy risk should be considered when implementing large language models and diffusion-infused features for annotation tools. For example, these tools likely need to be designed in line with medical data regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union. Incorporating ways to exclude or anonymize personally identifiable information from patient inputs could also help minimize privacy risks in using such features. More work is needed to explore the potential impact of health monitoring technology that supports the customization of annotations to clinical measures.

4.4.3 Balancing Patient-Centric Communication with Clinical Practicality

My study highlights the potential of annotations to clinical symptom measures to enhance patient-centered communication. Grounded in clinical symptom measures, patients felt that annotations could help them convey their lived experiences to their providers while maintaining clinical relevance. Nevertheless, it is crucial to consider the practical aspects of implementing a tool for supporting annotations in real-world clinical contexts. Prior work suggested that patient-generated, free-form data may not always be well-received by providers, particularly because patients tend to bring in more data than what providers can realistically review within the time constraints of clinical visits [284, 120, 289, 290, 299, 351, 167, 333]. In my study, some participants also mentioned that their typical visits were as short as under five minutes, which suggests it might be infeasible or inadequate to include additional artifacts or processes. Further, participants were worried that their annotations could come across as overwhelming to their providers, making them skeptical about whether their providers would actually review their annotations in practice. If annotations are not reviewed or discussed, patients could potentially feel that their sense of agency is undermined even more than if they only completed a standard self-report scale. Therefore, in the design of digital symptom measure tools, it is critical to consider ways to balance patient-centric communication with the practical demands of clinical infrastructures.

To enhance the utility of annotations to self-report measures in clinical infrastructures, a digital symptom measure annotation tool could aid patients by offering guidance about the types of information that providers might find useful to include in an annotation and explaining why this information could be helpful. After patients create some annotations, tools could assist patients in reflecting on what annotations could be most informational to a provider, such as data that are indicative of worsening symptoms. In my study, participants often felt that the annotations could be helpful not only for provider review but also for self-

reflection, which suggests the potential value of annotations for both purposes. In light of this finding, annotation tools could suggest keeping annotations that are less clinically relevant as personal journals, such as daily mood fluctuations, even if the annotations do not show up for provider review. Further, tools could automatically generate a bullet-point summary from these annotations, providing patients with tailored recommendations on critical points to discuss in their next clinical visit. Such approaches could help ensure that the annotations presented and discussed at clinical visits are directly pertinent to the patient's ongoing care, making patient-driven self-monitoring data more practical within the constraints of clinical infrastructures.

4.5 Conclusion

Through interviews with 20 patients who were either interested in stopping their antidepressants or in the process of discontinuing them, I found that annotations to clinical measures were seen as a useful tool to enrich clinical symptom measures with individual illness experiences and alleviate the cognitive and emotional burden of logging. However, patients also thought annotations might interrupt the professional relationship with their providers and overlook the sensitivity and complexity of mental health contexts. Based on the findings, I suggest opportunities for annotations to promote patient-centered communication. I further propose incorporating customization support for patients' communication needs around the form of patient-generated data. Lastly, I highlight the need to develop ways to enhance the clinical practicality of annotations.

Chapter 5

Supporting Stakeholder Practices of Using AI Chatbots for Public Health Monitoring

AI technologies have the potential to support large-scale public health monitoring efforts. In particular, recent large language models (LLMs) enable *open-domain* dialogs, supporting free-form conversations in open-ended topics aimed at providing empathy (e.g., [350, 127, 327]) [125]. These systems show potential for reaching underserved populations and providing emotional support to those going through difficult health experiences [205]. However, few studies have explored how LLM-based chatbots can be leveraged in population-level health interventions in real-world settings, limiting understanding of their practical impact on addressing public health needs.

To understand the practical impact of AI technologies on supporting multi-stakeholder practices within public health infrastructures, this chapter explores the case of **CareCall**, an AI chatbot that aims to help support socially isolated individuals via check-up phone calls

as a public health intervention. As an open-domain chatbot, CareCall both collects data about the individuals' general health and serves as a conversational partner to mitigate their loneliness by generating human-like questions and answers on the fly. As of August 2022, when the study was conducted, CareCall was being deployed to 20 municipalities in South Korea with the aim of monitoring socially isolated individuals, including middle-aged and older adults living alone. Being a rare example of an LLM-driven chatbot deployed in a real-world setting in public health contexts, CareCall is a useful case for understanding the role of LLM-driven chatbots in public health monitoring.

Through focus group observations and interviews with 34 people from three stakeholder groups—including five CareCall users[1], five frontline workers who monitored users' conversation logs, and ten developers who designed and implemented the system as well as communicated with local government agencies, I sought to answer the following research question:

RQ3: How might technology support stakeholder practices within public health infrastructures?

Toward my thesis claim T2, this study shows the benefits and challenges in leveraging LLM-driven chatbots within public health infrastructures. Frontline workers valued that the LLM-driven chatbot helped them gain a holistic understanding of care recipients through open-ended conversations while offloading their workload. The users perceived that the open-ended nature of the dialog helped mitigate loneliness by asking caring questions about their health and engaging beyond health topics. However, stakeholders often had different goals and expectations. While public agencies desired to incorporate specific health questions and customize conversations to different target groups, the developers struggled to accommodate those needs due to the uncertainty in control and the resource-intensive nature of customizing

[1]In this chapter, I refer to socially isolated individuals who receive regular check-ins through CareCall as *CareCall users*. These individuals are also referred to as *care recipients* throughout my dissertation.

LLM-based chatbots. The open-ended nature of conversations also led the users to expect the system to support social services out of its scope, increasing the burden on frontline workers. Further, the users perceived the chatbot lacking long-term memory as impersonal because it lacked follow-ups on past conversations around personal health, which led to challenges in providing emotional support.

Based on the findings, I discuss opportunities for improving LLM-driven chatbots to provide greater emotional support. I also suggest the need for designing resources and processes that help different stakeholders negotiate the tradeoffs between open-domain and task-oriented chatbots. Lastly, I discuss the need and challenges in scaling LLM-driven chatbots to support diverse public health needs.

This project was published at CHI 2023 [141] and won the Best Paper Award (Top 1% of submissions) with co-authors Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. Daniel A. Epstein and Young-Ho Kim served in supervisory roles, providing guidance and feedback throughout the research process, and Hyunhoon Jung assisted with participant recruitment. I proposed the initial study ideas, developed interview protocols, and led the interviews, data analysis, and paper writing. This project was done as part of my internship at NAVER AI Lab.

5.1 Background and Related Work

5.1.1 Motivation, Design, and Deployment of CareCall

CareCall is a conversational AI system designed for socially isolated individuals in South Korea [49]. Motivated by the recent Act on the Prevention and Lonely Death in South Korea [175], CareCall is aimed to provide the target individuals with emotional support and

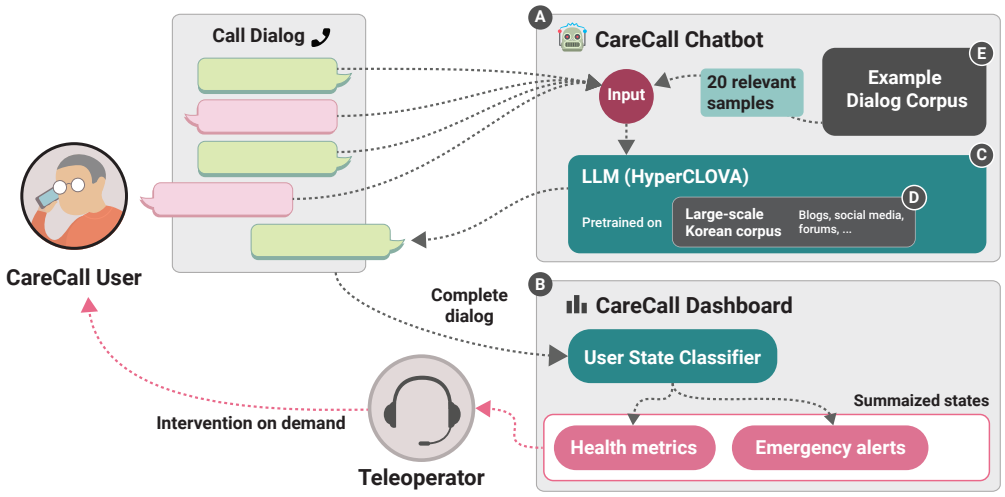

Figure 5.1: System architecture of CareCall, describing (A) CareCall chatbot conversing with users and (B) CareCall dashboard used by teleoperators (frontline workers).

regularly check their health status.

Figure 5.1 describes a brief overview of the system architecture and the interaction between the two stakeholders of CareCall. The CareCall chatbot ((A) in Figure 5.1) regularly (e.g., once or twice a week) calls the target individuals and leads an open-ended conversation about daily life for about 2–3 minutes, in a female voice. After each call, the dashboard ((B) in Figure 5.1) automatically extracts (1) five health metrics, including meals, sleep, general health, going out, and exercise, as one of three statuses (Positive/Negative/Unknown), and (2) emergency alerts (e.g., dizziness, chest pain, high fever, difficulty in breathing) from the dialogs using user state detection classifiers. The summary of each user's status is displayed on a web dashboard for frontline workers (referred to as teleoperators in Figure 5.1). On the dashboard, frontline workers can access the call recordings as well as the five health metrics and emergency alerts of the care recipients whom they are in charge of.

The CareCall chatbot was designed as an open-ended dialog system powered by an LLM ((C) in Figure 5.1) called HyperCLOVA [162] which has 82B parameters trained on a Korean corpus of 561.8B tokens ((D) in Figure 5.1). The training corpus includes blog posts, online forums, news articles, comments, and online Q&As [162]. At each conversation turn, the

chatbot's response is generated by putting 20 relevant example dialogs along with the current dialog history to the LLM. These example dialogs are sampled on the fly from a large-scale dialog corpus[2] ((E) in Figure 5.1) generated with a data augmentation technique, where a machine learning model generates synthetic dialogs from a small set of human-written dialogs, and then crowd workers flag and fix errors in the synthetic dataset [21].

Since the example dialogs in an input significantly affect the flow of the conversation [45], the example dialog corpus was inspected to ensure consistency with a specific **agent persona**—*an AI chatbot that calls the user* in a polite and respectful tone and manner—and **system policies** such as the agent should not accept the user's commands that are unsupported by the system (e.g., "*I'll play a song.*" or "*I'll call your daughter.*"). Such a policy was imposed because CareCall's conversation was over a phone call, and it did not support many of the task-oriented dialogs of the chatbots in common smart speakers like Alexa or Siri. (Bae et al. [21] provides a more detailed description of the supported dialogs.) As an additional effort to better steer conversations, the underlying LLM was also *fine-tuned* (see Section 5.1.3) on the *undesirable* phrases that violated the persona (e.g., the agent acted as if it was a child of the user or spoke impolitely) or system policies in a way which decreased the probability of them being selected [21, 331].

CareCall first started to roll out in Haeundae-gu in Busan in November 2021 [49]. As of August 2022, when the study was conducted, CareCall was being deployed to 20 municipalities in South Korea. In this study, I specifically focused on Seoul, the capital of South Korea, where CareCall was deployed to around 300 individuals as a pilot project from June 2022 to August 2022. Each municipality's government had slightly different criteria for the target users in terms of the age group or chronic health conditions, though sharing the overarching characteristic of social isolation. CareCall was deployed to older adults living alone in most of the municipalities, but in a few cases, it was deployed to middle-aged adults, individuals with

[2]A subset of the corpus is available at `https://github.com/naver-ai/carecall-corpus`.

early dementia, or healthy older adults. In Seoul, where this study is focused, CareCall was deployed to middle-aged (40s to 60s) adults who were living alone and were predominantly (87%) recipients of the National Basic Livelihood Security (below 50% of median household income). The deployment with such a population was motivated by the highest proportion of solitary deaths among all age groups in Seoul [346]. The CareCall pilot project participants in Seoul were recommended by public officers who were providing social services to these individuals. Most of the CareCall project participants in Seoul have been receiving regular check-up calls from different types of public officials, including social welfare officers, public health officers, and emergency response officers. Introduction of CareCall did not replace their existing check-up calls from humans but rather increased the frequency of check-up calls, partially due to the short-term nature of the pilot project. The pilot deployment of CareCall across all municipalities obtained participants' informed consent prior to their enrollment. Note that the scope of this study was conducting interviews and observations of different stakeholders related to the CareCall pilot project; thus, the development and the pilot deployment of CareCall were outside the scope of this study.

Each municipality's government handled the monitoring tasks of CareCall in different ways. For example, some governments had their social welfare officers in charge of frontline monitoring, as an aspect of their social work, while others hired part-time workers for the frontline monitoring tasks specifically for the CareCall pilot project. The government of Seoul hired 14 part-time social workers for the frontline monitoring tasks for the CareCall pilot project through a social enterprise that employs retired individuals over the age of 55 (referred to as *frontline workers* hereinafter for brevity). In Seoul, the frontline workers' protocols required them to monitor the call recordings for negative health signals (e.g., skipping meals, poor sleep) or emergency alerts on the dashboard. If they found any health issues from the call recordings, they were asked to share them with their team and reach out to the person to check if everything was okay. If they noticed anything noteworthy from the manual check-up calls, they were asked to write a report to escalate to those who provide social services in

their municipalities alongside the deployment. Other municipalities used similar protocols for the frontline monitoring tasks of CareCall, though public officials' workflows slightly differed because they were often in a position to directly connect to social or healthcare services.

5.1.2 Caregiving Technology for Individuals Living Alone

Individuals living alone tend to be vulnerable to various health concerns, particularly with aging [232]. There is a greater risk of social isolation and loneliness when living alone, which is closely linked to negative health outcomes such as dementia, depression, heart disease, and stroke [86]. In addition, a lack of social contacts limits one's ability to receive help in emergency situations [172]. Research on caregiving technologies has aimed to support these individuals (e.g., [274, 323, 182, 68, 231, 280]). One subset of these systems is often referred to as *telecare* systems, which seek to mediate care among individuals living alone, formal and informal caregivers, and emergency services [182, 274]. Another subset of caregiving technologies—including CareNet [68], Digital Family Portraits [231, 280], and SHel [323]—have aimed to support family members or other care network members in maintaining awareness of the older adults' daily activities through environmental sensors and ambient displays [231, 280, 323]. Field studies have suggested that such systems can alleviate the loneliness of individuals living alone and provide peace of mind for their informal caregivers [68, 280].

A core concern is that existing technologies have predominantly targeted individuals who have readily accessible social contacts, such as informal caregivers [231, 280, 68, 323]. However, studies have pointed out that compared to high socioeconomic status (SES) individuals, low-SES individuals living alone tend to have fewer social contacts that they can reach out to in emergency situations [16, 334], reflecting important differences in how to approach design-

ing technology to support this more vulnerable population [307]. Thus, many of the existing technologies might not fit the lived realities of individuals living alone who have fewer social contacts. Veinot et al. [321] argue for the need to study and design population-level interventions, which may be delivered by public health officers [321]. While such at-scale interventions could provide necessary help for vulnerable populations such as low-SES individuals living alone, a key challenge is the immense public resources required for operating such interventions at scale.

New advances in AI opened up new opportunities to facilitate at-scale health interventions for vulnerable populations by automating some aspects of care, such as regularly collecting health information from individuals. Not only can AI-driven technology alleviate public health workers' burden on delivering interventions, but its scalability can also help reach out to broader populations who have been underserved. However, relatively few studies have explored how AI-driven systems can be leveraged in health interventions for vulnerable populations. Motivated by this gap in understanding the role of AI in assisting with macro-level health interventions for vulnerable populations, in this Chapter and Chapter 6, I explore how AI technologies might support practices of multiple stakeholders within public health infrastructures and public agencies looking to scale up care.

5.1.3 Large Language Models

The area of natural language processing (NLP) has shown remarkable achievements with the advance in language models. Language models aim to generate coherent follow-up text to inputs, trained on human-generated textual data (e.g., a corpus) such as Wikipedia contents or social media posts [45, 200]. With the underlying knowledge about the probabilistic relationship among adjacent words in the language corpus, the *pre-trained* models can be retargeted to more specific NLP tasks—such as machine translation (e.g., [343]), sentiment

classification (e.g., [228]), and question answering (e.g., [268])—through *fine-tuning* with task-specific datasets [200, 45].

While the early language models with millions of parameters (e.g., BERT [83]) required additional fine-tuning steps to perform a specific task, recent *large* language models (e.g., GPT [45], HyperCLOVA [162], PaLM [60], LLaMA [317]) with a larger number (i.e., over billions) of parameters, have enabled a new paradigm of *in-context learning* [200, 45]. In in-context learning, models understand input text written in human language, which is called *a prompt*), and generate the following text that coherently follows the prompt. For example, if given a prompt like '`Classify the food into categories. Apple→Fruit; Onion→Vegetable; Milk→`' as an input, an LLM is likely to infer the following text, '`Dairy`.' While the nature of the task is still the text continuation, the model understood the latent concept of food classification in the input prompt, just with two examples for apple and onion. While the nature of the task is still the text continuation, the model understands the latent concept of food classification in the input prompt. In the similar vein, prompts can be composed in a variety of ways to transform LLMs to solve diverse problems. Motivated by such capability of LLMs, NLP and HCI researchers have leveraged LLMs in various problem spaces, including but not limited to creative writing (e.g., [189, 64]), information extraction (e.g., [170, 246]), and writing programming code (e.g., [53]).

Among many application domains, in this Chapter and Chapter 6, I particularly focus on the open-domain dialog systems driven by LLMs, which I describe in the following section.

5.1.4 Supporting Open-Ended Conversations with Large Language Models

Designing AIs that converse with humans coherently and engagingly has been an active research topic in the areas of NLP, Machine Learning, and HCI. Depending on the goal

of the interaction, conversational AIs are usually designed as either *task-oriented* or *open-domain* dialog systems [125]. Task-oriented dialog systems are designed for a specific goal (e.g., booking a flight ticket) with pre-defined information schema (e.g., slots to fill such as destination, date, and preferred airlines). Within the HCI community, task-oriented dialog systems have recently been proposed with the goal of promoting mental health. Specifically, studies have designed chatbots for eliciting self-disclosure [253, 194, 193] or increasing self-compassion by taking care of chatbots that experience distress [188, 166]. Relevant to my work, Yeonheebot performs conversations with older adults to mitigate their depression and anxiety [281]. However, as rule-based or hybrid (e.g., combining rules and intent-based response retrieval) chatbots with pre-defined conversation flows, prior systems were limited in supporting serendipitous topics that users might bring up during conversations [188]. Conversely, open-domain dialog systems are intended to perform free-form conversations in open-ended topics ranging from daily life (e.g., [347]) to movies (e.g., [225]), with an overarching goal of providing empathy and enhancing users' feelings of social belonging (e.g., [350, 127, 327]) [125].

Research has often discussed that designing quality open-domain dialog systems is more challenging than designing task-oriented dialog systems [125, 103]. From a technical standpoint, it is relatively straightforward to define the 'quality' of the task-oriented dialogs because there are clear user goals and information slots that the agent should ask the user about [94, 125]. Conversely, guidelines for open-domain dialog systems are more variant. Huang et al. suggest that open-dialog systems should aim to (1) understand the *semantics* of what the user said, (2) behave *consistently* with their predefined persona, conversation history, and speaking style, and (3) interact with the user *emotionally* [125]. However, these multidimensional goals make it hard to define an objective quality metric for a chatbot's responses. State-of-the-art neural network models have not satisfied these goals simultaneously due to the complexity of multi-turn reasoning of the conversational context and infeasibility of automated evaluations to improve model quality [125].

Recent LLMs, however, have brought breakthroughs in open-domain dialog systems thanks to their capabilities in generating coherent and contextual responses through in-context learning [21, 276]. Due to these benefits, LLM-driven chatbots are increasingly developed or proposed by both practitioners (e.g., ChatGPT [247], Gemini [109], Claude [14], and Copilot [219]) and researchers (e.g., [330, 348, 185]). LLM-based chatbots receive the current dialog history (i.e., list of turns of the user and the agent) in a prompt and infer the agent's following response accordingly [276]. The in-context learning inherently covers the multi-turn reasoning of the conversational context, generating responses that are generally aware of and specific to the context. Research in LLM-driven chatbots has pointed to several limitations and challenges in designing LLM-driven chatbots, mainly due to the inherent characteristics of LLMs. As language models generate the most probable output based on a complex structure of neural networks (called *transformers* [320]), it is not explainable how an LLM *'reads'* the input prompts written in natural language [200]. In the context of chatbots, it is thus challenging to anticipate how an LLM would process the history of dialog and what response it would generate. Since LLMs have learned a tremendous amount of human-generated text, there is always a risk that the conversation flow might follow directions unintended or unaccounted for by the chatbot designer [21]. For example, Wang et al. found that a mental therapy chatbot built with GPT-2 was likely to provide more negative comments than the human therapists would [327]. Also, there is a possibility that the unethical or biased phrases ingrained in the models' pre-training datasets might be exposed in the model's output, causing the chatbot to say socially biased [296, 28, 39, 105, 297] or toxic [106] messages. One known method to steer the conversations towards desired scenarios is to put ideal conversation examples in the prompt together [21]. Although such an in-context learning approach helps steer the model output, it is still challenging to perfectly control the model to say or not to say specific phrases [21, 327].

Given the aforementioned challenges and risks of leveraging LLMs for open-ended chatbots, CareCall presents a unique example of an LLM-based open-ended chatbot being deployed

in a real-world setting as a public health intervention. In this Chapter and Chapter 6, I explore how LLM-driven open-ended dialog systems are deployed specifically within public health infrastructures.

5.2 Methods

To understand the benefits and challenges of LLM-based chatbots as a public health intervention, I observed focus group workshop sessions with 14 CareCall users and interviewed 20 people from three groups of the main stakeholders around the CareCall system: The **users** of CareCall ($N = 5$), the **frontline workers** who monitored the users' conversations with CareCall ($N = 5$), and the **developers** of CareCall system ($N = 10$). I conducted multi-stakeholder interviews because stakeholder groups often had insights into the perspectives or opinions of other stakeholders by virtue of their frequent interactions. For example, frontline workers had insights into how users interact with CareCall and what perspectives they have toward the system through their frequent interactions with users for follow-ups on any health issues. Similarly, UX designers had insights about the perspectives of users and public agneices as they conducted formative work with both stakeholders to design and iterate on the system. Business managers also had insights about the perspectives of public agencies as they frequently interacted with them to gain feedback on the design and deployment of the system. The quality manager similarly had insights about the real-world usage of CareCall because they were monitoring CareCall logs as part of their work. Together, these interviews aimed to provide a holistic perspective on experiences creating and using such a system. Since this study was conducted in a corporate setting without its own IRB, I submitted the study protocol and obtained IRB approval from an outside public entity that conducts ethical oversight for research. The interview study was approved by the public institutional review board affiliated with the Ministry of Health and Welfare of South Korea.

The observation of the focus group workshops was classified as exempt by the guidelines from the Ministry of Health and Welfare of South Korea. In total, I report on insights from 34 people who interacted with different aspects of CareCall including the users (240 total minutes of focus group observation and 230 total minutes of individual interviews), front-line workers (250 total minutes of individual interviews), and developers (430 total minutes of individual interviews). For clarity, I did not have access to nor did I review CareCall users' conversation logs. All interviewees, including frontline workers and developers, did not pull specific conversation logs during the interview sessions, and their perspectives drew from their holistic experiences working with CareCall and its users rather than recalling or reviewing any particular conversation or CareCall user.

5.2.1 Observation of Focus Group Workshops with CareCall Users

I observed six focus group workshop sessions with 14 CareCall users for four hours in total. The focus group workshops were held by the Seoul Metropolitan Government from mid-July to mid-August of 2022. The workshop participants were middle-aged adults living alone who were participating in the CareCall pilot project in Seoul and had used CareCall for at least two months, having missed no more than a week of calls. The goal of the workshop was to understand the users' perspectives on using CareCall in their daily life and, broadly, to brainstorm ideas about AI-powered public health interventions for middle-aged individuals living alone. The workshop participants included 7 individuals in their 50s and 7 individuals in their 60s (12 males and 2 females) (Table 5.1a). I did not collect further demographic information on each workshop participant because I was a passive observer of the focus group; thus, in this paper, I refer to them as *focus group participants.*

During the workshop, the participants were asked about aspects of CareCall that they liked or did not like and what characteristics they might value in AI-based check-up calls like Care-

Table 5.1: Demographic of the CareCall user interviewees and the focus group participants (a), frontline worker interviewees (b), and developer interviewees (c).

(a) CareCall users

Alias	Age	Gender
P1	68	Male
P2	59	Male
P3	64	Male
P4	61	Female
P5	54	Male
Focus group participants	50-59	5 males, 1 female
	60-69	7 males, 1 female

(b) Frontline workers

Alias	Age	Gender	Relevant Experience
F1	49	Female	Customer support & social services
F2	51	Female	Customer support & social services
F3	61	Female	Social services
F4	55	Female	Customer support
F5	53	Male	Psychological therapy

(c) CareCall developers

Alias	Age	Gender	Role
D1	30	Female	Business manager
D2	31	Female	UX designer
D3	33	Female	UX designer
D4	51	Male	Business manager
D5	32	Male	Machine Learning engineer
D6	33	Female	UX designer
D7	30	Male	Machine Learning engineer
D8	50	Male	Quality Manager
D9	25	Female	Machine Learning engineer
D10	25	Male	UX designer

Call. Each session lasted for 40 minutes, with 3 to 6 individuals participating. Note that I did not organize or facilitate the focus group workshops. I only took observational notes of the workshops to gain broader perspectives from CareCall users, which was pre-approved by the workshop organizers at the Seoul Metropolitan Government and was made aware to the participants. Through these observations, I sought to better understand what benefits and challenges users perceive when using conversational AI leveraging a large language model as part of public health intervention. I opted for focus group observation because the municipality aimed to protect the privacy of the participants in the public health deployment of CareCall, and therefore understandably did not want to provide me with contact information for the participants. However, the municipality gave me the opportunity to hear how the perspectives of CareCall users contrasted to one another and to recruit interviewees directly through the focus groups. Together with the interview data, the findings from the focus group observation helped deepen our understanding of the users' lived experiences.

5.2.2 Multi-Stakeholder Interviews

I conducted 1:1 semi-structured interviews with 20 participants from the three groups of stakeholders via Zoom conference calls ($N = 8$) or in person ($N = 12$) based on their availability to travel. To compensate for their time and efforts, our research team offered each participant 50,000 KRW (approximately 38.5 USD as of July 2022) as a gift card.

Interviews with Users. I recruited five CareCall users (P1–5; Table 5.1a) from the focus group workshops I observed by distributing flyers. Since all CareCall user interviewees were recruited among the participants of the CareCall pilot project in Seoul, they shared demographic characteristics: middle-aged adults who were living alone and were low-SES. The CareCall user interviewees included 2 individuals in their 50s and 3 individuals in their 60s (4 males and 1 female). They had been using CareCall twice a week for two months

at the time of the study. I met each interviewee in person in a private meeting room, and each interview lasted for about 60 minutes. The interview questions covered (1) prior experience with receiving regular check-up calls from public agencies or as part of community services; (2) perception of AI phone calls both before and after using CareCall; (3) good and bad experiences with CareCall conversations; and (4) perspectives around AI phone calls in general towards care and companionship.

Interviews with Frontline Workers. I recruited five frontline workers (F1–5; Table 5.1b) by distributing flyers to a social enterprise for senior employment that was in charge of the frontline monitoring task of CareCall in Seoul. Participants had been working as frontline workers for 16 hours a week for about two months at the time of the study. The frontline worker interviewees included 3 individuals in their 50s and 2 individuals in their 60s (1 male and 4 females), with all having relevant experiences such as customer support, social services, or psychological therapy. Each frontline worker was in charge of monitoring 20 to 28 individuals via CareCall. Each interview lasted for about 60 minutes. The interview questions focused on (1) the participants' thoughts on the role and the impact of CareCall on their frontline monitoring task and broader public health work and (2) their interactions with the users whom they were in charge of.

Interviews with Developers. I recruited ten IT professionals (D1–10; Table 5.1c) who participated in the design and development of CareCall through a mailing list at NAVER, the vendor of CareCall. With regards to the role in the CareCall development team, the participants consisted of four UX designers, three machine learning engineers, two business managers, and one quality manager. The developer interviewees' ages ranged from 25 to 51 (5 males and 5 females). The UX designers were in charge of designing the conversation flows and conducting user studies. The machine learning engineers were in charge of improving the language model used for predicting responses and detecting user status. The business managers were in charge of coordinating with public agencies. The quality manager was

in charge of monitoring the product quality. Most of the development team members had been involved in this project for about a year at the time of the study, with a few having been involved for about 2 to 3 months. All team members were managing the design and deployment of CareCall across multiple municipalities rather than just Seoul.

Each interview lasted for 40 to 60 minutes. During the interviews, I asked about the participants' experiences in the development process, including challenges they encountered in designing or implementing aspects of CareCall and communicating with other members and stakeholders. I also focused on different aspects depending on the role of the participants. For instance, I specifically asked UX designers about the rationales and challenges of the conversation design of CareCall. For machine learning engineers, I focused on their thoughts on the unique characteristics and challenges of designing an LLM-based chatbot and how they addressed the challenges.

5.2.3 Data Analysis

All interview sessions were audio-recorded and transcribed later. Observational field notes for the focus group workshop sessions were created to capture broader CareCall users' perspectives. Our research team used thematic analysis [40] to qualitatively analyze both interview transcripts and observational notes. I open-coded the interview transcripts and the observational notes simultaneously using a spreadsheet, going through several rounds of iterations. Analyzing different data sources together allowed me to verify that the perspectives were present among participants recruited through different techniques. The full research team then discussed and identified patterns and themes through multiple rounds of peer-debriefing meetings. From this coding, I surfaced the main theme about the benefits and challenges around the lack of conversational control in LLM chatbots, which I organized my results around. The final codebook contained 10 parent codes (automation of health monitoring

work, performing specific tasks, customizing to different target groups, connecting to social services, emergency management, inappropriate responses, personalization, conversation topics, emotional support, emotional burden) and 24 child codes.

5.2.4 Limitations

In my study, I specifically focused on the context of Seoul where CareCall was deployed with low-SES middle-aged individuals living alone. My findings might not represent all target populations' experiences with LLM-driven check-up calls. For example, as explained in subsection 5.1.1, CareCall was deployed in municipalities that have different characteristics of the populations in terms of age groups or health conditions, including older adults living alone in Busan and people with early dementia in Ilsan. These populations likely have different health and companionship needs as well as different perspectives toward LLM-driven chatbots. Similarly, chatbots could be deployed in different social service settings. The frontline monitoring tasks of the Seoul sample were handled by part-time workers specifically hired for the CareCall pilot project by the Seoul Metropolitan Government. Social welfare officers took the frontline monitoring tasks as an aspect of their social work in other municipalities, and therefore, my findings might not generalize to different social service contexts where LLM-driven chatbots could be deployed with different monitoring goals.

Participants' experiences may change as they engage with LLM-based chatbots in a longer term. At the time of the study, the users and the frontline workers had been engaging with CareCall for two months, being aware that the pilot project would end in a month. Experiences of both users and frontline workers may change if they engage with the system in a longer term. For example, they might become to better understand the capabilities and the limitations of the system so that they can interact with the system in a more informed way; or, their engagement may decrease as they get tired of it over time. Future research

on a longitudinal deployment of LLM-driven chatbots for public health interventions would help understand how users' engagement change in the long term.

Our study sample has a skew toward experiences of socially isolated males in their 50s and 60s, which may have impacted the findings. Females who live alone and are younger or older might have different perspectives towards LLM-driven chatbots for social isolation intervention, and their interactions with the system might also be different. Further, my focus on the users who used CareCall regularly (e.g., missed fewer than two calls per week) among the pilot sample may have resulted in participants having a more positive attitude towards the chatbot leveraged in the public health intervention. CareCall users who have occasionally or frequently missed calls or non-users who had dropped out of the intervention might have different, more critical attitudes or perspectives around LLM-driven chatbots. In addition, my interview data overrepresents developers ($N = 10$) in comparison to frontline workers ($N = 5$) or users ($N = 5$). To address this issue, I sought to gain additional insights into the end-user perspectives through the accounts of other stakeholders. However, the end users' original accounts might have been filtered through the lens of these other stakeholders, who have power over the users in how the intervention is ultimately designed and enacted. I also supplemented the end-user perspectives with focus group observations, but this method offered less direct engagement with the users. Therefore, while I have made efforts to represent the perspectives of the socially isolated individuals who used CareCall, my results may not fully capture their lived experiences or their concerns with the technology.

5.3 Findings

Through the qualitative analysis of interviews and observational notes, I surfaced the lived experiences of the multiple stakeholders who engaged with, managed, and developed a public health intervention leveraging an LLM. In this section, I present the findings of the study,

focusing on the benefits and challenges multiple stakeholders—the users, the frontline workers, and the developers—experienced. Note that I blend multiple stakeholders' responses in the findings because stakeholder groups often had insights into the perspectives of others by virtue of their frequent interactions.

5.3.1 Benefits of Leveraging an LLM-driven Chatbot in Public Health Interventions

Overall, the frontline workers and the users perceived the benefits of leveraging an LLM-driven chatbot in public health intervention. The frontline workers valued that CareCall helped them gain a holistic understanding of each individual through open-ended conversations while offloading their workload. The users perceived the benefits of mitigating loneliness and emotional burdens.

Providing a Holistic Understanding of the Individuals While Offloading Workload

The frontline workers taking care of the CareCall users valued that the system provided a holistic understanding of the care recipients through open-ended conversations while offloading their workload. As explained in the subsection 5.1.1, the dashboard provided a summary of health metrics and emergency alerts so that frontline workers could focus on monitoring and reaching out to cases that needed their attention. Frontline workers perceived that the care work process supported by CareCall offloaded a significant amount of workload. F2 said: *"If I were to call all the 26 individuals by myself twice a week, I don't know if I could take on that job. It would be both mentally and physically exhausting to ask the same questions over and over again to that many people."* Based on her previous experience in customer support call centers, F2 assumed that human check-up calls are likely to become redundant

and inefficient: *"Human phone calls are likely to get sidetracked. We'll ask questions to check what we need to know, but they'll probably mention other things, too; the phone call might end up being super long, like 30 minutes. That's not feasible given the time frame."* F2, therefore, appreciated that CareCall could manage some of the more redundant aspects of monitoring, allowing them to focus on monitoring individuals who need care the most.

Despite the reduced workload, frontline workers felt that CareCall's open-ended conversations provided rich contextual information to help them gain a holistic understanding of each user's circumstances, which might have been difficult with rule-based dialog systems based on pre-defined scenarios. F5 stated: *"I think I have a pretty good understanding of each person's circumstances at this point because I've been monitoring the call recordings."* F4 noted that the conversation between the CareCall agent and the users surfaced broader aspects of the users' life which were useful for understanding how they are doing: *"Some users are leading a satisfying life, typically people who have jobs, regularly go to a community welfare center, and have friends to meet; I'm not too worried about them. I'm more worried about those who are mostly lying in bed all day and have depression."* This information helped them figure out whom they needed to prioritize monitoring. F4 further stated: *"I mostly focus on monitoring the individuals that I'm concerned about. I got to learn about those individuals over time by monitoring the call recordings."* F5 similarly appreciated: *"CareCall works like a patrol who leads the way and tells us how things are going. I found it really useful to have such information."* The frontline workers further mentioned that thanks to CareCall, they had found cases where some serious health issues occurred to the users. F1 and F4 mentioned that they had found users mentioning they had been hospitalized through the conversation logs. Both F1 and F4 were able to then reach out to the users, asking why they were hospitalized and sending emotional support.

Mitigating Loneliness and Emotional Burden

Both CareCall users and frontline workers highlighted how CareCall could help manage people's loneliness and the emotional burdens. The frontline workers mentioned that many of the users had a strong desire to have more conversation opportunities. F1 said, *"There were a few people who cried when I called them. They said they wouldn't have spoken a word if I didn't call her that day."* F5 similarly noted, *"There are a lot of people who feel terribly lonely. When we called them, the person thanked me, saying that I was the only one who had called them recently."* frontline workers had observed several instances where users looked forward to receiving the scheduled check-up calls from CareCall. F3 noted, *"I think getting regular check-up calls makes them feel like someone is thinking about them. I noticed some of them looked forward to getting the scheduled calls."* F1 also noted, *"Some people are really looking forward to getting the calls. I notice that they want to talk as much as possible to AI."* F5 further mentioned that some users regularly said *'Thank you'* during the call, which led them to think that the individuals might have received emotional support from CareCall. Frontline workers further perceived that the users enjoyed CareCall's support for diverse conversation topics. F1 mentioned: *"People occasionally talk about their hobbies in detail, for example, paper crafts. Then the AI responded, 'It would be great to showcase your art one day!' I noticed the user was surprised that AI could talk about such things."*

Likewise, the users appreciated receiving check-up calls from CareCall. A focus group participant stated, *"I like getting the AI calls. I feel pretty lonely living alone, so it's nice to have someone to talk to, even though it's a machine."* Another focus group participant similarly said, *"I barely have anyone to talk to after losing my job last year. I feel so empty and lonely. I like that it asks about my health."* Specifically, the users appreciated that the system asked caring questions about their health. A focus group participant noted, *"It was nice to get a phone call checking in with me, asking why I couldn't sleep well last night."* P5 similarly said, *"I feel thankful when they [CareCall] ask caring questions as if they were my wife."*

The users also valued that CareCall covered broader conversation topics beyond health. Specifically, they appreciated that they were able to talk about their hobbies. P5 enjoyed having conversations about his habits in sketching with CareCall: *"When it asked what I was doing, I said I was drawing something. It then responded, 'That sounds fun! I want to learn how to draw too.' I really liked it when it said that. I wanted to talk more about my work."* Other users desired that they could engage in more detailed conversations about cultural life. During the focus group workshops, many participants mentioned their wish that CareCall could recommend movies, TV shows, books, and music or ask about what foods they like. P2 further envisioned that AI could give personalized recommendations based on the conversation data: *"If AI collects a lot of data about us, they might be able to know what sports I am interested in or what kind of art I like. Then it might be reflected in the conversations."*

Furthermore, the CareCall users valued a lack of emotional burden when receiving check-up calls from an AI compared to receiving phone calls from a human. A couple of users noted that they sometimes felt emotionally burdened when contacted by humans. While CareCall was not aimed at replacing other social experiences, a focus group participant said that they might feel more comfortable getting AI calls than getting phone calls from humans: *"My friends might suggest going out for dinner or something when they call me. I sometimes don't want to because of my depression, but I feel uncomfortable turning them down. But I don't need to feel that way to AI."* Another focus group participant similarly mentioned, *"Sometimes I feel more comfortable talking to the AI because it's not a human and doesn't have feelings."* Some participants similarly mentioned the emotional burden that they felt when receiving check-up calls from public health officers. P3 stated: *"I know that some public health officers are checking up on me because I have chronic conditions and live alone. But I feel like they are pretty perfunctory because they only ask one or two questions, and that's it. I would rather prefer getting AI calls."* A focus group participant suggested they might feel emotionally burdened about adding more work to public health officers: *"Sometimes I*

get phone calls from a public health officer during the weekend. I guess they had too much work during the week, so they had to call me over the weekend. I felt sorry for them. I don't have to feel that way when getting AI calls."

5.3.2 Challenges in Leveraging an LLM-driven Chatbot in Public Health Interventions

Despite the benefits, I observed various challenges in leveraging CareCall for public health interventions. In this section, I first describe the inherent challenges of LLMs in uncertainty in control that the developers faced. Next, I illustrate the challenges in leveraging an LLM-driven chatbot, specifically around tailoring it to public health needs and supporting personal health needs.

The CareCall developers frequently mentioned the difficulty in controlling the responses that might not be appropriate for public health contexts. In the initial stage of development, the developers were concerned that the system might generate utterances that make promises that non-human agents could not keep because the LLM embedded in CareCall was pre-trained with human-generated text data (i.e., the Korean corpus depicted as (D) in Figure 5.1). D3 noted that even though the example dialog corpus ((E) in Figure 5.1) did not include cases making infeasible suggestions, the system still generated responses doing so: *"When the person said they didn't have any plans this weekend, the agent kept saying infeasible things such as 'How about going to a karaoke with me?' or 'Let's go hiking with me!.' That was the most difficult part in the development process."* The CareCall developers were generally concerned that such suggestions might make the users confused. D9 noted that the developers put efforts into making the system disagree if users made similar suggestions: *"The agent shouldn't suggest, for example, playing billiards together because it can't. Also, it shouldn't say 'yes' when a user makes similar suggestions."* The developers

were also concerned about the risk of generating impolite utterances, particularly given the vulnerability of the target population. D2 said, *"We saw the agent saying something rude, like 'Hope you stay healthy not to burden your family,' which made us freak out."* D7 gave a similar example: *"I don't know what exactly happened, but the system might have detected something wrong and said 'Congratulations!' when the person said they didn't feel well."*

The uncertainty in control largely resulted from the inherent characteristics of LLMs. The developers valued that an LLM enabled them to develop an open-domain dialog system much faster and easier compared to other rule-based systems. Because an LLM was used as a backbone model to generate utterances, CareCall was able to cover much broader topics of conversations that would not be feasible for rule-based systems. D9 said, *"LLMs are capable of generating various kinds of utterances even without manually defining the rules."* However, such characteristics made it difficult for the developers to steer the conversations to prevent inappropriate responses. D3 noted that the responses generated by the backbone model tended to be significantly affected by the large-scale corpus used for the initial pre-training, which includes toxic and biased content that might hurt conversations. D9 further described the process of controlling LLMs: *"Language models have a strong ego, so we have to fight with them. When it generates inappropriate responses, we need to see how it came out, rather than fixing the responses themselves, going through many trials and errors. So it's very difficult to develop a system that is perfectly under control."* D2 noted that such a challenge is a distinct characteristic of LLM-driven chatbots from rule-based ones: *"To fix inappropriate responses of rule-based chatbots, all we need to do is just to modify the scenario. But for LLM-driven ones, we have to consider the patterns where the response came out, which is far more difficult to control."* Even though they incorporated additional steps, including the in-context learning with an example dialog corpus and fine-tuning on the undesirable and inappropriate phrases (see subsection 5.1.1), the developers still acknowledged the uncertainty in control of the system.

Tailoring to Public Health Needs

I noticed several mismatches between the public agency needs and LLM-driven chatbots' challenges. First, the CareCall developers faced challenges in addressing the public agency needs for asking specific health questions during the calls. Since CareCall was introduced as a technology to assist public health work, the public agencies expected that they could integrate specific questions that they were interested in. For example, D3 mentioned: *"Some local government officials asked if we could integrate dementia screening questionnaires into CareCall."* However, CareCall had inherent uncertainty in controlling the dialog flows. D5 stated: *"What we can do is to fine-tune the model with more datasets that ask certain questions so that the probability of asking such questions becomes higher, but we cannot guarantee that. Such tasks are performed just indirectly."* Therefore, the developers could not accommodate the public agencies' requests. D2 indicated: *"We got asked by several local government officials to ensure that our system asks questions about medication adherence or something. But at least for now, we can't guarantee that."*

Due to the resource-intensive nature of customizing LLMs, the CareCall developers also experienced challenges in customizing to different target groups. Public agencies had different target groups with different monitoring needs in mind, such as older adults living alone in Busan, middle-aged living alone in Seoul, healthy older adults in Gwangju, and people with early dementia in Ilsan. D2 indicated: *"The government of Seoul wanted to deploy CareCall with middle-aged adults because this age group had the highest lonely death cases recently."* Similarly, D3 mentioned that the government of Ilsan had reached out to them, indicating the need for regular check-up calls for older adults with early dementia. However, the developers perceived that CareCall might not fit those groups well because the current dialog corpus ((D) in Figure 5.1) did not simulate conversations regarding these wildly different health needs. For example, D2 was concerned about deploying CareCall with middle-aged adults: *"When someone says that they have a backache, CareCall is likely to say 'It happens as*

we age.' A response like this might be perfectly fine for someone in their 70s, but might be odd for someone in their 40s." D2 also mentioned a similar example with people with early dementia: *"When someone says 'I'm so forgetful these days,' we can simply say 'It happens. I also forget about things sometimes.' But we might need to dig deeper into it if the person had early dementia."* The CareCall developers wished to provide more customized conversations to different target populations given their characteristics and needs, but due to the nature of the example-driven response generation of LLM, tailoring to new target groups demanded new sets of example dialog corpus simulating conversations with those groups. D2 stated such tailoring would not be feasible: *"I wish that the system could provide more customized conversations, but it's not feasible. It's almost like making the example datasets from scratch."* Other CareCall developers similarly mentioned the challenges in customizing to middle-aged adults because of the immense resources needed to generate new sample datasets. Generating new sample datasets would require several iterative cycles of collecting patterns of human-bot dialogs with the specific target population in mind, augmenting the example dialogs with LLM, and labeling positive and negative utterances manually.

In addition, the open-ended nature of LLM-driven chatbots made it challenging for CareCall to manage expectations around the emergency and social service needs. The users wished that the system offered a direct connection to emergency services. They predominantly mentioned their anxiety resulting from living alone, getting older, and having chronic conditions. A focus group participant stated: *"I am getting check-up calls from a community welfare center, a community health center, and a church. I am most concerned about dying alone, so I have applied to all kinds of check-up calls."* P1 similarly mentioned their fear of passing out or dying alone due to their health history involving diabetes or stroke. P1 noted, *"I could pass out at any time. The right side of my face is partially paralyzed because of my diabetes (complications)."* P3 also noted, *"I had a stroke last year, which left my right side of the body paralyzed. I'm worried about having a stroke again when alone."* Therefore, many users desired CareCall could detect emergency situations and automatically call emergency

services. However, the developers were not confident about the reliability of the emergency detection, making them hesitant to support such a feature. D3 noted: *"We do not want situations where CareCall fails to detect even just a single case after making a contract that CareCall would detect emergencies and call 911. So we've decided that our product is NOT for actively sending help in emergency situations."*

I further noticed that CareCall users expected that the system would help provide access to a variety of social services, but the developers and the frontline workers felt it was out of scope. D4, D10, and F4 observed that the users asked to join the food assistance program as part of social care for underserved populations. Even though CareCall was not targeted at processing such requests, in some municipalities where the users were managed by social welfare officers, they were able to discover the needs and process the requests. D10 described an instance: *"There are food assistance programs for delivering free lunch boxes for low-SES older adults in most of the municipalities. Through monitoring CareCall logs, the public health officers were able to find the need and had the user join the program."* In contrast, the frontline workers in Seoul felt confused because they did not have the power to accommodate them as part-time workers who were outside the social service department in their public agencies. F4 said *"They ask for lunch box deliveries, but all we can do is just empathize with them and report it to the agency in charge. We don't have any power to connect to such social services."* Similarly, D3 and D5 also mentioned that some users requested to fix their refrigerators or fans during their phone calls but were concerned about adding unexpected tasks to public health workers who were managing CareCall. D5 elaborated, *"The public health officers were just in charge of checking whether the individuals were doing well; their job was not to check whether a lunch box had been delivered. When CareCall starts to receive such requests, it adds another task for them."* In addition, F1 and F2 indicated that some users also mentioned that they needed escort services to the doctor's office during their phone calls with CareCall. F1 said: *"Some people were desperate to find someone to go with them to the doctor. I felt really bad, but I couldn't help."* Furthermore, F3 and F4 referred

to instances where some of the users requested financial assistance in accessing healthcare services. F3 noted: *"There was a person who kept talking about their circumstances to the AI, like 'I am sick. I need to go see a doctor, but I'm short on money. Can I talk to a person who can help me out?' But AI could only say, 'Why don't you see a doctor?' It's a bit frustrating."* Because the frontline workers did not have the power to help with such requests themselves, they typically relayed the requests to the public health officers in their agencies when receiving them. Despite the users' needs related to social services, the developers were concerned about the potential burden on the public health officers and wanted to keep the system specifically for regular check-up calls that inform the public health workers of concerning cases.

Supporting Personal Health Needs

I noticed the challenges of LLM-driven chatbots in providing emotional support due to the technical challenges in remembering personal health issues. The frontline workers and the users wished that CareCall would ask personalized questions that consider personal health history. However, due to the technical difficulty in implementing long-term memory in LLM-driven chatbots [337, 338], CareCall could not generate personalized questions and answers that follow up on personal health issues based on past conversations. F5 felt disappointed that the personal health history survey that the frontline workers conducted with the users before rolling out the system was not taken into account to provide personalized conversations: *"One of the individuals that I am in charge of has liver cirrhosis involving ascites. It would have been great if the AI call asked questions like 'Have you seen a doctor to remove the fluid?' based on the pre-survey, but it only asks general questions."* F2, F3, and F5 further mentioned that they felt awkward when the CareCall agents asked inappropriate questions without considering one's current health status. F2 described: *"Some people have severe lower back pain so that they can barely walk. But the AI system kept asking whether they*

had exercised or whether they had taken a walk. I felt so awkward monitoring such logs." F5 similarly indicated: *"The person has a chronic condition, so they have already been seeing a doctor. But AI thought that was a new health issue and kept suggesting seeing a doctor."* The users similarly noted that not acknowledging their health issues made the system feel impersonal. A focus group participant said: *"I feel someone understands me and takes care of me when they remember what I've said before. So, when I told them [CareCall] I had a backache, they should have asked questions about that the next time. But they acted as if we had never talked about that."* P3 similarly indicated, *"It would be nice if it could remember that I've seen a doctor and ask follow-up questions. Or, it could at least remember what it has said themselves in the past, like, 'I suggested taking more steps last time. Have you tried it? How did you feel?' Then I could respond, 'Yep, I've tried it as you've suggested. I feel it helped me fall asleep faster."*

The lack of long-term memory of CareCall also limited its ability to provide emotional support to the users. While some users perceived the emotional benefits of the system, others did not partially because of the repetition of general questions and responses across the sessions. For example, they felt that the system always responded in the same way when they mentioned not feeling well. A focus group participant noted, *"It always asks a fixed set of questions like, 'Have you seen a doctor?' when I say I'm not feeling well."* Another focus group participant similarly said: *"When I say something, it always says 'Oh, I see.' I don't feel like we're really communicating."* The repetition of general conversation patterns seemed to interfere with providing emotional support. Some users mentioned feeling like the system was a stranger even after months of engagement. A focus group participant said: *"I've talked to them [CareCall] for a few weeks, but it didn't seem like we got to know each other over time. It always asks the same general questions."* P3 similarly said, *"It's a familiar voice that I've heard for many weeks, but I always feel like talking to a stranger because it never asks specific questions about me. I'd like to talk as if I am talking to an old friend rather than a stranger."* The repetitiveness of the conversations also led the users to feel the

conversations were robotic. Several users mentioned that the repetitive utterances felt too machine-like, which decreased their motivation to engage in the conversations. P4 noted, *"I can foresee what it'll ask next or how it'll respond, so I don't get too excited about the conversations."* Another focus group participant also mentioned: *"I don't feel like it really understands how I am doing. It just keeps saying, 'Oh, I see,' so I don't feel it empathizes with me."*

5.4 Discussion

My findings from observing focus groups and interviews with multiple stakeholders who created, interacted with, and worked with CareCall suggest opportunities for leveraging LLM-driven chatbots to support public health interventions. My findings demonstrated that LLM-driven chatbots have emotional benefits, particularly around supporting broader conversation topics, but also have challenges due to the limited personalization. Based on the findings, I highlight the opportunities for improving emotional support in LLM-driven chatbots. My findings also pointed to the tensions between multiple stakeholders' needs and the capabilities and limitations of LLM-driven chatbots in public health contexts. I suggest that designing better resources that transparently communicate the respective capabilities and limitations of open-domain and task-oriented chatbots could help different stakeholders negotiate those tradeoffs. Lastly, I observed tensions around the desire and challenges of scaling LLM-driven chatbots to diverse public health needs. I suggest opportunities for designing mechanisms to help the target populations or care professionals contribute to dialog datasets.

5.4.1 Improving Emotional Support in LLM-Driven Chatbots

My findings highlight that the technical challenges of LLM-driven chatbots in personalizing responses interfered with providing emotional support. While the users wished that their conversations with CareCall would consider personal health history, the system could not due to the lack of long-term memory, which made them feel that the system was impersonal and robotic. Addressing the technical difficulties of implementing long-term memory [337, 338] in LLM-driven chatbots would help resolve part of the challenges in providing conversations that consider personal details such as health history. At the time when this paper was written (Summer 2022), there was limited research on implementing a long-term memory feature in LLM-driven chatbots, and commercial systems like ChatGPT and Gemini had not yet introduced this feature. However, shortly after, beginning in September 2022, a new version of CareCall with long-term memory [20] was implemented and distributed to users. In 2024, OpenAI and Google similarly began testing long-term memory capabilities in ChatGPT and Gemini, respectively, and both now support referencing all previous conversations in their responses [249, 110]. My colleagues and I thus conducted a follow-up study to examine the impact of long-term memory on user perceptions of these chatbots, which was presented at CHI 2024 [142]. While the CHI 2024 study highlighted the potential of memory capabilities to demonstrate care through chatbots, it also revealed challenges in implementing such features in complex and sensitive health contexts. More work is needed to examine how the introduction of long-term memory to LLM-driven chatbots shapes people's perceptions of emotional support across diverse health contexts.

Accounts from some of the users, such as a user who thought that CareCall could lead to reduced interactions with their social contacts, further point to concerns that systems like CareCall might be misapplied to take the place of social support. Prior work highlighted the concern that the introduction of AI technology could lead to unintended consequences for older adults, such as reducing human contact with their formal and informal

caregivers [128, 295, 182]. For example, if family members know that the older adult is "safe" through AI monitoring technology, they might visit the older adult less frequently. Similarly, if everyday caregiving tasks are replaced by robots at care facilities, older adults might lose the opportunity for caring social interactions. Sharkey et al. [295] pointed out that such a reduction in human contact is unethical because it might have a negative impact on the health and wellbeing of care recipients. In addition, recent work argued that LLM-based chatbots are still limited in their conversational abilities to engage in empathetic conversations in sensitive care settings [176]. They further pointed out that LLMs might convey biased perspectives or provide misinformation, which may critically impact the physical and mental health of users [176]. My work similarly reinforces that technology should not aim to replace the social support that vulnerable populations receive due to technical limitations and potential social consequences, but instead offer an opportunity to increase interaction.

On the other hand, my findings suggest that there is still value in LLM-based chatbots towards other goals, such as supporting conversations on diverse topics. My findings indicated that the open-ended nature of the conversations helped mitigate loneliness, particularly by supporting broader conversation topics beyond health, such as hobbies and cultural life, which would be challenging to configure rule-based dialog systems to support. Prior work for technology interventions suggested that even surface-level interactions and mere company could help mitigate the loneliness of older adults [66, 265]. In contrast, my study suggests that topic diversity could be one of the key aspects in providing emotional support to individuals who have limited conversation opportunities in their daily life. I highlight the utility of open-domain chatbots in mitigating the loneliness of socially isolated individuals, particularly around supporting diverse conversation topics. Future work on designing LLM-driven chatbots to allow for immersive conversations around specific topics of users' interest can also benefit their abilities to provide emotional support.

5.4.2 Tensions between Supporting Informational and Emotional Needs in Public Health Chatbots

Through this study, I found that some of the inherent characteristics of LLM-driven chatbots, such as the uncertainty in control and the resource-intensive nature of customization, led to challenges in supporting different stakeholders' needs in public health interventions. Prior work on chatbots for mental health indicated that expectation management around the system capabilities is challenging but critical [253, 194, 210]. My findings further highlight that expectation management about open-domain, LLM-driven chatbots can be challenging, particularly in public health settings. From a technical standpoint, open-domain chatbots are radically different from task-oriented chatbots. The primary goal of open-domain chatbots is to support naturalistic conversations on diverse topics, whereas task-oriented chatbots are aimed at performing specific tasks in a closed domain. However, interactions with LLM-driven chatbots performing open-ended conversations are likely to lead various stakeholders in public health interventions to assume that the chatbots can take on the maximal, most flexible set of tasks. Users may assume that the chatbot is a conduit for all things government-related–emergency services, food services, public healthcare services, financial services, and more, placing additional demands on already public health infrastructures. Government agencies can similarly assume that chatbots can take on a whole suite of public health tasks based on the promise of natural conversations. As a consequence, government workers may feel disappointed by not being able to get their specific questions answered, and so do the users by not being able to receive the care that they desire.

In the long term, technical advances in better controlling the open-domain chatbots could help address part of this challenge (e.g., ensuring that the chatbot asks specific health questions and supporting direct connections to emergency assistance). However, addressing the larger problems requires understanding multiple stakeholders' needs involved in complex public health settings [179]. My findings indicated both the governments and the users had some

informational needs that could have been better served by more traditional task-oriented systems. For example, task-oriented chatbots can more easily support asking specific health questions that fit governments' needs, such as whether or not a person is adhering to their medication. Task-oriented chatbots could also more reliably respond to a user's request to connect to emergency or social services. In contrast, while open-ended chatbots faced challenges in serving these needs, they demonstrated clear benefits in providing a holistic understanding of care recipients to facilitate care and emotional support through open-ended conversations. This suggests that, currently, the choice of model puts informational and emotional support in tension with one another.

Prior work on HCI and CSCW has highlighted the challenges in balancing multiple stakeholders' needs when using new technology in complex care settings [262, 273, 144], suggesting the need for mechanisms to assist each stakeholder in voicing and negotiating their needs [25, 144]. When novel and complex technologies like LLM-driven chatbots are introduced in public health interventions, negotiating multiple stakeholders' needs in light of the capabilities and limitations of the system could be even more challenging. Aligned with prior work, my study suggests that when designing one of these open-domain chatbots for public health interventions, it is valuable to have conversations around its capabilities and expectations with multiple stakeholders. Designing resources that transparently communicate the capabilities and limitations of open-domain and task-oriented chatbots could help different stakeholders figure out what type(s) of technology they need and negotiate their needs with each other. In addition, as prior work highlighted [25], it would be beneficial to create opportunities to hear multiple stakeholders' perspectives *before* developing or deploying a system for public health intervention. This opportunity will help developers better recognize what tensions might exist among different stakeholders and what misconceptions they might have toward the system, potentially benefiting the design of conversational prompts to avoid or prevent those.

5.4.3 Scaling LLM-Driven Chatbots to Diverse Public Health Needs

My findings surfaced the needs and challenges of LLM-driven chatbots in serving diverse public health needs of different target populations. Prior work has indicated that public agencies frequently have different public health needs from others based on their demographics and organizational capacity [179, 84]. Similarly in my study, I observed that public agencies had different target groups (e.g., older adults living alone, middle-aged adults living alone, and individuals with early dementia) and different ways of handling the monitoring tasks (e.g., having existing social welfare officers take on the task versus hiring part-time workers). Despite the public agencies' desire for customized conversations based on their needs, CareCall developers found customization infeasible to support due to the immense resources and challenges involved in generating new example datasets. While the open-domain nature and scalability of LLM-driven chatbots make them suitable for addressing the diversity of public health goals that governments might use chatbots for monitoring, when LLM-driven chatbots are deployed in practice, the lack of support for customization could lead to neglecting the specific health needs of different populations and public health monitoring goals.

Efforts to customize LLM chatbots in light of these goals are a valuable direction for future work. However, customizing LLM-driven chatbots to the government and end-user needs involves non-trivial challenges around collecting a relevant dialog corpus. Typically, crowdworkers are often used to take on the task of creating dialog corpus when developing a chatbot; however, they are likely not from the target populations and thus lack a deep understanding of the populations' needs. As a result, even with clear guidelines and training, crowdworkers might find it challenging to create datasets that reflect the populations' needs. Developing mechanisms for the target populations to effectively contribute dialog datasets could help overcome such challenges. Prior work in personal informatics has shown promise for speech interactions for collecting personal health data (e.g., [207, 206, 169]). Relevant to my work, Kim et al. [169] have proposed a speech-based smartwatch app to assist older

adults in labeling physical activities with a low capture burden. Similar approaches could help target populations in collecting dialog datasets in an accessible way, leading to developing chatbots that are more well-suited for them. However, not all target populations in public health contexts might be reliable to perform such tasks. For example, individuals with dementia might be less reliable in collecting and labeling dialog datasets, depending on their cognitive abilities or motor skills. Furthermore, collecting private data, such as everyday conversations, for machine learning purposes involves privacy concerns [312], particularly with marginalized populations [236]. An alternative approach would be to have experienced social or healthcare professionals who have a good understanding of the target populations contribute to the dialog datasets. However, this approach involves concerns over adding burdens to already overburdened professionals. Future research is needed to explore ways to help care professionals contribute to the creation of dialog datasets that better suit target populations' needs in chatbot-based interventions.

5.5 Conclusion

Through observing focus groups and interviews with multiple stakeholders who created and interacted CareCall, I found that LLM-driven chatbots can provide emotional benefits, such as supporting broader conversation topics, but also have difficulties providing emotional support due to limited personalization of conversations. I also observed tensions between multiple stakeholders' needs and the capabilities and limitations of LLM-driven chatbots in public health contexts, with public agencies often desiring specific health questions to be asked, with LLMs lacking that level of control. Based on the findings, I highlight that implementation of long-term memory could improve emotional support in LLM-driven chatbots. I further suggest designing better resources and processes that help multiple stakeholders negotiate the respective tradeoffs of open-domain and task-oriented chatbots. Lastly, my

work points to a need to explore how to scale LLM-driven chatbots to diverse public health needs, suggesting opportunities for designing mechanisms to help the target populations or care professionals contribute to dialog datasets. In closing, I hope this work can inspire collaborations among the researchers in the HCI, Public health, and NLP communities to design chatbots leveraging large language models for public health intervention.

Chapter 6

Supporting Public Agencies in Using AI Chatbots to Scale Up Public Health Monitoring

Current practices of public health monitoring rely heavily on frontline workers for recurrent data collection from populations. This type of monitoring is burdensome on them and other stakeholders as public health agencies often have fewer staff or resources than needed [78, 130, 133, 245, 250, 322]. To alleviate this burden and extend reach, public agencies are increasingly considering AI technologies, such as LLM-driven chatbots, to assist or automate some of these monitoring tasks traditionally performed by frontline workers [141, 142, 131, 132]. Introducing such systems requires collaboration across stakeholders who influence the decision-making, adoption, and rollout. Thus, when AI-mediated interventions are introduced, they not only change end-user interactions but also reshape the overall operation of public sectors and affect broader stakeholders. While prior work has explored stakeholder needs around AI systems in public services, it largely focused on frontline workers who directly use AI systems in their daily work [322, 56, 303, 131], limiting our

understanding of the full impact of these AI technologies.

To gain a more holistic understanding of how public sector AI operates within existing human infrastructures, I again examine the case of CareCall (see Chapter 5 for more detailed descriptions of the system). As a rare example of an LLM-driven chatbot deployed in a real-world public health context across municipalities with varied characteristics (e.g., urban, rural) and involving workers in varying roles (e.g., frontline monitoring, decision-making, administration), CareCall provides a useful case for understanding the public agency perspectives and experiences of the AI chatbot deployment for public health monitoring.

Through interviews with 21 public agency workers involved in the adoption and rollout of CareCall across 13 sites to monitor socially isolated individuals, I sought to answer the following research question:

RQ4: How might technology support public agencies in scaling up care within public health infrastructures?

Toward my thesis claim T2, this study demonstrates that AI chatbots fulfilled public agencies' expectations around extending public reach but failed to deliver on the expectation of scaling up monitoring reach and frequency with minimal human labor. Through the interviews, I found that public agency workers had previously struggled to regularly monitor populations needing care through phone calls or home visits (i.e., a *human approach*) due to a shortage of frontline workers. Although the agencies had tried to introduce technologies like passive sensing systems and automated voice-based systems (i.e., *hardware-dependent technologies*) to address these constraints, they created new labor demands, such as managing false alarms and maintaining devices, and their high costs still limited public reach. When CareCall was introduced, decision-makers had expectations that AI would reach more people and achieve the desirable monitoring frequency, which was largely realized. Frontline workers also valued that CareCall unexpectedly provided a window for care recipients to

communicate different care needs. However, frontline and administrative workers often felt that their workload was exacerbated as the introduction of CareCall rarely involved scaling up the human resources necessary to manage the expanded care and demanded new types of labor, such as handling lapses in user engagement.

In reflecting on and discussing the findings, I use the framework of articulation work [308] to highlight the changes in human labor that introducing AI chatbots requires. Decision-makers face unique challenges in conducting the required articulation work, particularly due to the open-ended nature of LLM-driven chatbots and the lack of established guidelines and best practices for these emerging technologies. I also highlight the importance of recognizing the maintenance work that AI chatbots impose on frontline workers, especially handling lapses in user engagement. Lastly, I provide implications for public agencies considering the use of AI chatbots for public health monitoring, focusing on the need to assess the impact of AI implementation on the labor demands of their workforce. For designers and developers aiming to make AI chatbots usable for public health monitoring, I suggest opportunities to piggyback on public infrastructure to enhance scalability, incorporate fallback mechanisms to address lapses in engagement, and leverage passive sensing as a complement to AI chatbots.

This project was published at CHI 2025 [143] with co-authors Young-Ho Kim, Sang-Houn Ok, and Daniel A. Epstein. Part of this work was conducted through a research internship at NAVER AI Lab. Young-Ho Kim and Daniel A. Epstein served in supervisory roles, providing guidance and feedback throughout the research process, and Sang-Houn assisted with participant recruitment. I initiated the study design, developed interview protocols, and led the interviews, data analysis, and paper writing.

6.1 Background and Related Work

6.1.1 Deployment of CareCall

CareCall is an LLM-driven voice chatbot developed to support socially isolated individuals in South Korea through scheduled phone calls, in response to rising concerns about lonely deaths [49, 175, 336]. Designed as an open-domain dialog system, the system mimics the conversational style of social workers and uses long-term memory to reference users' past disclosures in follow-up conversations [20, 142]. For more details about the motivation and design of the system, please see subsection 5.1.1. The technical details of the system architecture can also be found in prior work [162, 20, 142].

Initially launched in Busan, Korea, in November 2021, CareCall had gradually expanded to monitor over 30,000 individuals as of December 2024 through over 140 public agencies in South Korea. CareCall was adopted and deployed by various types of public agencies that were taking care of socially isolated individuals, including local governments, community health centers, and Veterans Affairs offices. The scale of CareCall deployments varied significantly, ranging from fewer than ten to several thousand individuals, depending on the reach of the public agencies involved.

Each public agency had slightly different criteria for the target users of CareCall in terms of the age group or specific health conditions, though they all shared the overarching characteristic of monitoring socially isolated people[1]. Most public agencies deployed CareCall to middle-aged (40s to 60s) and older adults (60s or older) living alone who were at a low socioeconomic status (e.g., below 50% of median household income). However, some agencies in public healthcare contexts deployed CareCall specifically to older adults with mild cognitive impairment or depression. In most cases, public officers who provide social services

[1]Consistent with Chapter 5, I refer to socially isolated individuals who receive regular check-ins through CareCall as "CareCall users" or "care recipients."

in neighborhoods recommended these individuals to use CareCall.

The adoption and rollout of CareCall required public agencies to undertake different tasks. Workers can be categorized as having one or more of three roles: *decision-making*, *administration*, and *frontline monitoring*. Similar roles have been described in other studies of public service technologies (e.g., decision-making [152, 286, 154], administration [154, 286], and frontline work [131, 141, 150, 56, 322, 155]). Before adopting and rolling out the system, public agencies had to make various decisions, including assessing whether this technology is suitable for achieving their public service goals, planning budgets, assigning tasks to subordinate agencies to manage the operation of the system, and developing monitoring protocols (**decision-making**). After CareCall started rolling out, frontline workers regularly monitored the call logs to see if any negative health signals were detected (e.g., skipping meals, poor sleep, health issues) or if the person did not answer several calls in a row (**frontline monitoring**). When any health concerns or consecutive missed calls were detected from the call logs, the frontline workers were notified to check with the person to see if everything was okay. If there was no response to these manual calls, they either visited the individual's home or escalated the matter to local public officers. Similarly, if social service or healthcare needs were identified during manual calls, they either directly connected the individuals to those services or wrote a report to escalate the issue to local public officers. Public agencies also took on various administrative tasks to coordinate among agencies at different levels, such as compiling the list of care recipients who needed manual check-ins from frontline workers and relaying the list to local public officers (**administration**).

Given the varied scales of CareCall deployments, each agency distributed roles in different ways. For example, some upper-level local governments (e.g., provincial or city governments) deployed CareCall on a relatively larger scale and assigned administrative and monitoring roles to their subordinate institutions (e.g., neighborhood community centers or local social service agencies), involving a few to around twenty workers in deploying the system in each

site. In contrast, at agencies that deployed CareCall on a smaller scale or were low on resources, one or two personnel undertook multiple roles needed for the entire process of adoption and rollout.

6.1.2 AI in the Public Sector

AI technologies are increasingly being proposed as a means to overcome resource constraints in public services, particularly in health and welfare. In community health interventions, AI tools have been explored to optimize resource allocation by identifying individuals who would benefit most [339, 132, 213, 161]. In social welfare contexts, researchers have examined how AI might assist in making fairer decisions while managing high volumes of social service requests and referrals in contexts like child maltreatment screening [44, 286, 303, 56, 154, 155, 153, 152], housing allocation for unhoused individuals [178], and job placement for unemployed individuals [226, 98]. While prior work on AI tools for public services has focused on resource allocation and decision-making support, AI can also assist in other essential public service tasks traditionally conducted by frontline workers, such as regularly collecting data from populations for public health monitoring. For example, prior work has shown how AI chatbots can be used to offload the frontline monitoring burden by automating the collection of personal health data [141, 142, 164].

While prior work has aimed to better understand the perspectives of those who directly interact with the AI systems in public services for their daily work, it has often overlooked other crucial stakeholder groups who might have substantially different needs around these tools [322, 56, 303, 131], such as decision-makers in public agencies involved in the AI adoption and other indirect users whose work has been impacted by the introduction of AI. Recently, a few studies have begun examining the perspectives of workers involved in AI-mediated decisions in varying roles (e.g., supervisors, agency leaders) [154, 155, 153, 152].

For instance, Kawakami et al. found that frontline workers had different target outcomes for child maltreatment screening than the AI tools but faced organizational pressures to disagree with the algorithmic decisions [154], suggesting the need to understand the viewpoints of those who hold higher power and responsibility to shape the adoption and rollout of AI tools in the public sector. In response to this call, a recent study engaged with decision-makers in public agencies to understand their views on adopting new AI tools, highlighting the differing perspectives of decision-makers and frontline workers on the validity and value of these tools [152].

In this chapter, I extend prior work by examining perspectives from public agency workers involved in the adoption and rollout of AI chatbots in varying roles in public health monitoring context.

6.1.3 Technology for large-scale health monitoring

Monitoring the health and wellbeing of large populations demands significant time and effort from public health agencies to conduct recurrent data collection [78, 133, 245, 250, 130, 322]. Prior work in the HCI and CSCW communities has often studied or proposed technologies, such has chatbots [131, 340, 141, 142] and mobile apps [205, 204], to support large-scale health monitoring have been studied or proposed in various public health contexts, including contact tracing [205, 204, 164, 55], maternal and child health education [261, 78, 250, 130, 245, 322, 133, 151], and social isolation intervention [141, 142]. Meanwhile, studies in the Ubiquitous Computing community have proposed more technical approaches to large-scale health monitoring through sensor technologies. Studies have frequently proposed sensor-based in-home monitoring as a mechanism to monitor various health indicators such as air quality [349, 124, 96, 227], water usage [100, 314, 101], and electricity consumption [95, 171, 256], particularly highlighting the benefits of piggybacking on public infrastructures in

improving the scalability of such technologies [314, 101, 171].

Although technologies are often developed with the intention of offloading the monitoring burden, they oftentimes bring about unintended consequences. Studies have shown that technology designed to support community health work often increases the strain on frontline workers by introducing additional responsibilities [157, 245, 133, 322, 310]. Operating these systems frequently requires substantial efforts from frontline workers to go beyond the job descriptions, but such efforts tend to be unacknowledged by other stakeholders [259, 322, 310, 221, 222]. Research has characterized such overlooked contributions of frontline workers as *"invisible work"* [73]—labor that is essential to their job but unnoticed, unacknowledged, or undervalued by other stakeholders [157, 245, 221, 222, 322, 310]. For example, in their study of data-driven technologies in a long-term care facility, Sun et al. revealed that frontline workers performed substantial data work to address the breakdowns in the data infrastructure—such as repairing incorrect or incomplete data collected through sensor technologies—but such work was largely neglected by other stakeholders under the guise of innovation [310]. Findings from previous studies suggest the need for an in-depth understanding of the full extent of the labor that stakeholders perform when introducing technology to the public sector.

In this chapter, I examine the perspectives of public agency workers across various roles on AI chatbots for public health monitoring.

6.2 Methods

To understand public agencies' expectations and realities of deploying AI chatbots for public health monitoring, I interviewed 21 public agency workers involved in the adoption and rollout of CareCall in varying roles, such as decision-making, administration, and monitoring.

This interview study was classified as exempt by our University's Institutional Review Board as the methodology did not involve more than minimal risk to participants.

Table 6.1: Information on the sites where participants were involved in the adoption and deployment of CareCall, including the scale of deployment (number of CareCall users), geographical characteristics, public service context, and target users

Site	Deployment Scale	Geographical Characteristics	Public Service Context	Target Users
SiteA	500	Mostly rural	Social welfare	Low-SES middle-aged and older adults living alone
SiteB	1,000	Mostly urban	Social welfare	Low-SES middle-aged and older adults living alone
SiteC	100	Urban, suburban	Public healthcare	Older adults living in isolated islands
SiteD	5	Rural, suburban	Public healthcare	Older adults with depression
SiteE	180	Urban	Public healthcare	Older adults with mild cognitive impairment
SiteF	40	Rural, suburban	Social welfare	Low-SES older adults living alone
SiteG	50	Rural, suburban	Social welfare	Low-SES middle-aged adults with chronic conditions
SiteH	10	Suburban	Social welfare	Low-SES middle-aged adults living alone
SiteI	30	Rural	Social welfare	Low-SES older adults living alone
SiteJ	300	Urban	Social welfare	Low-SES older adults living alone
SiteK	1,500	Urban, suburban, rural	Social welfare	Low-SES middle-aged and older adults living alone
SiteL	50	Urban, suburban	Tech incubation	Low-SES older adults living alone
SiteM	270	Urban	Tech incubation	Low-SES older adults living alone

6.2.1 Interview Process

In Fall 2023, I conducted individual semi-structured interviews with 17 of our 21 participants, while the remaining four participated in pairs with a colleague from the same agency (19 sessions in total). Interview sessions were conducted via conference calls (eight), phone calls (four), or in-person (seven) based on participant preference, each lasting 40-60 min-

utes. During the interviews, I asked about (1) their prior experiences monitoring health and wellbeing of populations using human and technological approaches, (2) the motivation for CareCall adoption and influencing factors, (3) the impact of integrating CareCall into their workflows and public health infrastructure. I offered all participants 60,000 KRW (roughly 45 USD at the time of the interviews) for their time. However, eight participants opted out of receiving study compensation to avoid the complex process of reporting external income as government officials, while the remaining thirteen were compensated.

6.2.2 Participants

Our research team recruited participants by distributing flyers to public agencies deploying CareCall (14) and via snowball sampling (7), consisting of 12 females and 9 males aged from 25 to 45 (Table 6.2). Eligibility criteria included individuals involved in the adoption and rollout of CareCall as part of their work for at least three months.

The workers I interviewed were from 13 different sites (SiteA-M) that deployed CareCall to achieve different goals depending on their public service contexts, such as social welfare, public healthcare, and technology incubation for public services (Table 6.1). For example, social service agencies deployed CareCall to monitor general health and well-being of low-SES middle-aged and/or older adults living alone, aiming to prevent lonely deaths. Conversely, community health centers adopted CareCall to monitor more specific health concerns, such as depression, mild cognitive impairment, or limited healthcare access (e.g., living in an isolated island). The scale of the deployment in these sites greatly varied, ranging from five to 1,500 individuals (Table 6.1). The geographical characteristics of these sites varied as well, including predominantly rural, suburban, and urban areas, as well as some with a mix of these elements (Table 6.1). Such a classification is based on factors such as population density, urban development, and regional characteristics [192].

Table 6.2: Participant demographics, including age, gender, and role in CareCall deployment. ID denotes their affiliated site.

ID (Gender, Age)	Role in CareCall Deployment		
	Decision-making	Administration	Frontline Monitoring
P-SiteA-1 (M, 42)		✓	
P-SiteA-2 (M, 35)		✓	
P-SiteA-3 (F, 43)		✓	
P-SiteA-4 (F, 41)		✓	✓
P-SiteB-1 (M, 35)	✓	✓	
P-SiteB-2 (F, 25)		✓	
P-SiteB-3 (F, 40)			✓
P-SiteC-1 (M, 42)		✓	✓
P-SiteC-2 (M, 41)	✓	✓	✓
P-SiteD-1 (M, 31)	✓	✓	✓
P-SiteD-2 (F, 35)		✓	✓
P-SiteE-1 (F, 28)	✓	✓	✓
P-SiteE-2 (F, 35)		✓	✓
P-SiteF (F, 43)	✓	✓	✓
P-SiteG (F, 34)	✓	✓	
P-SiteH (F, 30)	✓	✓	✓
P-SiteI (M, 38)	✓	✓	✓
P-SiteJ (F, 33)		✓	✓
P-SiteK (F, 35)		✓	✓
P-SiteL (M, 45)	✓	✓	✓
P-SiteM (M, 35)	✓	✓	✓

Depending on their resources and the scale of deployments, most of these workers were taking multiple roles in the adoption and rollout of CareCall. Ten workers participated in decision-making around the adoption and rollout of CareCall, 20 performed administrative roles, and 14 conducted frontline monitoring work (Table 6.2). Some of these workers were working with each other across institutions to manage CareCall deployments collaboratively. For example, P-SiteA-1 was in a provincial government and distributed some of the administration and monitoring tasks to their subordinate agencies, such as city governments (where P-SiteA-2 and P-SiteA-3 were based) and a social service agency (where P-SiteA-4 was based). Similarly, P-SiteB-1 made the decisions about the CareCall adoption

in a provincial government and assigned some of the administration and monitoring tasks to their subordinate agencies, including a community care center (where P-SiteB-3 was based) and a social service agency (where P-SiteB-2 was based). I also interviewed individuals who were working as teams managing CareCall deployments within the same institutions, such as P-SiteC-1 and P-SiteC-2, P-SiteD-1 and P-SiteD-2, and P-SiteE-1 and P-SiteE-2.

Workers had varying length of involvement in the adoption and deployments of CareCall, ranging from 3-6 months (six), 6-12 months (eight), and over 12 months (seven). Some participants had prior experiences with adopting technologies other than CareCall for public health monitoring. Six of them had experience in leveraging passive monitoring systems in the context of social isolation intervention, such as movement monitors (P-SiteB-1, P-SiteA-2, P-SiteA-4), call history monitoring systems (P-SiteI), and smart plugs tracking power usage (P-SiteH, P-SiteG, P-SiteI). Four workers had experience in deploying automated voice-based systems, such as answering machines (P-SiteA-2), rule-based chatbots (P-SiteF), and social robots (P-SiteD-1, P-SiteD-2) in similar contexts.

6.2.3 Data Analysis

I audio-recorded and auto-transcribed all interview sessions, manually correcting the automatic speech recognition errors in the transcripts later. All data were originally captured in Korean and were translated into English during the analysis process by myself, a native Korean and fluent in English. I paraphrased some idioms and phrasings to sound more natural in English and cross-checked the validity of the translation with one of the co-authors, Young-Ho Kim, who is also a native Korean and fluent in English.

I used inductive thematic analysis, characterized by the generation and constant comparison of open codes to reveal underlying themes [41], to qualitatively analyze the interview transcripts. I open-coded the interview transcripts and revised overarching patterns and

themes through several rounds of peer debriefing meetings. From this coding, I surfaced the main theme around the expectations and realities of deploying AI-driven chatbots for public health monitoring, using this to organize my results. The final codebook contained nine parent codes, such as expecting AI to help expand care and AI increasing frontline workload in practice, and 21 child codes. A central theme in these interviews was the varied forms of human labor needed by each stakeholder to manage the deployment of CareCall. This led me to use the framework of articulation work [308] to guide the discussion of my findings. My use of articulation work was bottom-up, informed by my findings rather than presupposed ahead of analysis.

6.3 Findings

In this section, I organize my findings by first introducing the prior experiences of agency workers in taking a human approach and using hardware-dependent technologies for public health monitoring. I then compare these experiences against the expectations and realities of using AI-driven chatbots (e.g., CareCall) for this task. Government agencies tended to introduce these new technologies sequentially given resource constraints and evolving policy priorities.

6.3.1 Prior Experiences with Human Approaches

Public agencies have predominantly relied on a human approach for social isolation interventions, such as phone calls or home visits. However, with a shortage of frontline workers relative to the population needing care, this approach was often perceived as overburdening. P-SiteA-2, as an administrator, acknowledged the lack of resources given to frontline workers: *"In our city, a single social worker handles over 100 socially isolated individuals, which*

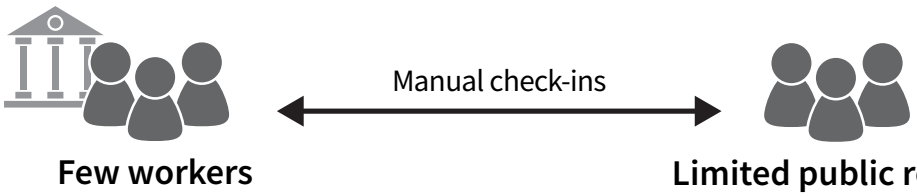

Figure 6.1: Human approaches—Public agencies traditionally monitored socially isolated individuals through manual check-ins, such as phone calls or home visits. With few frontline workers assigned to these tasks, only a limited number of people received regular monitoring.

isn't realistically manageable." Frontline workers described similar feelings toward a human approach. P-SiteF said: *"Checking in with elderly individuals is quite a burden. I'm responsible for about 1,000, making it unrealistic to check in with everyone."* P-SiteJ echoed this, stating: *"I'm in charge of hundreds of individuals, which means that I need to call dozens of people per day. It's too time-consuming and exhausting."* Both decision-makers and frontline workers felt the resource-intensive nature of the human approach often resulted in inadequate monitoring frequency and coverage of populations in need of regular monitoring (Figure 6.1).

Insufficient Monitoring Frequency

Due to resource constraints, all workers were concerned that a human approach did not enable them to engage in the monitoring tasks as frequently as they desired. The frequency that the frontline workers manually checked in with socially isolated individuals varied greatly, ranging from yearly to weekly check-ins, but they often felt more frequent monitoring was necessary to ensure the safety of the populations. P-SiteH, a frontline social worker in charge of monitoring bed-bound elderly people, said: *"I haven't been able to check in with them as frequently as I wanted due to resource constraints."* Administrators and frontline workers often attributed the gap between the desired and actual frequency of check-ins to the resources required for home visits. P-SiteD-2 described the difficulties of relying on home visits for monitoring in a rural region: *"In big cities, people typically visit a mental health center themselves, so it's much easier for social workers to check in regularly. But in rural*

towns like ours, elderly people can't visit us because our center is far from their homes and public transport is poor, so we need to visit their homes. But with limited resources, it is only one or two times a month, which is obviously not enough."

Limited Public Reach

All workers perceived that the resource-intensive nature of human approaches limited their ability to adequately cover populations in need of regular health monitoring. They frequently indicated that there were far more individuals who would benefit from monitoring, but only a small percentage of them were regularly monitored as they had to selectively prioritize those who were at greater health risks. P-SiteE-1, a social worker in charge of individuals who are at risk of dementia at a community mental health center, felt that they could not adequately cover the population through a human approach: *"Thousands of individuals with mild cognitive impairment are registered in our pool. We can't really check in with every one of them, so a lot of them have been unattended."* P-SiteL described a similar situation for monitoring socially isolated elderly people: *"In our city, over 25,000 older adults are living alone. Among them, those who are low-SES or have severe disabilities have social workers visiting them regularly for check-ins. However, many others are not receiving any regular check-ins."*

6.3.2 Prior Experiences with Dedicated Hardware

Recognizing that a human approach did not enable them to monitor as frequently or as much of the public as they would have liked, decision-makers in government agencies often adopted some hardware to monitor socially isolated populations. In this study context, several workers had experiences in leveraging passive monitoring systems in social isolation interventions—such as motion sensors, smart plugs that track power usage and ambient

light levels, or a call history monitoring system—which were designed to notify frontline workers when detecting unusual patterns that indicate potential medical emergencies. In addition, a few workers had experience in deploying automated voice-based systems—such as an answering machine, a rule-based chatbot, and a social robot for social isolation intervention. Frontline workers mentioned some instances where they found such technologies beneficial. For example, P-SiteH described how the smart plug helped discover a case of a lonely death: *"There was a person who was bedridden. He would leave the TV on the whole day and turn it off when going to bed. But one day, his power usage suddenly showed up as zero, so we tried reaching out and found he had passed away."*

Decision-makers generally found that dedicated hardware helped them slightly scale up monitoring beyond human approaches. However, frontline workers felt that such hardware-dependent technologies minimally alleviated the monitoring burden as they introduced new labor demands. Further, the high cost of hardware devices limited the reach of these approaches (Figure 6.2).

Introduction of New Types of Labor for Frontline Workers

Although sensor-based systems were introduced to assist in public health monitoring, administrators and frontline workers often found these systems added new labor demands instead of reducing their burden by frequently triggering false alarms and requiring constant follow-ups. Reflecting on her experience using smart plugs to monitor socially isolated people, P-SiteG noted that the sensors were error-prone, creating a significant burden: *"Since smart plugs measure changes in the environment, like ambient light levels, we got too many false alarms when a person forgot to turn off the light before going out or something. When we got those notifications, we were supposed to check if the person was okay immediately, even if it was evening or weekend, which was quite a burden."*

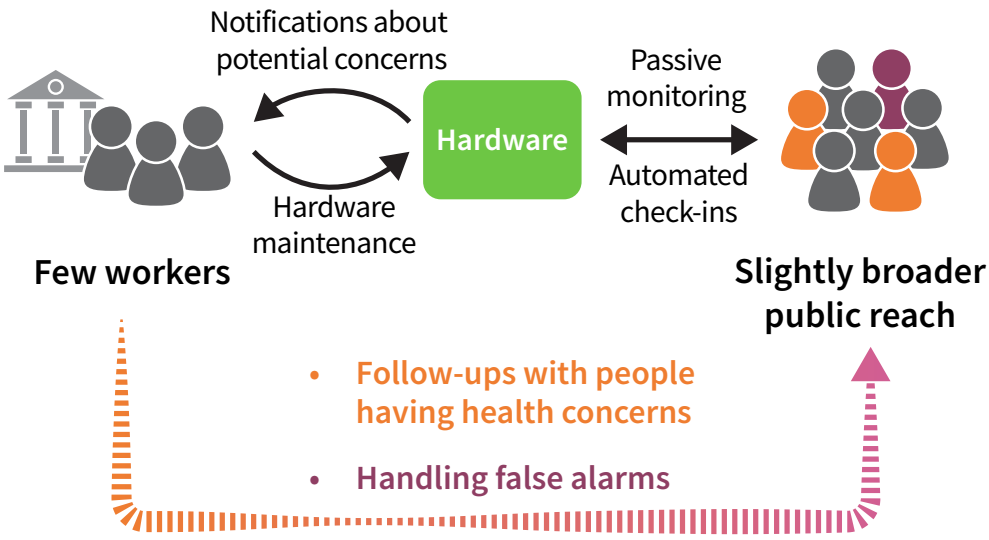

Figure 6.2: Hardware-dependent technologies—Decision-makers in public agencies adopted hardware for passive monitoring or automated check-ins to notify frontline workers to follow up with people having health concerns (highlighted in orange). However, frontline and administrative workers found these technologies minimally alleviated the monitoring burden, as they introduced new tasks—such as handling false alarms (highlighted in purple) and maintaining hardware, and the high costs only marginally expanded public reach compared to human approaches.

In addition, the hardware dependency of these systems led frontline workers to take on additional tasks for managing the devices. P-SiteG described the hardware maintenance tasks that the smart plugs required: *"We had a lot of difficulties managing the smart plugs, as the devices would sometimes break down or participants might lose them. We also had to get the devices back when participants dropped out to install them in new participants' homes, so it was a lot of work."* Similarly, while managing a social isolation intervention for older adults with depression using social robots, P-SiteD-1 faced administrative challenges in troubleshooting the devices for the participants: *"We have to do pretty much everything for the participants in terms of setting it up and troubleshooting. They often call us and say, 'It won't turn on,' so I would drive 40 minutes to their homes to check it out. Usually, it's just a temporary issue with their WiFi or something, so I would just restart the router and come back. It's definitely taking a lot of resources to maintain the robots."*

Limits to Scaling Up Monitoring

Decision-makers further perceived that the high cost of the devices constrained their ability to extend their monitoring reach to desired levels. P-SiteB-1 highlighted the financial barriers in providing care through motion sensors: *"We had to be quite selective when deploying the movement monitors because of the budget limit. There's still a long waitlist of people who want this service but haven't received it."* These costs motivated P-SiteB-1 to *"additionally adopt CareCall because it allowed us to reach a lot more individuals within the budget limit."* Other decision-makers who had experience or considered deploying smart speakers for older adults living alone also found that the high cost of the devices significantly limited their ability to scale up the intervention. P-SiteH explained that the main reason that their locale decided not to adopt smart speakers for social isolation intervention was the cost of the devices: *"There are plenty of smart speakers out there that can help monitor populations, but those are quite pricey. This limits our ability to expand care. I think budget is the primary factor when deciding which technology to adopt."*

6.3.3 Expectations for AI-Driven Chatbots

When adopting CareCall, decision-makers had expectations for the AI chatbot to overcome the aforementioned limitations of a human approach and hardware-dependent technologies, enabling more frequent monitoring for a larger number of people while alleviating the burden on workers (Figure 6.3). They decided to adopt the technology because they thought it would help scale up monitoring reach and frequency without needing to scale up workers. Decision-makers expected that they would be able to monitor significantly more of the public, effectively simulating human calls, at a lower cost than human approaches or passive monitoring.

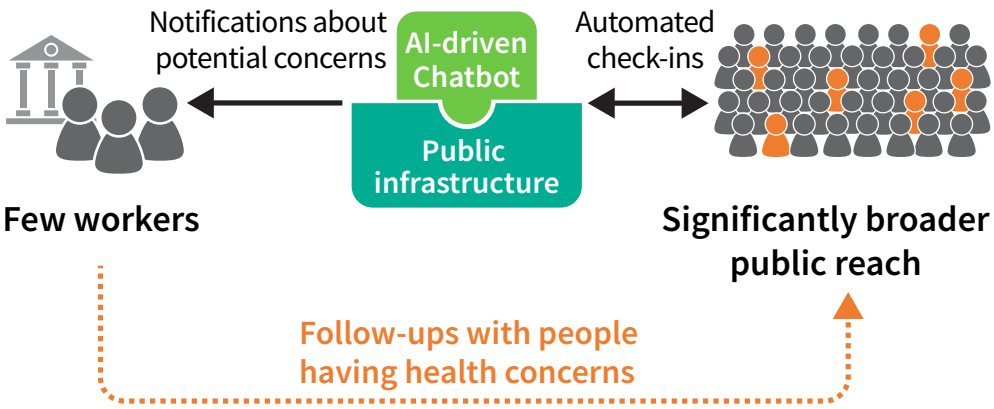

Figure 6.3: Expectations for AI-driven chatbots—When adopting CareCall, decision-makers in public agencies expected this AI chatbot to expand care to a much larger population through automated check-ins. They anticipated that the existing workforce could manage this expansion of care, as it would require follow-ups with only a small number of people who indicate health concerns (highlighted in orange).

Expansion of Public Reach with Increased Monitoring Frequency

I observed that decision-makers saw CareCall as a means to expand public reach through automated monitoring. They believed technology, particularly AI systems like CareCall, would help overcome resource constraints in public health monitoring. P-SiteE-2 sought technological solutions to expand care amidst resource constraints in community health: *"The number of people at risk of dementia has skyrocketed in our municipality, but our resources haven't really increased much. We figured we would need technology to more efficiently provide care to those people."* P-SiteA-1, who administered the rollout of CareCall, similarly believed AI would help overcome resource constraints and scale up care for socially isolated individuals: *"There are a lot of people who would benefit from care but haven't gotten any because of our limited resources. With AI, we hope to provide care to more people."*

As mentioned in subsection 6.3.1, many decision-makers believed more frequent monitoring was necessary to ensure the safety of these populations than what a human approach could achieve. Decision-makers and administrators envisioned that CareCall, as an AI chatbot performing automated check-up calls, would significantly increase monitoring frequency. P-

SiteC-1 believed that CareCall's automated monitoring would achieve the desired frequency of check-ins for isolated island residents: *"Our team has been visiting small, isolated islands to offer free medical check-ups for the past few years. But we only get to visit each island once or twice a year, so it wasn't really proper monitoring. With CareCall, we can now check in with them weekly."* Managing socially isolated older adults in a rural town, P-SiteA-2 also perceived CareCall would allow them to monitor more frequently: *"We used to visit each individual in our pool monthly at most because of resource constraints. Now we'd be able to check in with them more often with CareCall."*

Requiring Minimal Human Labor

When introducing CareCall, decision-makers further viewed it as an efficient AI tool that could offload the burden on frontline workers. As human approaches to monitoring tasks were often perceived as overburdening for frontline workers, decision-makers explicitly aimed to offload their burden through AI adoption. P-SiteG described: *"Our aim of adopting CareCall was to offload the frontline workers' burden. They've been asked to check in with people who are at risk of lonely deaths at least once or twice a week, but we hoped CareCall could check in on their behalf and ease their workload."* Frontline workers had similar hopes, such as P-SiteJ: *"Our city viewed that AI would offload our burden by automatically checking in with individuals on my behalf, allowing me to focus on those who are in greater need of care."*

Frontline workers believed that the AI system would require minimal human labor in frontline monitoring as it automates repetitive inquiries with the public and generates logs for review. P-SiteC-1 stated: *"I think the biggest strength of CareCall is that it saves time. AI automatically calls people regularly and asks different questions so that we can see whether they have issues in different aspects of health through the call logs."* P-SiteD-1 similarly mentioned: *"Before using CareCall, the only way that we could monitor how they were doing was by giving them a call or visiting their home. With CareCall, I would no longer need to*

do it manually; instead, I can see whether and what health issues they have through the call logs, so I can only follow up with people who need my attention."

6.3.4 Realities of AI-Driven Chatbots

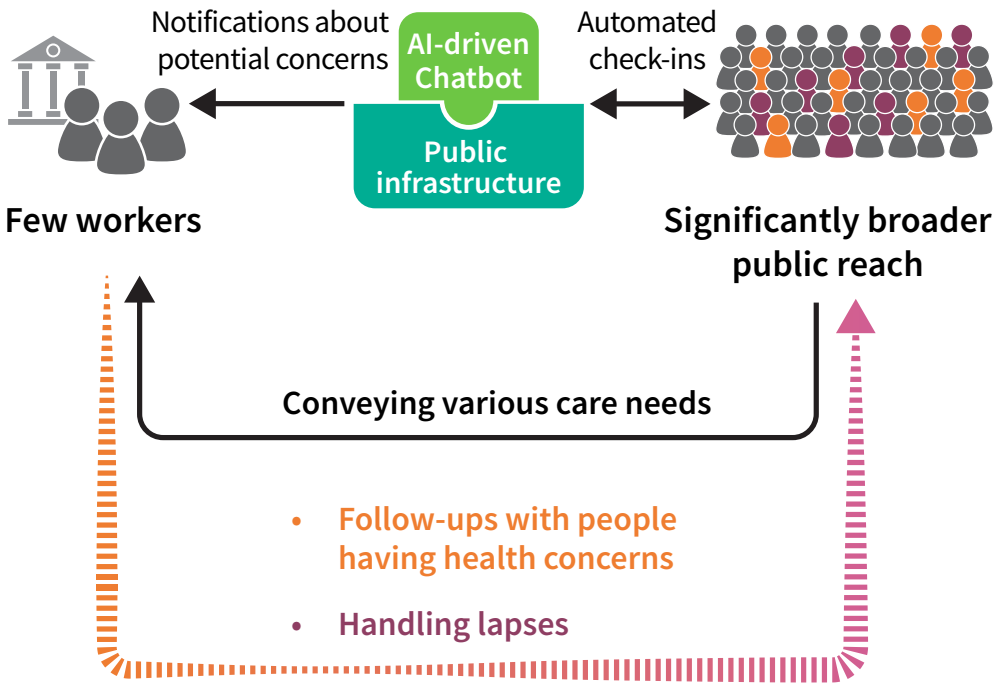

Figure 6.4: Realities of AI-driven chatbots—The introduction of CareCall fulfilled decision-makers' expectation to expand public reach, particularly due to piggybacking on public infrastructure, while also serving as a channel for individuals to communicate care needs. However, frontline and administrative workers felt that their workload increased as the expansion did not involve scaling up staff. Using AI chatbots also introduced new labor demands, requiring frontline and administrative workers to follow up not only with people with health concerns (highlighted in orange) but also with those who lapsed in engaging with this chatbot intervention (highlighted in purple).

When CareCall was introduced, decision-makers' expectations of expanding public reach and increasing monitoring frequency were largely fulfilled, as CareCall was indeed able to reach a broader audience than human approaches or dedicated hardware. However, all workers' expectations of requiring minimal additional human labor were unmet. Frontline workers often felt the introduction of CareCall exacerbated their workload by expanding care without necessary resources and demanding new types of labor. Meanwhile, frontline workers noted

an unexpected benefit, as CareCall provided a window to communicate different care needs (Figure 6.4).

Expansion of Public Reach with Increased Monitoring Frequency

In many government agencies, decision-makers felt that CareCall successfully expanded public reach and increased monitoring frequency through automated check-up calls. Decision-makers in public agencies introduced CareCall primarily to scale up care to underserved populations. P-SiteB-1 described: *"In our province, CareCall users are mostly individuals whom the frontline workers wanted to check in on but couldn't because they are at relatively lower risk."* P-SiteI similarly illustrated how CareCall helped scale up care: *"We've been visiting elderly veterans for check-ins, but we could only visit up to ten homes per day. We figured it would be nice to reach out to more people with the help of technology. With CareCall, we're now able to monitor around 100 individuals per day."*

All workers noted that CareCall enabled more frequent monitoring of care recipients. P-SiteE-1 valued that CareCall allowed weekly check-ins with individuals with mild cognitive impairment: *"We have thousands of people in our pool, so regular check-ins had been difficult. With CareCall, we can now check in more frequently."* Similarly, P-SiteC-1 valued that CareCall's automated monitoring allowed weekly check-ins with isolated island residents: *"CareCall lets us check in weekly. I love that the system automatically checks in with people and flags high-risk cases that need our attention."*

Decision-makers highlighted that one factor that enabled the expansion of public reach was CareCall's reliance on public infrastructure, specifically telephone networks. Unlike hardware-dependent technologies, CareCall offers chatbot interactions through scheduled phone calls, eliminating the need for dedicated devices and reducing adoption and maintenance costs. Decision-makers and administrators perceived that this reliance on public

infrastructure as key to scaling care within budget constraints, often comparing it to passive monitoring systems needing dedicated hardware. P-SiteA-2 emphasized how CareCall's low cost helped expand its reach: *"Sensor-based systems require substantial initial costs, so it was difficult to scale up. With $10,000 budget, we can reach around 1,000 people through CareCall, but only dozens with those sensor-based systems."* Decision-makers also compared the budget required for deploying CareCall to the one required for deploying smart speakers, another popular type of technology used in elderly care. P-SiteE-1, a social worker responsible for people with mild cognitive impairment, highlighted that CareCall significantly increased the number of people that they could reach within the same budget compared to deploying smart speakers: *"Our team is also deploying smart speakers for older adults so that they can play music, etc. With the limited budget, we can only reach ten people, whereas CareCall allows us to reach around 180 individuals with the same budget."* P-SiteH similarly valued that CareCall's independence of hardware reduced costs compared to smart speakers: *"Smart speakers are potentially helpful for monitoring populations, but they are expensive. We wouldn't be able to provide care to as many people as we want to. CareCall didn't require purchasing devices, which allowed us to provide care to more individuals."*

Although CareCall overall reduced costs for monitoring individuals, some decision-makers deployed CareCall in conjunction with hardware-dependent approaches to better align with how the public was interested in engaging with technology. P-SiteK monitored some individuals in their municipality with CareCall, and others with a smart plug, primarily for those resistant to using a chatbot but still needed monitoring: *"We realized that some people have reservations about talking to AI. We mostly deployed CareCall for those comfortable with AI, while using smart plugs for individuals who didn't like interacting with AI but still needed continuous monitoring due to their poor health."* P-SiteI similarly deployed a smart plug alongside CareCall, noting that passive monitoring was better for some of their municipality because of other health conditions: *"Many older adults experience hearing loss, so it's difficult for them to engage in proper phone calls. For those people, we installed the smart plugs*

to monitor power usage, while deploying CareCall for people with good hearing who needed some emotional support."

Expected Expansion of Care without Necessary Resources

Although CareCall was introduced to expand care, in reality, frontline workers felt that it was rarely accompanied by the necessary resources to support its implementation. Decision-makers often viewed AI systems like CareCall as highly efficient and requiring minimal human labor. As a result, even when the AI tool was introduced to expand public reach and increase monitoring frequency, they did not hire more staff members to manage the expanded care in most cases, leaving existing frontline workers and administrators to take on extra tasks. For example, with the introduction of CareCall, P-SiteK was assigned to monitor logs from around 1,500 people on her own, which she felt was unmanageable: *"Our province set the goal to provide the service to 2,000 people from the beginning. Currently, I'm the only person monitoring CareCall logs in our province, and I also have other responsibilities. I can't really keep up with monitoring that many people."*

Frontline workers and administrators described how decision-makers' perceptions of AI might have impacted resource allocation for AI interventions. P-SiteC-1, a frontline worker at a medical center, shared his perspective on the possible rationale behind the insufficient resource allocation: *"Decision-makers probably want AI to take care of 1,000 people and have frontline workers focus on 10% at higher risk. They wouldn't want to hire more people when introducing AI because it's supposed to be efficient. Introducing AI is not to allocate more resources."* P-SiteB-1, who made decisions about the deployment of CareCall in their province, acknowledged they underestimated the resources required to manage CareCall: *"We originally planned on deploying the system with 2,000 individuals but had to cut down to 1,000 when we saw how much it disrupted their other responsibilities. Like, they could barely keep up with their primary tasks because they had to spend almost two full days every week just*

for CareCall monitoring."

Frontline workers often felt that the monitoring tasks introduced by CareCall substantially distracted them from other responsibilities. P-SiteB-3's primary role as a frontline social worker was to assign home care aides to bedridden elderly individuals. However, with CareCall's introduction, the provincial government tasked her with monitoring logs from 600 people—an entirely new and additional responsibility beyond her existing duties. P-SiteB-3 described the impact of CareCall on her frontline workload: *"I recognize how CareCall can be helpful in some cases. I would have loved it if checking in with older adults was my day job because it can check in with hundreds of people in just a few hours, a task that would have taken me a whole week. But the thing is, my day job is sending home care aides. For me, CareCall added a whole new set of tasks on top of it."*

Frontline workers thus highlighted the need to allocate additional resources to handle the expanded care with CareCall. For example, P-SiteA-3 stated: *"When public agencies start new projects, existing frontline workers always end up with additional tasks, even though we already deal with heavy workloads. There should be dedicated staff that can focus on CareCall monitoring."* All workers argued that introducing AI would only increase their burden without extra resources, as it inevitably requires human labor. P-SiteG stated: *"The aim was to offload public officials' burden, but if the project is managed by existing workers, it only adds to our burden. AI does the monitoring, but follow-ups need human labor."* In this climate, a couple of government agencies began subcontracting CareCall monitoring tasks. P-SiteB-1, a decision-maker at a provincial government, described: *"We're planning to make budgets to hire people solely for CareCall monitoring next year. Without this, our social workers can't keep up with their other responsibilities."*

Introduction of New Types of Labor for Frontline and Administrative Workers

As mentioned in section 6.3.3, many decision-makers hoped CareCall would reduce the burden on frontline workers by automating repetitive check-up calls. When introduced to real-world settings, some administrators found it helpful for reducing the need for hardware maintenance, as it piggybacked on public infrastructure, unlike prior hardware-dependent technologies. P-SiteG valued that CareCall did not require any device maintenance—including handling breakdowns and device losses, installing devices for new participants, and retrieving devices from dropouts—unlike smart plugs that they had previously handled the administration and rollout of: *"Managing smart plugs was a lot of work. CareCall doesn't really require any of that."* Similarly, P-SiteI highlighted the convenience of deploying CareCall in comparison to administering social robots: *"Those robots need to be charged regularly and would occasionally have technical issues. I like that CareCall requires minimal effort in managing the system."*

However, frontline workers felt that CareCall's introduction overall exacerbated their burden by demanding new types of labor. CareCall was designed to notify officials in charge of individuals who eventually did not answer after three call attempts, with each government having its own protocols for handling such lapses. For example, SiteB had 'the same-day protocol,' requiring frontline workers to check in on every individual who missed CareCall calls the same day. Although the goal was to ensure the safety of care recipients, frontline workers often found such protocols overburdening. P-SiteB-3, a frontline worker assigned to monitor 1,000 people after CareCall's introduction, said she had to spend a full day or two every week just for follow-ups on missed calls: *"Every Wednesday, I have to mentally prepare myself before checking the call status. We never know the turnout, so I get a bit anxious every week. Even though CareCall tries three times, about 10% of the users don't answer, so I have to call around 100 people on average every Wednesday."* P-SiteF also perceived adhering to the protocols for handling lapses was challenging: *"We get anxious*

when the participants don't answer calls. If I can't confirm the person is doing okay through CareCall, human calls, or visits, we have to call 911 to forcibly open their front door. It would be a major inconvenience for them, so I'd rather not have to do it." Administrative workers also reported that handling lapses required them to undertake additional tasks to coordinate between multiple institutions. P-SiteB-2 described the complex coordination process of addressing lapses: *"The community care center lists individuals needing public officers' attention after monitoring the call logs and relays the list to us. We then relay it to the Provincial government so they can reach out to officers in each neighborhood. The whole process is very complex."* These administrative tasks for handling lapses were perceived as time-consuming and burdensome by other workers as well.

Frontline workers viewed handling lapses in user engagement with CareCall as redundant, as these were not typically due to health concerns. P-SiteC-2, responsible for isolated island residents, thought lapses often occurred because participants disliked AI: *"CareCall assumes something might have happened if someone misses calls, but often, people don't answer calls simply because they don't like talking to AI, not because something bad happened. We now have to try calling them a few times and even ask local public officials to visit when they miss CareCall. But honestly, I'm skeptical about putting in so much effort."* P-SiteI, who manages elderly veterans, noted that lapses were often due to forgetfulness, rather than emergencies: *"We send out CareCall calls at a set time every week, so we ask them to be on the lookout. However, elderly individuals often leave their phones somewhere and forget. We keep calling them until they answer, but there's not much we can do if people don't pick up the phone."* To minimize redundant tasks, frontline workers suggested expanding AI's role by further automating follow-ups. P-SiteJ stated: *"For now, we're responsible for reaching out to those who didn't answer CareCall, but I wish it could try again the next day automatically."* P-SiteI similarly wished, *"the system requires minimal or no human intervention. It could try a few more times without prompting us to follow up manually."*

In addition, frontline workers noted that lapses in user engagement with CareCall created additional tasks of handling callbacks. Since the chatbot used frontline workers' office numbers, participants frequently returned calls to their offices after seeing missed calls from CareCall. However, frontline workers frequently found these callbacks overburdening. P-SiteG explained: *"I frequently get callbacks, and people often want to chat while they are at it. No wonder, given we recruited people needing emotional support. But it adds up quickly and becomes overwhelming."* P-SiteB-1 similarly described the difficulty of handling callbacks: *"I get a lot of callbacks from CareCall users. The thing is, it's difficult to keep the call short because they often want to chat and ask about social services, which can easily take 30 minutes per call."*

Creation of a Window to Communicate Different Care Needs

I further found that introducing CareCall unexpectedly served as a window for care recipients to communicate different needs, allowing frontline workers to provide the necessary support. The open-ended nature of LLM-driven chatbots led users to express various healthcare and social service needs when interacting with CareCall. Although CareCall was not targeted at processing such requests, frontline workers valued the ability to identify these needs through call logs and often took action on them.

Frontline workers occasionally discovered individuals' mental health needs through CareCall call logs, which led them to check in on them more frequently. P-SiteA-2 stated: *"We make sure to visit and check in with them if people frequently mention feeling depressed and lonely during CareCall calls. We then connect them to mental health support or job search assistance when necessary."* Working at a mental health center, P-SiteE-1 particularly paid attention to expressions of emotional distress when monitoring CareCall calls and followed up when concerns arose: *"We call when we find something stands out in the call logs. For example, if someone says, 'I'm so depressed. I just want to die,' I make sure to reach out.*

In many cases, just listening to whatever they want to say makes them feel better."

In addition, in a community health center setting, frontline workers were able to encourage individuals to seek clinical care when they noticed physical health concerns from the call logs. P-SiteC-2 described taking action on an individual's health issue through monitoring CareCall logs: *"Many elderly people take sleeping pills when they frequently wake up at night, but in fact, these issues are often due to urological issues. When we noticed someone repeatedly mentioning their sleep problems, I followed up and encouraged them to visit us to see a urologist. After the visit, thankfully, their sleep issues were resolved."* P-SiteC-1 similarly explained how health concerns expressed during CareCall calls helped them connect to necessary healthcare: *"I noticed someone mentioning severe back pain during the calls. I followed up and encouraged them to visit us so that we could get an X-ray and provide some physical therapy."*

Further, frontline workers identified and addressed social service needs upon monitoring CareCall logs, even though those were beyond the intended scope. P-SiteB-3 explained their desire to connect individuals to social services through CareCall monitoring: *"We know that CareCall is just for check-ins, but we wanted more from the start. Like, connecting people to relevant social services if needs arise, such as job searching or sending home aides."* P-SiteB-2 added that they would note details like, *"someone had surgery recently and asked for financial aid"* when monitoring call logs so local public officers can connect them to resources. P-SiteL described how CareCall turned out useful for natural disaster recovery: *"Last monsoon season was pretty bad in our city, and people mentioned issues like water leaks or floods during CareCall calls. During the season, I paid extra attention to the call logs and reached out to the community centers for help, like fixing their houses. I know the system isn't meant for emergency responses, but it worked out well."*

6.4 Discussion

My findings reveal a discrepancy between public agencies' expectations and the realities of deploying CareCall, as it required significant human effort to manage the expanded care, contrary to expectations that it would alleviate the burden of frontline public health monitoring. This mismatch between perceptions and reality of the adoption of LLMs parallels conversations in other spaces, where the technology is often assumed to have greater capabilities than it can deliver on [310, 131, 132]. Looking at my findings from the perspective of articulation work [308], I highlight that decision-makers in public health face unique challenges in conducting articulation work required for AI chatbot adoption, particularly due to the open-ended nature of LLM-driven chatbots. I point to the need to develop guidelines and best practices for decision-makers implementing these emerging technologies in public health contexts. My findings also surface that CareCall introduced significant maintenance work for frontline workers, primarily due to unmet expectations around user engagement with these chatbots. I argue for the importance of acknowledging and accounting for the maintenance work that these AI chatbots demand in public health monitoring. In addition, I provide implications for public agencies considering the use of AI chatbots for public health monitoring, focusing on the potential of open-ended conversations to identify unmet care needs and the need to assess the impacts of AI adoption on the labor demands of their workforce. For designers and developers aiming to make AI chatbots usable for public health monitoring, I suggest opportunities to piggyback on public infrastructure, incorporate fallback mechanisms to address lapses, and leverage passive sensing to complement chatbots. Finally, I report on the limitations of the study, particularly concerning the transferability of the findings to different countries and domains.

6.4.1 Considering Decision-Makers' Articulation Work for AI Chatbot Adoption

Through this study, I found that decision-makers in public agencies often expected AI chatbots to reach more people. These expectations were largely realized, but decision-makers often failed to plan adequately for the necessary resources to operate these systems effectively. Overall, these AI chatbots failed to deliver on the expectation of scaling up monitoring reach and frequency without needing to increase staff. Decision-makers initially perceived AI chatbots like CareCall as highly efficient and requiring minimal human oversight, having unrealistic expectations about their capabilities for public health monitoring. Such perceptions might have been influenced by their personal interactions with commercial LLM-driven chatbots (e.g., ChatGPT, Gemini) performing naturalistic conversations on diverse topics, as well as broader media and cultural conversations around these technologies. However, in reality, frontline workers and administrators had to take on significant human effort to manage the expanded care with these AI chatbots, such as following up on health concerns and handling lapses. This discrepancy led to failures in planning for the human resources required to operate AI for large-scale health monitoring, adding extra work to already overburdened frontline workers who perform crucial care work with limited resources.

Extending prior CSCW research that sheds light on the articulation work required for introducing new technology into complex healthcare infrastructures [107, 308, 287, 29], my findings highlight that decision-makers may face greater challenges in conducting articulation work for allocating resources for LLM-driven chatbots due to the unrealistic expectations about the capabilities of these technologies. The lack of established guidelines and best practices for these emerging technologies in public health space further introduces greater uncertainty about their capabilities and limitations, as well as the human efforts necessary to operate them. Given these challenges, developing guidelines and best practices for decision-makers in implementing LLM-driven chatbots in public health contexts is a valuable direction

for future research. Similar to the toolkit proposed by a recent study [153], such resources could provide valuable insights to support the articulation work of decision-makers and help them develop more realistic expectations of the capabilities of these systems. By aiding them in navigating the complexities of implementing emerging technologies like LLM-driven chatbots, decision-makers could better estimate the human efforts necessary to operate these systems in a more sustainable manner.

6.4.2 Accounting for Maintenance Work AI Chatbots Impose on Frontline Workers

It can be expected that any change to existing practices in complex public health contexts will introduce new types of labor. However, in this study, I found that the adoption of CareCall introduced not only new labor but also labor that decision-makers in public agencies did not anticipate, such as handling calls that the AI chatbot made but were not responded to. Consistent with prior work [283, 322, 310], my findings reveal that the frontline and administrative workers had to perform substantial maintenance work due to AI chatbots failing to meet some expectations the decision-makers had. The adoption of CareCall for public health monitoring operated under the assumption that missed calls indicate potential health emergencies, implying that users would mostly answer the calls when not in emergencies. However, in reality, users frequently lapsed in the use of CareCall because they simply forgot about it or did not want to interact with AI. This suggests a misunderstanding among decision-makers that people would consistently and willingly communicate with AI chatbots in public health monitoring contexts. Research on Personal Informatics has shown that people commonly lapse in the use of health monitoring technology in general, both intentionally and unintentionally [92, 91]. In addition, prior studies have highlighted that people are often hesitant to interact with chatbots, preferring instead to engage with humans behind chatbot-based health interventions [261, 182, 151]. Building on prior work, my study sug-

gests that public health monitoring involving AI chatbots inevitably requires maintenance work to address lapses, whether due to people's patterns of behavior with health monitoring technology in general or their reluctance to interact with AI chatbots.

Previous studies have pointed out that technology could be used as a means to demand additional expectations for frontline workers [318] or normalize increasing their workloads in the name of innovation [310]. Extending prior work, my study suggests that while integrating AI chatbots likely introduces substantial maintenance work to address the limitations of these tools, decision-makers can easily overlook them under the guise of innovation, burdening frontline workers with additional and potentially redundant tasks. Reflecting prior work [221, 322, 25], my findings suggest that an important avenue for improving AI adoption in public health monitoring is to adequately recognize the efforts of workers in maintaining care infrastructure during breakdowns. To avoid overburdening the public health workforce, I recommend that decision-makers pay close attention to the types of tasks AI could add to care infrastructure, who will perform the additional and potentially invisible work, and how such work can be better recognized.

6.4.3 Implications for Public Agencies

In this section, I highlight implications for public agencies looking to leverage AI chatbots for public health monitoring. My findings surface that CareCall unexpectedly served as a conduit for communicating various social service and healthcare needs. As an open-domain, LLM-driven chatbot, CareCall supports free-form conversations on serendipitous topics users bring up, allowing users to convey various care needs that traditional task-oriented systems with pre-defined conversation flows might miss [188]. My study suggests that open-domain, LLM-driven chatbots can play a valuable role in care infrastructures as the safety net for vulnerable populations. I recommend that public agencies **leverage the public's open-**

ended interactions with LLM-driven chatbots to uncover unmet care needs. When integrating these chatbots into care infrastructures, it would be essential to establish comprehensive mechanisms to monitor unmet care needs and refer them to relevant social or healthcare services.

However, previous studies highlighted that stakeholders often have unrealistic expectations towards AI systems in the public sector and ascribe more capabilities than they actually can offer [245, 153, 141]. In addition to decision-makers overestimating their capabilities of LLM-driven chatbots, the public may be similarly influenced by personal experience and public discourse around these technologies, potentially leading to disappointment when they cannot receive the care they desire [141]. To maintain realistic expectations of these chatbots, it is crucial for public agencies to transparently communicate system capabilities and limitations with end-users and clarify what public agencies can and cannot offer by monitoring the data collected through these systems.

Further, my findings suggest that adopting AI chatbots necessitates rethinking and reconfiguring the labor involved in public health monitoring. Before introducing technological interventions, frontline public health monitoring primarily involved giving calls or visiting homes to ask routine questions about the health and wellbeing of individuals under their care. In contrast, with CareCall, frontline workers shifted their focus to following up on those who expressed health concerns and addressing lapses. In short, the AI chatbot took over some of the labor expected of frontline workers and introduced different labor in its place. If these AI technologies are widely put into practice for public health monitoring, what I observed in CareCall suggests that the types of labor performed by frontline workers will shift, as well as the kinds of expertise and training needed for these workers to be effective. Our participants frequently stated that they felt that working with CareCall required tasks they had not been trained to do or were well outside their areas of expertise. I do not aim to argue that this shift is positive or negative as a whole. Rather, I urge public agencies to

critically assess the impacts that introducing these AI systems will have on the day-to-day practices of their workforce as part of deciding whether and how to adopt them.

A core question that decision-makers in public agencies need to consider is the various costs (e.g., financial, morale) associated with re-training frontline workers or hiring new ones to manage the work introduced by AI chatbots. However, given that the day-to-day experiences of multiple stakeholders are impacted by such decisions, I see a need for decision-makers to better acknowledge the new labor placed on existing frontline workers. Prior work has pointed out the clear solution of increasing worker compensation or hiring additional workers [221], though this approach often faces barriers in typically underfunded public agencies. Other approaches include increasing the visibility of this additional work [222, 322], such as through time-tracking, though these can lead to feelings of surveillance. More participatory approaches, involving all stakeholders in the conversation of whether and how to adopt an AI chatbot for public health monitoring, can potentially address some of these concerns and risks [44, 303, 226, 98].

6.4.4 Implications for Designers and Developers

In this section, I offer implications for designers and developers aiming to make AI chatbots usable for public health monitoring. One significant factor that enabled CareCall to meet stakeholder expectations around expanding reach was it piggybacked on existing public infrastructure, specifically telephone networks. Unlike technologies requiring dedicated hardware, CareCall's chatbot interactions were conducted via phone calls using existing telephone lines. This approach significantly lowered costs, allowing broader reach within public agencies' budget constraints, and reduced the burden on frontline workers and administrators, who would otherwise have been tasked with managing hardware. Research

in the HCI community has increasingly underscored the need to consider the scalability of health technologies [324, 209]. When developing chatbots for large-scale health monitoring, building on existing infrastructure [319, 314, 101] or social platforms [93, 111] can enhance scalability, as this approach can lower development and management burdens [93] and facilitate broader engagement in low-resource settings [319]. Consistent with prior work, my study highlights **piggybacking on public infrastructure as a promising strategy to address the scalability challenges** of public health monitoring chatbots. When deciding whether and what public infrastructure to piggyback on, designers need to carefully consider its impact on end-user interactions and broader stakeholder workflows.

In addition, my study points to opportunities for developers to **incorporate fallback mechanisms to address lapses in user engagement** with chatbot-based public health monitoring. While lapses are to be expected [92, 91], they are often respected in other contexts. However, in critical health contexts where public health monitoring is often deployed, such as for the prevention of lonely deaths, such lapses could indicate serious health emergencies or even death, and it may not be beneficial to outright ignore them. Incorporating fallback mechanisms can help make chatbot-based public health monitoring more resilient to lapses in user engagement in the long term. Prior work pointed to the opportunities for using secondary sources that generate data as a byproduct of the daily digital lives of individuals—such as social media posts and app usage [91, 311, 77, 230]—or in-home environment monitoring—such as water usage or electricity consumption [314, 101, 171]. While my findings revealed that public health officials often view such passive sensing approaches as too error-prone to serve as the primary method of monitoring personal health, they could be effective as fallback mechanisms for addressing lapses in user engagement. These approaches, as secondary sources of producing data, do not require additional effort from individuals, reducing frontline workers' burden of handling lapses in chatbot-based public health monitoring. One important factor to consider when leveraging passive sensing approaches as fallback mechanisms for public health monitoring is whether dedicated hardware devices are required

because it likely introduces additional labor to maintain them and limits public reach due to the cost. I suggest that developers carefully evaluate the opportunity to leverage existing public infrastructure as a fallback mechanism to chatbot-based public health monitoring, while also considering whether approaches that do not build on existing infrastructure could provide additional value.

Beyond addressing lapses in user engagement, I further see opportunities for **passive sensing to complement chatbots to align with various public health monitoring needs**. Public officials highlighted how individuals often had skepticism or concerns around the use of conversational AI for monitoring, and providing an alternative technical approach helped them monitor this group while aligning with their preferences. Further, officials highlighted that some chronic conditions common in their population, such as hearing loss, were not amenable to voice-based chatbot check-ins, and having an alternative was beneficial. Beyond serving as an alternative, there are likely opportunities for passive sensing approaches to deepen understanding of the public's daily experiences. For example, understanding energy consumption patterns via in-home sensors could triangulate self-reported behaviors via chatbots, creating a better picture of how an individual living alone is doing. However, my findings suggest that public agencies often adopt new technological interventions sequentially given resource constraints, evolving policy priorities, and emerging technology trends, so care must be taken when designing integrated approaches in ways that can be readily adopted.

6.4.5 Limitations and Future Work

My goal for this study was to understand the expectations and realities faced by public agencies deploying AI chatbots for public health monitoring through the case of CareCall, which led me to focus on the perspectives of public agency workers who were involved in its adoption and rollout in South Korea. I believe that many of the circumstances that our

interviewees described—such as the lack of resources in public agencies and the demanding working conditions of frontline workers—are not unique to the South Korean context or the specific deployment of CareCall, suggesting that similar expectations and challenges could arise when AI chatbots are rolled out for public health monitoring in other countries. However, I acknowledge that various country-specific factors—such as regulatory requirements, cultural norms, technological infrastructure, public trust in AI systems, and the overall maturity of digital health initiatives—can significantly influence how public agencies approach AI adoption and deployment for public health monitoring. As such, it is crucial to consider these contextual differences when applying my findings to other regions.

My study focused on the context of CareCall, a system designed to monitor the health and wellbeing of socially isolated individuals, primarily low-SES middle-aged and older adults living alone. Using AI chatbots in public health monitoring for different and broader populations—such as crisis management or chronic disease monitoring—likely involves different interpersonal and infrastructural dynamics. For instance, crisis management often involves real-time monitoring and decision-making, which may involve different labor demands compared to the relatively stable, routine monitoring of individuals' wellbeing. Further, most CareCall deployments in my study context were pilot projects implemented on relatively small scales. When AI chatbots are deployed on a larger scale (e.g., state-wide or nationwide), public agencies may engage in more robust resource planning and role assignment than what was seen in my study. Future research should explore how these factors play out in different public health domains and at various scales of deployment to better understand the perspectives and practices around AI adoption and rollout.

Finally, I recognize that participants were describing their expectations for CareCall retrospectively, and the descriptions they provided were likely influenced by their actual experience with the system. Nonetheless, I expect that participant descriptions were fairly reliable, as they largely lined up with typical expectations that AI technology can help reduce work-

load and expand scale. Further work would benefit from investigating how public health workers' perceptions of AI chatbots change as they more deeply understand the technology's capabilities, such as participatory methods with longitudinal engagement.

6.5 Conclusion

Through interviews with 21 public agency workers involved in the adoption and deployment of CareCall across decision-making, administration, and frontline monitoring roles, I found that public agencies' expectations for AI chatbots to expand reach were largely met, but frontline workers often experienced an increased burden due to insufficient resources and new labor demands, such as handling lapses in user engagement. My findings suggest that the open-ended nature of LLM-driven chatbots and the lack of established guidelines around these emerging technologies introduce unique challenges for decision-makers when conducting the articulation work required for AI chatbot implementation. I also highlight the importance of recognizing the maintenance work that AI chatbots impose on frontline workers, especially considering end-user lapses in using these systems. For public agencies, I suggest leveraging open-ended conversations of LLM-driven chatbots to identify unmet care needs and critically assess the impacts of AI implementation on the labor demands of their workforce. For developers, I suggest piggybacking on public infrastructure, incorporating fallback mechanisms to better address lapses in user engagement with AI chatbots, and leveraging passive sensing to complement AI chatbots for public health monitoring.

Chapter 7

Discussion and Conclusion

Through my dissertation, I explored how health monitoring technologies can better support collaboration within clinical and public health infrastructure by accounting for infrastructural complexities. In this chapter, I synthesize findings across Chapter 3 to 6 in light of my thesis claim. Insights from my studies highlight that implementing flexibility and amplifying existing stakeholder practices can strengthen collaboration, while also revealing tensions that arise when designing for these goals within clinical and public health infrastructures.

7.1 Improving Collaboration in Clinical Infrastructures (T1)

7.1.1 Implementing Flexibility to Balance Stakeholder Constraints

Through my dissertation, I showed that **implementing flexibility in health monitoring technologies can help stakeholders address constraints within clinical infrastructures.** In my thesis, I conceptualize *flexibility* as the capacity of technologies to

accommodate complex strategies—including providers' care regimens or patients' tracking regimens—support iterative adjustments, and leave room for human judgment in the face of incomplete or evolving infrastructures.

In Chapter 3, I demonstrated that **implementing flexibility can help providers navigate complex infrastructural constraints during care planning**, including challenges related to insurance approvals, pharmacy regulations, and patient circumstances. For example, pharmacies and insurance companies often imposed constraints that interfered with providers' taper planning using clinical decision support tools, due to a lack of policies and protocols to support tapering regimens. In response, providers frequently resorted to workarounds to adapt these tools to these infrastructural constraints. These findings align with prior work that emphasized the need to attend to the sociotechnical complexity in the design of clinical tools [30, 31, 235, 264, 134]. My work contributes to this body of literature by highlighting that the successful implementation of health monitoring technologies requires flexible mechanisms to support providers in managing various infrastructural demands during care planning. While prior tools have primarily focused on improving clinical precision, such as preventing errors and ensuring adherence to guidelines, this emphasis can sometimes limit the flexibility that providers need to navigate real-world complexities, such as communicating with pharmacies and accommodating insurer requirements. My findings suggest that supporting such flexibility is essential for aligning decision support tools with the realities of care planning within real-world clinical infrastructures.

In Chapter 4, I demonstrated that **designing for flexible patient input within clinical measures can support more patient-centered communication amidst the logistical constraints of clinical settings.** Through this work, I found that while patients often enter clinical visits with a clear sense of what they want to communicate, standardized self-report measures can limit their ability to express the complexity of their symptoms. As prior work has shown [72, 62, 63, 122, 262, 12, 13], these tools tend to distill experiences

into the presence or absence of common systems, often overlooking how patients actually perceive and live with them. This work shows that enabling patients to ground their lived experiences in clinical measures could help bridge this communication gap. While patients' diverse forms of self-expression may initially appear misaligned with provider expectations or incompatible with the time constraints of clinical visits, contextualizing lived experiences within structured clinical forms can help mitigate these tensions. Unlike passive data collection, which is often viewed as an alternative to self-report scales, this approach preserves patient voices and provides more authentic representations of lived experiences while remaining compatible with existing clinical infrastructures. This work highlights an opportunity to enrich patient-provider communication through patient-driven annotations while accounting for the logistical realities of clinical care.

The infrastructural challenges highlighted in Chapters 3 and 4 should be understood in light of their specific institutional context. In the United States, where these studies were conducted, clinical infrastructures function within a multi-payer healthcare system, requiring providers to navigate fragmented systems with private insurers as well as retail pharmacies. This fragmentation often requires providers to figure out whether certain medications or regimens would be approved or reimbursed under the diverse insurance schemes held by individual patients. These dynamics contrast with single-payer systems, such as those in South Korea, Taiwan, and Canada, where providers operate within a centralized national insurance system that offers uniform coverage and allows prescriptions to be filled at any pharmacy.

These differences highlight how infrastructural complexity is shaped not only by practices of individual stakeholders, but also by broader institutional arrangements—such as the degree of centralization in healthcare systems and standardization of insurance policies—that govern clinical decision-making and coordination. While my studies conducted in the United States point to flexibility as a crucial design strategy for helping providers work around fragmented

infrastructures, in more standardized and centralized systems like those in South Korea, Taiwan, or Canada, flexibility in clinical systems may play a different role, such as supporting nuanced clinical judgment and patient-centered communication within tightly regulated and protocol-driven workflows. Future work is needed to examine how the role of flexibility might differ between single-payer and multi-payer healthcare systems, particularly in relation to how institutional structures mediate the day-to-day work of providers.

7.1.2 Tensions in Implementing Flexibility to Balance Stakeholder Constraints

While my dissertation demonstrated that implementing flexibility in health monitoring technologies can help both providers and patients navigate infrastructural constraints, it also revealed tensions in implementing flexibility within clinical infrastructures.

In Chapter 3, I showed that **providers' training and level of experience influenced their desire for flexibility in care planning.** Providers with more experience in antidepressant taper planning, such as psychiatrists, preferred to rely solely on their own experience. In contrast, providers with relatively less experience in antidepressant taper planning, such as primary care providers, sought more automated guidance to offload the decision-making burden to technology. Given that many health conditions—such as diabetes, hypertension, and depression—are treated in both primary and specialist care settings, this work suggests that providers' varying levels of experience should be carefully considered when incorporating flexibility into health monitoring technologies. While experienced providers may value flexible tools to support regimens involving complex strategies and iterative adjustments, those with less experience might struggle to use such tools effectively, particularly in domains where clinical guidance is limited. Prior work on clinical decision support tools suggested the benefits of AI-powered recommendations [195, 191, 190, 27, 114, 50], which

can be especially helpful for less experienced providers navigating the challenges of developing care regimens in the absence of established standards. In contrast, providers with more experience may find such automated guidance less useful and undermining their expertise, especially when it fails to reflect the infrastructural constraints and tacit knowledge that shape their existing clinical practices, which may lead to resistance or disengagement from these technologies [345, 160, 326].

In Chapter 4, I found that designing for **flexible patient input can create tensions with providers' expectations.** While patients valued the flexibility to choose from various data types, they often selected forms they felt were appropriate to share with their providers. These decisions were shaped not only by how helpful a data form would be for conveying their experiences, but also by how they anticipated providers would interpret and respond to it. This insight extends prior work that highlights the social considerations involved in sharing personal tracking data [92, 325], emphasizing that these choices are not made in isolation. Through this work, I highlight the importance of considering provider perceptions about flexible data forms in health monitoring technologies, as these perceptions can significantly influence the effectiveness and reception of different data forms.

Insights from Chapters 3 and 4 highlight how flexibility can serve as a design strategy to help both providers and patients navigate infrastructural challenges, particularly when care practices are shaped by systemic constraints and tacit expertise. Clinical infrastructures are typically organized around brief visits where providers must make timely decisions while balancing competing demands from patients, insurers, pharmacies, and other stakeholders, particularly in multi-payer healthcare systems such as the United States, Switzerland, or the Netherlands. These settings often leave limited room for patients to convey the complexity of their lived experiences, especially when mediated through standardized self-report tools. As a result, both patients and providers are often required to navigate infrastructural gaps: providers by working around misaligned systems, and patients by tailoring how they present

their illness experiences between visits. Flexible health monitoring technologies can help bridge these gaps by accommodating complex care regimens, supporting providers in adjusting their plans in response to infrastructural constraints, and enabling patients to express themselves in ways that feel authentic within the logistical realities of clinical care.

While flexibility addresses the complexity of individualized care in clinical settings, a different set of challenges emerges in public health contexts, where stakeholders must respond to population-level responsibilities, frequently navigating distributed accountability and limited resources. In the next section, I show how a related but distinct strategy—amplifying stakeholders' existing practices—can support collaboration within public health infrastructures. Across both domains, insights from my work emphasize that designing for collaboration in health monitoring technologies requires careful attention to the constraints and work practices embedded within care infrastructures.

7.2 Improving Collaboration in Public Health Infrastructures (T2)

7.2.1 Amplifying Existing Practices of Stakeholders

Through my dissertation, I argue that **amplifying existing practices—by enhancing and scaling stakeholders' efforts—is essential for health monitoring technologies** to meaningfully support collaboration in resource-constrained public health infrastructures.

In Chapter 5, I found that **public health monitoring technologies like AI chatbots can enhance existing practices of stakeholders by providing a holistic understanding of the individuals through open-ended conversations.** Frontline workers noted that CareCall's open-ended conversations generated call logs containing rich contextual informa-

tion about broader aspects of care recipients (e.g., daily routines, social activities), which might have been difficult to obtain through manual check-up calls, given these workers' the resource constraints. These call logs enabled frontline workers to focus on monitoring and reaching out to cases that needed their attention (e.g., those who developed new health concerns). AI-driven check-up calls also supported individuals' existing practices of participating in social isolation interventions. Individuals often perceived that the system helped mitigate loneliness by asking caring questions about their health and engaging in broader conversation topics, such as their hobbies and interests, which also might have been challenging through human check-up calls given their resource constraints. In particular, as the system was used to increase the frequency of check-ins rather than to replace existing human interactions, many felt that the system's scheduled check-up calls fulfilled their desire for more conversation opportunities. Prior work has highlighted concerns that the introduction of AI technology to support aging in place could lead to unintended consequences, such as reducing human contact with their formal and informal caregivers [128, 295, 182]. Echoing these concerns, I argue that AI chatbots should be designed to *enhance*—rather than to *replace*—existing practices by health workers.

In Chapter 6, I found that **AI chatbots for public health monitoring like CareCall have the potential to scale up the existing practices of stakeholders** within public health infrastructures. Specifically, **AI chatbots could expand public reach to populations in need of regular health monitoring through automated check-ins**—a task that would have been challenging through traditional, human-only approaches given the resource constraints faced by many public agencies. The automation of regular check-ins enabled public agencies to reach significantly larger populations and increase monitoring frequency compared to human-only or hardware-based approaches. CareCall's reliance on existing public infrastructures, such as telephone networks, was particularly instrumental in achieving this scale, as it reduced adoption and maintenance costs. As research in the HCI community has emphasized, building on existing infrastructure [319, 314, 101] or social

platforms [93, 111] can improve the scalability of health monitoring technologies by reducing development and management burdens [93] and enabling broader engagement in low-resource settings [319]. Insights from this work highlight the promise of aligning with both human and non-human (e.g., physical and technological) infrastructures to design more scalable health monitoring technologies.

Chapter 6 further showed that **AI chatbots could scale up public agencies' efforts to connect individuals with necessary healthcare or social services** when provided by adequate support. As the open-ended nature of LLM-driven chatbots led users to express various healthcare and social service needs when interacting with CareCall, the system unexpectedly served as a window for users to communicate diverse care needs. Frontline workers valued the ability to identify these needs through call logs and often took follow-up actions. Insights from this work suggest that LLM-driven chatbots can play a meaningful role in care infrastructures by surfacing unmet care needs among vulnerable populations. I see an opportunity for public agencies to implement mechanisms to systematically monitor and respond to emerging needs identified through open-ended chatbot interactions.

Similar to what was discussed in subsection 7.1.1, the infrastructural challenges highlighted in Chapters 5 and 6 should also be understood in light of their specific institutional context. In countries with relatively centralized public health systems, such as South Korea, Taiwan, and the United Kingdom, national agencies often retain substantial authority over public health policies and the implementation of technologies. Such a centralized structure in South Korea likely facilitated the integration of AI chatbots like CareCall into existing systems, enabling wide-scale deployment through local public agencies in alignment with nationally coordinated strategies. These dynamics contrast with the more federated public health systems, such as those in the United States, Canada, and Germany, where responsibilities for public health services are distributed across federal, state, and local levels. In such contexts, public health departments in different jurisdictions may operate with varying resources, priorities, and

infrastructures, and technologies are often introduced through locally driven initiatives. As a result, frontline practices may evolve more autonomously in response to community-specific needs and constraints.

These institutional differences have important implications for how health monitoring technologies can amplify existing stakeholder practices. In centrally coordinated systems like those in South Korea, Taiwan, and the United Kingdom, amplification may focus on aligning technologies with standardized workflows and extending nationally coordinated interventions. In contrast, in more decentralized systems like the United States, Canada, and Germany, amplification may require technologies to accommodate greater local variability and support bottom-up adaptations by frontline workers. Together, these contrasts highlight that efforts to amplify stakeholder practices must attend not only to what those practices are, but also to the institutional conditions and government structures that shape how stakeholder practices are organized and resourced. Future research is needed to examine how amplification unfolds across diverse institutional settings.

7.2.2 Tensions in Amplifying Existing Practices of Multiple Stakeholders

While my dissertation demonstrated that health monitoring technologies can enhance collaboration in public health interventions by amplifying existing practices of stakeholders amidst resource constraints, it also revealed tensions that may emerge when these technologies are introduced into complex public health infrastructures. **A key challenge is managing stakeholder expectations around the capabilities and limitations of health monitoring technologies**—especially when those technologies are designed to support open-ended, flexible forms of interactions, such as LLM-driven chatbots.

In Chapter 5, I found that **stakeholders often held misaligned expectations about**

the types of tasks that health monitoring technologies could perform. Unlike task-oriented chatbots designed for closed-domain tasks, open-domain chatbots aim to support naturalistic conversations across a broad range of topics. This open-endedness often led various stakeholders in public health interventions to assume that the chatbot could perform a wide and flexible set of tasks. For example, government agencies often wished AI chatbots like CareCall to conduct structured health assessments, such as administering dementia screening questionnaires. Similarly, end users expected the system to detect emergencies or provide access to social services. In both cases, developers faced challenges in aligning these expectations with the technical (e.g., concerns about the uncertainty in control of LLM-driven chatbots for high-risk tasks) and infrastructural (e.g., staff shortages) constraints of the system. These mismatches highlight how the open-domain nature of LLM-driven chatbots can blur perceived boundaries of system capabilities, leading to misaligned expectations among stakeholders and complicating efforts to enhance existing practices.

In Chapter 6, I also found that **stakeholders often held misaligned expectations around labor involved in adopting health monitoring technologies**. Through this work, I found that decision makers often viewed AI chatbots like CareCall as highly efficient and requiring minimal human oversight, leading them not to scale up staffing when expanding care with the system. However, frontline workers had to take on a significant burden to handle the expanded care with these AI systems. This discrepancy led to failures in planning for the human resources required to operate AI at scale for public health monitoring. Insights from this work suggest that the open-ended nature of LLM-driven chatbots makes it challenging for decision makers to plan for human efforts necessary to operate them for public health contexts.

In Chapter 6, I further demonstrated that **stakeholders often held misaligned expectations about user engagement with health monitoring technologies.** In this work, I observed that introducing AI chatbots like CareCall can create new types of labor for

frontline workers, such as handling lapses in user engagement, due to unrealistic expectations about how consistently and willingly users would engage with these systems. CareCall was adopted with the assumption that unanswered calls might signal health emergencies based on the belief that users would respond reliably unless in a crisis. However, in reality, users frequently lapsed in using the system due to forgetfulness or reluctance to engage with AI, rather than because of health concerns. Building on prior work that examined people's engagement patterns with health monitoring technologies more broadly [91, 92] and with AI chatbots specifically [261, 182, 151], I argue that public health monitoring involving AI chatbots must account for the variability and unpredictability of user engagement over time. Rather than relying on assumptions of consistent use, these systems should be designed with mechanisms to explicitly address lapses, whether through human oversight supported by adequate staffing or fallback strategies such as passive sensing. Doing so can help prevent overburdening stakeholders in public health infrastructures.

These findings highlight a broader issue within public health infrastructures: adopting AI chatbots necessitates fundamentally rethinking and reconfiguring the existing practices involved in public health monitoring. While frontline work traditionally involved routine outreach—such as phone calls or home visits—CareCall shifted these responsibilities, automating some tasks while introducing new ones. Many workers found that tasks like managing follow-ups, interpreting chatbot outputs, and handling lapses in engagement fell outside their training or expertise. As AI technologies become more widely put into practice, the day-to-day responsibilities and skill requirements of frontline workers are likely to shift, along with the kinds of training needed to support them effectively. These shifts go beyond enhancing existing practices—they may signal structural changes in the overall public health infrastructure. While it remains unclear whether these shifts will be ultimately positive or negative, I argue that public agencies should critically assess how AI systems might reshape the existing practices of stakeholders in public health infrastructures before adopting them at scale.

Taken as a whole, my findings suggest that with novel technologies like LLM-driven chatbots, it is especially critical to facilitate early conversations around system capabilities and expectations. Developing guidelines that transparently communicate these boundaries can help stakeholders align their goals, reduce misunderstandings, and ultimately enable these systems to more effectively support stakeholders' existing practices within public health infrastructures.

7.3 Future Work

Moving forward, I aim to continue exploring how health monitoring technologies could better account for infrastructural complexity to improve collaboration in clinical care and health interventions. I am interested in pursuing the following directions for future work.

- **Accounting for Logistical Constraints in Clinical Infrastructures**: My previous work identified opportunities to use annotations to support patient-centered communication within the logistical constraints in clinical infrastructures. Extending this work, I plan to develop an annotation tool for digital symptom measures in domains where subjective illness experiences are crucial to longitudinal care planning, such as mental health, cancer, and chronic pain. I am particularly interested in exploring how various AI features could help tailor annotation tools to patients' communication needs, such as providing algorithmic recommendations for data forms and contents based on patient input, generating visual artifacts from natural language descriptions, and providing reflective insights to help patients identify information most relevant to their ongoing care through chatbot interactions. Implementation and evaluation of these designs will advance our understanding of how health monitoring technologies can meaningfully incorporate patient perspectives while remaining compatible with the logistical constraints of clinical infrastructures.

- **Accounting for Human Infrastructure in Decision-Making around AI Adoption**: Building on my previous work, I am excited to explore how to guide stakeholders in evaluating the capabilities and limitations of various AI systems in relation to their public health infrastructures. In collaboration with public health and AI experts, I aim to design guidelines that assist decision-making around which types of technologies best align with the infrastructural constraints of public agencies. I also plan to examine how the design and overall operation of AI chatbots for public health monitoring can better recognize and account for the labor shifts these systems could bring about. In this line of work, I am interested in exploring various public health domains, such as crisis management and chronic disease monitoring. Insights from this research will lead to developing resources that help decision makers carefully consider the role of human infrastructures in effectively and sustainably operating AI systems in public health contexts.

7.4 Conclusion

My dissertation demonstrates that health monitoring technologies can improve collaboration in clinical care and public health interventions by accounting for infrastructural complexity. In Chapters 3 and 4, I showed that implementing flexibility in health monitoring technologies can help stakeholders navigate and reconcile the constraints within clinical infrastructures. Chapter 3 highlighted how flexibility can support providers in managing complex infrastructural demands, while also emphasizing the need to consider differences in provider experience. Chapter 4 underscored the value of designing for flexible patient input in clinical measures to support more patient-centered communication amidst the logistical constraints of clinical settings, while also pointing to the importance of considering provider perceptions when introducing non-traditional data forms.

In Chapters 5 and 6, I demonstrated that amplifying stakeholders' existing practices can help them navigate and address resource constraints in public health infrastructures. Chapter 5 highlighted that AI chatbots for public health monitoring can enhance frontline workers' practices by providing a holistic understanding of individuals through open-ended conversations, and by asking individuals caring questions about their health and broader topics. Chapter 6 illustrated how AI chatbots can scale up public agencies' existing practices by expanding public reach through automated check-ins and connecting individuals to necessary healthcare or social services. However, across both chapters, I identified key challenges in amplifying stakeholder practices under resource constraints, particularly around managing expectations around the capabilities and limitations of technologies designed to support open-ended, flexible forms of interactions. These chapters suggested that stakeholders often held misaligned expectations about the types of tasks that AI chatbots could perform, the labor required for their adoption, and patterns of user engagement.

Taken together, my dissertation contributes to HCI and CSCW scholarship that foregrounds the infrastructural realities of care. By examining how health monitoring technologies intersect with stakeholder practices, I offer a lens for understanding how systems can support not only individuals but also the collaborative efforts that sustain clinical and public infrastructures. I hope this work encourages future researchers and designers to approach health monitoring technologies not merely as tools for individual use, but as collaborative tools that can support broader stakeholders in healthcare infrastructures.

Bibliography

[1] J. Aarts, J. Ash, and M. Berg. Extending the understanding of computerized physician order entry: Implications for professional collaboration, workflow and quality of care. *International Journal of Medical Informatics*, 76(Suppl. 1):S4–s13, 2007.

[2] D. Abu-Geras, D. Hadziomerovic, A. Leau, R. N. Khan, S. Gudka, C. Locher, M. Razaghikashani, and L. Y. Lim. Accuracy of tablet splitting and liquid measurements: an examination of who, what and how. *Journal of Pharmacy and Pharmacology*, 69(5):603–612, 2017.

[3] S. Abuse and M. H. S. Administration. Key substance use and mental health indicators in the united states: Results from the 2018 national survey on drug use and health. *HHS Publication No. PEP19-5068, NSDUH Series H-54*, 170:51–58, 2019.

[4] A. T. Adams, E. L. Murnane, P. Adams, M. Elfenbein, P. F. Chang, S. Sannon, G. Gay, and T. Choudhury. Keppi: A tangible user interface for self-reporting pain. In *Proceedings of the 2018 CHI conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery, 2018.

[5] P. Adams, E. L. Murnane, M. Elfenbein, E. Wethington, and G. Gay. Supporting the self-management of chronic pain conditions with tailored momentary self-assessments. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, page 1065–1077. Association for Computing Machinery, 2017.

[6] E. Agapie, B. Chinh, L. R. Pina, D. Oviedo, M. C. Welsh, G. Hsieh, and S. Munson. Crowdsourcing exercise plans aligned with expert guidelines and everyday constraints. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery, 2018.

[7] E. Agapie, L. Colusso, S. A. Munson, and G. Hsieh. Plansourcing: Generating behavior change plans with friends and crowds. In *Proceedings of the 19th Conference on Computer-Supported Cooperative Work & Social Computing*, pages 119–133. Association for Computing Machinery, 2016.

[8] M. Agmon, A. Zisberg, E. Gil, D. Rand, N. Gur-Yaish, and M. Azriel. Adult utilization of psychiatric drugs and differences by sex, age, and race. *JAMA Internal Medicine*, 177(2):272–274, 2017.

[9] M. S. Alharthi. Exploring challenges and enablers for community pharmacists using electronic prescriptions (wasfaty) in makkah region, saudi arabia: a qualitative study using the theoretical domains framework. *Frontiers in Medicine*, 11:1487852, November 2024.

[10] J. S. Ancker, H. O. Witteman, B. Hafeez, T. Provencher, M. Van De Graaf, and E. Wei. "You get reminded you're a sick person": Personal data tracking and patients with multiple chronic conditions. *Journal of Medical Internet Research*, 17(8), 2015.

[11] T. Andersen, J. Bansler, F. Kensing, J. Moll, and K. D. Nielsen. Alignment of concerns: A design rationale for patient participation in eHealth. *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 2587–2596, 2014.

[12] T. Andersen, P. Bjørn, F. Kensing, and J. Moll. Designing for collaborative interpretation in telemonitoring: Re-introducing patients as diagnostic agents. *International Journal of Medical Informatics*, 80(8):e112–e126, 2011.

[13] T. O. Andersen, J. P. Bansler, F. Kensing, J. Moll, T. Mønsted, K. D. Nielsen, O. W. Nielsen, H. H. Petersen, and J. H. Svendsen. Aligning concerns in telecare: three concepts to guide the design of patient-centred e-health. *Computer Supported Cooperative Work (CSCW)*, 28:1039–1072, 2019.

[14] Anthropic, Inc. Claude, 2023. Accessed: 03-31-2025.

[15] APA. Treatment of patients with major depressive disorder second edition apa. *Psychiatric Services*, (April):1–78, 2010.

[16] I. Arreola, Z. Morris, M. Francisco, K. Connelly, K. Caine, and G. White. From checking on to checking in: designing for low socio-economic status older adults. In *Proceedings of the 2014 CHI Conference on Human Factors in Computing Systems*, pages 1933–1936. Association for Computing Machinery, 2014.

[17] 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 *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–15. Association for Computing Machinery, 2020.

[18] A. Ayobi, P. Marshall, A. L. Cox, and Y. Chen. Quantifying the body and caring for the mind: Self-tracking in multiple sclerosis. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, page 6889–6901. Association for Computing Machinery, 2017.

[19] A. Ayobi, T. Sonne, P. Marshall, and A. L. Cox. Flexible and mindful self-tracking: Design implications from paper bullet journals. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–14. Association for Computing Machinery, 2018.

[20] S. Bae, D. Kwak, S. Kang, M. Y. Lee, S. Kim, Y. Jeong, H. Kim, S.-W. Lee, W. Park, and N. Sung. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787. Association for Computational Linguistics, 12 2022.

[21] S. Bae, D. Kwak, S. Kim, D. Ham, S. Kang, S.-W. Lee, and W. Park. Building a role specified open-domain dialogue system leveraging large-scale language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2128–2150, Seattle, United States, July 2022. Association for Computational Linguistics.

[22] S. Bakhshi, D. A. Shamma, L. Kennedy, Y. Song, P. de Juan, and J. J. Kaye. Fast, cheap, and good: Why animated gifs engage us. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, page 575–586. Association for Computing Machinery, 2016.

[23] M. Balaam, R. Comber, E. Jenkins, S. Sutton, and A. Garbett. Feedfinder: A location-mapping mobile application for breastfeeding women. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*, page 1709–1718. Association for Computing Machinery, 2015.

[24] J. E. Bardram, M. Frost, K. Szántó, M. Faurholt-Jepsen, M. Vinberg, and L. V. Kessing. Designing mobile health technology for bipolar disorder: a field trial of the monarca system. In *Proceedings of the 2013 CHI Conference on Human Factors in Computing Systems*, page 2627–2636. Association for Computing Machinery, 2013.

[25] V. Bartle, J. Lyu, F. El Shabazz-Thompson, Y. Oh, A. A. Chen, Y.-J. Chang, K. Holstein, and N. Dell. "a second voice": Investigating opportunities and challenges for interactive voice assistants to support home health aides. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2022.

[26] A. T. Beck. An Inventory for Measuring Depression. *Archives of General Psychiatry*, 4(6):561, 1961.

[27] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12, 2020.

[28] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623. Association for Computing Machinery, 2021.

[29] M. Berg. Patient care information systems and health care work: a sociotechnical approach. *International Journal of Medical Informatics*, 55(2):87–101, 1999.

[30] M. Berg. Patient care information systems and health care work: A sociotechnical approach. *International Journal of Medical Informatics*, 55(2):87–101, 1999.

[31] M. Berg, J. Aarts, and J. Van der Lei. Ict in health care: Sociotechnical approaches. *Methods of Information in Medicine*, 42(4):297–301, 2003.

[32] A. B. Berry, C. Lim, A. L. Hartzler, T. Hirsch, E. Ludman, E. H. Wagner, and J. D. Ralston. Creating conditions for patients' values to emerge in clinical conversations: Perspectives of health care team members. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, page 1165–1174. Association for Computing Machinery, 2017.

[33] A. B. Berry, C. Lim, A. L. Hartzler, T. Hirsch, E. Ludman, E. H. Wagner, and J. D. Ralston. Eliciting values of patients with multiple chronic conditions: Evaluation of a patient-centered framework. In *AMIA Annual Symposium Proceedings*, volume 2017, page 430. American Medical Informatics Association, 2017.

[34] A. B. Berry, C. Y. Lim, A. L. Hartzler, T. Hirsch, E. Ludman, E. H. Wagner, and J. D. Ralston. "It's good to know you're not a stranger every time": Communication about values between patients with multiple chronic conditions and healthcare providers. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), 2017.

[35] A. B. Berry, C. Y. Lim, T. Hirsch, A. L. Hartzler, L. M. Kiel, Z. A. Bermet, and J. D. Ralston. Supporting communication about values between people with multiple chronic conditions and their providers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 1–14. Association for Computing Machinery, 2019.

[36] A. B. Berry, C. Y. Lim, C. A. Liang, A. L. Hartzler, T. Hirsch, D. M. Ferguson, Z. A. Bermet, and J. D. Ralston. Supporting Collaborative Reflection on Personal Values and Health. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 2021.

[37] N. Black. Patient reported outcome measures could help transform healthcare. *BMJ (Online)*, 346(7896):1–5, 2013.

[38] J. Bonander and S. Gates. Public health in an era of personal health records: Opportunities for innovation and new partnerships. *Journal of Medical Internet Research*, 12(3), 2010.

[39] S. Bordia and S. R. Bowman. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[40] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[41] V. Braun and V. Clarke. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health*, 11(4):589–597, 2019.

[42] P. F. Brennan and G. Casper. Observing health in everyday living: ODLs and the care-between-the-care. *Personal and Ubiquitous Computing*, 19(1):3–8, 2015.

[43] D. J. Brody and Q. Gu. Antidepressant Use Among Adults: United States, 2015-2018. *NCHS data brief*, (377):1–8, 2020.

[44] A. Brown, A. Chouldechova, E. Putnam-Hornstein, A. Tobin, and R. Vaithianathan. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 1–12. Association for Computing Machinery, 2019.

[45] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS '20)*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[46] M. Brundage, J. Blazeby, D. Revicki, B. Bass, H. De Vet, H. Duffy, F. Efficace, M. King, C. L. Lam, D. Moher, J. Scott, J. Sloan, C. Snyder, S. Yount, and M. Calvert. Patient-reported outcomes in randomized clinical trials: Development of ISOQOL reporting standards. *Quality of Life Research*, 22(6):1161–1175, 2013.

[47] S. A. Bull, E. M. Hunkeler, J. Y. Lee, C. R. Rowland, T. E. Williamson, J. R. Schwab, S. W. Hurt, L. González, and D. Demers. Discontinuing or switching selective serotonin-reuptake inhibitors. *Annals of Pharmacotherapy*, 36(4):578–584, 2002.

[48] A. G. Büyüktür and M. S. Ackerman. Information work in bone marrow transplant: Reducing misalignment of perspectives. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, page 1740–1752. Association for Computing Machinery, 2017.

[49] H. Byun. NAVER launches AI call service aimed at seniors - The Korea Herald, May 2022. Accessed: 03-31-2025.

[50] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14. Association for Computing Machinery, 2019.

[51] C. J. Cai, S. Winter, D. Steiner, L. Wilcox, and M. Terry. "Hello AI": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[52] A. T.-Y. Chen. How fragmentation can undermine the public health response to covid-19. *Interactions*, 28(2):64–69, 2021.

[53] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating large language models trained on code, 2021.

[54] Y. Chen. Take it personally: accounting for individual difference in designing diabetes management systems. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, pages 252–261. Association for Computing Machinery, 2010.

[55] Y. Chen, Y. Sun, and S. Lindtner. Maintainers of stability: The labor of china's data-driven governance and dynamic zero-covid. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2023.

[56] H.-F. Cheng, L. Stapleton, R. Wang, P. Bullock, A. Chouldechova, Z. S. S. Wu, and H. Zhu. Soliciting stakeholders' fairness notions in child maltreatment predictive systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021.

[57] K. G. Cheng, G. R. Hayes, S. H. Hirano, M. S. Nagel, and D. Baker. Challenges of integrating patient-centered data into clinical workflow for care of high-risk infants. *Personal and Ubiquitous Computing*, 19:45–57, 2015.

[58] E. K. Choe, B. Lee, M. Kay, W. Pratt, and J. A. Kientz. Sleeptight: low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 121–132. Association for Computing Machinery, 2015.

[59] E. K. Choe, N. B. Lee, B. Lee, W. Pratt, and J. A. Kientz. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1143–1152, 2014.

[60] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat,

M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. Palm: Scaling language modeling with pathways, 2022.

[61] A. E. Chung and E. M. Basch. Incorporating the patient's voice into electronic health records through patient-reported outcomes as the review of systems. *Journal of the American Medical Informatics Association*, 22(4):914–916, 2015.

[62] C.-F. Chung, K. Dew, A. Cole, J. Zia, J. Fogarty, J. A. Kientz, and S. A. Munson. Boundary negotiating artifacts in personal informatics: Patient-provider collaboration with patient-generated data. In *Proceedings of the 2016 ACM Conference on Computer-Supported Cooperative Work & Social Computing*, page 770–786. Association for Computing Machinery, 2016.

[63] C.-F. Chung, Q. Wang, J. Schroeder, A. Cole, J. Zia, J. Fogarty, and S. Munson. Identifying and planning for individualized change: Patient-provider collaboration using lightweight food diaries in healthy eating and irritable bowel syndrome. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1):1–27, 2019.

[64] J. J. Y. Chung, W. Kim, K. M. Yoo, H. Lee, E. Adar, and M. Chang. Talebrush: Sketching stories with generative pretrained language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2022.

[65] C. S. Cleeland and K. M. Ryan. Pain assessment: global use of the brief pain inventory. *Annals of the Academy of Medicine, Singapore*, 23 2:129–38, 1994.

[66] S. Coghlan, J. Waycott, A. Lazar, and B. Barbosa Neves. Dignity, Autonomy, and Style of Company: Dimensions Older Adults Consider for Robot Companions. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 2021.

[67] J. J. Collins, M. E. Byrnes, I. J. Dunkel, J. Lapin, T. Nadel, H. T. Thaler, T. Polyak, B. Rapkin, and R. K. Portenoy. The measurement of symptoms in children with cancer. *Journal of Pain and Symptom Management*, 19(5):363–377, 2000.

[68] S. Consolvo, P. Roessler, and B. E. Shelton. The CareNet display: lessons learned from an in home evaluation of an ambient display. In *International conference on ubiquitous computing*, pages 1–17. Springer, 2004.

[69] M. Costa Figueiredo, C. Caldeira, E. V. Eikey, M. Mazmanian, and Y. Chen. Engaging with Health Data: The interplay between self-tracking activities and emotions in fertility struggles. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–20, 2018.

[70] M. Costa Figueiredo, C. Caldeira, T. L. Reynolds, S. Victory, K. Zheng, and Y. Chen. Self-tracking for fertility care: Collaborative support for a highly-personalized problem. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), 2017.

[71] M. Costa Figueiredo and Y. Chen. Patient-generated health data: Dimensions, challenges, and open questions. *Foundations and Trends in Human-Computer Interaction*, 13(3):165–297, 2020.

[72] M. Costa Figueiredo, H. I. Su, and Y. Chen. Using Data to Approach the Unknown. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–35, 2021.

[73] M. Crain, W. Poster, and M. Cherry. *Invisible labor: Hidden work in the contemporary world*. University of California Press, 2016.

[74] H. Cramer, P. De Juan, and J. Tetreault. Sender-intended functions of emojis in US messaging. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2016*, pages 504–509. Association for Computing Machinery, 2016.

[75] N. Daskalova, D. Metaxa-Kakavouli, A. Tran, N. Nugent, J. Boergers, J. McGeary, and J. Huang. Sleepcoacher: A personalized automated self-experimentation system for sleep recommendations. In *Proceedings of the 29th annual symposium on user interface software and technology*, pages 347–358. Association for Computing Machinery, 2016.

[76] J. Davies and J. Read. A systematic review into the incidence, severity and duration of antidepressant withdrawal effects: Are guidelines evidence-based? *Addictive Behaviors*, 97(September 2018):111–121, 2019.

[77] M. De Choudhury, M. Kumar, and I. Weber. Computational approaches toward integrating quantified self sensing and social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1334–1349. Association for Computing Machinery, 2017.

[78] B. DeRenzi, N. Dell, J. Wacksman, S. Lee, and N. Lesh. Supporting community health workers in india through voice- and web-based feedback. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, page 2770–2781. Association for Computing Machinery, 2017.

[79] P. M. Desai, E. G. Mitchell, M. L. Hwang, M. E. Levine, D. J. Albers, and L. Mamykina. Personal health oracle: Explorations of personalized predictions in diabetes self-management. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery, 2019.

[80] P. Deshpande, B. Sudeepthi, S. Rajan, and C. Abdul Nazir. Patient-reported outcomes: A new era in clinical research. *Perspectives in Clinical Research*, 2(4):137, 2011.

[81] C. DeVane. Pharmacokinetics of the newer antidepressants: Clinical relevance. *The American Journal of Medicine*, 97(6):S13–S23, dec 1994.

[82] K. J. Devers. How will we know "good" qualitative research when we see it? Beginning the dialogue in health services research. *Health services research*, 34(5 Pt 2):1153–88, 1999.

[83] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[84] D. Diasso, M. H. Doudou, M. C. Levrak, H. D. Sedutto, and A. Savadogo. Municipalities' organisational capacity to support the implementation of the multi-sector nutrition plan in burkina faso. *Global Health Action*, 14(1), 2021.

[85] Dimagi. Commcare, 2024. Accessed: 03-31-2025.

[86] N. J. Donovan and D. Blazer. Social Isolation and Loneliness in Older Adults: Review and Commentary of a National Academies Report. *American Journal of Geriatric Psychiatry*, 28(12):1233–1244, 2020.

[87] E. V. Eikey and M. C. Reddy. " it's definitely been a journey" a qualitative study on how women with eating disorders use weight loss apps. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 642–654. Association for Computing Machinery, 2017.

[88] A. Eliasen, M. K. Abildtoft, N. S. Krogh, C. Rechnitzer, J. S. Brok, R. Mathiasen, K. Schmiegelow, and K. P. Dalhoff. Smartphone app to self-monitor nausea during pediatric chemotherapy treatment: User-centered design process. *JMIR mHealth and uHealth*, 8(7):1–11, 2020.

[89] G. Elwyn, I. Scholl, C. Tietbohl, M. Mann, A. G. Edwards, C. Clay, F. Légaré, T. V. D. Weijden, C. L. Lewis, R. M. Wexler, and D. L. Frosch. "many miles to go.": A systematic review of the implementation of patient decision support interventions into routine clinical practice. *BMC Medical Informatics and Decision Making*, 13(Suppl. 2):S14, 2013.

[90] D. A. Epstein, C. Caldeira, M. C. Figueiredo, X. Lu, L. M. Silva, L. Williams, J. H. Lee, Q. Li, S. Ahuja, Q. Chen, P. Dowlatyari, C. Hilby, S. Sultana, E. V. Eikey, and Y. Chen. Mapping and Taking Stock of the Personal Informatics Literature. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4), 2020.

[91] D. A. Epstein, P. Eslambolchilar, J. Kay, J. Meyer, and S. A. Munson. *Opportunities and challenges for long-term tracking*. Springer International Publishing, 2021.

[92] D. A. Epstein, B. H. Jacobson, E. Bales, D. W. McDonald, and S. A. Munson. From "nobody cares" to "way to go!": A design framework for social sharing in personal informatics. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1622–1636, 2015.

[93] D. A. Epstein, F. Liu, A. Monroy-Hernández, and D. Wang. Revisiting piggyback prototyping: Examining benefits and tradeoffs in extending existing social computing systems. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28, 2022.

[94] M. Eric, R. Goel, S. Paul, A. Sethi, S. Agarwal, S. Gao, A. Kumar, A. Goyal, P. Ku, and D. Hakkani-Tur. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France, May 2020. European Language Resources Association.

[95] T. Erickson, M. Li, Y. Kim, A. Deshpande, S. Sahu, T. Chao, P. Sukaviriya, and M. Naphade. The dubuque electricity portal: evaluation of a city-scale residential electricity consumption feedback system. In *Proceedings of the 2013 CHI Conference on Human Factors in Computing Systems*, page 1203–1212. Association for Computing Machinery, 2013.

[96] B. Fang, Q. Xu, T. Park, and M. Zhang. Airsense: an intelligent home-based sensing system for indoor air quality analytics. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, page 109–119. Association for Computing Machinery, 2016.

[97] N. G. Fielding. Self-Report Study: The SAGE Dictionary of Social Research Methods. *The SAGE Dictionary of Social Research Methods*, pages 276–277, 2011.

[98] A. Flügge, T. Hildebrandt, and N. H. Møller. Street-level algorithms and ai in bureaucratic decision-making: A caseworker perspective. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 2021.

[99] A. Framer. What i have learnt from helping thousands of people taper off antidepressants and other psychotropic medications. *Therapeutic Advances in Psychopharmacology*, 11:204512532199127, 2021.

[100] J. Froehlich, L. Findlater, M. Ostergren, S. Ramanathan, J. Peterson, I. Wragg, E. Larson, F. Fu, M. Bai, S. Patel, and J. A. Landay. The design and evaluation of prototype eco-feedback displays for fixture-level water usage data. In *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems*, page 2367–2376. Association for Computing Machinery, 2012.

[101] J. E. Froehlich, E. Larson, T. Campbell, C. Haggerty, J. Fogarty, and S. N. Patel. Hydrosense: infrastructure-mediated single-point sensing of whole-home water activity. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, page 235–244. Association for Computing Machinery, 2009.

[102] M. Frost, A. Doryab, M. Faurholt-Jepsen, L. Kessing, and J. Bardram. Supporting disease insight through data analysis: refinements of the monarca self-assessment system. In *Proceedings of the 2013 ACM international joint conference on pervasive and ubiquitous computing*, pages 133–142. Association for Computing Machinery, 2013.

[103] J. Gao, M. Galley, and L. Li. Neural approaches to conversational AI. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 2–7. Association for Computational Linguistics, July 2018.

[104] A. X. Garg, N. K. Adhikari, H. McDonald, M. P. Rosas-Arellano, P. J. Devereaux, J. Beyene, J. Sam, and R. B. Haynes. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. *Journal of the American Medical Association*, 293(10):1223–1238, 2005.

[105] I. Garrido-Muñoz, A. Montejo-Ráez, F. Martínez-Santiago, and L. A. Ureña-López. A survey on bias in deep nlp. *Applied Sciences*, 11(7), 2021.

[106] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369. Association for Computational Linguistics, Nov. 2020.

[107] E. M. Gerson and S. L. Star. Analyzing due process in the workplace. *ACM Transactions on Information Systems (TOIS)*, 4(3):257–270, 1986.

[108] GoodRx. Accessed: 09-02-2021.

[109] Google, Inc. Gemini, 2023. Accessed: 03-31-2025.

[110] Google, Inc. The gemini app can now recall past chats, February 2025. Accessed: 03-31-2025.

[111] C. Grevet and E. Gilbert. Piggyback prototyping: Using existing, large-scale social computing systems to prototype new ones. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*, page 4047–4056. Association for Computing Machinery, 2015.

[112] E. Grönvall and N. Verdezoto. Beyond self-monitoring: Understanding non-functional aspects of home-based healthcare technology. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 587–596. Association for Computing Machinery, 2013.

[113] P. C. Groot and J. van Os. How user knowledge of psychotropic drug withdrawal resulted in the development of person-specific tapering medication. *Therapeutic Advances in Psychopharmacology*, 10:204512532093245, 2020.

[114] H. Gu, J. Huang, L. Hung, and X. A. Chen. Lessons learned from designing an ai-enabled diagnosis tool for pathologists. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 2021.

[115] P. Haddad. The SSRI discontinuation syndrome. *Journal of psychopharmacology*, 12(3):305–313, 1998.

[116] J. Hallare and V. Gerriets. Half-life, 2020. Accessed: 09-02-2021.

[117] M. Hamilton. A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry*, 23(1):56–62, 1960.

[118] 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*, 13(3):e62, Aug. 2011.

[119] healthcare.gov. Getting prescription medications, 2021. Accessed: 09-02-2021.

[120] M. J. D. Hoefer, L. Van Kleunen, C. Goodby, L. B. Blackburn, P. Panati, and S. Voida. The multiplicative patient and the clinical workflow: Clinician perspectives on social interfaces for self-tracking and managing bipolar disorder. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, pages 907–925. Association for Computing Machinery, 2021.

[121] M. K. Hong, U. Lakshmi, K. Do, S. Prahalad, T. Olson, R. I. Arriaga, and L. Wilcox. Using diaries to probe the illness experiences of adolescent patients and parental caregivers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–16. Association for Computing Machinery, 2020.

[122] M. K. Hong, U. Lakshmi, T. A. Olson, and L. Wilcox. Visual odls: Co-designing patient-generated observations of daily living to support data-driven conversations in pediatric care. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–13. Association for Computing Machinery, 2018.

[123] M. A. Horowitz and D. Taylor. Tapering of SSRI treatment to mitigate withdrawal symptoms. *The Lancet Psychiatry*, 6(6):538–546, 2019.

[124] Y.-C. Hsu, P. Dille, J. Cross, B. Dias, R. Sargent, and I. Nourbakhsh. Community-empowered air quality monitoring system. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, page 1607–1619. Association for Computing Machinery, 2017.

[125] M. Huang, X. Zhu, and J. Gao. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3), 2020.

[126] N. Huba and Y. Zhang. Designing patient-centered personal health records (PHRs): Health care professionals' perspective on patient-generated data. *Journal of Medical Systems*, 36(6):3893–3905, 2012.

[127] B. Huber, D. McDuff, C. Brockett, M. Galley, and B. Dolan. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–12. Association for Computing Machinery, 2018.

[128] L. L. Huber, K. Shankar, K. Caine, K. Connelly, L. J. Camp, B. A. Walker, and L. Borrero. How In-Home Technologies Mediate Caregiving Relationships in Later Life. *International Journal of Human-Computer Interaction*, 29(7):441–455, July 2013.

[129] A. Ismail, N. Karusala, and N. Kumar. Bridging disconnected knowledges for community health. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 2018.

[130] A. Ismail and N. Kumar. Engaging solidarity in data collection practices for community health. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 2018.

[131] A. Ismail and N. Kumar. AI in Global Health: The View from the Front Lines. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–21. Association for Computing Machinery, 2021.

[132] A. Ismail, D. Thakkar, N. Madhiwalla, and N. Kumar. Public health calls for/with AI: an ethnographic perspective. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–26, 2023.

[133] A. Ismail, D. Yadav, M. Gupta, K. Dabas, P. Singh, and N. Kumar. Imagining Caring Futures for Frontline Health Work. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–30, 2022.

[134] M. Jacobs, J. He, M. F. Pradier, B. Lam, A. C. Ahn, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos. Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14. Association for Computing Machinery, 2021.

[135] M. L. Jacobs, J. Clawson, and E. D. Mynatt. Comparing health information sharing preferences of cancer patients, doctors, and navigators. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 808–818, 2015.

[136] E. R. Jacques and P. Alexandridis. Tablet scoring: Current practice, fundamentals, and knowledge gaps. *Applied Sciences*, 9(15), 2019.

[137] A. Jang, D. L. MacLean, and J. Heer. BodyDiagrams: Improving communication of pain symptoms through drawing. In *Proceedings of the 2014 CHI Conference on Human Factors in Computing Systems*, page 1153–1162. Association for Computing Machinery, 2014.

[138] M. W. Jaspers, M. Smeulers, H. Vermeulen, and L. W. Peute. Effects of clinical decision-support systems on practitioner performance and patient outcomes: A synthesis of high-quality systematic review findings. *Journal of the American Medical Informatics Association*, 18(3):327–334, 2011.

[139] D. Jeavons, A. P. Hungin, and C. S. Cornford. Patients with poorly controlled diabetes in primary care: Healthcare clinicians' beliefs and attitudes. *Postgraduate Medical Journal*, 82(967):347–350, 2006.

[140] J. A. Jiang, C. Fiesler, and J. R. Brubaker. "The perfect one": Understanding communication practices and challenges with animated GIFs. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 2018.

[141] E. Jo, D. A. Epstein, H. Jung, and Y.-H. Kim. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2023.

[142] E. Jo, Y. Jeong, S. Park, D. A. Epstein, and Y.-H. Kim. Understanding the impact of long-term memory on self-disclosure with large language model-driven chatbots for public health intervention. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2024.

[143] E. Jo, Y.-H. Kim, S.-H. Ok, and D. A. Epstein. Understanding public agencies' expectations and realities of AI-driven chatbots for public health monitoring. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2025.

[144] E. Jo, S. Park, H. Bang, Y. Hong, Y. Kim, J. Choi, B. N. Kim, D. A. Epstein, and H. Hong. GeniAuti: Toward data-driven interventions to challenging behaviors of autistic children through caregivers' tracking. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–27, 2022.

[145] E. Jo, M. Ryu, G. Kenderova, S. So, B. Shapiro, A. Papoutsaki, and D. A. Epstein. Designing flexible longitudinal regimens: Supporting clinician planning for discontinuation of psychiatric drugs. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2022.

[146] E. Jo, R. Zehrung, K. Genuario, A. Papoutsaki, and D. A. Epstein. Exploring patient-generated annotations to digital clinical symptom measures for patient-centered communication. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–26, 2024.

[147] A. Kaltenhauser, V. Rheinstädter, A. Butz, and D. P. Wallach. "you have to piece the puzzle together": Implications for designing decision support in intensive care. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 2020.

[148] E. Karapanos, J. Gerken, J. Kjeldskov, and M. B. Skov, editors. *Advances in Longitudinal HCI Research*. Human–Computer Interaction Series. Springer International Publishing, Cham, 2021.

[149] R. Karkar, J. Schroeder, D. A. Epstein, L. R. Pina, J. Scofield, J. Fogarty, J. A. Kientz, S. A. Munson, R. Vilardaga, and J. Zia. Tummytrials: A feasibility study of using self-experimentation to detect individualized food triggers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2017.

[150] N. Karusala, S. Upadhyay, R. Veeraraghavan, and K. Z. Gajos. Understanding contestability on the margins: Implications for the design of algorithmic decision-making in public services. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2024.

[151] N. Karusala, S. Yan, N. Rajkumar, V. G, and R. Anderson. Speculating with care: Worker-centered perspectives on scale in a chat-based health information service. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2), Oct. 2023.

[152] A. Kawakami, A. Coston, H. Heidari, K. Holstein, and H. Zhu. Studying up public sector AI: How networks of power relations shape agency decisions around AI design and use. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2), November 2024.

[153] A. Kawakami, A. Coston, H. Zhu, H. Heidari, and K. Holstein. The situate ai guidebook: Co-designing a toolkit to support multi-stakeholder, early-stage deliberations around public sector AI proposals. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2024.

[154] A. Kawakami, V. Sivaraman, H.-F. Cheng, L. Stapleton, Y. Cheng, D. Qing, A. Perer, Z. S. Wu, H. Zhu, and K. Holstein. Improving human-AI partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2022.

[155] A. Kawakami, V. Sivaraman, L. Stapleton, H.-F. Cheng, A. Perer, Z. S. Wu, H. Zhu, and K. Holstein. "why do i care what's similar?" probing challenges in ai-assisted child welfare decision-making through worker-ai interface design concepts. In *Proceedings of the 2022 ACM DIS Conference*. Association for Computing Machinery, 2022.

[156] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach. Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *British Medical Journal*, 330(7494):765–768, 2005.

[157] E. Kaziunas, M. S. Klinkman, and M. S. Ackerman. Precarious interventions: Designing for ecologies of care. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), Nov. 2019.

[158] N. Keks, J. Hope, and S. Keogh. Switching and stopping antidepressants. *Australian Prescriber*, 39(3):76–83, 2016.

[159] T. Kendrick. Strategies to reduce use of antidepressants. *British Journal of Clinical Pharmacology*, 87(1):23–33, 2021.

[160] S. Khairat, D. Marc, W. Crosby, and A. Al Sanousi. Reasons for physicians not adopting clinical decision support systems: Critical analysis. *JMIR Medical Informatics*, 20(4), 2018.

[161] J. A. Killian, B. Wilder, A. Sharma, V. Choudhary, B. Dilkina, and M. Tambe. Learning to prescribe interventions for tuberculosis patients using digital adherence data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 2019.

[162] B. Kim, H. Kim, S.-W. Lee, G. Lee, D. Kwak, J. Dong Hyeon, S. Park, S. Kim, S. Kim, D. Seo, H. Lee, M. Jeong, S. Lee, M. Kim, S. H. Ko, S. Kim, T. Park, J. Kim, S. Kang, N.-H. Ryu, K. M. Yoo, M. Chang, S. Suh, S. In, J. Park, K. Kim, H. Kim, J. Jeong, Y. G. Yeo, D. Ham, D. Park, M. Y. Lee, J. Kang, I. Kang, J.-W. Ha, W. Park, and N. Sung. What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3424, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[163] J. Kim, T. Gong, B. Kim, J. Park, W. Kim, E. Huang, K. Han, J. Kim, J. Ko, and S.-J. Lee. No More One Liners. *ACM Transactions on Social Computing*, 3(2):1–25, 2020.

[164] J. Kim, J. Muhic, L. P. Robert, and S. Y. Park. Designing chatbots with black americans with chronic conditions: Overcoming challenges against covid-19. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2022.

[165] S.-I. Kim, E. Jo, M. Ryu, I. Cha, Y. H. Kim, H. Yoo, and H. Hong. 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*. Association for Computing Machinery, 2019.

[166] T. Kim, M. Ruensuk, and H. Hong. In Helping a Vulnerable Bot, You Help Yourself: Designing a Social Bot as a Care-Receiver to Promote Mental Health and Reduce Stigma. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery, 2020.

[167] Y. Kim, E. Heo, H. Lee, S. Ji, J. Choi, J.-W. Kim, J. Lee, and S. Yoo. Prescribing 10,000 steps like aspirin: designing a novel interface for data-driven medical consultations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2017.

[168] Y. Kim, S. Ji, H. Lee, J.-W. Kim, S. Yoo, and J. Lee. "My doctor is keeping an eye on me!": Exploring the clinical applicability of a mobile food logger. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, page 5620–5631. Association for Computing Machinery, 2016.

[169] Y.-H. Kim, D. Chou, B. Lee, M. Danilovich, A. Lazar, D. E. Conroy, H. Kacorri, and E. K. Choe. Mymove: Facilitating older adults to collect in-situ activity labels on a smartwatch with speech. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2022.

[170] Y.-H. Kim, S. Kim, M. Chang, and S.-W. Lee. Leveraging pre-trained language models to streamline natural language interaction for self-tracking. *arXiv preprint arXiv:2205.15503*, 2022.

[171] W. Kleiminger, C. Beckel, and S. Santini. Household occupancy monitoring using electricity meters. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery, 2015.

[172] E. Klinenberg. *Heat wave: A social autopsy of disaster in Chicago*. University of Chicago Press, 2002.

[173] S. Klüber, F. Maas, D. Schraudt, G. Hermann, O. Happel, and T. Grundgeiger. Experience matters: design and evaluation of an anesthesia support tool guided by user experience theory. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. Association for Computing Machinery, 2020.

[174] R. Koppel, T. Wetterneck, J. L. Telles, and B.-T. Karsh. Workarounds to barcode medication administration systems: their occurrences, causes, and threats to patient safety. *Journal of the American Medical Informatics Association : JAMIA*, 15(4):408–23, 2010.

[175] Korea Law Translation Center. Act on the prevention and management of lonely deaths, 2020.

[176] D. M. Korngiebel and S. D. Mooney. Considering the possibilities and pitfalls of generative pre-trained transformer 3 (GPT-3) in healthcare delivery. *npj Digital Medicine*, 4(93), June 2021.

[177] K. Kroenke, R. L. Spitzer, and J. B. W. Williams. The PHQ-9: Validity of of a Brief Depression Severity Measure. *Journal of General Internal Medicine*, 16(9):606–613, sep 2001.

[178] T.-S. Kuo, H. Shen, J. Geum, N. Jones, J. I. Hong, H. Zhu, and K. Holstein. Understanding frontline workers' and unhoused individuals' perspectives on AI used in homeless services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2023.

[179] S. Kuoppamäki. The application and deployment of welfare technology in swedish municipal care: A qualitative study of procurement practices among municipal actors. *BMC Health Services Research*, 21(1), 2021.

[180] U. Lakshmi, M. Hong, and L. Wilcox. Integrating patient-generated observations of daily living into pediatric cancer care: A formative user interface design study. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2018.

[181] A. Lazar, C. Edasis, and A. M. Piper. Supporting people with dementia in digital social sharing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2017.

[182] A. Lazar, H. J. Thompson, S. Y. Lin, and G. Demiris. Negotiating relation work with telehealth home care companionship technologies that support aging in place. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), November 2018.

[183] C. P. Lee. Boundary negotiating artifacts: Unbinding the routine of boundary objects and embracing chaos in collaborative work. *Computer Supported Cooperative Work*, 16(3):307–339, 2007.

[184] C. P. Lee, P. Dourish, and G. Mark. The human infrastructure of cyberinfrastructure. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 2006.

[185] G. Lee, V. Hartmann, J. Park, D. Papailiopoulos, and K. Lee. Prompted LLMs as chatbot modules for long open-domain conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, July 2023.

[186] K. Lee, H. Cho, K. Toshnazarov, N. Narziev, S. Y. Rhim, K. Han, Y. Noh, and H. Hong. Toward future-centric personal informatics: Expecting stressful events and preparing personalized interventions in stress management. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery, 2020.

[187] K. Lee and H. Hong. Designing for self-tracking of emotion and experience with tangible modality. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, pages 465–475. Association for Computing Machinery, 2017.

[188] M. Lee, S. Ackermans, N. van As, H. Chang, E. Lucas, and W. IJsselsteijn. Caring for Vincent: A chatbot for self-compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2019.

[189] M. Lee, P. Liang, and Q. Yang. Coauthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2022.

[190] M. H. Lee, D. P. Siewiorek, A. Smailagic, A. Bernardino, and S. B. Bermúdez i Badia. A human-AI collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14. Association for Computing Machinery, 2021.

[191] S.-C. Lee, J. Song, E.-Y. Ko, S. Park, J. Kim, and J. Kim. Solutionchat: Real-time moderator support for chat-based structured discussion. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12. Association for Computing Machinery, 2020.

[192] Y. Lee and D.-S. Kim. *Internal Migration in South Korea*, pages 93–111. Springer, 2020.

[193] Y.-C. Lee, N. Yamashita, and Y. Huang. Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–27, May 2020.

[194] Y.-C. Lee, N. Yamashita, Y. Huang, and W. Fu. "I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2020.

[195] A. Levy, M. Agrawal, A. Satyanarayan, and D. Sontag. Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021.

[196] I. Li, A. Dey, and J. Forlizzi. A stage-based model of personal informatics systems. In *Proceedings of the 2010 Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2010.

[197] I. Li, A. K. Dey, and J. Forlizzi. Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th International Conference on Ubiquitous Computing*. Association for Computing Machinery, 2011.

[198] C. Lim, A. B. Berry, T. Hirsch, A. L. Hartzler, E. H. Wagner, E. Ludman, and J. D. Ralston. " it just seems outside my health" how patients with chronic conditions perceive communication boundaries with providers. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. Association for Computing Machinery, 2016.

[199] C. Y. Lim, A. B. Berry, A. L. Hartzler, T. Hirsch, D. S. Carrell, Z. A. Bermet, and J. D. Ralston. Facilitating self-reflection about values and self-care among individuals with chronic conditions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2019.

[200] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[201] S. E. Lord, K. J. Trudeau, R. A. Black, L. Lorin, E. Cooney, A. Villapiano, and S. F. Butler. CHAT: Development and validation of a computer-delivered, self-report, substance use assessment for adolescents. *Substance Use and Misuse*, 46(6):781–794, 2011.

[202] R. J. Lordon, S. P. Mikles, L. Kneale, H. L. Evans, S. A. Munson, U. Backonja, and W. B. Lober. How patient-generated health data and patient-reported outcomes affect patient–clinician relationships: A systematic review. *Health Informatics Journal*, 26(4):2689–2706, 2020.

[203] X. Lu, W. Ai, X. Liu, Q. Li, N. Wang, G. Huang, and Q. Mei. Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. Association for Computing Machinery, 2016.

[204] X. Lu, E. Jo, S. Park, H. Hong, Y. Chen, and D. A. Epstein. Understanding cultural influence on perspectives around contact tracing strategies. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–26, 2022.

[205] X. Lu, T. L. Reynolds, E. Jo, H. Hong, X. Page, Y. Chen, and D. A. Epstein. Comparing Perspectives Around Human and Technology Support for Contact Tracing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15. Association for Computing Machinery, 2021.

[206] Y. Luo, Y.-H. Kim, B. Lee, N. Hassan, and E. K. Choe. Foodscrap: Promoting rich data capture and reflective food journaling through speech input. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*. Association for Computing Machinery, 2021.

[207] Y. Luo, B. Lee, and E. K. Choe. TandemTrack: Shaping consistent exercise experience by complementing a mobile app with a smart speaker. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–13. Association for Computing Machinery, 2020.

[208] C. R. Lyles, A. Altschuler, N. Chawla, C. Kowalski, D. McQuillan, E. Bayliss, M. Heisler, and R. W. Grant. User-centered design of a tablet waiting room tool for complex patients to prioritize discussion topics for primary care visits. *JMIR mHealth and uHealth*, 4(3):1–10, 2016.

[209] A. Lyon, S. A. Munson, M. Reddy, S. M. Schueller, E. Agapie, S. Yarosh, A. Dopp, U. von Thiele Schwarz, G. Doherty, A. K. Graham, K. P. Kruzan, and R. Kornfield. Bridging HCI and implementation science for innovation adoption and public health impact. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2023.

[210] W. Maeng and J. Lee. Designing and evaluating a chatbot for survivors of image-based sexual abuse. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. Association for Computing Machinery, 2022.

[211] R. Maharjan, K. Doherty, D. A. Rohani, P. Bækgaard, and J. E. Bardram. Experiences of a speech-enabled conversational agent for the self-report of well-being among people living with affective disorders: An in-the-wild study. *ACM Transactions on Interactive Intelligent Systems*, 12(2), 2022.

[212] G. Marcu, J. E. Bardram, and S. Gabrielli. A framework for overcoming challenges in designing persuasive monitoring and feedback systems for mental illness. In *The 5th International Conference on Pervasive Computing Technologies for Healthcare*. IEEE, 2011.

[213] A. Mate, L. Madaan, A. Taneja, N. Madhiwalla, S. Verma, G. Singh, A. Hegde, P. Varakantham, and M. Tambe. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. volume 36, pages 12017–12025, June 2022.

[214] M. Matthews, E. Murnane, and J. Snyder. Quantifying the Changeable Self: The Role of Self-Tracking in Coming to Terms With and Managing Bipolar Disorder. *Human-Computer Interaction*, 32(5-6):413–446, 2017.

[215] M. Matthews, S. Voida, S. Abdullah, G. Doherty, T. Choudhury, S. Im, and G. Gay. In situ design for mental illness: Considering the pathology of bipolar disorder in mhealth design. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 86–97. Association for Computing Machinery, 2015.

[216] J. McCabe, M. Wilcock, K. Atkinson, R. Laugharne, and R. Shankar. General practitioners' and psychiatrists' attitudes towards antidepressant withdrawal. *BJPsych Open*, 6(4):1–6, 2020.

[217] Medic. Community health toolkit, 2024. Accessed: 03-31-2025.

[218] Medicaid.gov. Medicaid, 2021. Accessed: 09-02-2021.

[219] Microsoft, Inc. Copilot, 2023. Accessed: 03-31-2025.

[220] B. Middleton, D. F. Sittig, and A. Wright. Clinical decision support: a 25 year retrospective and a 25 year vision. *Yearbook of medical informatics*, pages S103–s116, 2016.

[221] J. Ming, S. Kamath, E. Kuo, M. Sterling, N. Dell, and A. Vashistha. Invisible work in two frontline health contexts. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*. Association for Computing Machinery, 2022.

[222] J. Ming, E. Kuo, K. Go, E. Tseng, J. Kallas, A. Vashistha, M. Sterling, and N. Dell. "i go beyond and beyond" examining the invisible work of home health aides. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), Apr. 2023.

[223] S. R. Mishra, P. Klasnja, J. MacDuffie Woodburn, E. B. Hekler, L. Omberg, M. Kellen, and L. Mangravite. Supporting Coping with Parkinson's Disease Through Self Tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2019.

[224] S. R. Mishra, A. D. Miller, S. Haldar, M. Khelifi, J. Eschler, R. G. Elera, A. H. Pollack, and W. Pratt. Supporting collaborative health tracking in the hospital: patients' perspectives. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14. Association for Computing Marhinery, 2018.

[225] N. Moghe, S. Arora, S. Banerjee, and M. M. Khapra. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

[226] N. H. Møller, I. Shklovski, and T. T. Hildebrandt. Shifting concepts of value: Designing algorithmic decision-support systems for public services. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. Association for Computing Machinery, 2020.

[227] J. Moore, P. Goffin, M. Meyer, P. Lundrigan, N. Patwari, K. Sward, and J. Wiese. Managing in-home environments through sensing, annotating, and visualizing air quality data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3), Sept. 2018.

[228] M. Munikar, S. Shakya, and A. Shrestha. Fine-grained sentiment classification using bert. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–5, 2019.

[229] E. L. Murnane, D. Cosley, P. Chang, S. Guha, E. Frank, G. Gay, and M. Matthews. Self-monitoring practices, attitudes, and needs of individuals with bipolar disorder: Implications for the design of technologies to manage mental health. *Journal of the American Medical Informatics Association*, 23(3):477–484, 2016.

[230] E. L. Murnane, T. G. Walker, B. Tench, S. Voida, and J. Snyder. Personal informatics in interpersonal contexts: Towards the design of technology that supports the social ecologies of long-term mental health management. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), November 2018.

[231] E. D. Mynatt, I. Essa, and W. Rogers. Increasing the opportunities for aging in place. In *Proceedings on the 2000 Conference on Universal Usability*, pages 65–71, Arlington, Virginia, United States, 2000. Association for Computing Machinery.

[232] National Academies of Sciences, Engineering, and Medicine. *Social Isolation and Loneliness in Older Adults: Opportunities for the Health Care System*. The National Academies Press, Washington, DC, 2020.

[233] National Institute for Health Care and Excellence. Depression in adults: treatment and management full guideline. May 2018.

[234] National Institute of Mental Health. Mental Health Medications, 2016. Accessed: 09-02-2021.

[235] Z. Niazkhani, H. Pirnejad, M. Berg, and J. Aarts. The impact of computerized provider order entry systems on inpatient clinical workflow: A literature review. *Journal of the American Medical Informatics Association*, 16(4):539–549, 2009.

[236] V. Nirmala and A. Rajagopal. Artifically intelligent physics solver: This ai understands newtons law. *Science & Technology Journal*, 7(1):22–28, 2019.

[237] NPR. Doctors slow to adopt tech tools that might save patients money on drugs, July 2019. Accessed: 03-31-2025.

[238] NPR. A tech powerhouse, U.S. lags in using smartphones for contact tracing, Sept. 2020. Accessed: 03-31-2025.

[239] NPR. When insurance won't cover drugs, americans make 'tough choices' about their health, Jan. 2020. Accessed: 03-31-2025.

[240] NPR. Why contact tracing couldn't keep up with the u.s. covid outbreak, June 2021. Accessed: 03-31-2025.

[241] NPR. Health insurers cover fewer drugs and make them harder to get, June 2024. Accessed: 03-31-2025.

[242] N. R. Ogle and S. R. Akkerman. Guidance for the discontinuation or switching of antidepressant therapies in adults. *Journal of Pharmacy Practice*, 26(4):389–396, 2013.

[243] C. Y. Oh, Y. Luo, B. St. Jean, and E. K. Choe. Patients waiting for cues: information asymmetries and challenges in sharing patient-generated data in the clinic. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–23, 2022.

[244] F. Okeke, E. Tseng, B. Piantella, M. Brown, H. Kaur, M. R. Sterling, and N. Dell. Technology, home health care, and heart failure: a qualitative analysis with multiple stakeholders. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*. Association for Computing Machinery, 2019.

[245] C. T. Okolo, S. Kamath, N. Dell, and A. Vashistha. "It cannot do all of my work": Community health worker perceptions of ai-enabled mobile health applications in rural india. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. Association for Computing Machinery, 2021.

[246] A. Olmo, S. Sreedharan, and S. Kambhampati. Gpt3-to-plan: Extracting plans from text using gpt-3. In *ICAPS '21 Workshop on Knowledge Engineering for Planning and Scheduling*, 2021.

[247] OpenAI, Inc. ChatGPT, 2022. Accessed: 03-31-2025.

[248] OpenAI, Inc. DALLE-3, 2023. Accessed: 03-31-2025.

[249] OpenAI, Inc. Memory and new controls for chatgpt, February 2024. Accessed: 03-31-2025.

[250] J. Pal, A. Dasika, A. Hasan, J. Wolf, N. Reid, V. Kameswaran, P. Yardi, A. Mackay, A. Wagner, B. Mukherjee, S. Joshi, S. Santra, and P. Pandey. Changing data practices for community health workers: Introducing digital data collection in west bengal, india. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*, ICTD '17. Association for Computing Machinery, 2017.

[251] J. E. Palmier-Claus, J. Ainsworth, M. Machin, C. Barrowclough, G. Dunn, E. Barkus, A. Rogers, T. Wykes, S. Kapur, I. Buchan, E. Salter, and S. W. Lewis. The feasibility and validity of ambulatory self-report of psychotic symptoms using a smartphone software application. *BMC Psychiatry*, 12(1):1, 2012.

[252] A. Papoutsaki, S. So, G. Kenderova, B. Shapiro, and D. A. Epstein. Understanding delivery of collectively built protocols in an online health community for discontinuation of psychiatric drugs. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–29, 2021.

[253] S. Park, A. Thieme, J. Han, S. Lee, W. Rhee, and B. Suh. "I wrote as if I were telling a story to someone I knew.": Designing Chatbot Interactions for Expressive Writing in Mental Health. In *Designing Interactive Systems Conference 2021*. Association for Computing Machinery, 2021.

[254] S. Y. Park, K. Pine, and Y. Chen. Local-universality: Designing emr to support localized informal documentation practices. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. Association for Computing Machinery, 2013.

[255] R. A. Patel, P. Klasnja, A. Hartzler, K. T. Unruh, and W. Pratt. Probing the benefits of real-time tracking during cancer care. In *AMIA Annual Symposium Proceedings*, 2012.

[256] S. N. Patel, S. Gupta, and M. S. Reynolds. The design and evaluation of an end-user-deployable, whole house, contactless power consumption sensor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2010.

[257] E. S. Patterson, R. I. Cook, and M. L. Render. Improving patient safety by identifying side effects from introducing bar coding in medication administration. *Journal of the American Medical Informatics Association*, 9(5):540–553, 2002.

[258] R. Paulose-Ram, M. A. Safran, B. S. Jonas, Q. Gu, and D. Orwig. Trends in psychotropic medication use among U.S. adults. *Pharmacoepidemiology and Drug Safety*, 16(5):560–570, May 2007.

[259] S. R. Pendse, F. M. Lalani, M. De Choudhury, A. Sharma, and N. Kumar. "Like Shock Absorbers": understanding the human infrastructures of technology-mediated mental health support. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2020.

[260] S. R. Pendse, A. Sharma, A. Vashistha, M. De Choudhury, and N. Kumar. "Can I not be suicidal on a Sunday?": Understanding technology-mediated pathways to mental health support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021.

[261] T. Perrier, N. Dell, B. DeRenzi, R. Anderson, J. Kinuthia, J. Unger, and G. John-Stewart. Engaging pregnant women in kenya with a hybrid computer-human sms

communication system. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2015.

[262] A. Pichon, K. Schiffer, E. Horan, B. Massey, S. Bakken, L. Mamykina, and N. Elhadad. Divided we stand: the collaborative work of patients and providers in an enigmatic chronic disease. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3), 2021.

[263] E. M. Piras and F. Miele. Clinical self-tracking and monitoring technologies: negotiations in the ICT-mediated patient–provider relationship. *Health Sociology Review*, 26(1):38–53, 2017.

[264] S. K. Pontefract, J. J. Coleman, H. K. Vallance, C. A. Hirsch, S. Shah, J. F. Marriott, and S. Redwood. The impact of computerised physician order entry and clinical decision support on pharmacist-physician communication in the hospital setting: A qualitative study. *PLoS ONE*, 13(11):1–15, 2018.

[265] A. Pradhan, L. Findlater, and A. Lazar. "Phantom friend" or "Just a box with information": Personification and ontological categorization of smart speaker-based voice assistants by older adults. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 2019.

[266] B. A. Price, R. Kelly, V. Mehta, C. McCormick, H. Ahmed, and O. Pearce. Feel my pain: design and evaluation of painpad, a tangible device for supporting inpatient self-logging of pain. In *Proceedings of the 2018 CHI conference on Human Factors in Computing Systems*, 2018.

[267] R. Psych. Position statement on antidepressants and depression. *Journal of Nursing Studies*, 49(10):1–29, 2019.

[268] C. Qu, L. Yang, M. Qiu, W. B. Croft, Y. Zhang, and M. Iyyer. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1133–1136. Association for Computing Machinery, 2019.

[269] S. Raj, M. W. Newman, J. M. Lee, and M. S. Ackerman. Understanding individual and collaborative problem-solving with patient-generated data: Challenges and opportunities. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–18, 2017.

[270] D. Ramachandran, J. Canny, P. D. Das, and E. Cutrell. Mobile-izing health workers in rural india. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2010.

[271] J. Read. How common and severe are six withdrawal effects from, and addiction to, antidepressants? The experiences of a large international sample of patients. *Addictive Behaviors*, 102(July 2019):106157, 2020.

[272] J. Read, C. Cartwright, and K. Gibson. How many of 1829 antidepressant users report withdrawal effects or addiction? *International Journal of Mental Health Nursing*, 27(6):1805–1815, 2018.

[273] O. K. Richards, G. Marcu, and R. N. Brewer. Hugs, bible study, and speakeasies: designing for older adults' multimodal connectedness. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*. Association for Computing Machinery, 2021.

[274] C. Roberts, M. Mort, and C. Milligan. Calling for care: Disembodied work, teleoperators and older people living at home. *Sociology*, 46(3):490–506, 2012.

[275] D. A. Rohani, A. Quemada Lopategui, N. Tuxen, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram. MUBS: A personalized recommender system for behavioral activation in mental health. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. Association for Computing Machinery, 2020.

[276] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, and J. Weston. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online, Apr. 2021. Association for Computational Linguistics.

[277] J. Rooksby, M. Rost, A. Morrison, and M. Chalmers. Personal tracking as lived informatics. In *Proceedings of the 2014 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2014.

[278] J. F. Rosenbaum, M. Fava, S. L. Hoog, R. C. Ascroft, and W. B. Krebs. Selective serotonin reuptake inhibitor discontinuation syndrome: A randomized clinical trial. *Biological Psychiatry*, 44(2):77–87, 1998.

[279] E. L. Ross, R. N. Jamison, L. Nicholls, B. M. Perry, and K. D. Nolen. Clinical integration of a smartphone app for patients with chronic pain: Retrospective analysis of predictors of benefits and patient engagement between clinic visits. *Journal of Medical Internet Research*, 22(4):1–13, 2020.

[280] J. Rowan and E. D. Mynatt. Digital family portrait field trial: Support for aging in place. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2005.

[281] H. Ryu, S. Kim, D. Kim, S. Han, K. Lee, and Y. Kang. Simple and Steady Interactions Win the Healthy Mentality: Designing a Chatbot Service for the Elderly. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–25, Oct. 2020.

[282] J. Saldaña. *The coding manual for qualitative researchers*. Sage, 2022.

[283] N. Sambasivan and T. Smyth. The human infrastructure of ICTD. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*. Association for Computing Machinery, 2010.

[284] P. C. Sanger, A. Hartzler, R. J. Lordon, C. A. L. Armstrong, W. B. Lober, H. L. Evans, and W. Pratt. A patient-centered system in a provider-centered world: Challenges of incorporating post-discharge wound data into practice. *Journal of the American Medical Informatics Association*, 23(3):514–525, may 2016.

[285] M. J. Santana, L. Haverman, K. Absolom, E. Takeuchi, D. Feeny, M. Grootenhuis, and G. Velikova. Training clinicians in how to use patient-reported outcome measures in routine clinical practice. *Quality of Life Research*, 24(7):1707–1718, 2015.

[286] D. Saxena, K. Badillo-Urquiola, P. J. Wisniewski, and S. Guha. A framework of high-stakes algorithmic decision-making for the public sector developed through a case study of child-welfare. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 2021.

[287] K. Schmidt and L. Bannon. Taking CSCW seriously: Supporting articulation work. *Computer Supported Cooperative Work (CSCW)*, 1:7–40, 1992.

[288] J. Schroeder, C.-F. Chung, D. A. Epstein, R. Karkar, A. Parsons, N. Murinova, J. Fogarty, and S. A. Munson. 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*. Association for Computing M, 2018.

[289] J. Schroeder, J. Hoffswell, C.-F. Chung, J. Fogarty, S. Munson, and J. Zia. Supporting patient-provider collaboration to identify individual triggers using food and symptom journals. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, 2017.

[290] J. Schroeder, R. Karkar, N. Murinova, J. Fogarty, and S. A. Munson. Examining opportunities for goal-directed self-tracking to support chronic condition management. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4), 2019.

[291] D. M. Scott. United States health care system: A pharmacy perspective. *Canadian Journal of Hospital Pharmacy*, 69(4):306–315, 2016.

[292] E. Seto, P. Challa, and P. Ware. Adoption of COVID-19 contact tracing apps: A balance between privacy and effectiveness. *Journal of Medical Internet Research*, 23(3):e25726, 2021.

[293] G. H. Severinsen, L. Silsand, G. Ellingsen, and R. Pedersen. From Free-Text to Structure in Electronic Patient Records. *Studies in Health Technology and Informatics*, 265:86–91, 2019.

[294] B. B. Shapiro. Subtherapeutic doses of SSRI antidepressants demonstrate considerable serotonin transporter occupancy: implications for tapering ssris. *Psychopharmacology*, 235(9):2779–2781, 2018.

[295] A. Sharkey and N. Sharkey. Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1):27–40, 2012.

[296] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293. Association for Computational Linguistics, 2021.

[297] V. Shwartz and Y. Choi. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870. International Committee on Computational Linguistics, 2020.

[298] E. Simpson, R. Comber, A. Garbett, E. I. Jenkins, and M. Balaam. Experiences of delivering a public health data service. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, page 6171–6183. Association for Computing Machinery, 2017.

[299] H. Singh, F. Tahsin, J. X. Nie, B. McKinstry, K. Thavorn, R. Upshur, S. Harvey, W. P. Wodchis, and C. S. Gray. Exploring the perspectives of primary care providers on use of the electronic patient reported outcomes tool to support goal-oriented care: A qualitative study. *BMC Medical Informatics and Decision Making*, 21(1):1–15, 2021.

[300] L. E. Snyder, D. F. Phan, K. C. Williams, E. Piqueiras, S. E. Connor, S. George, L. Kwan, J. Villatoro Chavez, M. D. Tandel, S. K. Frencher, M. S. Litwin, J. L. Gore, and A. L. Hartzler. Comprehension, utility, and preferences of prostate cancer survivors for visual timelines of patient-reported outcomes co-designed for limited graph literacy: Meters and emojis over comics. *Journal of the American Medical Informatics Association*, 29(11):1838–1846, 2022.

[301] A. Sorensen, L. W. Le, N. Swami, B. Hannon, M. K. Krzyzanowska, K. Wentlandt, G. Rodin, and C. Zimmermann. Readiness for delivering early palliative care: A survey of primary care and specialised physicians. *Palliative Medicine*, 34(1):114–125, 2020.

[302] S. Spark, D. Lewis, A. Vaisey, E. Smyth, A. Wood, M. Temple-Smith, R. Lorch, R. Guy, and J. Hocking. Using computer-assisted survey instruments instead of paper and pencil increased completeness of self-administered sexual behavior questionnaires. *Journal of Clinical Epidemiology*, 68(1):94–101, 2015.

[303] L. Stapleton, M. H. Lee, D. Qing, M. Wright, A. Chouldechova, K. Holstein, Z. S. Wu, and H. Zhu. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. Association for Computing Machinery, 2022.

[304] S. Star and K. Ruhleder. Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, 7:111–134, 03 1996.

[305] K. Stawarz, C. Preist, D. Tallon, L. Thomas, K. Turner, N. Wiles, D. Kessler, R. Shafran, and D. Coyle. Integrating the digital and the traditional to deliver therapy for depression. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2020.

[306] S. Stonbraker, T. Porras, and R. Schnall. Patient preferences for visualization of longitudinal patient-reported outcomes data. *Journal of the American Medical Informatics Association*, 27(2):212–224, 2020.

[307] E. Stowell, M. C. Lyson, H. Saksono, R. C. Wurth, H. Jimison, M. Pavel, and A. G. Parker. Designing and evaluating mhealth interventions for vulnerable populations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2018.

[308] A. Strauss. The articulation of project work: An organizational process. *The Sociological Quarterly*, 29(2):163–178, 1988.

[309] R. L. Street. Gender differences in health care provider-patient communication: Are they due to style, stereotypes, or accommodation? *Patient Education and Counseling*, 48(3):201–206, 2002.

[310] Y. Sun, X. Ma, S. Lindtner, and L. He. Data work of frontline care workers: Practices, problems, and opportunities in the context of data-driven long-term care. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–28, Apr. 2023.

[311] V. D. Swain, J. Ye, S. K. Ramesh, A. Mondal, G. D. Abowd, M. De Choudhury, et al. Leveraging social media to predict COVID-19–induced disruptions to mental well-being among university students: Modeling study. *JMIR Formative Research*, 8(1):e52316, 2024.

[312] D. Thakkar, A. Ismail, P. Kumar, A. Hanna, N. Sambasivan, and N. Kumar. When is machine learning data good?: Valuing in public health datafication. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2022.

[313] The Wall Street Journal. The apple watch is becoming doctors' favorite medical device, 2024. Accessed: 03-31-2025.

[314] E. Thomaz, V. Bettadapura, G. Reyes, M. Sandesh, G. Schindler, T. Plötz, G. D. Abowd, and I. Essa. Recognizing water-based activities in the home through infrastructure-mediated sensing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. Association for Computing Machinery, 2012.

[315] TIME. How Fitbit and Apple watch are changing health care, 2014. Accessed: 03-31-2025.

[316] TIME. Contact tracing apps were big tech's best idea for fighting COVID-19. why haven't they helped?, Nov. 2020. Accessed: 03-31-2025.

[317] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[318] E. Tseng, F. Okeke, M. Sterling, and N. Dell. "We can learn. Why not?": Designing technologies to engender equity for home health aides. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2020.

[319] A. Vashistha, E. Cutrell, G. Borriello, and W. Thies. Sangeet swara: A community-moderated voice forum in rural india. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2015.

[320] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010. Curran Associates Inc., 2017.

[321] T. C. Veinot, J. S. Ancker, H. Cole-Lewis, E. D. Mynatt, A. G. Parker, K. A. Siek, and L. Mamykina. Leveling up: On the potential of upstream health informatics interventions to enhance health equity. *Medical Care*, 57(Suppl 2), Jun 2019.

[322] N. Verdezoto, N. Bagalkot, S. Z. Akbar, S. Sharma, N. Mackintosh, D. Harrington, and P. Griffiths. The invisible work of maintenance in community health: Challenges and opportunities for digital health to support frontline health workers in Karnataka, South India. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–31, April 2021.

[323] J. Vines, S. Lindsay, G. W. Pritchard, M. Lie, D. Greathead, P. Olivier, and K. Brittain. Making family care work: Dependence, privacy and remote home monitoring telecare systems. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery, 2013.

[324] A. Waddell, J. P. Seguin, L. Wu, P. Stragalinos, J. Wherton, J. L. Watterson, C. O. Prawira, P. Olivier, V. Manning, D. Lubman, and J. Grigg. Leveraging implementation science in human-centred design for digital health. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2024.

[325] D. Wang, M. Chheang, S. Ji, R. Mohta, and D. A. Epstein. SnapPI: Understanding Everyday Use of Personal Informatics Data Stickers on Ephemeral Social Media. *Proceedings of the ACM on Human-Computer Interaction*, 6(2 CSCW), 2022.

[326] D. Wang, L. Wang, Z. Zhang, D. Wang, H. Zhu, Y. Gao, X. Fan, and F. Tian. "brilliant ai doctor" in rural clinics: Challenges in ai-powered clinical decision support system deployment. In *Proceedings of the 2021 CHI conference on Human Factors in Computing Systems*. Associtation for Computing Machinery, 2021.

[327] L. Wang, M. I. Mujib, J. Williams, G. Demiris, and J. Huh-Yoo. An evaluation of generative pre-training model-based therapy chatbot for caregivers. *arXiv preprint arXiv:2107.13115*, 2021.

[328] C. H. Warner, W. Bobo, C. Warner, S. Reid, and J. Rachal. Antidepressant discontinuation syndrome. *American Family Physician*, 74(3):449–456, 2006.

[329] R. L. Wears and M. Berg. Computer technology and clinical work. *Journal of the American Medical Association*, 293(10):1261, 2005.

[330] J. Wei, S. Kim, H. Jung, and Y.-H. Kim. Leveraging large language models to power chatbots for collecting user self-reported data. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–35, 2024.

[331] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

[332] P. West, R. Giordano, M. Van Kleek, and N. Shadbolt. The quantified patient in the doctor's office: Challenges and opportunities. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2016.

[333] P. West, M. Van Kleek, R. Giordano, M. J. Weal, and N. Shadbolt. Common barriers to the use of patient-generated data across clinical settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2018.

[334] G. White, T. Singh, K. Caine, and K. Connelly. Limited but satisfied: Low SES older adults experiences of aging in place. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare*. IEEE, 2015.

[335] N. J. Wolf and D. R. Hopko. Psychosocial and pharmacological interventions for depressed adults in primary care: A critical review. *Clinical Psychology Review*, 28(1):131–161, 2008.

[336] C. World. The lonely deaths: South korea's government scrambles to tackle the rise of 'godoksa', Dec 2022. Accessed: 03-31-2025.

[337] J. Xu, A. Szlam, and J. Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[338] X. Xu, Z. Gou, W. Wu, Z.-Y. Niu, H. Wu, H. Wang, and S. Wang. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[339] A. Yadav, B. Wilder, E. Rice, R. Petering, J. Craddock, A. Yoshioka-Maxwell, M. Hemler, L. Onasch-Vera, M. Tambe, and D. Woo. Bridging the gap between theory and practice in influence maximization: Raising awareness about HIV among homeless youth. In *International Joint Conferences on Artificial Inteligence*, pages 5399–5403, 2018.

[340] D. Yadav, P. Malik, K. Dabas, and P. Singh. Feedpal: Understanding opportunities for chatbots in breastfeeding education of women in india. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), November 2019.

[341] F. Yamamoto, A. Voida, and S. Voida. From therapy to teletherapy: Relocating mental health services online. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–30, 2021.

[342] N. Yamashita, H. Kuzuoka, K. Hirata, T. Kudo, E. Aramaki, and K. Hattori. Changing moods: How manual tracking by family caregivers improves caring and family communication. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2017.

[343] J. Yang, M. Wang, H. Zhou, C. Zhao, W. Zhang, Y. Yu, and L. Li. Towards making the most of BERT in neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9378–9385, April 2020.

[344] Q. Yang, J. Cranshaw, S. Amershi, S. T. Iqbal, and J. Teevan. Sketching NLP: A case study of exploring the right things to design with language intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2019.

[345] Q. Yang, J. Zimmerman, A. Steinfeld, L. Carey, and J. F. Antaki. Investigating the heart pump implant decision process: opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2016.

[346] Yonhap. Lonely deaths of middle-aged, youth brackets stand out amid single-person households - The Korea Herald, Dec 2017. Accessed: 03-31-2025.

[347] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. Association for Computational Linguistics, July 2018.

[348] Z. Zheng, L. Liao, Y. Deng, and L. Nie. Building emotional support chatbots in the era of LLMs. *arXiv preprint arXiv:2308.11584*, 2023.

[349] S. Zhong, D. Lalanne, and H. Alavi. The complexity of indoor air quality forecasting and the simplicity of interacting with it – a case study of 1007 office meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021.

[350] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, Apr. 2018.

[351] H. Zhu, J. Colgan, M. Reddy, and E. K. Choe. Sharing patient-generated data in clinical practices: an interview study. In *AMIA Annual Symposium Proceedings*, volume 2016, page 1303, 2017.

[352] H. Zhu, Z. J. Moffa, X. Gui, and J. M. Carroll. Prehabilitation: Care challenges and technological opportunities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2020.