Designing Personal Health Technologies for Translating Population-Level Evidence into Personal Understanding

Ravi Karkar

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

James Fogarty, Chair

Sean A. Munson

Julie A. Kientz

Program Authorized to Offer Degree:
Computer Science & Engineering

University of Washington

Abstract

Designing Personal Health Technologies for
Translating Population-Level Evidence into Personal Understanding

Ravi Karkar

Chair of the Supervisory Committee:
Professor James Fogarty
Computer Science & Engineering

Growing interest in better tracking and managing health is leading a global market for self-monitoring technologies to reach an estimated $72 billion by 2022. This boom has contributed to a new potential to support people in collecting and interpreting data about their own health and well-being. However, there is often a mismatch between what technology currently delivers (e.g., step counts, sleep scores) versus what people expect from it (e.g., personalized health insights and recommendations). Current technologies often fall short of their potential to address individual health needs due to complex and interrelated challenges (e.g., in meeting individual needs, in data quality, in their integration into clinical practice).

Such personalized understanding of one's own health is vital for people with various chronic conditions where symptoms and their triggers generally vary across individuals (e.g., irritable bowel syndrome, migraine). To scaffold people's attempts to better understand their chronic conditions through collecting and interpreting personal health data, I first present a framework for self-experimentation. The framework provides guidelines on how tools can be designed to provide the necessary support in avoiding common pitfalls associated with tracking health data and in assisting people in answering health questions in a scientifically valid manner. Through the framework, I propose better design of tools aimed at assisting people in answering questions related to the individualized nature of their condition using personal health data.

To demonstrate a practical application of the framework, I next present my work in the

domain of Irritable Bowel Syndrome. I leveraged the self-experimentation framework to design TummyTrials. TummyTrials incorporates existing medical knowledge to assist people in conducting scientifically valid self-experiments in the context of their day-to-day life to determine which foods are triggering their symptoms. Through my research in TummyTrials, I demonstrate the feasibility of people conducting rigorous experiments when a system provides sufficient scaffolding to bridge associated expertise gaps.

I next explored supporting people with chronic liver diseases in gaining an understanding of their cognitive function. I developed Beacon, a platform that enables cirrhotic patients to self-measure critical flicker frequency, a measure of cognitive function that has been considered promising in this population but has previously been impractical to collect. Through Beacon, I am enabling a new data stream for patients and providers to gain a easy-to-measure proxy of cognitive functioning in a high-risk population.

Across these projects, the common theme is my pursuit of designing, developing, and promoting tools which leverage the existing population-level evidence in medicine to provide new value at a personal level. I synthesize design recommendations based on empirical findings from building the above tools to assist future tool builders in designing improved tools. I emphasize that realizing the full potential of new tracking technologies requires that designers and researchers scaffold the necessary domain expertise and the individual context to provide meaningful data experiences.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The footnote on the Google Scholar home page has always deeply resonated with me – *Stand on the shoulders of giants*. I feel that it perfectly encapsulates my journey over the past six years. My PhD was made possible through the guidance, contribution, and support of a multitude of people. I believe it takes a community to produce a good researcher and *DUB* provided me with that community at UW. Below I thank people without whom I would not be at this stage in my life.

I would like to start-off by thanking Gregory Abowd who not only introduced me to research in HCI but also set me on my path of pursuing a PhD. I am happy I heeded Edison Thomaz's advice when I was applying to graduate schools and reached out to James Fogarty as a potential advisor. James set the tone of the advisor-advisee relationship during the CSE visit days by wiping the floor with me in multiple bar games. Over the years I have come to appreciate his honest advising style and cherished a shared grumpy attitude towards many things. It would be an understatement to say I could not have made it this far without having James's mentorship and support. He has been and continues to be my biggest advocate and for that I am deeply grateful.

I consider myself to be very fortunate to have a dissertation committee of researchers who I consider to be role-models and whom I wish to emulate. Sean Munson has been a collaborator and a mentor since day one. He has been a stable foundation in my dissertation research where I can always count on him to engage with any problem and offer insightful suggestions. Sean has taught me not only how to be a good qualitative researcher but also how to be a well-balanced researcher by prioritizing a healthy work-life balance. Although I did not get to collaborate with Julie Kientz during the second half of my dissertation research, she has been an instrumental part of my PhD. Julie always brought in a diverse perspective and challenged me to think critically in my research. Among the few regrets that I have from my time at UW, the foremost is that I did not get to collaborate with Wanda Pratt or Anind Dey on a research project. Regardless, they have provided me with invaluable feedback and guidance on framing my research and career planning

for which I am grateful.

I would also like to express my gratitude toward my medical collaborators – Jasmine Zia, George Ioannou, Kara Walter, and Philip Vutien. Without their expertise and support this dissertation would not have been possible.

There are many past and present students at UW who have made my journey enjoyable. In no particular order, I thank Daniel Epstein for being the de facto person I turn to for any questions related to Personal Informatics, Elena Agapie for her continued support and the shared venting sessions, Jessica Schroeder for showing me what an organized researcher looks like, Fahad Pervaiz for introducing me to ICTD, Edward Wang for stepping in when I am way in over my head with electronics, Eric Whitmire for being a constant source of inspiration and an all-round debugger, Alex Mariakakis for always being there to help with research and joining in the food excursions, Aditya Vashistha for sharing his wisdom and support over coffee and badminton, Catie Baker for somehow convincing me to run a half marathon, Vincent Lee for the delightfully grumpy meals and the hikes, Matthew Kay for introducing me to uncertainity and Bayesian statistics, and Mayank Goel for inspiring me to pursue health research that can be translated to practice. I would also like to thank the Fogies for putting up with the incessant parade of practice talks and requests for feedback on drafts and for piloting studies – Raymond Fok, Liwei Jiang, Richard Li, Jesse Martinez, Alex Okeson, Anne Ross, Jina Suh, Amanda Swearngin, Xiaoyi Zhang, and Mingyuan (Jason) Zhong.

I am grateful to have a group of friends to not talk research with. Tushar Chaturvedi, Shrainik Jain, Madhurima Pore, Susmita Rishi, and Rishabh Shukla thank you for your continued companionship.

Finally, this endeavour would not have been possible without the continued support of my family. I would like to thank my parents Kantilal and Kalavati Karkar for encouraging me to pursue my PhD. I am deeply grateful to have the constant support of my brother Jay Karkar. My wife, Aesha Bhattacharya, has made the second half of my PhD much more enjoyable. She has been very patient in the face of my ever changing work hours and habits and has even occasionally stepped up to the role of a project manager to ensure that I finish my paper, job application, or dissertation on time.

This research could not have been possible without funding from the Intel Science and Technology Center for Pervasive Computing, Nokia Research, the National Science Foundation under awards IIS-1553167, IIS-1813675 and SCH-1344613, the Agency for Healthcare Research Quality under award 1R21HS023654, and the National Institute on Drug Abuse under award 1K99DA037276-01.

Chapter 1

INTRODUCTION

Current medical research, evidence, and understanding is primarily based on population-level group designs, such as epidemiological surveys, longitudinal studies, and randomized controlled trials. Such methods can provide a good understanding of the epidemiology, clinical course, and effects of specific treatments for certain medical conditions. Despite their merit, these traditional methods generally cannot address the question most relevant to any specific individual: *what is the likelihood that a treatment will have an effect on the symptoms of that individual?* This question is particularly important for people with chronic conditions, where individual variations are common and often lead to challenges in effective diagnosis and management.

Chronic conditions account for as much as 90% of health spending in the US, with 60% of US adults managing at least one condition [25]. Chronic conditions lead to greater difficulties with day-to-day activities and with social and cognitive functions that impact personal independence. Even among individuals with the same condition, the manifestations of symptoms, the underlying triggers, and associated challenges vary significantly. Much of the existing understanding of these conditions is based on population averages, leaving individuals struggling to understand and manage their own condition. Tracking technologies offer a potential solution by allowing people to collect data about themselves to better understand and manage their condition. However, individualized diagnosis and management of chronic conditions requires novel approaches to collecting and interpreting streams of personal health data, something not yet well-supported by existing tracking technologies.

People have access to a range of activity trackers that track different aspects of one's life, such as physical fitness [41,64,108], sleep [64,97,108], diet [14,43,120], smoking [2], and stress [132]. However, these commercial trackers primarily focus on supporting a high-level health goal (e.g., staying healthy, sleeping better) and often fail to help people answer specific questions they might have regarding their health or other aspects of their lives (e.g., does caffeine affect my abdominal

pain?). Existing tools generally support reviewing collected data over time or performing simple correlational analyses, both of which are often insufficient to answer specific questions people have about the different aspects of their health and well-being.

People with technical expertise can build their own solutions using commercial or custom tools to answer questions they have. However, research has shown that even such relative experts are prone to common pitfalls which can prevent them from finding the answer they were looking for, or worse, lead them to false conclusions. Three common pitfalls that such individuals often face when trying to answer specific questions about their health include: (1) tracking too many variables, (2) not tracking the appropriate triggers and context relevant to their condition, and (3) lacking scientific rigor in data collection and analysis [31].

In my research, I aim to empower and support individuals with chronic conditions through building end-to-end personal health technologies that assist in collecting and interpreting appropriate data toward meeting their individualized needs. My research focuses on human-centered approaches to collecting, interacting with, and using novel health data toward improving human well-being through personalized insights and recommendations. In addition to formalizing a framework to answer questions about personal health using novel data, my research contributes specific tools that enable easier and reliable collection and interpretation of streams of personal health data. In building these technologies I assist people in navigating around the pitfalls and scaffolding the necessary medical expertise through design.

1.1 Thesis Statement

My dissertation research demonstrates my thesis:

> *Novel personal health technologies can support collection and interpretation of personal data streams as part of translating available population-level evidence into personal understanding.*

By (1) leveraging existing *population-level evidence* in medicine and (2) identifying existing challenges that people have in determining what data to collect, how to collect it, and how to interpret it; I claim that we can design personal health technologies that support people in effective *collection and interpretation* of data towards gaining *personal understanding* of their

condition. Toward demonstrating my thesis, I contribute – a Framework for Self-Experimentation, the TummyTrials platform, and the Beacon platform.

1. People with chronic conditions seek indvidualized understanding of their health but this is not effectively supported by exising tools. The framework for self-experimentation provides guidelines on how to design tools that leverage existing population-level medical evidence to collect targeted personal health data and provide actionable personalized understanding
2. People with irritable bowel syndrome (IBS) have challenges in determining their individual triggers for their IBS symptoms. To support them, I built TummyTrials, an app that leverages existing population-level evidence supporting food as a common trigger and combines it with a self-experiment design that reduces tracking burden and potential confounds to enable people determine their individualized food triggers.
3. People with Cirrhosis do not have a way to detect decline in their cognitive functioning. To support them, I built Beacon, a platform that leverages population-level evidence that the critical flicker frequency (CFF) threshold decreases as a person's hepatic encephalopathy worsens to design a tool that is easy-to-use and enables self-assessment of their CFF.

Across these projects, the common thread is leveraging existing population-level evidence to support people in collecting and interpreting health data towards gaining personal understand of their chronic condition. The framework provides an actionable design process for tool builders, TummyTrials demonstrates an implementation of that approach within the context of IBS, and Beacon creates a new data stream by enabling collection and laying the groundwork for interpreting individual variation.

Throughout the dissertation, I use the words *tool*, *technology*, *device*, *platform*, and *system* interchangeably to mean software, hardware, or a combination of the two that provides an interface for people to use to engage with their health data.

1.2 Dissertation Overview

My dissertation makes three types of contributions: a theoretical contribution, two artifacts, and design implications based on empirical findings [176]. I first present a framework for self-experimentation which provides an actionable approach for building tools which support

running experiments to answer questions about personal health. I then present two artifacts in the form of specialized tools aimed at addressing the current and emerging needs of people with chronic conditions. Finally, I consolidate my learnings from the above projects and other relevant research to present a list of design considerations aimed at assisting future researchers and designers in creating better tools in the space.

In **Chapter 2**, I review necessary background, related work, and medical research as context for my dissertation. I first describe the research area of personal informatics, which encompasses technology related to tracking and has been the foundational motivation behind my dissertation research. I next situate my contributions by summarizing relevant literature on health research within HCI that focuses on collecting and interpreting data and translating population-level evidence to personal understanding. Finally, I establish the medical contexts for the two artifacts that I contribute in my dissertation - contexts of irritable bowel syndrome (IBS) and end-stage liver disease (cirrhosis). I describe the two conditions, what makes them difficult to diagnose and support, and discuss the current standard of diagnosis and care.

In **Chapter 3**, I introduce the framework for self-experimentation. Although the primary focus of my dissertation research is on building personal health technologies, I first need to understand people's existing practices, challenges therein, and desired outcomes of those practices. One prevalent practice that exists in this space is that of searching for personalized understanding through empirical evidence collected in the form of health data. With the framework, I aim to better support this practice though proposing guidelines to design tools that assist in discovering individual variations towards improved personal understanding. As part of my dissertation's demonstration of my thesis, in this chapter:

1. I examine how *population-level evidence* can be leveraged to design targeted tools to support individuals with specific chronic conditions;
2. I examine the need of *personalized understanding* among individuals with chronic conditions in the form of answering questions about the individualized nature of their health; and
3. I provide guidelines for designing tools that support the *collection and interpretation* of data collected through self-experimentation to answer questions individuals have about their health.

I thus propose self-experimentation as a more rigorous form of self-tracking that motivates tools

to assist in discovering individual variations towards improved personal understanding. The aim of the support is to assist people in (1) narrowing down the question to determine the minimum data needed; (2) using effective study design to maintain rigor while reducing burden; and (3) providing analyses to support effective decision making. I conducted three focus groups and an online survey (60 respondents) to evaluate the feasibility and understandability of the framework. This research is one of the pioneering works within HCI proposing the use of a self-experimentation or n-of-1 approach toward building and evaluating health and wellness tools.

In **Chapter 4**, I introduce the first artifact contribution – TummyTrials. TummyTrials is aimed at assisting people with irritable bowel syndrome who have challenges with determining individualized triggers for their symptoms. As part of my dissertation's demonstration of my thesis, in this chapter:

1. I leverage existing *population-level evidence* that suggests food as the most common trigger for IBS symptoms and that there is no clinical test which can determine the specific trigger for a given individual;
2. I examine IBS patient desire for *personalized understanding* of their condition by conducting self-experiments to determine which foods are triggering their symptoms and how the lack of appropriate tools creates barriers for this endeavour; and
3. I demonstrate how TummyTrials supports the patient's desire by scaffolding in-the-wild *collection and interpretation* of relevant data through designing an end-to-end guided self-experiment that is low-burden and addresses common confounds.

TummyTrials is thus an app that leverages the framework for self-experimentation to guide a person in designing, conducting, and analyzing self-experiments to determine their individualized food triggers for IBS. I designed, developed, and evaluated TummyTrials as a platform that scaffolds the necessary medical expertise and supports people in learning about the individualized nature of their IBS by conducting experiments in their day-to-day life. I investigated the feasibility of TummyTrials for supporting people in collecting and interpreting data through in-the-wild self-experiments by recruiting 15 IBS patients and asking them to undergo a 12-day self-experiment to determine if a specific food triggered their IBS symptoms. Participants found that TummyTrials reduced the overall burden of remembering and tracking, provided structure to a process that can otherwise be trial and error, and provided accountability.

In **Chapter 5**, I introduce the second artifact contribution – Beacon. Beacon is a system which enables people with cirrhosis to self-assess their current cognitive functioning in the form of their critical flicker frequency (CFF) threshold. As part of my dissertation's demonstration of my thesis, in this chapter:

1. I leverage existing *population-level* evidence that promotes the use of a CFF threshold as an early indicator and that a threshold of <39Hz is indicative of decline in cognitive performance among cirrhotic patients;
2. I examine the unmet desire among the provider, patient, and care-givers for *personalized understanding* in the form of an objective measure of a patient's cognitive functioning for timely treatment as there is currently no way for them to get such a measure; and
3. I demonstrate how Beacon scaffolds the *collection* of CFF measurement, thus laying groundwork for future work examining individual *interpretation* of CFF variation.

I thus developed the Beacon platform as a novel portable device which enables self-measurement of CFF as part of understanding and managing a chronic condition, something that is currently not well-supported through existing personal health monitoring tools. To establish the performance of Beacon compared to the existing Lafayette research device, I ran a study with 41 healthy adults ranging from 18 to 99 years of age. I found that Beacon performs on par with the existing device, achieving a Pearson's correlation of 0.88. I am currently running a comparative study among the patient population and preparing for a longitudinal deployment study aimed at understanding the day-to-day variations in cognitive performance among the patient population. Toward my vision of adopting a data driven approach to monitoring cognitive function among cirrhotic patients, Beacon provides an alternative to current occasional clinical screenings (i.e., once every three to six months), instead providing an important step toward enabling daily at-home self-monitoring.

In **Chapter 6**, I take a step back to reflect on other relevant projects I have been involved in during the research that forms my dissertation. These additional projects have helped shape my perspectives on building better personal health technologies and provide additional context for the design implications I discuss in Chapter 7.

Finally in **Chapter 7**, I discuss how the presented projects provide evidence in support of my thesis statement. Additionally, I discuss broader design implication stemming from the empirical results of evaluating the artifacts I built and the additional projects I discussed in the prior chapter.

The goal of the design implications is to help future tool builders better understand the needs of their audience and build better tools. I then conclude by briefly describing a few research directions I plan to pursue in the near future.

Chapter 2

BACKGROUND, RELATED WORK, AND RESEARCH CONTEXT

My dissertation explores the design of tools which leverage new forms of health data to provide personalized understanding. In this chapter, I first provide relevant background information on health and health-tracking research within the broader HCI community. Next, I focus on related work in collecting and interpreting health data which allows me to situate my dissertation research within the broader literature. Finally, I provide the necessary medical context in the conditions for which I develop specialized tools. The details I present in this chapter represent the state of research in the domain when I started working on my dissertation. When I later discuss my research outcomes, I also briefly summarize research that built on top of and followed my work.

2.1 Background

Personal informatics research within HCI is often concerned with understanding the practices and design of tools to support people in their personal tracking. Researchers have proposed frameworks such as the stage-based model [113] and the lived informatics model [59] that describe how individuals engage with personal data, including the processes they follow and their motivations for self-tracking. Different people track data about themselves for different reasons (e.g., to help track a target or a goal, for documentary purposes with no clear intention of using the data, to answer questions about themselves) [148]. Understanding these differences in motivations for tracking is crucial towards designing tools to effectively support individuals.

In medicine, the rise of cheap mobile health sensing platforms like physical trackers and smartphone apps have resulted in a new paradigm of *healthcare in the pocket* or mHealth [102]. Researchers have recognized the significance of self-tracked data, which is increasingly seen as an important contributor to improvements in both medical care and self-management [119,122,126,179]. People who self-track to answer specific questions about their health often endeavor to use data as a means to manage a condition, find triggers, or identify relationships pertaining

to their health or other aspects of life [31]. Patients with various chronic conditions (e.g., diabetes, irritable bowel syndrome, migraine, asthma, hypertension) often track related data (e.g., glucose, bowel movements, migraines, and other symptoms) using various devices and apps [34, 66, 71, 85, 119, 120, 140, 152]. Research has also examined how to support people in maintaining control of their lives despite intermittent symptoms, as with intermittent fatigue in multiple sclerosis (e.g., [5, 177]). Such research often points towards the need for personalization in the tools built to support patients in their self-tracking endeavors.

Major technology companies like Apple and Google have responded to the increasing interest in and value of self-tracked health data and begun offering their own platforms towards that need [4, 70]. Existing devices and apps in health-related domains, such as physical fitness (e.g., [41, 64, 108, 115]), sleep (e.g., [64, 97, 108]), diet (e.g., [14, 45, 120]), smoking (e.g., [2]), and stress (e.g., [132]), often focus on supporting a high-level health goal (e.g., staying healthy, sleeping better). Tools designed to support such health goals often fail to help people answer specific questions they might have regarding their health or other aspects of their lives. These tools generally support reviewing collected data over time or performing simple correlational analyses, both of which are often insufficient to answer specific questions people have about causal relationships between variables.

Researchers have examined practices of people with technical expertise who track various aspects of their life to better understand the common challenges and opportunities for future tools to better support such endeavors. Choe et. al. highlighted three common pitfalls self-trackers often face when trying to answer specific questions about their health: (1) tracking too many variables, (2) not tracking the appropriate triggers and context relevant to their condition, and (3) lacking scientific rigor in data collection and analysis [31]. In my dissertation, I build on this line of research to design and develop tools which support a person in avoiding these common pitfalls and achieve their goal of personalized understanding of their condition.

2.2 Related Work

HCI researchers have a long history of examining technology to understand and support health and well-being, and recent growth in the topic area has resulted in addition of *Health* as a subcommittee

at CHI, the premier conference in HCI. There are many different ways of looking at this vast body of research. Given my focus on collection and interpretation of personal health data, I survey relevant literature to better situate my research.

Collecting personal health data has become more accessible with the rise in adoption of smartphones and wearables. Within HCI, health researchers have explored repurposing existing sensors or creating new sensors and tools to enable collection of health data. Smartphones have provided researchers with ubiquitous access to on-person sensors such as accelerometers, microphones, and cameras. Researchers have leveraged such sensors and repurposed the signals to capture various health data. Examples of such research include using the built-in accelerometers to measure step count [42] and heart rate [81]; microphone to infer physiological state such as sleep [75] and cough [111,143], and psychological state such as stress [116]; and using the built-in camera to capture biosignals such as heart rate [42], blood volume changes [87], and hemoglobin levels [174]. Another key focus of this category of research is on exploring how specialized medical tests currently limited to research or the clinic can be made more broadly accessible. Examples of such research include SymDetector for detecting sound-based respiratory symptoms [165], contactless sleep apnea detection using smartphones [134], SpiroSmart and SpiroCall for enabling spirometry tests over a phone call [110], PupilScreen for on-field assessment of traumatic brain injury [123], categorizing skin lesions as malignant melanoma or benign moles using camera [172], and Face2Gene[2] for diagnoses rare diseases by analyzing deformities in facial structure [62]. My dissertation contributes to this line of research by presenting tools which enable the collection of health data streams towards enabling new mechanisms for understanding individual variation among people with chronic conditions.

The influx of these health data streams comes with a new set of challenges around supporting people in interpreting the data. People can be overwhelmed by new devices, sensors, and measures which are becoming commercially available, and people need support in making sense of different stages of the tracking journey, including what data to track, how to track, and how to act on it. Within Personal Informatics [113], researchers have focused on developing new approaches to support existing medical practice by augmenting the new forms of data and proposing clinical integration. This has led to the development of both (1) various models and frameworks which improve our understanding of the needs of individuals and (2) new tools to better support

individuals in their pursuits. For example, the stage-based model of personal informatics explores tracking journeys and the barriers people face at different stages [113], lived informatics model of personal informatics examines the challenges in integrating the self-tracking in everyday life [59], and the sensemaking framework examines how individuals interpret self-monitoring data towards chronic self-management [122]. Leveraging such models and frameworks and current medical evidence, researchers have also contribtued various systems to support individuals. For example, systems have examined and supported tracking practices around specific needs in mental health [9, 16], diabetes [38, 61, 121], glucose monitoring [72], and sleep apnea [50]. Finally, researchers have also explored how traditional measures can be re-imagined to better support individual health goals in areas like food journaling [15, 32, 33, 35, 44, 58, 153], weight measurement [98], and blood pressure [99]. Having access to these data streams may also change how we perceive such data. My dissertation contribution to this line of research by developing tools which incorporate and build on top of the existing understanding of how to effectively support people in collecting and interpreting personal health data by scaffolding the necessary medical expertise in the design itself.

2.2.1 Moving from Population-Level Evidence to Personalized Understanding

A majority of medical research and evidence is based on population-level understanding of health conditions. Within HCI, researchers have been able to leverage such evidence to design technologies to promote healthy behaviors. Ambient display technology such as ShutEye [13] is a good example of this type of research as it presents individuals with recommendations, based on population-level evidence, according to the time of day while not requiring any active data collection. Similar research often presents designs to promote established healthy behaviors such as physical activity, healthy sleeping routine, and overall healthy lifestyle [101]. Researchers combine the existing understanding of behavior change and emerging forms of interactive technology to promote healthy behavior change [39, 121]. For example, UbiFit Garden leveraged step count data to present a *live garden* on a mobile home screen to encourage physical activity [40], SmartQuit leveraged the Acceptance and Commitment Therapy to provide people with techniques to aid smoking cessation [22, 178], and Playful Toothbrush taught proper and thorough brushing skills

to kindergarten kids through motion tracking and virtual display [29].

However, there is a major limitation of such a macro population-level only understanding. The recommendations and guidelines stemming from them are generic. This leads to a gap between what the recommendations are proposing and what might work for a particular individual. To bridge this gap and to build tools that take existing population-level evidence and provide personal understanding I leverage a particular type of self-tracking – self-experimentation. Within HCI, self-experimentation is viewed as a subset of self-tracking in which individuals investigate the questions they have about their health in a rigorous and scientifically valid manner. As more people use sensors and tools to self-track with the goal of taking greater control of their health, a growing number of self-trackers will need better support for such self-experimentation.

Where self-tracking is a broad, well-established practice that encompasses myriad goals, self-experimentation is a method to support self-trackers as they work to test causal relationships, rather than just observe correlations. Although a novel approach in HCI, self-experimentation has been used for decades in psychology and medicine research as a valid experiment design [127]. Self-experiments are n-of-1 experiments in which an individual is their own control, highlighting that individual's response to an intervention rather than an average of many responses. Understanding individual variation and treating individual needs (i.e., personalized medicine [88]) is important in medicine and clinical science. Although personalized medicine historically emphasized genetics and pharmacology, the term is increasingly applied to other areas of health and disease [166], emphasizing the importance of personalized health.

Self-tracking methods that do not involve self-experimentation can support identification of correlations between variables and may help inform a self-experiment that can rigorously assess causation. Importantly, an experience-sampling study that asks someone to record several variables over the course of time, with no control, is self-tracking but not a self-experiment. By contrast, in a self-experiment, an individual varies one or more factors in a controlled manner, with the intent of making causal inferences about the effect of those factors [114, 145]. Medicine has long used single-case designs to determine the relationship between individualized causes and symptoms or to determine the best treatment for an individual patient [10, 11, 28, 109]. HCI researchers have recently noted the potential for technology to support individuals in conducting and analyzing self-experiments [138], independently or in collaboration with health providers [12].

The goal of a self-experiment or an n-of-1 intervention is not to gain generalizable knowledge, but to produce personal insights for an individual [135]. This focus on the individual rather than the population is what makes n-of-1 an ideal methodology for conditions where clinical diagnoses may not be accurate (e.g., relationships between symptoms and triggers not clearly identifiable through lab tests), or where the condition fluctuates over time (e.g., blood glucose level variation over time). In such conditions, an individualized approach is needed for successful interventions.

In my research, I explore the application of self-experimentation in the context of chronic conditions. Self-experiments offer an ideal approach for understanding individual variation, which is vital for effective chronic care.

2.3 Research Context

In my research I aim to inform how tools can help patients collect data and analyze it, thereby supporting patients in being equal stakeholders in their own health and well-being. I examine existing medical research to design tools that scaffold leveraging self-tracked data to assist patients in gaining a personalized understanding of their condition. To ground my research and demonstrate the efficacy of my approach I focus on two specific medical domains: irritable bowel syndrome (IBS) & end-stage liver disease (cirrhosis).

2.3.1 Irritable Bowel Syndrome

IBS is a chronic functional disorder characterized by episodic abdominal pain with diarrhea and/or constipation despite normal blood tests, X-rays, and colonoscopies. It affects 20% of the U.S. population and is one of the top 10 reasons people seek primary care [57]. Potential triggers for IBS symptom flare-ups include certain foods, eating behaviors, stress, sleep disturbances, and menstruation, with foods as the most common trigger [80]. IBS symptoms are often caused or worsened by specific foods, but different foods can be problematic for different individuals. People with IBS report a lower quality of life and consume 50% more healthcare resources than non-IBS counterparts [105, 128].

Potential triggers for IBS symptom flare-ups include certain foods, eating behaviors, stress, sleep disturbances, and menstruation, with foods as the most common trigger [76, 79]. Traditional IBS

medications have only marginal therapeutic gains of 7-15% over placebo [30]. The most promising elimination diets (e.g., lactose, fructose, gluten) surpass traditional IBS medications in their effectiveness, when both modes of therapies were compared to placebo [69, 74]. An elimination diet process can last up to six months [125, 129]. Patients find elimination diets frustrating because they are high burden, are unintuitive, and lack sufficient instructions to successfully complete and interpret [19, 20, 74, 163]. If only certain foods need to be tested, the process can be cut down from a few months to a few weeks. However, even with a reduced number of foods, successfully identifying a trigger is not guaranteed.

Current methods for identifying individualized trigger foods generally include food journals and elimination diets. However, studies have shown that journals are burdensome, difficult to maintain, and unreliable [19, 20, 74, 163]. Additionally, clinicians are not trained to analyze such food journals and their interpretations have a high degree of interobserver variability [34, 175]. Fortunately, complete elimination of all known possible trigger foods is not necessary for most people. An individual's response to specific foods is variable, with a given food triggering bowel symptoms in some people but not others [84]. This means, patients have specific foods which trigger their individualized symptoms, thus reducing the number of foods to be tested. However, even with a reduced number of foods, successfully identifying a trigger is not guaranteed.

IBS is a useful domain for understanding the potential for personal health technologies because patients struggle to manage their condition, particular triggers are highly individualized [131, 162], symptoms tend to be experienced within a short time window of consuming the trigger food [96, 162], and the current identification process is lengthy, tedious, and frustrating [84]. I aim to improve both process and outcome, aiding IBS patients in effectively determining whether a particular food is a trigger while minimizing the impact of data collection burden on their daily life.

2.3.2 Chronic Liver Diseases

End-stage chronic liver disease, including cirrhosis of the liver, is a major source of morbidity and mortality that is both preventable and currently underestimated [21]. In the United States alone, there are 3.9 million adults diagnosed with liver disease [27]. According to the Centers for Disease

Control and Prevention, cirrhosis was the 12th-leading cause of death in the United States in 2013, accounting for more than 36 thousand deaths [103].

Hepatic Encephalopathy

Up to 80% of patients with cirrhosis will develop a spectrum of neurocognitive impairments known as hepatic encephalopathy (HE) [170]. HE fluctuates over time and occurs in a spectrum of severity ranging from minimal hepatic encephalopathy (MHE) to advanced/overt hepatic encephalopathy (OHE). OHE can be life-threatening, often leading to coma or death. It is important to identify HE at its earliest stage before it progresses to more severe form. An added benefit of early diagnosis is relatively easy treatments exist for controlling MHE [7, 144, 161]. Since HE results in cognitive impairment, it is difficult for a person to self-diagnose. In addition, given the fluctuating nature of the chronic condition, an objective self-tracking measure would be ideal for the patients to get a better understanding of their condition at home.

Critical Flicker Frequency

MHE is often undiagnosed, in part because of the complicated and time-consuming nature of the major available screening tests. Two most common categories of tests are: (1) psychometric screening tests (tests measuring mental capabilities and behavioral style e.g., number connection test or NCT) and (2) neurophysiological screening tests (tests measuring function of the nervous system e.g., stroop test). Abnormalities in psychometric tests can be caused by many other conditions that affect attention and concentration, such as sleep deprivation/insomnia, medications, alcohol, and drug use. In addition, performance in these tests is strongly affected by effort, training, age, interaction with the test administrator, literacy, numeracy, and education level [91, 160]. These limitations make psychometric tests such as PHES less desirable for early diagnosis of MHE, leaving neurophysiological tests as the more reliable alternative.

The critical flicker frequency (CFF) test is a neuropsychological test that has shown promising results in early MHE diagnosis through limited clinical and research studies over the past decades [146]. CFF is defined as the frequency (in Hz) at which an intermittent light stimulus stops strobing to the observer [100, 167]. Multiple studies have established that a healthy CFF of

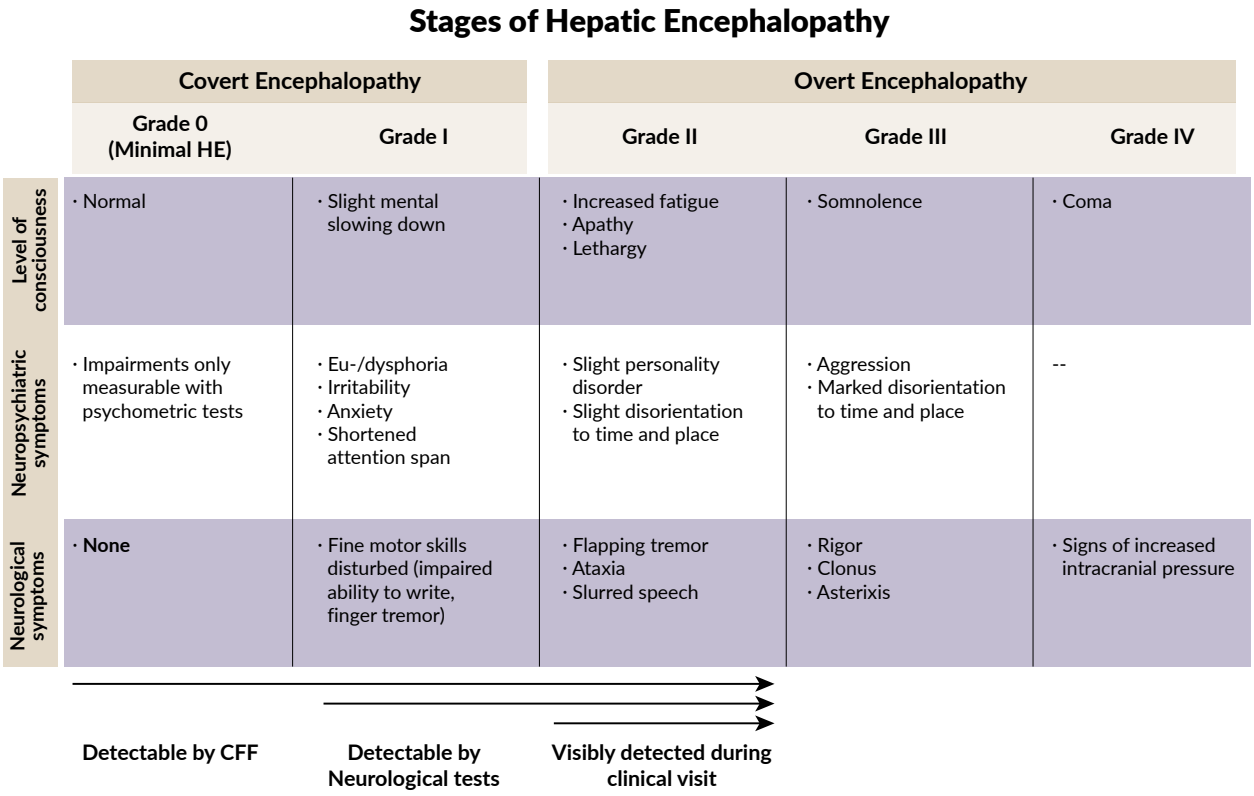
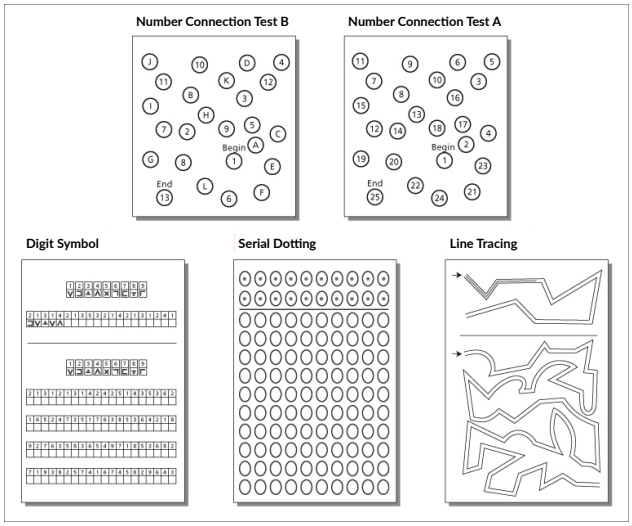


Figure 2.1: The stages of hepatic encephalopathy (HE). Current clinical screening practices only diagnose HE at Grade 1 or above. Our goal with *Beacon* is to identify Grade 0 HE (MHE) in individuals.

40-45 Hz is reduced to <39 Hz in people with MHE [3, 100, 147, 158, 159, 167].

CFF is a promising measure which can provide fine-grained understanding of the condition for cirrhotic patients with HE. I propose a reframing of CFF from a screening test to a self-tracked measure to help cirrhotic patients monitor the stability of their condition at home. By measuring their CFF at home patients will have a more up-to-date understanding of their chronic condition and may lead to timely interventions in case of a sudden worsening of the condition.

Psychometric Hepatic Encephalopathy Score



Test	Description
Number connection test A (NCT-A)	Randomly dispersed numbers are to be connected with each other in serial order as quickly as possible.
Number connection test A (NCT-B)	Randomly dispersed numbers and letters are to be connected in alternating series (1-A-2-B...) as quickly as possible.
Digit-symbol	Digits from 1 to 9 are assigned respective symbols. Under each digit the corresponding symbol is to be written within a given time.
Serial dotting	Draw a dot inside each circle as quickly as possible.
Line tracing	A given line is to be traced as quickly as possible.

Figure 2.2: Psychometric Hepatic Encephalopathy Score (PHES) consists of a battery of psychometric tests recommended for diagnosing minimal hepatic encephalopathy (example of the tests on the left, descriptions on the right). In practice, clinicians only use a subset of the tests for diagnostic purposes.

Chapter 3

A FRAMEWORK FOR SELF-EXPERIMENTATION IN PERSONALIZED HEALTH

Current medical research, evidence, and understanding is primarily based on population-level group designs, such as epidemiological surveys, longitudinal studies, and randomized controlled trials. This makes it difficult to answer questions at an individual level, including questions which on the surface may seem trivial (e.g., *how will this treatment affect me?*). Although self-tracking technologies provide a promising avenue for such personalized understanding, the current state of the field makes it difficult for even the most technologically adept people to get the desired value from such tools [31].

To address this need and to support better design of tools in this space, I developed a framework for self-experimentation. I view self-experimentation as a subset of self-tracking, but one that provides much-needed improvement in methodological rigor for certain types of questions. The framework provides recommendations on what kind of data to collect and what kind of analysis to conduct to provide the necessary findings. This chapter details components of this framework and presents a case study: a mobile application supporting self-experimentation for people with irritable bowel syndrome (IBS). I also discuss opportunities and challenges with this framework and identify areas of future work.

As part of my dissertation's demonstration of my thesis, in this chapter:

1. I examine how *population-level evidence* can be leveraged to design targeted tools to support individuals with specific chronic conditions;
2. I examine the need of *personalized understanding* among individuals with chronic conditions in the form of answering questions about the individualized nature of their health; and
3. I provide guidelines for designing tools that support the *collection and interpretation* of data collected through self-experimentation to answer questions individuals have about their health.

In addition to leading the team of multi-disciplinary researchers to develop the framework, I was involved in the following aspects of this research: conception and design of the framework; design

and development of new visual analysis tools; designing, conducting, and analyzing data from the focus group study; designing, deploying, and analyzing the results from the survey; designing the mobile application.

The research described in this chapter was done in collaboration with Jasmine Zia, Roger Vilardaga, Sonali R. Mishra, James Fogarty, Sean A. Munson, and Julie A. Kientz and was published in *JAMIA* 2016 [96]. This research has also resulted in several workshop publications, the most notable of which is "Opportunities and Challenges for Self-Experimentation in Self-Tracking", presented at *UbiComp* 2015 [92].

3.1 Need for a Self-Experimentation Approach

Individuals can currently access countless websites, applications, and sensing technologies intended to improve personal health (e.g., Fitbit, MyFitnessPal, Weight Watchers, RunKeeper). However, these technologies have neither realized widespread, sustained adoption nor their promised health benefits [53]. Although there is broad interest in using technology to track health information, people often lack scientific rigor in their analyses [31]. Analysis is also frequently done without consultation of healthcare providers. People seek answers to specific health questions, such as "does caffeine impact my sleep quality?". However, current tools generally support only data collection, not a systematic approach to answering such questions. For example, self-tracked data may suggest associations between sleep quality and caffeine, but then not support determining if caffeine is causing poor sleep quality or if a person is instead consuming more caffeine because they are tired. Such uncertainty may prevent people from making lifestyle changes that can lead to improved health outcomes (e.g., eliminating caffeine).

Our aim with the framework is to support everyday people in successfully conducting self-experiments to understand causes of their symptoms and possibly take effective action. Although self-experiments may be more complex than simple self-tracking, it is our hope that this framework can reduce the burdens and challenges of tracking through targeted data collection while also providing more rigorous and concrete answers to specific health questions.

Technology for self-experimentation fits into a larger process of personalized health (Figure 3.1a). Traditional self-tracking methods and correlational approaches, such as food journals or

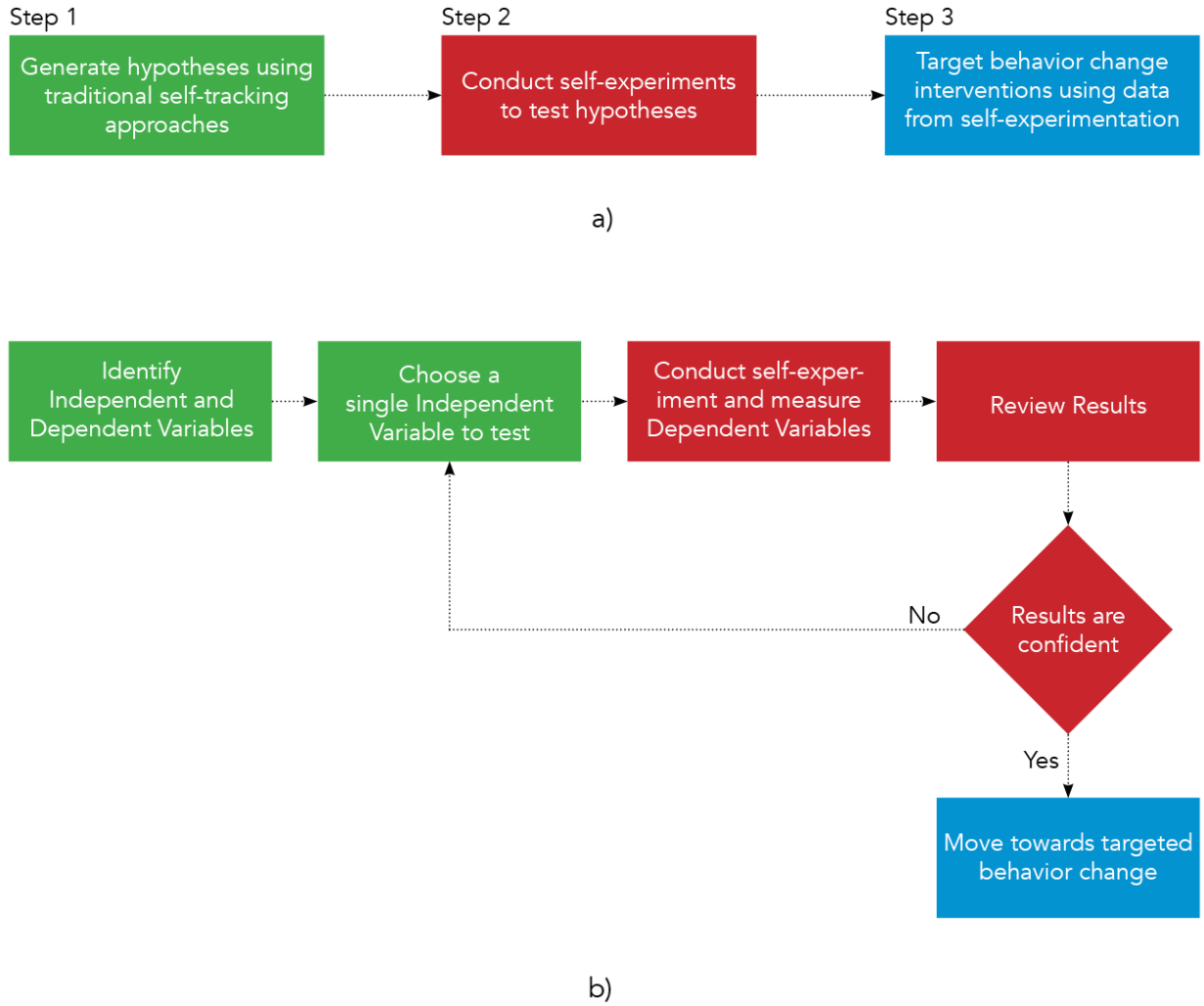


Figure 3.1: (a) Overall process of personalized health framework. (b) Expanded view of the self-experimentation process of our framework.

fitness trackers, can be used to generate hypotheses (Step 1). Self-experimentation technology then robustly tests those hypotheses (Step 2). A person can then use findings to target the most appropriate health behavior change to address their needs (Step 3).

This chapter focuses on Step 2 of this process (Figure 3.1b). Self-experiments begin with identifying hypotheses an individual wants to test (e.g., "does caffeine impact my sleep quality?"), then proceed with systematically testing hypotheses until results can support a person in making a decision about their health (e.g., "should I eliminate caffeine from my diet?"). This process

includes defining *independent variables* (e.g., causes, triggers) and the *dependent variables* they may affect (e.g., symptoms, health outcomes). A person then conducts a multi-day self-experiment where they are randomly assigned to either apply the independent variable (e.g., drinking at least 100mg of caffeine) or not (e.g., avoiding caffeine). Dependent variables are measured throughout the experiment (e.g., subjective sleep quality on a 5-point scale). Data is then analyzed and visualized using techniques suitable for single-case study designs, yielding a confidence value for the self-experiment (e.g., "there is strong confidence ($P<.05$) that drinking caffeine reduces sleep quality by half"). A person can then review results to determine if they are compelling enough to make health behavior changes.

3.2 Self-Experiment Study Designs and Analysis Methods

Although SCDs are traditionally used in the early stages of an intervention's development to determine its feasibility and estimate its potential effects, our framework proposes using SCDs at the later stages of the research cycle continuum, after randomized controlled trials are conducted and general guidelines for successful treatments are proposed (Figure 3.2) [49]. Using SCDs at the later stages can bridge the gap between average treatment effects for groups and specific treatment effects for individuals. SCDs can thus improve health interventions by providing more definitive, rigorous, and actionable guidance to individuals.

However, traditional SCDs have several limitations. First, according to some methodologists, their internal validity is questionable because decisions about the stability of a baseline (Phase A) before implementation of an intervention (Phase B) are not based on randomization [104]. Second, statistical inference with these methods has been challenging until most recently (e.g., see [117, 130, 141, 157] for most recent developments in quantitative analysis of SCDs), so visual analysis has been the primary method of evaluating data from single-case experiments [78, 89]. Third, SCDs are criticized for not providing population-based estimates for effect size of an intervention. Finally, SCDs in clinical settings will typically generate measurements during face-to-face contact with patients (e.g., weekly counseling sessions), producing a very limited number of observations.

These limitations can be overcome by applying randomization tests to SCDs, where a random

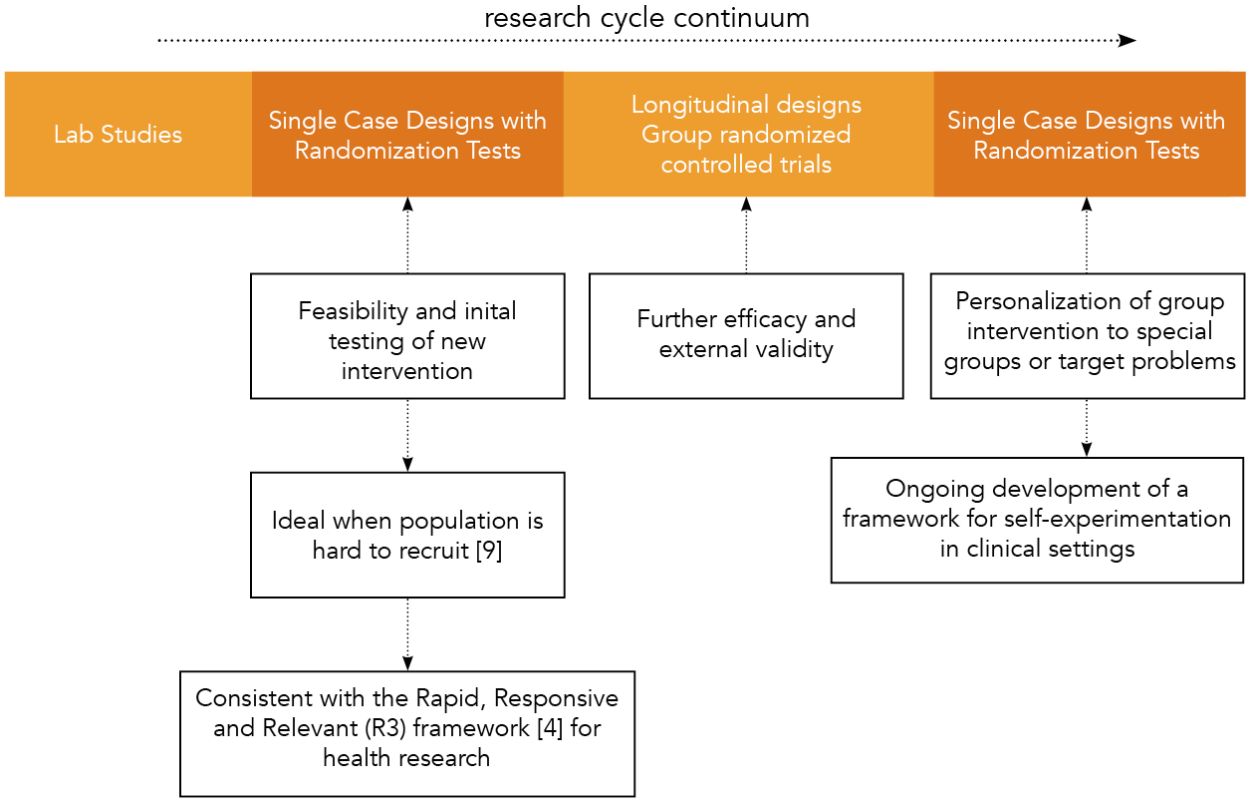


Figure 3.2: Single case designs' role within the research cycle continuum.

assignment is given to a population of occasions rather than a population of individuals [54, 82]. For example, in a traditional group design, each individual who belongs to a group of 100 people is assigned to treatment A or B. Instead, in a SCD with randomization tests, each measurement in a group of 100 measurement occasions is assigned to treatment A or B. After the data are obtained, a permutation procedure can be used to create all possible combinations of treatment exposures (A, B) and outcome measures and render a p-value indicating the probability of the null hypothesis (i.e., no differences between treatment A and B exposures). This procedure, originally envisioned by Fisher in the 1930s, ensures the same internal validity as group experiments and eliminates the statistical assumptions of parametric tests (i.e., normally distributed sample, homoscedasticity, independence of errors) [63]. The result is a statistical test that can be used in both individuals and groups of individuals. Its primary limitation remains that it does not provide external validity.

This limitation is not applicable in our case, as we use it to personalize known group-based or clinical guidelines to specific individuals. Lastly, advances in mobile technology now allow more frequent and ecologically valid measurements [169], reducing challenges regarding the need for face-to-face contact. Thus, SCDs with randomization tests overcome many limitations of traditional SCDs.

In the framework we propose, the choice of a specific SCD depends on: (1) the nature of dependent variables, (2) the lag effect of independent variables, (3) the statistical power of the design, and (4) effective human-centered design of the technology. For example, completely randomized SCDs, a form of alternation design [135], can be combined with randomization tests to provide the highest level of statistical power, but they require an independent variable with quick effects and the absence of carry over effects [137]. In the case study section, we describe the preparation of a study that required completely randomized SCDs, which is the best fit for IBS.

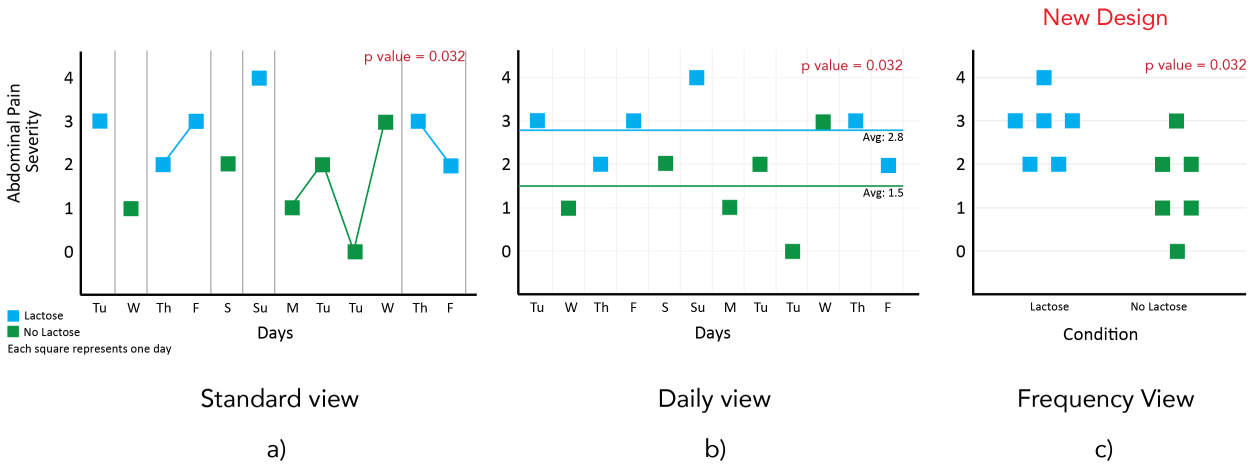


Figure 3.3: Example visualizations for a completely randomized single-case design of 12 observations showing a statistical significant effect of lactose on an individual's abdominal pain. (a) Standard view. (b) Proposed visualization with daily view. (c) Proposed visualization with frequency view.

SCDs are often visually analyzed to look for trends and infer relationships, but a completely randomized SCD can make it challenging to apply visual analysis in identifying trends. Conditions (e.g., lactose versus no lactose) can be distributed in any order, with no fixed phase lengths. It can

therefore be difficult to find patterns across varying phase lengths. To overcome this we designed a new visualization, inspired from a violin plot [83], that removes the temporal information and instead focuses on the distribution of the dependent variable (e.g., abdominal pain) across different conditions (Figure 3.3c).

3.3 Defining Eligibility of Health Conditions and Questions

We believe self-experimentation can be applied broadly across many health conditions, though some are better suited than others. Analyzing literature across different fields of health, we identify absolute and desired requirements of health conditions appropriate and ideal for this self-experimentation for personalized health framework (Table 3.1, 3.2).

Absolute Requirements	**Case Study: Irritable Bowel Syndrome (IBS)**
People must be uncertain about the effect of the independent variable on the dependent variable(s)	Most people with IBS are uncertain about their personalized food triggers [84]
Independent variables must be: A. Controllable and actionable B. Well-specified	People with IBS are able to consume or not consume certain foods, including their amounts, during specific times of the day
Dependent variables of condition being tested must: A. Recurrent episodes or flare-ups B. Follow the application of the independent variable within a defined period C. Be quantifiable and measurable	A. IBS symptoms (abdominal pain, bloating, diarrhea and constipation) vary in daily intensity and/or flare-up intermittently for set periods of time B. The time period between eating a trigger food and onset of symptoms typically occurs within a 3-hour window for most people with IBS [60] C. IBS symptoms can be quantified and measured using self-reported scales
Independent and dependent variables must not result in any serious health risks (immediate and/or long-term)	Exposing people with IBS to potential triggers will not result in any life-threatening consequences and/or immediate or long-term serious health risks [26]

Table 3.1: Table of absolute requirements for health conditions to which our self-experimentation framework can be applied.

We also identify classes of independent variables and assess their eligibility for our framework (Table 3.3).

Desired Requirements	Case Study: Irritable Bowel Syndrome (IBS)
Independent variables should be: A. Low-burden for people B. Reliably applied in the same way C. Easy to apply frequently	A. Altering people's diet is relatively low-burden B and C. It can be done frequently and with minimal daily variation.
Dependent variables should be: A. Tolerable for people to endure B. Easy to measure frequently C. Measured consistently over time	A and B. Symptoms can be measured frequently and consistently over time with well-defined scales. C. Although unpleasant, enduring provoked symptoms should generally be tolerable for most people with IBS.
Duration of the independent variable's effect on dependent variables should be defined	The exact duration of trigger foods on IBS symptoms is unknown at this time but assumed to be no more than 3 days (unpublished focus group)

Table 3.2: Table of desired requirements for health conditions to which our self-experimentation framework can be applied.

3.4 Case Study: IBS Self-Experimentation

IBS is an appropriate domain for exploring the potential of self-experimentation as an approach towards improving personalized understanding. The unique challenges in assisting people with IBS determine their individual triggers are highlighted in Section 2.3.1.

3.4.1 Designing a Mobile Application for Self-Experimentation in IBS

IBS patients and their providers need more efficient and effective methods to determine individualized food triggers for symptom reduction and improved quality of life. Following a human-centered design approach, we created a mobile application prototype that can support self-experimentation for people with IBS. Although we believe self-experimentation can be applied more broadly across many conditions, working with a specific population and concrete variables helps ground our design toward a real solution.

To design the application, we assembled a team of researchers in medicine, behavioral psychology, computer science, information science, and human-centered design to work with people who experience IBS through an iterative human-centered design process [118]. Through multiple rounds of iteration, we generated process flows, storyboards, and prototypes of a mobile application.

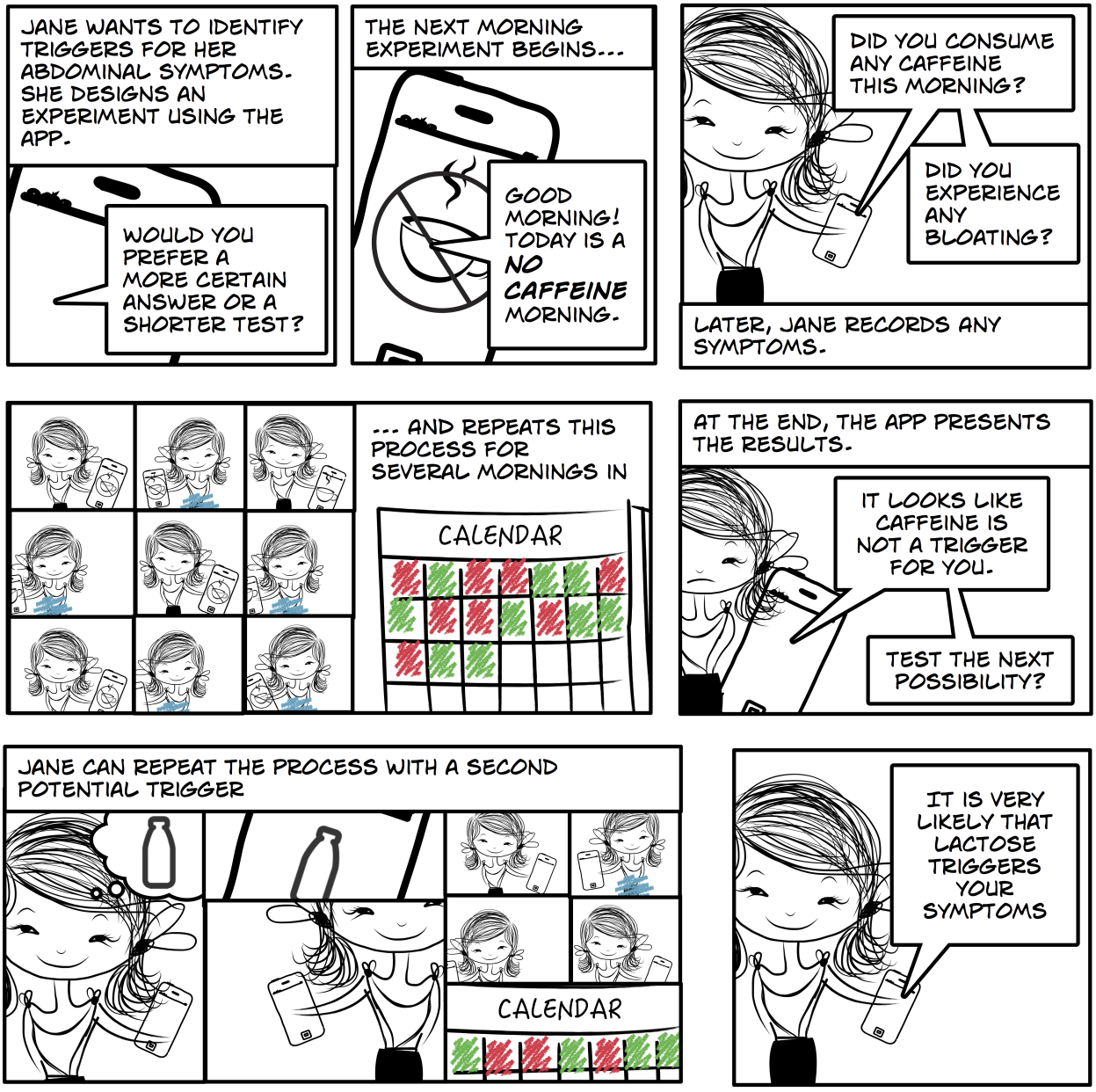


Figure 3.4: Storyboard depicting Jane's journey leveraging a self-experiment to determine the impact of caffeine on her bloating.

Our initial application focuses on the set-up and deployment of self-experiments (Step 2 in Figure 3.1a). In our process, individuals generate a list of hypotheses that narrow which trigger

foods to test (i.e., independent variables). The application uses relevant medical knowledge, the individual's personal experience, and/or the expertise of a medical provider to help generate testable hypotheses (storyboard in Figure 3.4). The individual then configures a self-experiment by setting their personalized symptoms (i.e., dependent variables). They also choose a start date, a study duration, and a time for daily reminders to enter symptoms.

When self-experimentation begins, the application shows the individual a schedule of which days to avoid and which days to consume the experimental trigger food, following a completely randomized SCD as previously described. The application shows sample meals for trigger and non-trigger food days (i.e., experimental and control). The individual is instructed to eat an otherwise consistent meal during each day, varying only the food component being tested. During each day of the experiment, the individual is also instructed to log peak symptom severity using a subjective scale at a defined time after eating the test meal (e.g., 4 hours). At the end of the experiment, findings are summarized and interpreted (Figure 3.5). The results include a P value on the likelihood the trigger food is causing IBS symptoms by chance. The individual may choose to re-run the experiment with a different possible trigger, or may choose to share the results with their medical provider for recommendations on how to avoid this trigger.

3.5 Preliminary Evaluation: Human-Centered Design Research and Findings

To learn about past experiences of identifying potential gastrointestinal food triggers and assess the feasibility of implementing our self-experimentation framework, we conducted focus groups and an online survey. We recruited participants from primary care and gastroenterology clinics associated with an academic center (University of Washington Medical Center, Seattle, WA, United States). Survey respondents were also recruited through social networks.

3.5.1 Focus Groups

We conducted three focus groups with people with IBS, totaling 6 participants (5 female, 1 male). To gather reactions to the self-experimentation process and how self-experimentation fits with participant priorities, focus groups included walkthroughs of the overall process using our flowchart (Figure 3.1b) and storyboard (see Figure 3.4) [52, 73, 168]. To elicit reactions to our interface

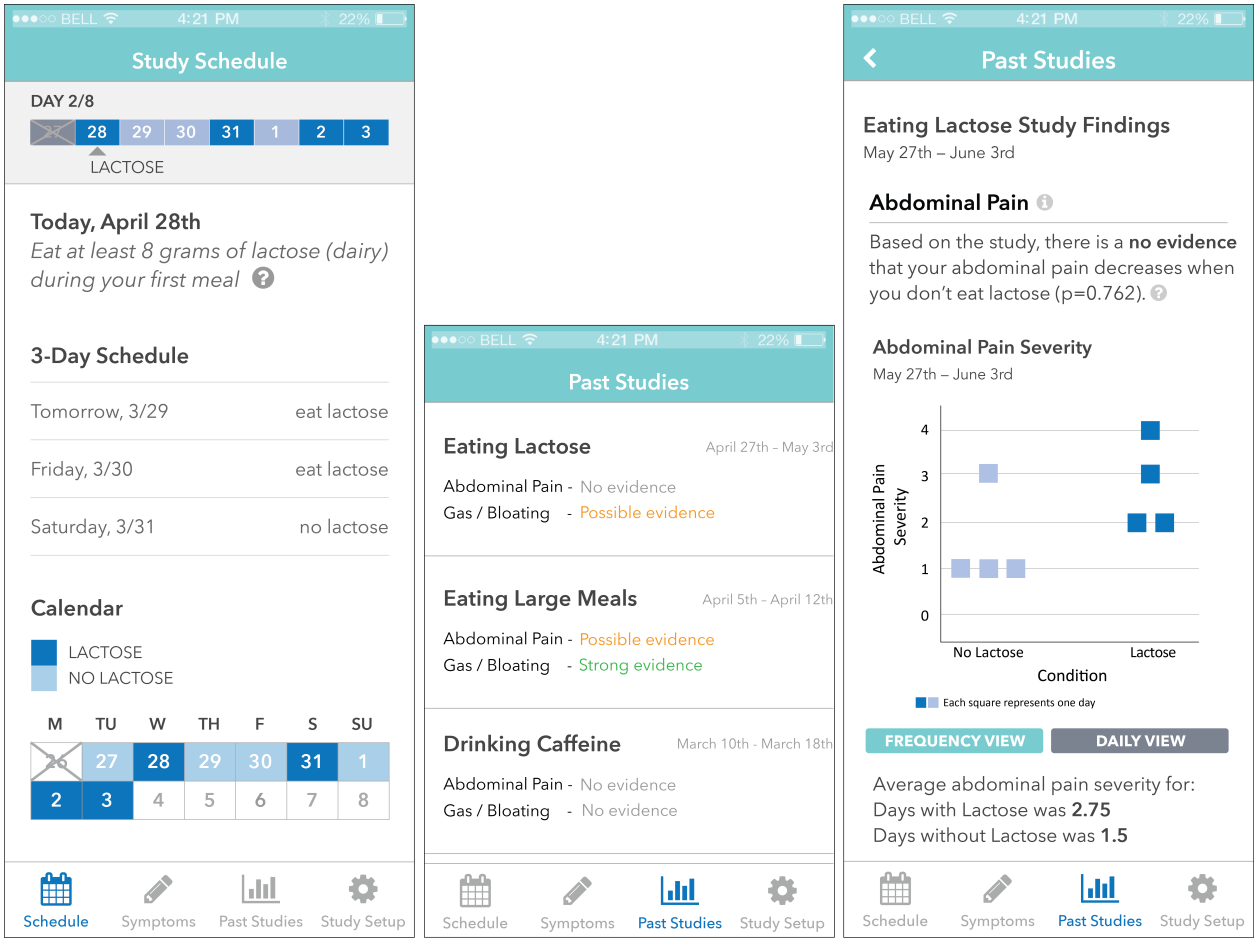


Figure 3.5: Screenshots of pages from our mobile app prototype. From left to right: study schedule, result summary of past studies, and results of a study.

designs, focus groups included a click-through of a prototype of our mobile application (Figure 3.6) [18, 77]. Questions focused on feasibility of the framework, understandability of the process, interface, and visualizations, and overall usability and usefulness of the application.

3.5.2 Survey

We complemented the focus groups with an online survey targeted toward people experiencing any type of gastrointestinal food intolerances, which commonly overlaps with the diagnosis of

IBS. Questions were internally tested for face validity and focused on specific topics raised during focus groups, such as how to rate symptoms and trade-offs in design (e.g., shorter studies vs more confident results) and on factual information about food intolerances (e.g., "What food intolerance symptoms do you experience? You can select more than one option"). We received 60 responses to the survey (53 female, 7 male), with 75% aged between 25 and 45 and 90% having a B.S. degree or higher.

Prior Experiences

Most focus group participants described feeling *"overwhelmed and frustrated"* with the trial-and-error process they used to help determine possible IBS triggers. They instead wanted more guidance during this process. Most had tried elimination diets as recommended by health providers, friends, family, and/or online research. Participants were concerned about labeling certain foods as triggers if their symptoms were more delayed and/or subtle, and many were also concerned about findings being confounded by nonfood factors (e.g., mood, stress, food preferences).

Overall Reactions to the Framework

During focus groups, participants expressed excitement by pointing out that this was a better approach to finding out their triggers than what they had been trying so far. In their view, part of their excitement was due to the fact that the proposed design was showing them their data and not just presenting a vague number. This was in reference to the design showing the visualization with their responses as data points instead of an abstract construct that many health applications present (e.g., fuel level in Nike Fuel Band). This in turn boosted their confidence in the framework. Participants stated they could trust the results because they were rigorous and based on their own data, rather than an average of other people's data. One participant mentioned liking that the application was *"honest about its limitations and shortcomings."*

Participants appreciated that our process gave them the decisional authority about whether or not to eliminate certain foods after weighing the magnitude of the effect, the confidence level, and their food preferences. Participants preferred this over a simple recommendation (e.g., "eliminate lactose"). The application acted like a "decision assistant" instead of a "decision maker." Most

also said the application would give them more “credibility” and empower them to talk with their medical providers because self-experiments were conducted in a scientifically rigorous manner.

Meal Choice

Participants requested an app with more detailed sample meals for both trigger and nontrigger days, as they feared erroneously consuming trigger foods on a nontrigger day. Focus group participants also questioned the assumption that trigger foods result in symptom onset within 4 hours. The majority of survey respondents (80%) did, however, agree that their symptoms occurred within 4 hours after eating. Participants were also concerned that trigger foods could result in symptoms lasting longer than 24 hours. They were also wary of foods being categorized either too broadly or narrowly (e.g., testing “nuts” vs “almonds”). These concerns suggest additional design opportunities such as experimenting with different symptom onset periods or iteratively categorizing foods.

Seventy percent of participants in our survey selected breakfast as their preferred test meal. Focus group participants shared part of their rationale behind this preference: they thought breakfast would be the most feasible mealtime to eat the same type of food and drink for the duration of the study, unless the trigger food was an unconventional breakfast food (e.g., alcohol). Further exploration should be conducted to determine how this influence might affect the framework’s efficacy.

Recording Symptoms

Focus group participants were concerned about inconsistent symptom ratings over time. The symptom scales allowed *“too much room for interpretation.”* They expressed the desire to customize symptom scales to be personally meaningful. We explored this issue further in the survey. We first asked the survey participants to generate their own custom scale with labels, then offered them a selection of four scales to choose from: 7-point with labels, 5-point with labels, 5-point with only endpoint labels, and 3-point with labels (Figure 3.6). Participants developed custom scales ranging from 4 to 10 points. Fifty-seven percent preferred one of our predefined scales over their custom scale whereas 25% preferred their custom scale. Of the 57%, 47% preferred scale 6b.

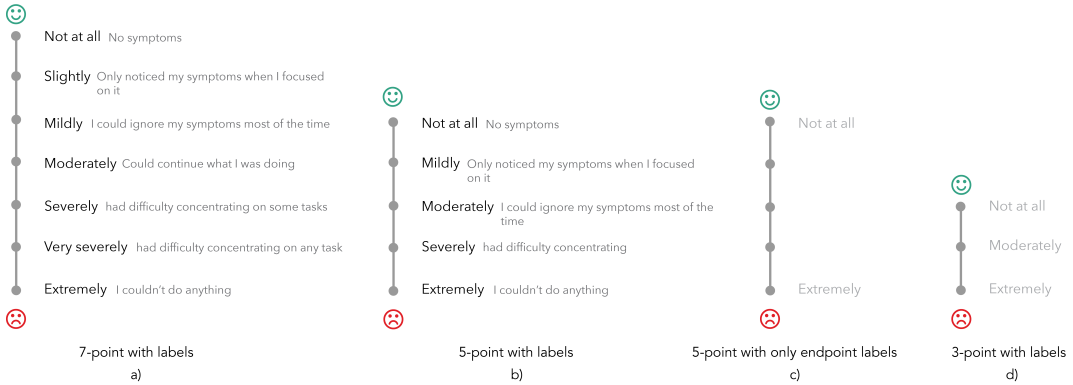


Figure 3.6: Selection of predefined scales shown during the online survey. From left to right: 7-point scale with labels, 5-point scale with labels, 5-point scale with only endpoint labels, and 3-point scale with labels.

Duration of Study and Meal Choice

Although focus group participants felt they could adhere to self-experiments, most preferred a shorter 8-day study, even when told that a longer 12-day study would likely result in higher confidence levels. In the survey, 83% said they would be "extremely likely" or "likely" to complete an 8-day study without giving up or missing days, compared to 67% for a 12-day study.

Understanding Results

All focus group participants preferred the frequency visualization over the standard and daily visualization designs (Figure 3.3). However, participants also wanted access to the standard or daily visualizations. Participants felt the calendar associated with these could help them recall details of individual measures and events leading up to them. For all visualizations, participants valued seeing individual data points, as each represents a specific measure to which they could relate. Because data corresponded to each individual, and not to a group of other IBS patients, participants felt more trust that summary statistics described their situation. Participants with stronger backgrounds in statistics preferred interpreting the graphs over the summary statistics because the graphs allowed them to better identify and understand the impact of outliers. All participants expressed a desire for an easy way to share and communicate results with their healthcare providers.

Independent Variable	Example Health-Related Question	Eligibility
Dietary	Does drinking coffee give me heartburn?	Eligible
Actionable Behavior	Does exercising in the morning after waking up give me more energy later in the day?	Eligible
Medications/Supplements		
As needed	Does this inhaler reduce my cough?	Eligible
Daily	Is my antidepressant improving my mood?	Ineligible
Environmental		
Controllable	Does elevating the head of my bed reduce my heartburn?	Eligible
Uncontrollable	Do I get headaches every time it rains?	Ineligible
Illness/pathogen	Is my mood worse when I'm coming down with the flu?	Ineligible
Combination of variables	Do I sleep better if I both exercise early in the day and stop working one hour before bedtime?	Eligible

Table 3.3: Classes of independent variables and their eligibility for the self-experimentation framework.

3.6 Discussion and Future Challenges

The results from our focus groups and survey both provide insight into the feasibility of our self-experimentation framework and the design of our mobile application. Despite this evidence of initial feasibility, our self-experimentation framework is still in its initial stages of development and must overcome a number of future hurdles. First, the feasibility and clinical efficacy of applications using our framework must still be evaluated across a larger sample of individuals with IBS and then other health conditions. I provide initial evidence of efficacy in the next chapter. Second, self-experimentation is more difficult to understand than simple journaling. As opposed to traditional journaling, where an individual is essentially noting down everything they are eating, in a self-experiment they are asked to consume and avoid certain foods on certain days. In our focus groups, it was certainly more complicated to explain the self-experimentation process and rationale to participants than it would have been to simply tell them to record all of the food that they eat. However, after participants understood the process, they believed our framework would be much easier to adhere to than traditional methods, especially with the guidance of the

mobile application. Third, a lack of control over confounds can bias and diminish the effects of self-experimentation. Unlike laboratory or more controlled clinical studies, our framework has an emphasis on "real-world" deployment that may diminish its scientific rigor.

Despite these hurdles, self-experimentation has the potential to improve existing tracking methods across multiple health conditions and even non-health domains. Our self-experimentation framework can be applied to any condition or situation that meets the absolute requirements (Table 3.1). We can extend our framework to other health conditions with possible dietary triggers such as gastroesophageal reflux disease, migraines, and gout. Our self-experimentation framework can also be applied in mental health and addiction, where the specific skills or behavioral strategies that help individuals meet their goals varies by person. Insomnia is another possible application of our framework, wherein individuals identify behavioral or environmental variables impacting their sleep. Our framework can even be applied to nonhealth conditions such as increasing work productivity, testing whether a new habit saves energy, or evaluating if a skincare product acts as advertised.

Much work remains to be done in expanding to new domains. The degree of customization needed for each new health condition can be significant and may not be a straightforward process. Additional challenges can arise in adapting to conditions where the desired requirements are not met (e.g., conditions where the duration of the independent variable's effect is not defined). We ultimately hope to support people and their healthcare providers with customizable experiments, but first need to explore how to support people in planning valid self-experiments and understanding trade-offs between study design choices. Our initial work on defining absolute and desired requirements (Tables 3.1 and 3.2) and eligible independent variable types (Table 3.3) has helped establish the trajectory for this expansion, but more work is needed on how to translate these issues to everyday people.

3.7 Summary

In this chapter, I presented the background and conceptual foundation of a new framework for self-experimentation with person-generated health and wellness data. I described a design framework, its advantages over other proposed methods, the conditions in which this framework is

applicable, a case study applying this framework to a specific health condition, and initial reactions among people with IBS and gastrointestinal food intolerances. The framework proposed here combines advanced statistical methods and SCDs to assist individuals in their own process of self-experimentation. The framework presents a way of leveraging the existing population-level evidence in medicine to design tools which collect personal health data to provide personal understanding to the individual.

This framework is among the pioneering research in HCI which proposes the use of n-of-1 or self-experimentation as an approach to designing personal health technologies. It has not only contributed to my long-term research agenda of leveraging population-level understanding to provide personalized value, but has directly influenced my research on the TummyTrials platform which I present in the next chapter. In addition to influencing TummyTrials, there are a number of notable research projects within HCI which were influenced by the framework. My colleague Alexandra Okeson is improving the statistical modeling that we proposed here in her research on actionable Bayesian analysis for evolving health goals [1]. Sleepcoacher [50] is a platform which adapted an n-of-1 approach to improve sleep quality through personalized understanding of the impact of different interventions on one's sleep quality. Researchers have also explored adopting the n-of-1 approach to facilitate better behavior change interventions [112]. In my discussion in Chapter 7, I revisit how my understanding and research has in-turn been informed and shaped by such work.

Chapter 4

TUMMYTRIALS: SUPPORTING SELF-EXPERIMENTATION TOWARD INDIVIDUALIZED TRIGGER IDENTIFICATION AMONG IRRITABLE BOWEL SYNDROME PATIENTS

IBS is complex chronic condition where symptoms range across individuals and their underlying triggers also vary (i.e., as discussed in section 2.3.1). At a population-level, we have medical evidence that suggests food as the most common trigger for IBS symptoms. However, traditional lab tests are often inconclusive, and so people frequently resort to trial-and-error to figure out the cause of their symptoms. This is an error-prone process with a likelihood that a person might reach an erroneous conclusion based on spurious correlations. I developed TummyTrials to support people with IBS to determine their individual food-based trigger. TummyTrials supports low-burden collection of relevant symptom and trigger data and presents analysis for personalized interpretation of the findings. TummyTrials is based on the framework of self-experimentation I presented in the previous chapter, leveraging current medical understanding and an n-of-1 study design to support people in running scientifically valid experiments in their daily life.

In this chapter I highlight the design process involved in customizing the framework for the context of IBS, present findings from the feasibility study I ran to evaluate the efficacy of people to reliably undergo a self-experiment, and discuss opportunities and challenges in designing future tools to support such personalized discovery in chronic conditions.

As part of my dissertation's demonstration of my thesis, in this chapter:

1. I leverage existing *population-level evidence* that suggests food as the most common trigger for IBS symptoms and that there is no clinical test which can determine the specific trigger for a given individual;
2. I examine IBS patients' desire for *personalized understanding* of their condition by conducting self-experiments to determine which foods are triggering their symptoms and how the lack of appropriate tools creates barriers for this endeavour; and
3. I demonstrate how TummyTrials supports the patient's desire by scaffolding in-the-wild

collection and interpretation of relevant data through designing an end-to-end guided self-experiment that is low-burden and addresses common confounds.

In addition to leading the team of multi-disciplinary researchers involved in this project, my contributions include: designing early mockups and low-fidelity prototypes of the platform; designing and developing the TummyTrials platform; and designing, deploying, and analyzing the results from the feasibility study.

This research was done in collaboration with Jessica Schroeder, Daniel A. Epstein, Laura R. Pina, Jeffrey Scofield, James Fogarty, Julie A. Kientz, Sean A. Munson, Roger Vilardaga, and Jasmine Zia. It was presented at CHI 2017 and was awarded a Best Paper Honorable Mention [95].

4.1 Formative Design Research

TummyTrials is based on our framework for self-experimentation in personal health, as well as a formative and iterative design process with input from existing medical literature, domain experts, and people with IBS [96]. The goal of TummyTrials is to provide an effective and low-burden approach for people experiencing IBS to systematically test potential food-based symptom triggers to inform decisions on whether they might reduce those triggers in their everyday diet. TummyTrials is designed to be used when a person has one or more hypotheses regarding personal food-based triggers. Hypotheses may rely on intuition or experience. They may be formed in consultation with a medical provider considering triggers that are common in the broader population, or through analysis of a food and symptom journal. Regardless of how a hypothesis is formed, TummyTrials aims to guide a person through a self-experiment testing that hypothesis.

TummyTrials builds on top of the existing research presented in Chapter 3. The mockups created in prior research along with the findings from the survey and the focus group with IBS patients informed the design of the platform we present below. Figure 4.1 provides an overview of the overall design process which spanned a full year.

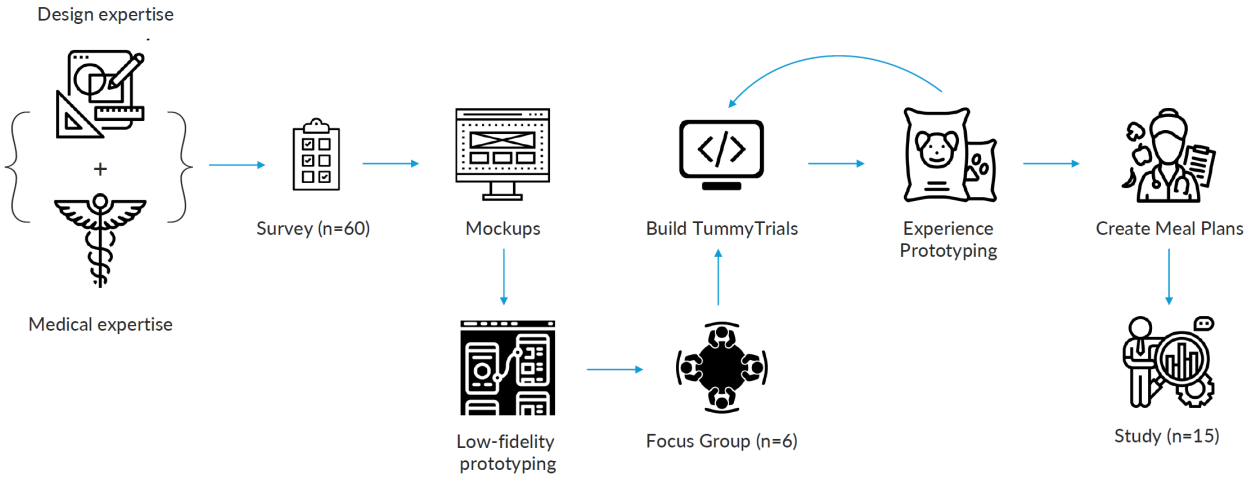


Figure 4.1: Overall design process of the TummyTrials platform.

4.2 Designing Self-Experiments in IBS

4.2.1 Self-Experiment Setup

TummyTrials uses a wizard design to guide a person through setting up their self-experiment. To configure a self-experiment, a person must select: (1) one or more symptoms the person is experiencing and wants to track (Figure 4.2A), (2) the trigger food to test, (3) the start date and trial duration, (4) times of day to receive TummyTrials reminders, and (5) food and drink preferences for breakfast in each experimental condition.

TummyTrials currently supports seven symptoms as dependent variables (abdominal pain, bloating or gas, hard passage of stool, loose passage of stool, infrequent bowel movements, frequent bowel movements, bowel urgency) and four trigger foods as independent variables (caffeine, gluten, sorbitol, lactose). Prior focus group we conducted with IBS patients suggest these are the most common symptoms and trigger foods for patients with IBS (Section 3.5).

Patients have reported that the onset of symptoms generally occurs within a short duration after consuming a trigger food, typically under three hours [96]. Due to the extended fasting period that occurs during sleep, we chose breakfast for the experimental manipulation, with a person then not consuming other food during the time that symptoms might be expected to manifest. This

combination of fasting before and after consuming the potential trigger food is thus intended to remove potential confounds that could otherwise be introduced by other meals. It also reduces the burden of experimentation by limiting it to the morning (i.e., consuming breakfast, fasting for the potential onset period, and reporting symptoms).

A person's daily self-experimentation therefore consists of: (1) eating breakfast in accordance with the day's assigned experimental condition (i.e., avoiding or consuming the experimental trigger food), (2) fasting for three hours (with drinking water permitted), and (3) monitoring their symptom during the fasting period. After the three hours have passed, TummyTrials prompts the person to report their peak symptoms during the fasting period. After reporting symptoms, a person can continue eating and drinking as normal, whatever foods they please.

A person is asked to eat a consistent breakfast, changing only per the manipulation. We worked with a dietitian to develop a sample menu for each potential trigger food, including menus both for days when the person should consume the experimental trigger and for days when they should avoid the experimental trigger. Our initial choice of independent variables was therefore limited, but the menus were intended to help patients keep other aspects of their diet consistent to avoid confounding their experiment. Sample menus were provided for a variety of food preferences (i.e., bagel / bread / English muffin / toast, cereal, muffin, waffle / pancake, yogurt) and drink preferences (i.e., coffee / espresso, energy drink, juice, milk, soda, specialty drink, tea, water). For example, if a person conducting an experiment with lactose as a potential trigger chooses cereal and milk as their menu preference, TummyTrials will suggest consuming 6 oz. of cow's milk with cereal on experimental days versus consuming 6 oz. of lactose-free milk with cereal on control days.

The natural extended fasting period that occurs during sleep allows us to consider the gastric system as reset daily. TummyTrials treats each day as an independent sample, and experiments use a completely randomized alternating treatment design [55, 82]. This design allows a shorter duration study; there are no minimum phase length requirements as in a more traditional AB single-case design. For A days defined as those where a person consumes their trigger, and B days defined as those when they avoid it, a TummyTrials experiment over n days includes n / 2 A days and n / 2 B days that are randomly distributed. For example, a 12-day study will include 6 random days a person consumes their trigger food at breakfast and 6 random days when they avoid it.

A person chooses the start date for their experiment based on what fits best in their lifestyle and their plans. Menstruation can potentially trigger IBS symptoms [90], so we encouraged patients to wait until their current cycle completed before beginning an experiment. A person can choose the number of days in a trial, required to be an even number of at least 6 days. People are instructed that longer studies provide more certain results, and we set the default to 12 days as a balance between study duration and experimental power. Informed by prior work showing that timely reminders and notifications improve compliance [17], TummyTrials allows a person to configure four reminder times: (1) an initial reminder of the day's experimental condition (i.e., whether to avoid or consume the trigger food), (2) reporting breakfast compliance, (3) reporting fasting compliance and symptom severity, and (4) an evening reminder that is delivered only if the person has not yet reported their symptom severity.

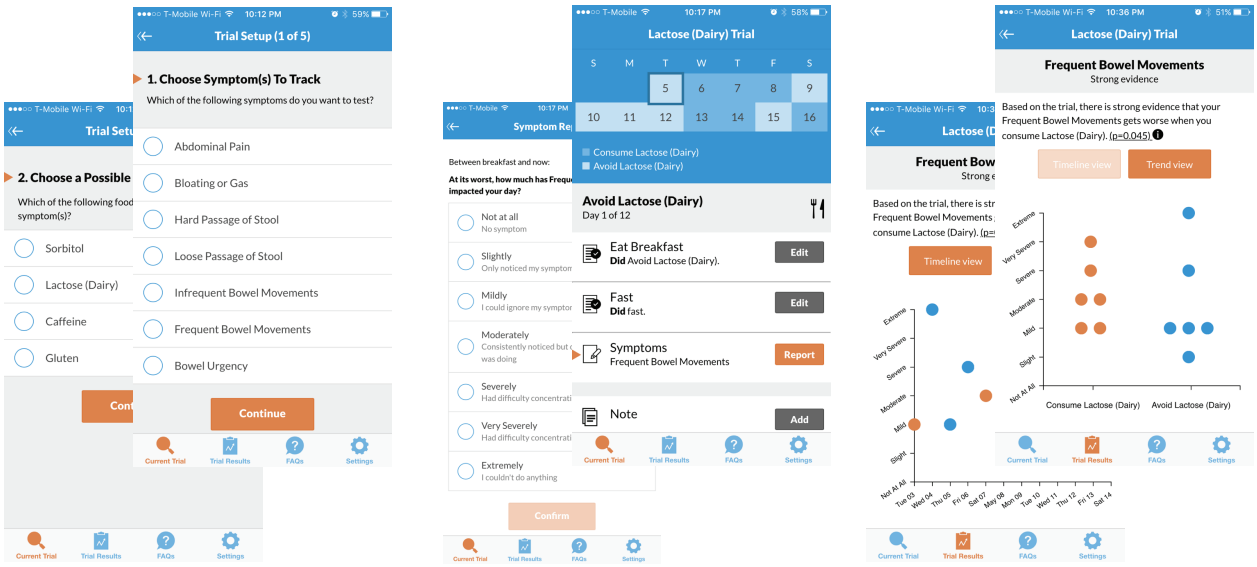


Figure 4.2: TummyTrials supports scientifically valid self-experimentation for identifying individualized food triggers. In this case, the self-experiment is aimed at answering whether lactose (dairy) is causing the person to experience more frequent bowel movements. The screen-shots **(left)** highlight the self-experiment set up process, **(middle)** show the data collection during the self-experiment, and **(right)** show the result of the self-experiment.

4.2.2 Self-Experiment Execution and Data Collection

We designed TummyTrials to be low burden relative to current standards of care: elimination diets and food / symptom journaling. We sought to minimize what patients must record to receive results. During a self-experiment, a person only reports: (1) breakfast compliance (whether they avoided or consumed the trigger as instructed, a Yes/No question), (2) fasting compliance (whether they avoided eating or drinking for three hours following breakfast, a Yes/No question), and (3) peak symptom severity (at its worst, how much impact each symptom had on daily activities, a 7-point scale; Figure 4.2C). TummyTrials provides an optional notes section to record any additional information a person wants to add (Figure 4.2B).

If a person chooses to not begin a self-experiment immediately (e.g., delaying due to menstruation), TummyTrials sends a reminder two days and one day prior to the experimental start date (e.g., so a person can plan to buy any needed groceries). After the self-experiment begins, the person sees a screen with a calendar for the experiment at the top (Figure 4.2B), giving an overview of the entire self-experiment and highlighting the "avoid" (control) or "consume" (experimental) condition for each morning. The person can review reports for prior days and the food plan for the current morning. A daily checklist shows which of the day's compliance and symptom reports have been completed and which still need to be completed. This process is repeated for the duration of the self-trial (e.g., 12 days). A person can abandon the scheduled self-experiment, either to end self-experimentation early or start over. TummyTrials also provides a FAQ with expert answers to questions about IBS and about TummyTrials functionality.

4.2.3 Self-Experiment Results Review

Upon completing a self-experiment, TummyTrials generates a results page for each symptom a person tracked (Figure 4.2D). Visual analysis is the traditional approach to analyze single-case designs [24]. However, the use of randomization to address confounds renders a standard timeline visualization used in visual analysis ineffective due to irregular phase lengths and misleading implications of the area under a trend line [96]. We therefore do not plot trend lines, instead illustrating the data in a timeline plot and a trend plot (Figure 2), as proposed in [96]. The trend plot provides an overview of symptom severity in the manipulation and control conditions, allowing

for a quick and easy visual analysis. A timeline view can be toggled by clicking a button, which animates the dots into chronological order. This transition is intended to reinforce that the same data is in both views.

We also determine the confidence of the experimental result by calculating a p value using randomization tests with the R SCRT package [156]. TummyTrials provides a one-sentence summary based on this analysis: "Based on the self-trial there is (strong/possible/weak/no) evidence that your (symptom) gets worse when you consume (trigger)." The strength of the evidence is bucketed with the following cut-offs: $p<.05$, $.05<p<.10$, $.10<p<.20$, and $p>0.20$. Figure 4.2D is an example of strong evidence ($p=0.045$) while Figure 4.3 is an example of no evidence ($p=0.65$). Although p values over .05 are rarely considered evidence in scientific literature, we relaxed the thresholds traditionally used in population-based research. This lower threshold supported a wider range of feedback more consistent with the purpose of self-experimentation.

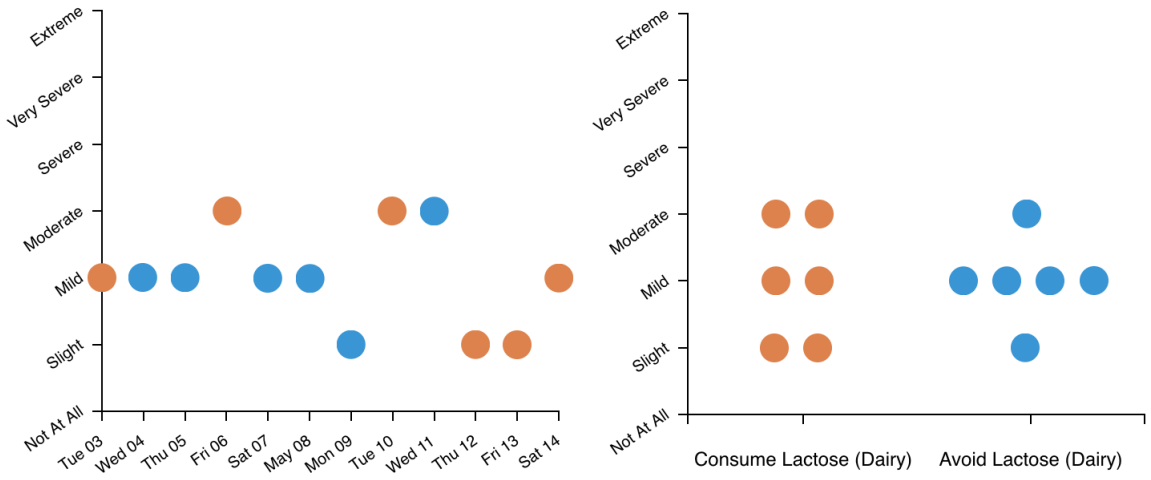


Figure 4.3: TummyTrials visualizes self-experimentation both as a timeline (left) and by trend in experimental condition (right).

4.3 Feasibility Study Design

We conducted a feasibility study to assess the practicality, usability, and user burden of TummyTrials while gathering participant feedback in a primarily qualitative study. This is a best practice for

evaluating early-stage health technologies [101]. Our recruitment methods and study protocol were reviewed and approved by our university Institutional Review Board.

Study participants received guidance from the researchers as to what hypotheses they might test and how to interpret the results of the self-experiment. This guidance is consistent with current practices in patient-provider consultation (e.g., in the context of an elimination diet or a food and symptom journal), where a provider may give instructions, ask a patient to keep a record, and collaboratively review the record. Our goal was to determine whether TummyTrials can successfully support people in completing a self-experiment and discover any challenges people encounter throughout it.

Patients were asked to avoid testing known (diagnosed or strongly suspected) triggers. The purpose of TummyTrials is to support a person in testing a hypothesis where there is uncertainty. Testing a known trigger would therefore have undermined validity and risked unnecessary flares in patient symptoms. Participants were encouraged to test a trigger from the list which they felt might be a trigger for them, but which they were not certain about.

For the purpose of the study, duration was fixed at 12 days. Participants chose their own start date (e.g., to schedule around menstruation or to avoid vacations)

4.3.1 Procedure

The study was divided into three parts: (1) a screening and intake interview, (2) completing the 12-day self-experiment, and (3) an exit interview. Interviews were conducted at a local hospital. Compensation was pro-rated, to a maximum of $175, based on participation in study milestones (intake questionnaire and interview, study participation, exit questionnaire and interview). To receive full compensation, participants were required to use TummyTrials for two days and to share their data for analysis, whether or not they had otherwise complied. Compensation was therefore not linked to TummyTrials experimental compliance.

Prior to the intake interview, we asked participants to complete the IBS symptom severity scale (IBS SSS) [65], which is commonly used in clinical trials. During the intake interview, we asked participants about any prior attempts to determine their triggers and gave an overview of the self-experimentation process. We installed TummyTrials and asked them to configure their

first self-experiment. Participants answered several questions regarding their expectations of the research study and the self-experiment. After using the app to complete a 12-day self-trial, participants completed the IBS SSS again, the System Usability Survey (SUS) [23], the User Burden Scale (UBS) [164], and a questionnaire we developed specifically for the study. We then conducted a semi-structured exit interview on participant experiences. Interviews were recorded and transcribed by a professional service. An audio recording error lost the second half of P2's interview, and P11 did not consent to be recorded. In these cases, the interviewer took detailed notes and recreated a transcript as best as possible immediately after the interview.

Quantitative analysis consisted of calculating participant compliance (days they reported breakfast compliance, fasting compliance, and symptom), analyzing usability and user burden using the SUS and UBS scales, and measuring the change in IBS symptoms during the study. For qualitative data, we took a bottom-up approach where the entire research team divided the interview transcripts, read them, and extracted notes containing portions relevant to the research questions. Two members of the research team created an affinity diagram with these notes [18], iterating on themes with the rest of the research team through discussion.

4.3.2 Recruitment

We recruited participants by emailing 1100 randomly selected patients with food intolerances resulting in gastrointestinal symptoms from a list of a patients in a local medical system acquired under a HIPAA waiver. Of 190 patients who replied, we filtered to 41 eligible participants based on those who owned an iPhone, were between 18 and 70 years of age, and met the Rome IV IBS criteria [139], a validated screener for IBS. We excluded participants with medical conditions that might impact IBS. Of 41 eligible patients, 18 enrolled for the study. We report on data from 15 participants (Table 4.1), as 3 scheduled or deferred their experiments outside of the study window. 5 participants reported being Asian and 10 reported being White. 4 participants reported having a bachelors, 7 masters, 2 doctorates, 1 trade school, and 1 associates. Our recruitment approach may have oversampled people who are more receptive to technology and from higher socio-economic groups. A majority of participants were women, but IBS patients are more likely to be women [26].

ID	Age	Gender	Stats Experience	Trigger	Symptoms
1	20s	Female	College courses	Sorbitol	3, 4, 6, 7
2	30s	Female	College courses	Lactose	2, 3, 5
3	30s	Female	College courses	Sorbitol	1, 2, 4, 7
4	20s	Female	College courses	Caffeine	4
5	50s	Female	College courses	Lactose	2, 3
6	20s	Male	Professionally	Caffeine	1, 2
7	30s	Female	Professionally	Lactose	1, 2, 3, 4, 5, 6, 7
8	50s	Female	College courses	Caffeine	2, 4, 6
9	50s	Female	None	Caffeine	1, 2, 3, 5, 6*
10	50s	Female	College courses	Lactose	4, 6, 7
11	30s	Male	College courses	Lactose	1, 2
12	60s	Male	Professionally	Lactose	1, 7
13	40s	Female	Professionally	Sorbitol	1, 2
14	30s	Male	High School course	Lactose	1, 2, 3, 4, 6
15	40s	Male	College course	Lactose	1*, 4, 6, 7

Table 4.1: Participant Summary. Symptoms Tracked are (1) Abdominal Pain, (2) Bloating or Gas, (3) Hard Passage of Stool, (4) Loose Passage of Stool, (5) Infrequent Bowel Movement, (6) Frequent Bowel Movement, (7) Bowel Urgency, (*) Substituted.

4.4 Feasibility Study Results

We conducted semi-structured interviews to provide flexibility in probing points raised by participants, and therefore not every participant was asked every question. We provide counts where the question was answered by all participants.

4.4.1 Overall Experience and Compliance

Participants had positive experiences with self-experimentation and TummyTrials. Compared to their prior attempts to identify triggers, participants appreciated the structure and support: *"I would say that, it provided the structure, it provided the discipline and it provided the reminders" (P10)*.

Participants and generally tested foods they doubted were triggers but wanted to verify. Consistent with their expectations, most did not find evidence that the tested food was a trigger. As we will discuss, our experiment and analysis were designed for one-sided analysis (i.e., to detect if something is a trigger rather than to rule it out). However, many participants interpreted "no evidence" of a food worsening their symptoms as proof that the food was not a trigger (e.g., *P1,*

"I'm glad they didn't show any evidence because it means I can eat more things"). Although most participants were unsurprised by their self-experiment results, they still saw value in the process. P12 said *"when I ended [the trial] on Saturday, I said to my wife, 'This was an exercise worth really doing.' I said, 'For my own edification because I suffer from this.'"*

Usability and User Burden

13 participants reported using the app was less burdensome than their prior attempts to identify triggers, such as food diaries and elimination diets: *"It definitely took a lot of that strain away of trying to remember all of this stuff that you're supposed to be paying attention too, because it's all in the app" (P2)*. The usability and user burden ratings supported these results. On the System Usability Survey (SUS), participants reported a mean of 83, median of 87.5, and standard deviation of 9.3, well above the suggested threshold of 68 [150].

Results from the User Burden Scale (UBS) indicate most participants did not find TummyTrials burdensome, though some improvements could be made to further reduce user burden from the perspective of several participants. The mean (x), median (M), and standard deviation (s) within each subscale of the scale were as follows: difficulty of use (x=0.73, M=0, s=1.1, Grade=C), physical (x=0.2, M=0, s=0.56, Grade=C), time and social (x=0.5, M=0, s=1.35, Grade=B), mental and emotional (x=0.47, M=0, s=0.9, Grade=B), privacy (x=0.8, M=0, s=1.42, Grade=C). Although the official grades of B and C place us within the 15%-45% and 45%-85% of apps evaluated in the UBS validation process [69], the fact that every scale had a median of 0 and a high standard deviation corresponds to most participants not reporting any burden. For those participants that reported a higher user burden, we later describe their qualitative feedback on that burden.

Effect on Symptom Severity

Participation in TummyTrials neither aggravated nor alleviated participant IBS symptom levels. The mean change in pre- and post-IBS SSS scores was 2.7 (median: 18, standard deviation: 71), a difference that is neither statistically nor clinically significant across participants. A difference of greater than 50 points is considered clinically significant according to the IBS-SSS scale [65]. 3

participants reported a significant improvement in their scores (P6: 55, P8: 74, and P2: 79) while one reported a negative change (P12: -223). P12 however, felt *"the study had nothing to do with it."* He reported he had experienced a particularly good ten days before the study, but had already felt symptoms returning at the time he began the study.

Compliance

TummyTrials asked participants to self-report whether they followed the experimental condition for the day, whether they fasted afterwards, and their symptom severity. Of 15 participants, 12 reported 100% compliance for the 12-day period. Participants reported that their adherence was improved by both the self-experimentation process (i.e., with its fixed duration and clear rules), and by support from the TummyTrials app (i.e., with its reminders and reporting features):

> *P2: This held me accountable and it required me to keep track of it which is always a challenge ... With Trial and Error there's nothing holding me accountable, so I appreciated that.*

Reasons for non-compliance varied. P2 had one day where she reported breakfast compliance but did not return to report fasting or symptoms. P3 did not comply for four days, three of which she attributed to full-day kickball practice. However, she did report her breakfast compliance on all four days (i.e. did not provide symptom reports). P11 did not report symptoms for five days, including two when his phone was in a repair shop and two during a weekend trip. He reported breakfast compliance only one of the five days.

Log data (e.g., page visits, session length, session count) was collected and analyzed, but it contributed no exceptional or informative patterns. We note our design is intended to minimize a need for engagement, and we believe compliance data and qualitative results better represent usage.

4.4.2 Self-Experiment Set Up

TummyTrials supported most gastrointestinal symptoms that participants wanted to track. However, some wanted to track non-gastrointestinal symptoms. Two participants therefore substituted

an existing symptom to track something currently not supported (e.g., P15 used the entry for Abdominal Pain as a placeholder to instead report migraines). Initial development of TummyTrials for this feasibility research prioritized four possible triggers. Participants wanted a larger selection of triggers (e.g., raw vegetables, fried foods, spicy foods, alcohol, fructose). The breadth of requested triggers aligns with IBS literature, which suggests a wide variety of triggers [129, 162]. Some participants were unsure which triggers to test and wanted to work with their provider to decide:

> *P5: I mean helping to choose by talking to the dietitian, identifying possible triggers, and, then, saying this could be the trigger. Let's use that with the app.*

A minority of participants preferred not to test certain foods they particularly enjoyed or relied upon, as they did not want to discover such foods as a trigger. *P7 said, "Sometimes I don't want to try things that I don't want to lose in my diet."*

14 participants were happy to limit the self-experiment to breakfast. However, some expressed a desire to test food triggers in other meals (i.e., lunch, dinner). Some participants preferred not to disrupt their morning routine, suggesting that dinner might be a more desirable option. Others wanted to test a particular food that is typically unsuitable for breakfast (e.g., beer, wine). A few participants felt avoiding or eating the trigger food for breakfast was not rigorous enough. They would be more confident in the results if the experimental condition were applied in meals for the entire day.

> *P2: I think I would have needed to avoid it longer not just for breakfast. . . . Like I said, the window for fasting and avoiding the food should be longer even as uncomfortable as that may be.*

TummyTrials provided guidance and sample menus for avoiding or consuming each supported trigger. Participants appreciated the concrete guidance, reporting this reduced the burden compared to their previous attempts to identify triggers. In cases where none of the available food options were agreeable to the participant, the gastroenterologist on the team worked with the participant customize the food menu.

The final step of scheduling a self-experiment was deciding the length of the experiment and when to start it. Although 13 were happy with the 12-day duration, we received mixed feedback from some participants who preferred either a shorter experiment or a longer one. To avoid confounds due to disruptions in their routine, we requested that participants not undergo the self-experiment while their schedule was in flux (e.g., travel, a deadline). Participants echoed this recommendation when asked if they would schedule another self-experiment. They gave examples of times they would not want to complete a self-experiment, such as when needing a break from the experimental regimen, due to an upcoming vacation, or for work-related concerns.

> *P10: The idea of having to eat the exact same thing for breakfast every day for 12 days is a challenge. It was doable, but it kind of made it so I thought I'd have to really be strategic about picking a time to do this again if I wanted to test another group.*

4.4.3 Conducting the Self-Experiment

Daily Reports

All participants were satisfied with the provided reminders. However, some wanted additional or more salient reminders if they had not reported by the evening. Participants were particularly frustrated when they remembered to comply with breakfast, but later forgot to log symptoms or fasting compliance, and wanted to avoid this situation.

Participants understood the instructions to report their breakfast compliance and then peak symptoms during the three hour fasting window. Most followed the instructions, but a few knowingly appropriated the report to log peak symptoms over the entire day, though this could be confounded by a later meal. *P4 describes: "(I) would wait until the three hours and then I would report it. Then, if anything else changed throughout the day, I would go back put a note or change it."*

Because we were interested in checking daily compliance, for this study, TummyTrials enforced a strict cutoff time for reporting symptoms. Participants were not able to report symptoms after midnight passed. Participants who struggled with compliance found this frustrating, and reported opening the app post-midnight or the next day.

> *P2: I think I may have missed (reporting). By the time I went back to do it was after*

> *midnight so it had already switched to the next day. Because I was up late and I couldn't go back and do it.*

A commonly criticized aspect of TummyTrials was the scale used to report symptoms during the self-experiment. Feedback on the seven-point symptom scale ranged from suggesting changing the wording of the existing scale to adding different measures for tracking (e.g., using the *fit* of clothes as a scale for measuring the impact of bloating) . Participants reported wanting to track the number of bowel movements they had, the acuteness and duration of abdominal pain, and number of days since their last bowel movement. Two participants wanted to convert bowel urgency to a binary yes/no response instead of the seven-point scale. Some participants were confused about how to interpret levels on the scale and wanted more detailed descriptions about what they should be reporting. However, there seemed to be no common consensus as to the "best" option. P2 wanted to use the fit of her clothes as a measure of her bloating.

> *P2: How are your clothes fitting? Sometimes when you're bloated your clothes fit awful. Maybe there's a self-esteem portion in there too. For me there's a huge correlation between bloating, not going to the bathroom so being constipated, and my self-esteem and self-image.*

Self-Experiment Design

For the study, we adopted a completely randomized alternating treatment design (ATD), which treats each day as an independent sample based on natural fasting and sleep serving to reset a person's gastrointestinal system. However, participants suffering from constipation-related symptoms (e.g., infrequent bowel movement, bloating, gas) reported feeling the time period was not enough to reset. They instead reported a buildup period or a delayed reaction as long as three days after consuming their potential trigger, which might indicate a need to develop different designs for IBS-D and IBS-C (i.e., IBS associated with diarrhea versus IBS associated with constipation). One possibility is longer phases (e.g., a minimum of two avoid or consume days per phase). Two participants also mentioned that their metabolism had been clinically evaluated and found to be longer than average, which likely delayed their symptom reaction time.

A couple of participants suggested a traditional AB design of six continuous avoid days and six continuous consume days, suggesting this would be easier (though we note that it would also provide less power and potentially introduce confounds):

> *P8: I honestly think, having done an elimination diet before, that you could use this and say for even 10 days you're going to eliminate this and then the next, and then the next, and then really get a good set of information. ... (Randomization) was harder to manage.*

Randomization also sometimes produced sequences in which participants had three to four consecutive days in the same condition or sequences where the majority of days in a condition fell on either a weekday or weekend. P6 thought such long streaks could confound results:

> *P6: If you see I had 3 days in a row with no caffeine, maybe that helped my stomach settle.*

Similarly, having a condition mostly on weekdays or weekends could confound results if a participant's routine differed in ways that affected their symptoms. Potential examples include different weekday and weekend eating routines, different amounts of stress, or different amounts of exercise.

A few participants reported experiencing carry-over effects from the previous night's dinner: *"I have recognized, for me, that sometimes my symptoms are showing overnight" (P7)*.

Many participants reported eating the same breakfast for 12 days in a row was boring. Although they understood the importance of consistency to avoid confounds, a couple unintentionally broke the protocol by occasionally eating a non-standard breakfast. For example, P8 was supposed to have toast and decaf on avoid caffeine days: *"I think I pretty much didn't (have any toast). I had bananas on two of the days and I had ... One day I had eggs. And bacon because I had family in town and I made it. One day I had cantaloupe."*

Finally, a few participants said their motivation to complete the self-experiment declined after they felt their symptoms did not differ between avoid and consume days. They suggested ending a trial early when this happened. P5 felt *"Yeah, my motivation had waned ... and it was obvious to me that I probably had figured out."*

Impact on Social Life

All participants reported negligible impact on their day-to-day social life. In particular, they felt experimental manipulation of breakfast reduced the burden: *"Breakfast is probably the best option. It has the least social impact, don't need to eat a lot, and low number of food items keeps it simple" (P11)*. Still, people in shared living situations developed workarounds for potential conflicts. This was particularly true when testing caffeine. P8 told her husband she was not drinking coffee some mornings, and *"he said he was going to go to work early, and he did, in case I was grouchy."* In another example, P3 and her husband went for brunch and found it *"difficult"* to have the brunch they wanted while complying with the trial. As a workaround, she worked with restaurant to customize her omelet and brought her own pear and avocado. Overall, after participants explained their needs and the experiment to friends, family, or even restaurant staff, they were met with support. P3 also ate breakfast one morning with her kickball team. After she explained the experiment, the team was curious and supportive, and even helped her comply.

4.4.4 Interpreting the Results

At the end of the self-experiment, participants received a separate result for each tracked symptom, including two parts (Figure 1D): the sentence summarizing evidence of the food being a trigger and an interactive visualization showing their data. Participants varied in how they interpreted their results, which also led to a variety of planned follow up actions.

Three participants received 'possible evidence' ($.05 < p < .1$) on symptoms they were tracking, and only P9 received 'strong evidence' on one of her five symptoms. By far, 'no evidence' was the most common result for all participants, and they commonly interpreted 'no evidence' as an actionable result. To most, 'no evidence' meant the food is not a trigger and they can consume it without exacerbating their symptoms. Interviews also found participant interpretation often differed from the summary result provided by TummyTrials analysis. We observed this in both directions of possible interpretation.

Study shows food is a trigger - Does not believe food is a trigger. P9 had 'Strong evidence' that caffeine affects her hard passage of stool, but had 'No evidence' for the other four symptoms she was tracking. She interpreted her overall result to be no evidence and said she would need the

visualized data points for avoid and consume days to be quite far apart in the trend visualization to be otherwise convinced of the results.

Study does not show food is a trigger - Believes food is a trigger. P6 believed caffeine was a trigger for him despite a lack of evidence from the self-experiment. He attributed the lack of evidence to taking a proton pump inhibitor to treat acid reflux (i.e., Lansoprazol), which masks his symptoms.

Participants expressed varying expectations when asked to explain what they would need to see in the Trend Plot to view their results as significant. Although some considered a consistent difference of one to two points between the avoid and consume days to be significant for them to take action, others wished to see the consume days consistently near the Very Severe to Extreme levels to be certain of any effect.

Future Actions

Because most results presented no evidence a tested food was a trigger, most participants planned to continue consuming the food (i.e., if they were already eating it) or to introduce it (i.e., if they had been avoiding it). They often found these results to be a relief, though some were still skeptical. For example, P14 had been warned about dairy by his naturopath, and so despite a "no evidence" result, he planned to introduce dairy *"not aggressively. . . maybe gradually."*

After reviewing results, a few participants wanted to understand the relationship between quantity of food and symptom severity. P7 was aware that caffeine is a trigger for her, and explained how she manages the tradeoff: *". . . is being more awake worth potentially having stomachache? Which matters more to me at this particular moment?"*

We asked participants if and how they would be interested in sharing their self-experimentation results with their providers (e.g., dieticians, gastroenterologist, physician). Most were planning to share the results during their next appointment.

> *P2: Probably on a routine visit. If my results were more severe and more definitive then I might make an appointment right away and say, "Oh my gosh I need to do this. Or this is the finding." I think with technology now though, I think it would be really cool just to like send the person the results.*

A few participants also expressed a desire for a collaborative self-experiment process where the provider is involved in all the stages and keep track of their progress.

4.5 Discussion

4.5.1 Low Burden and High Compliance

TummyTrials was designed to scaffold the self-experimentation process, and it was successful in doing so. Participants described it as a low burden experience and they achieved high compliance rates relative to food diaries.

Although participants completed their self-experiments, they sometimes faced challenges in terms of flexibility and monotony. Some people may benefit from alternate study designs that further reduce impact on the participant's life, such as by allowing them to designate gap days in the experiment in advance. Although not a barrier to compliance, many participants discussed the monotony of eating the same breakfast for 12 days in a row. For some, the duration was a barrier to quickly moving to a new trial and trigger. Gap days, or a range of meals with similar content but differing tastes and textures, might mitigate this barrier.

Participants were particularly frustrated when they complied with the condition but forgot to log symptoms before the next day. Future work should determine the longest that symptom reports are valid. For example, pain reports are considered valid for three days [86]. Some sacrifice in rigor may be justified to improve compliance rates and reduce frustration.

Some participants reported the app helped them stay honest in reporting their symptoms. While this may be true in the current study, designers and researches should be attuned to possible changes if the stakes of the self-experiment change (e.g., if a provider is actively involved, if the outcome may have impact on the treatment they are receiving).

4.5.2 Tension between Scientific Rigor and Lived Experience

As a proof of concept, we chose a completely randomized ATD with 12 days of observations. The statistical power of this type of design relies on the number of observations collected for that specific case or individual. As the number of observations (in this case, days) increases, the number of permutations increases as well, and this leads to increased statistical power. However,

the experiment's ability to detect an effect is also dependent upon the size of the effect. A food trigger that has a small effect on IBS symptoms will require an ATD with a larger number of days to be detected. Conversely, a food trigger with a very large effect on IBS symptoms can be reliably observed with a short ATD experiment. Combinations of these factors (e.g., observations, effect size) can lead to errors in hypothesis testing. With a small number of observations, Type 2 errors are more likely (i.e., an incorrect conclusion of "no effect").

Therefore, some of the "failed" experiments reported in this study may have been the result of the limiting design of our application. Participants, however, saw "no evidence" results as a success. The potential trigger foods, in the quantities they ate them, did not lead to personally meaningful changes in symptoms, so they felt comfortable continuing to eat them.

As self-experimentation moves forward, designers and researchers will need to develop techniques that help people create experiments with a level of rigor appropriate for their own questions and constraints. For example, one person might want a longer experiment to test for small effects, while another might want a shorter experiment in which they consume amounts of foods that could have large effects. Someone else may not care to test triggers that are likely to have only small effects or might determine effect sizes small enough to not matter and thus not design studies to detect effects smaller than that. Finding ways to scaffold this process of designing more flexible experiments, and engaging clinical expertise as necessary, remains an open challenge.

Designing and completing the right SCD for each individual problem is a complex process. Participants can quite easily falter along the way, but also showed an ability to improvise. For example, participants who overslept shifted their breakfast time, fasting period, and reporting time, a reasonable workaround. In other cases, participants changed their experiment in ways that present a greater threat to the validity of results, such as reporting symptoms outside of the fasting window or symptoms that extended beyond a single day.

Therefore, an important take away from this study is that when directly applying SCDs from the lab into the wild, self-experimentation systems should be designed such that they: (1) "are prepared to fail and designed for failure", including incorporating flexibility in the design to have tolerance for missing or corrupted data and ensuring common failure points are accounted for in the design to ensure adherence to the methodological requirements of the self-experiment; (2) take advantage of the wide range of SCD methods so that particular users can choose designs appropriate

to their individual situations; and (3) provide people with enough of a scientific understanding about system design choices so they can appropriately weigh the different sources of evidence and SCD rigor, from there further advancing their self-understanding of food triggers.

Finally, while the TummyTrials self-experiments were designed to answer an "all or nothing" question (e.g., is this food a trigger for my symptoms?), people do not usually think in such binary terms. Although we could answer if the food was a trigger, some people were more interested in questions of the form "How much food can I consume and still manage my symptoms?" and "Am I willing to increase my symptoms by X amount if I consume more of the food?". The results of this study suggest a "threshold" testing approach that helps people predict the consequences of eating a certain amount of food would be of additional value.

4.5.3 Supporting Post-Outcome Steps

Although feasibility is important, the main outcome of a self-experiment is to support action or behavior change stemming from the result. In the study, we found hints of confirmation bias within some study participants. As with P9, participants are prone to glean what they expect to from the data. This bias might indicate a need to present a more comprehensive result section rather than just showing evidence or lack thereof. Results can also be modelled to be pathways to the possible next steps. For example, if the participant is still doubtful, they could re-test the same food for higher confidence. If they are confident in the result, a system could suggest the possibility of testing for a threshold. If they are still not confident of the result, a system could prompt them to consult their physician to ensure experimental validity.

We did not substantially explore opportunities for patient-provider collaboration in the self-experimentation process. Many participants mentioned the desire to involve their providers at various stages of the self-experiment. If an interface design enabled the provider to assist in creation, monitor the self-experiment, and collaboratively go over the results, the process might have a more significant impact.

4.5.4 Toward a General-Purpose Self-Experimentation App

The TummyTrials app was designed specifically to help people with IBS to identify food triggers. However, people may wish to investigate other questions across a variety of domains using a similar systematic process [96]. We believe the self-experimentation framework is applicable across many domains, consisting of the choice of an independent and one or more dependent variables, support for the self-experiment process, reminders to report compliance and enter data, with analysis and visualization of the results. In our previous work, we describe the various absolute and desired requirements for applying the framework to a domain [96]. However, a substantial amount of expertise was required to design a self-experiment which maximized potential for a statistically significant result, minimized confounds, identified appropriate measures, and chose hypotheses that were most likely to have an impact on health outcomes. Although having a completely customizable platform for self-experimentation may be possible, there is a risk that it would result in people conducting many self-trials that do not reach meaningful results. Incorporating advice from domain experts would minimize this risk. Experts can design valid self-experiments for different questions that people can choose from as a starting point. The process can be simplified by choosing among dependent variables, such as by creating a curated library of validated measures from which people could select to improve the quality of the experimental designs. This content can include subjective self-report measures like those used in TummyTrials, but also more objective measures imported from automated sensing approaches.

4.6 Summary

My research aims to support people with chronic conditions to get a personalized understanding of the unique nature of their condition. As I noted in the introduction to this chapter, TummyTrials leverages population-level evidence that food is the most common trigger for IBS to design self-experiments (Chapter 3) that scaffold collection and interpretation of relevant data to provide people with a personalized understanding of which foods are triggering their symptoms. Although TummyTrials was limited in scope (it supported limited food items for testing), it demonstrated the feasibility of patients undergoing valid self-experiments as a diagnostic approach to better understand their chronic condition. It also surfaced the tension arising from the lived experience of

conducting self-experiments at home. Designing for the uncertainties of the lived experience will be vital component for future systems aimed at supporting in-the-wild experimentation.

TummyTrials was among the first tools to demonstrate the identification of individualized symptom triggers by effectively supporting an n-of-1 process in everyday life. Although this was done in the context of supporting people with IBS, the approach is generalizable to other conditions requiring such individualized understanding. It was instrumental in securing an NIH R01 (#LM012810) to improve and expand the self-experimentation approach to other chronic conditions. The research team has also seen interest from several medical researchers, startups, and venture capitalists seeking to incorporate the approach in their work. Based on my learnings from this research, I discuss my thoughts on the future direction of n-of-1 research that I would be interested in pursuing in chapter 7.

Chapter 5

BEACON: SUPPORTING SELF-TRACKING TOWARD INDIVIDUALIZED COGNITIVE ASSESSMENT IN CHRONIC LIVER DISEASE PATIENTS

People with chronic liver disease or cirrhosis are at high risk of developing hepatic encephalopathy (HE) which leads to cognitive impairments (i.e., as discussed in Section 2.3.2). In early stages, people with undiagnosed HE are more prone to driving accidents, falls, and other complications which progress to severe personality disorders and coma as the condition progresses. However, given the *hidden* nature of the condition, it is difficult to diagnose in a timely manner. Research in medicine has shown that critical flicker frequency (CFF) threshold is capable of enabling early diagnosis of HE [146]. At a population-level, we know that a CFF threshold of $<$39Hz is indicative of early stage HE among cirrhotic patients. However, literature also acknowledges factors such as age, vision impairments, and fatigue many influence a person's measured CFF [107]. What is needed is a personal CFF baseline for an individual which enables them to monitor if their condition is stable, improving, or worsening. To support individuals with cirrhosis get a personalized understanding of their CFF I developed Beacon, a device that enables self-assessment of CFF at-home.

In this chapter, I highlight the iterative design process of Beacon; the formative design study and comparative study, both with healthy participants, to establish performance against the existing Lafayette device; findings from a focus with hepatologist; and the ongoing validation study with cirrhotic patients. Figure 5.1 provides an overview of the research.

As part of my dissertation's demonstration of my thesis, in this chapter:

1. I leverage existing *population-level* evidence that promotes the use of a CFF threshold as an early indicator and that a threshold of $<$39Hz is indicative of decline in cognitive performance among cirrhotic patients;
2. I examine the unmet desire among the provider, patient, and care-givers for *personalized understanding* in the form of an objective measure of a patient's cognitive functioning for

timely treatment as there is currently no way for them to get such a measure; and

3. I demonstrate how Beacon scaffolds the *collection* of CFF measurement, thus laying groundwork for future work examining individual *interpretation* of CFF variation.

In addition to leading the team of multi-disciplinary researchers involved in this project, my contributions include: designing and building early prototypes of the platform; designing and administering the studies; designing and leading the focus group; and continuing to guide the long-term direction of the research project towards clnical validation and adoption.

This research was done in collaboration with Rafal Kocielnik, Xiaoyi Zhang, Jasmine Zia, George N. Ioannou, Sean A. Munson, and James Fogarty. This line of research has lead to workshop submission at UbiComp 2016 [93], poster presentations at the Hepatology conference [171, 173], a journal publication at IMWUT [94], and a patent application.

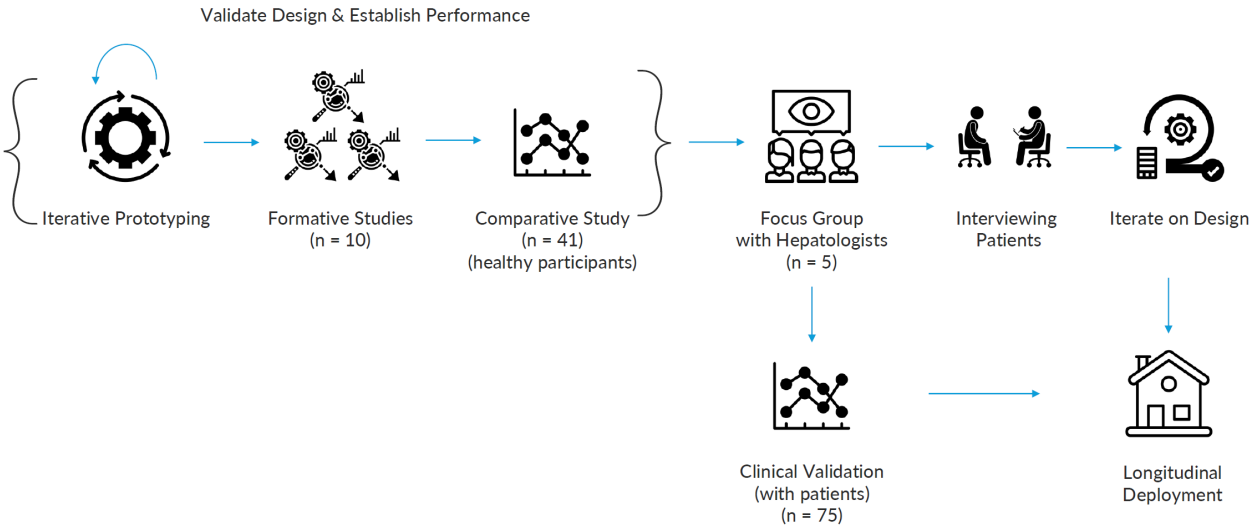


Figure 5.1: Overall design process of the Beacon platform.

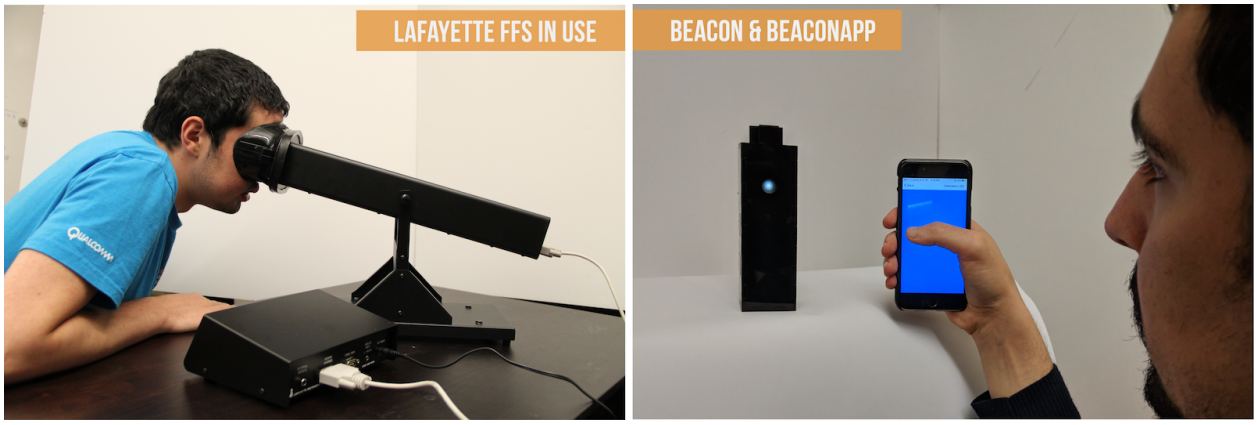


Figure 5.2: Beacon is a portable, inexpensive, and self-administrable alternative for measuring critical flicker frequency (CFF). **(left)** The existing gold standard device, Lafayette Flicker Fusion System; **(right)** Beacon.

5.1 Design Goals & Principles

Design Goal

With *Beacon*, we sought to create a portable, inexpensive, and self-administrable device for measuring the CFF threshold, thereby increasing access to the measure by clinicians and patients. We selected the Lafayette FFS, a medical gold standard device, as the reference device for us to model and compare performance against while building and testing *Beacon*. We adopted an iterative design process to arrive at the current version of the *Beacon* prototype.

Design Principles

Our design principles for *Beacon* were guided by current barriers to the adoption of the CFF measure in clinics. The key barriers we focused on addressing include: (1) limited access to testing devices due to cost constraints, and (2) the need for specialized training to use the device [8].

These led to our initial design principles that *Beacon* should be: (1) affordable enough enable adoption as a widely-used screening device, and (2) easy to use (i.e., requiring minimal training to operate). We also envisioned an additional use case for *Beacon*, enabling at-home CFF measurements for patient self-monitoring. This further emphasizes the above principle of

minimal training, as people without medical training will be using the device. It also leads to the additional design principle that *Beacon* should have: (3) a portable form factor to allow easy at-home *self-administering*. These design principles are in addition to the core requirement that the *Beacon* prototype should perform as well as the existing medical devices that measure CFF.

5.1.1 Factors Affecting CFF Measurement

The CFF measure of an individual is not only affected by the neurological symptoms an individual might be experiencing, but also by the apparatus design and set up [106, 107]. We discuss some of the key factors that can affect an individual's CFF measurement, which therefore need to be addressed in a design, and how current medical devices like Lafayette FFS address these factors.

The key factors affecting an individual's CFF measure:

1. The *intensity of the light source* used in the device is directly proportional to the measured CFF value (i.e., the brighter the light source, the higher the CFF value measured using that light).
2. The *intensity of the ambient light* of the room has an inverse relationship with the measured CFF value (i.e., the brighter the room, the lower the CFF value measured in that room).
3. The *viewing angle* also affects the CFF measure, with an ideal viewing angle being close to 0°. Peripheral vision has a greater sensitivity to movement, and hence a higher CFF value may be measured at other viewing angles.

Measuring CFF also requires choice of an appropriate *threshold detection algorithm*. A common algorithm used in CFF research is the *method of limits*, which is borrowed from psychophysical research and focuses on the influence and relationship between stimuli and the sensation and perception of these stimuli by an individual [67]. In the method of limits, a stimulus (i.e., light in the case of CFF) is presented and increased or decreased until it is perceivable by an individual. The primary parameter to be tuned in the method of limits is the *step rate* (i.e., the rate of change in the stimulus's ascent or descent). The design trade-off between step rate and effective CFF measurement is that of *reaction time* and *fatigue*. If the step rate is high (e.g., 2Hz/sec), even a small delay in reaction time would amount to a large error in the measured CFF. If the step rate is too low (e.g., 0.1Hz/sec), there will be an adverse impact of fatigue, which will likely compound during consecutive CFF measures.

5.2 Formative Design Research

We now discuss how these factors are addressed by current medical devices like the Lafayette FFS and by *Beacon*. Systems like the Lafayette FFS encase the light source in a viewing chamber (Figure 5.3A) and have the individual use a mask (Figure 5.3B) with the device. The viewing chamber allows even a low intensity light to be clearly visible in the dark chamber, and the mask ensures that ambient light is completely blocked out (Figure 5.3 right image), thereby controlling for the effect of both variables. The viewing chamber also has two circular cut-outs for the eyes, which are aligned with the light sources, ensuring that the viewing angle is close to 0°.

Figure 5.3: Left: The Lafayette Flicker Fusion System (Lafayette FFS) has three components: a viewing chamber with the light stimulus (A, B), a controller (C), and a software program to record results (D). The clicker (E) is connected to the viewing chamber and records a person's input (i.e., indicating they see a flickering light or a fused light). Right: To use the Lafayette Flicker Fusion System, a person presses their face against the mask covering the viewing chamber so as not to allow any outside light. The person then focuses on the light inside the viewing chamber and uses the clicker to record their input.

We evaluated the feasibility of using existing capabilities of a modern phone to measure CFF and faced prohibitive limitations in choosing an appropriate light source, which deterred us from further pursuing that direction. Although many phones include an LED used as a flash, no modern phones provide the API access necessary to generate frequency modulations to administer a CFF measurement. Additionally, standard phone displays do not have a high enough refresh rate to generate the necessary frequencies at 50% duty cycle.

The goal of our first prototype was to create a less expensive CFF device which can be operated

using a phone (i.e., replacing the controller and computer set up with an application on the phone). Initially, by still using a viewing chamber design, we addressed the three factors in the same manner as Lafayette FFS. We began with a cardboard prototype to understand the set up we would need to build *Beacon*, such as what components are required and how the components might be housed (Figure 5.4 B). Upon finalizing the initial electronics (i.e., LED, microprocessor, power source), we upgraded to an acrylic box that allowed us to manipulate the distance between the light source and the eyes by sliding the LED inside the viewing chamber along pre-cut slots (Figure 5.4 C). It was during these design explorations that we realized that we could make the device portable by removing the viewing chamber. This led us to the first portable prototype of *Beacon*, which housed the LED and held the circuit inside it (Figure 5.4 D). However, the removal of viewing chamber would require examining the three factors described above to ensure the device still functions as intended. We conducted the formative studies described in section 5.3.1 to determine the value of the intensity of the light source and ambient light. We also added four calibration lights, red LEDs, as guiding lights to assist an individual in self-aligning themselves horizontally and vertically to be centered with the white light (Figure 5.4 E & 5.5).

The phone application, *BeaconApp*, connects with the RFduino over Bluetooth and allows an individual to self-administer the CFF measurement. It can be used to run the protocol to measure CFF (i.e., the method of limits), turn the calibration lights on or off, and view results. To start the measurement and record input, a person can press anywhere on the screen (i.e., a person does not look at the screen while measuring CFF).

Towards our design principles, we offer details about the affordability, portability, and ease-of-use of the current *Beacon* prototype. The total cost of the prototype, including the microprocessor, LED, diffuser, wiring, and acrylic for the case is less than $50 USD. Our current prototype's external dimensions are 140 x 40 x 40 mm. For a device like Lafayette FFS, the process of setting up the device includes plugging in cables across the viewing chamber, controller, and computer, starting the components and software, and then starting the measurement process. In contrast, the set up process for *Beacon* includes turning the device on and launching *BeaconApp* on the phone to begin taking a measure.

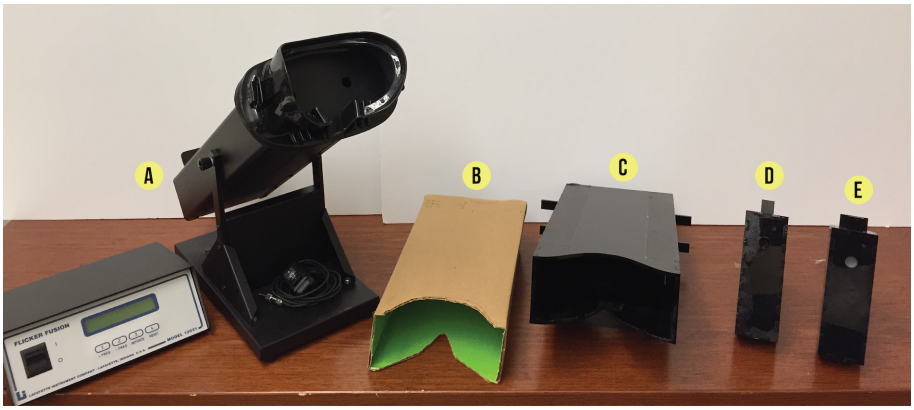


Figure 5.4: Stages of the iterative design process for *Beacon*. Left: (A) Lafayette FFS, the golden standard device that informed our design of *Beacon*. Right: (B) Cardboard prototype, (C) Acrylic prototype with adjustable distance between light source and eyes, (D) portable prototype, and (E) portable prototype with guiding lights for horizontal and vertical alignment.

5.2.1 Implementation Details

Beacon is built using off-the-shelf components and software with minor modifications. We use a RFduino microprocessor (RFD22102), selected because of it's small form factor and two-way communication API allowing for rapid prototyping of the mobile application to control the device. We selected a wide range white LED (C503D-WAN-CCBEB151 - luminous intensity from 28cd to 64cd) paired with a milky white diffuser in front of it to be the light source. The RFduino generates 50% duty cycle square waves (i.e., where pulse remains high for half of the period and low for the other half of the period) ranging from 25.0Hz to 55.0Hz at a step rate of 0.5Hz/sec (implemented as 0.1Hz/0.2sec) for the ascending and descending phase of CFF measurement.

Although past experiments have used a wider range of frequencies, most commonly between 20.0Hz and 60.0Hz, we selected 25.0 to 55.0Hz because (1) we expect all participants (including healthy individuals) to fall comfortably between those end points based on past medical research in MHE, and (2) a tighter range means a shorter run time which leads to less fatigue. Similarly, past experiments have used step rates ranging from 0.1Hz/sec to 3Hz/sec [56, 100]. We selected 0.5Hz/sec as a compromise in the trade-off between accuracy and fatigue.

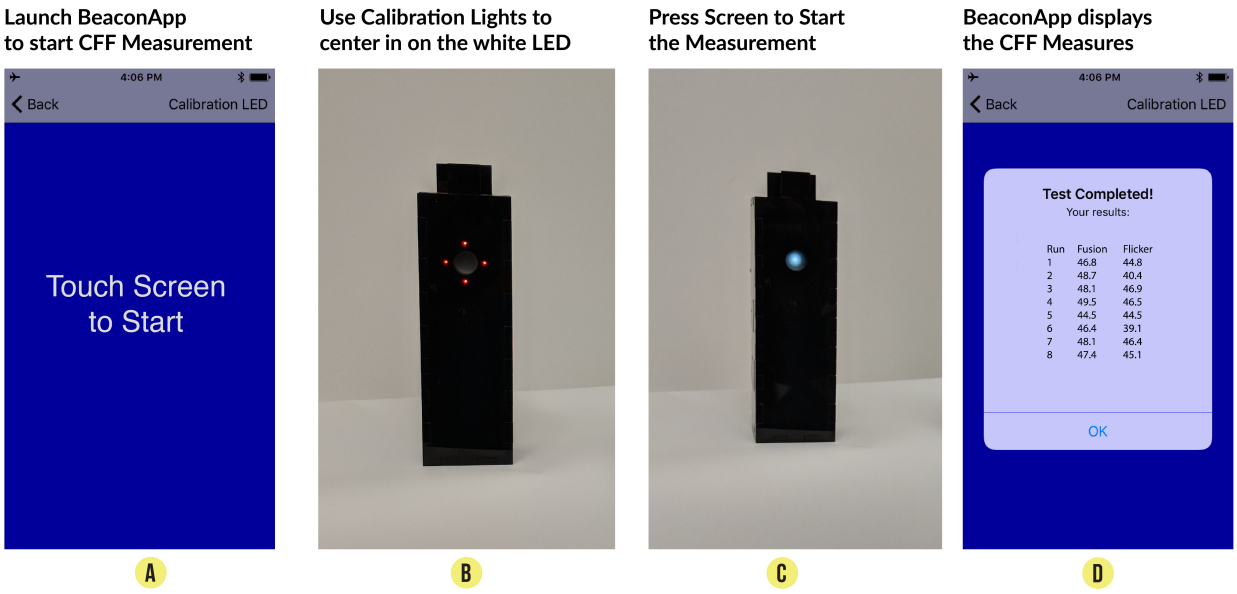


Figure 5.5: (A) To begin the process of a CFF measurement, a person launches *BeaconApp*. They can press the 'Calibration LED' button on the top right of screen to initiate calibration step, or skip the step and start the CFF measurement by pressing anywhere in the blue region. (B) To ensure an appropriate viewing angle, *Beacon* uses two sets of two red LEDs forming a target around the central white LED. During the calibration phase, the red LEDs are visible through a tiny holes only when looking straight at them (i.e., at a viewing angle close to 0°). (C) The CFF measurement is conducted using the method of limits, with the LED starting the ascending step at 25.0Hz and increasing at a fixed rate of 0.5Hz/sec. A person presses on the anywhere on the screen when they see it as no longer flickering. The same process is repeated in the descending step, decreasing from an initial 55.0Hz. (D) After a person has repeated the ascending and descending steps a predefined number of times (8 times in image), they are presented with their results of each step. The CFF is calculated as a mean of all the steps.

5.2.2 Using Beacon & BeaconApp

We now briefly illustrate how *Beacon* and *BeaconApp* work through a scenario following a person who needs their CFF measure taken, either at the clinic or at home.

The person would sit in a dimly lit room in a chair next to a table. They place *Beacon* in front of them, on the table, at eye height. They then turn on *Beacon*, and open the associated *BeaconApp* on their phone. Parameters for the measurement are preset per clinician instructions (i.e., minimum and maximum frequencies of the sweep, step rate, number of repetitions). Administering a measurement then follows these steps:

- Turn on the red calibration lights to ensure the person's view is aligned with the center (Figure 5.5B).
- Press on the *BeaconApp* screen (Figure 5.5A), which automatically turns off the calibration lights and then begins the CFF measurement procedure.
- The central white LED starts at the lowest frequency of the sweep (i.e., perceived as flickering) and then increases at a fixed step rate until the person presses the application screen to indicate that they see it as fused (i.e., no longer flickering) (Figure 5.5C).
- Upon pressing the screen again, the central white LED starts at the highest frequency of the sweep (i.e., perceived as fused) and then decreases at the same step rate until the person presses the application screen to indicate that they see it as flickering.
- The ascending and descending measure are repeated several times per clinician instructions.

Voice prompts are provided by *BeaconApp* at each step of this procedure, guiding a person through the process without requiring they look at the phone application. Upon completing the final repetition, a person sees a screen which shows them their CFF measure from this session (Figure 5.5D).

5.3 Establishing Comparative Performance Among Healthy Participants

A multi-disciplinary team of researchers across computer science, human-centered design, and medicine have contributed to the design and development of *Beacon*. Although our end goal with *Beacon* is for it to be used as a clinical decision support tool, we first need to evaluate whether its performance is comparable to existing CFF measurement devices. We first aimed to demonstrate

that *Beacon* performs on par with an existing gold standard device (Lafayette FFS) using healthy study participants. In addition to comparing the two devices, we also conducted a focus group with hepatologists regarding their current practices for diagnosing HE and possible opportunities and challenges for *Beacon*. This focus group complemented the expertise of the medical researchers on the team. In particular, we were able to solicit additional expert perspectives on the opportunities afforded by *Beacon*'s portable and easy-to-use form factor, including the prospect of patients using *Beacon* for at-home self-monitoring of their HE.

We adopt a mixed-methods approach to informing and evaluating the design of and potential opportunities for *Beacon*. We conducted three studies, with each study informing the next:

1. Two *formative data collection studies* examining the impact of the intensity of the light source and the intensity of ambient light on CFF measurement with our current *Beacon* prototype.
2. A *comparative study* examining CFF measurement with our current *Beacon* prototype relative to CFF measurement with the Lafayette FFS.
3. A *focus group* with hepatologists to understand current practices regarding screening and treatment of hepatic encephalopathy and their perspectives on the potential opportunities and challenges of integrating of a device like *Beacon* into their clinical care.

The formative and comparative study consisted of self-reported healthy individuals above 18 years of age. We recruited healthy participants because the goal of the studies was to compare the designs of the devices and not to perform a medical diagnosis. As noted in our introduction, the effectiveness of CFF as a diagnostic test for MHE is already proven, with a reported accuracy of 80%, with a sensitivity and specificity of 65% and 91% [159]. Our recruitment methods and study protocols were reviewed and approved by our University's Human Subjects Division under IRB number 00001222, *Evaluating a Novel Flicker Frequency Device*.

5.3.1 Formative Studies to Configure Beacon's Parameters

Due to *Beacon*'s portable design, we cannot use the same values of intensity of the light source and intensity of ambient light that have been used in prior medical research (i.e., similar to the designs which consisted of a viewing chamber). We therefore conducted two formative studies to examine the impact of these two parameter on the measurements of CFF by *Beacon*. For

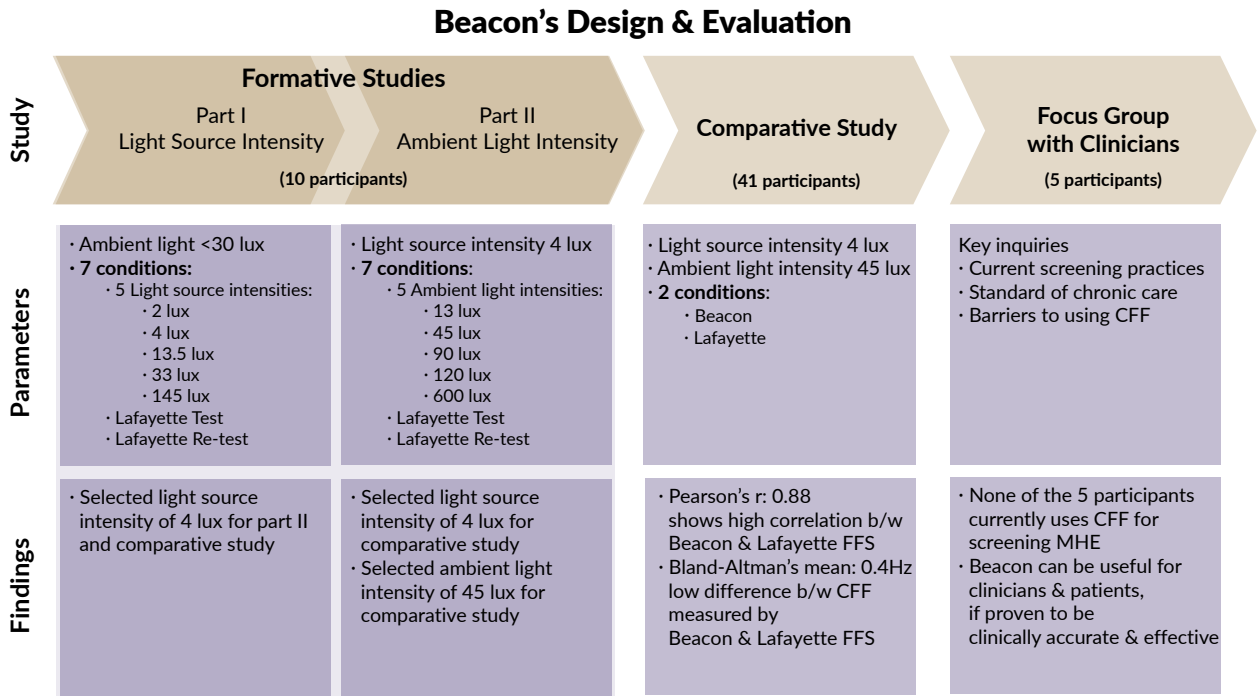


Figure 5.6: Series of studies used to evaluate the design of and potential opportunities for *Beacon*. Formative studies assisted in selecting appropriate values of light source intensity and ambient light intensity for *Beacon*. Comparative study provided evidence of *Beacon*'s comparable performance to Lafayette FFS in measuring CFF of healthy individuals. Focus group provided information about current screening practices and potential impact of *Beacon* in screening and treating patients with HE.

each intensity parameter, we selected five different values to examine their impact on the CFF measurement. Conducting a fully balanced study including two parameters with five levels each would be overly burdensome and fatigue-inducing for participants. To circumvent this problem, we examined each parameter independently and utilized the findings from the first to inform the design of the second. We initially measured the impact of light source intensity while controlling the ambient light intensity. We then used a fixed value of the light source intensity while varying ambient light intensity. We recruited the same participants for both parts of these formative studies.

Part 1: Impact of Light Source Intensity on Measured CFF

Study Parameters. To examine how varying the intensity of the light source impacts the CFF measured by *Beacon*, we collected data using **five levels of intensity — 2 lux, 4 lux, 13.5 lux, 33 lux, and 145 lux**. The intensity levels are not evenly spaced because (1) visual perception is not linear, and (2) we selected them subjectively to cover a wide range from dim to extremely bright. Approximately 15 different intensities were visually inspected in a dimly lit room, to emulate conditions we might expect in the wild, and from among them five we chose 5 for this formative study. Light intensities were measured using a light meter sensor (Amprobe LM-200LED) pressed against the diffuser of *Beacon* in a completely dark room. The different intensities were generated using the same white LED by varying the value of the resistor.

We configured both *Beacon* and the Lafayette Flicker Fusion System to use the previously described method of limits algorithm, with a range between 25.0Hz and 55.0Hz, and a step rate fixed at 0.5Hz/sec.

Procedure. The study was conducted in an empty office room on the University campus. Participants first filled out a pre-task form consisting of their legal name, age, gender, and contact information. Participants were then introduced to the Lafayette FFS and *Beacon*. Because learnability of the system was not an aim of this study, a researcher guided each participant through measuring their CFF using both devices for at least two ascending and two descending measures

until they were comfortable and confident with the measurement process. The study began after the participant reported they understood how the CFF measure worked and the devices functioned. The ambient light of the room was measured by putting the sensor next to *Beacon* and facing the wall. The ambient light was less than 30 lux across all 10 participants with a mean of 16.9 lux, median of 15.35 lux and standard deviation of 5.44 lux.

To mitigate any carryover effect (e.g., fatigue, learning), the study included **seven conditions** administered in a completely randomized order. Five of the conditions were the varying intensities of the *Beacon* light source, and the other two conditions were the Lafayette FFS (i.e., Lafayette Test and Lafayette Retest). We included the collection of two measures with the Lafayette FFS to observe any variance of measured CFF within the same participants. For each of the seven conditions, participants were asked to do six ascending and six descending runs of the method of limits, yielding **12 measures per condition**. Although CFF measurements can be conducted with six to eight runs, we decided to collect 12 to account for possible outliers in the measurements due to lapses in attention, delays in response time, or some other unforeseen reasons. Participants recorded their input using either a study phone with *BeaconApp* installed and paired with *Beacon* or the switch associated with the Lafayette FFS (Fig. 5.3 C). Participants were instructed to take as many breaks as they needed during the study. The study session lasted between 60-75 minutes.

Recruitment. We recruited 10 healthy participants through campus mailing lists. Of the 10, six reported themselves as male and four as female. Participants were between the ages of 18 and 36 years, with a mean of 27.1 years, a median of 25.5 years, and standard deviation of 4.9 years. Participants were provided with a $15 gift card as a compensation for participating in the study.

Data Cleaning. Collecting 12 measures per condition allowed us to discard extreme measures which may have been caused by reaction time, confusion, or some other factor. This use of a *trimmed* or *truncated mean* is common to obtain a more robust statistic. For each of the seven conditions we discarded the four most extreme CFF measurements (i.e., two lowest and two highest), then averaged the remaining values per participant. We combined the two Lafayette FFS measures to form the Lafayette_Avg condition to use as a baseline comparison value.

Results. The goal of this part of the formative study was to examine the impact of light source intensity on the measured CFF and select an appropriate value for *Part 2* of the formative study. We examine the mean CFF value per condition (i.e., across participants) to understand the impact of light source intensity. We found that 145 lux (highest intensity) had a mean CFF of 43.01Hz, and 2 lux (lowest intensity) had a mean of 36.96Hz. The trend in Figure 5.7, shows that measured CFF is directly proportional to the *Beacon*'s light source intensity and is in alignment with prior research and our understanding [56]. The difference between the CFF measured using 145 lux and 2 lux intensities is 6.05Hz, or 15.7% of the Lafayette FFS's average mean (38.49Hz).

We also observed that there was low variability within the conditions, indicating consistent measures. Examining the variation within the seven conditions, we observed low CFF variability between participants ($\sigma_{\bar{x}}$= std error = $\frac{\sigma}{\sqrt{n}}$: 0.58Hz to 0.91Hz). We observed slightly lower variability for Lafayette device, test, re-test, and average ($\sigma_{\bar{x}}$: 0.56Hz to 0.60Hz) as compared to *Beacon* with its five intensities ($\sigma_{\bar{x}}$: 0.74Hz to 0.91Hz).

For *Part 2* of the formative study, we wanted to select a value of the light source intensity which would produce CFF measures closest to the Lafayette FFS. We observed the intensity of 4 lux had a mean CFF of 38.85Hz and standard error of 0.91Hz, which was closest to Lafayette_Avg mean of 38.49Hz and standard error of 0.60Hz. Based on these observations, we selected light source intensity of 4 lux as the study parameter for *Part 2* of the formative study.

Part 2: Impact of Ambient Light Intensity on Measured CFF

Study Parameters. To examine how varying the intensity of ambient light impacts the CFF measured by *Beacon*, we collected data using **five levels of intensity — 13 lux, 45 lux, 90 lux, 120 lux, and 600 lux**. The different ambient light intensities were achieved by combining and manipulating the settings of multiple light bulbs capable of three different brightness settings (GE LED A21 3-Way Lamp). During the study, the participant sat on a chair at a table adjacent to a wall of the room, facing the wall, with all of the light sources behind the participant. We selected the ambient intensities based on how easy it would be to replicate the lighting conditions in a home or

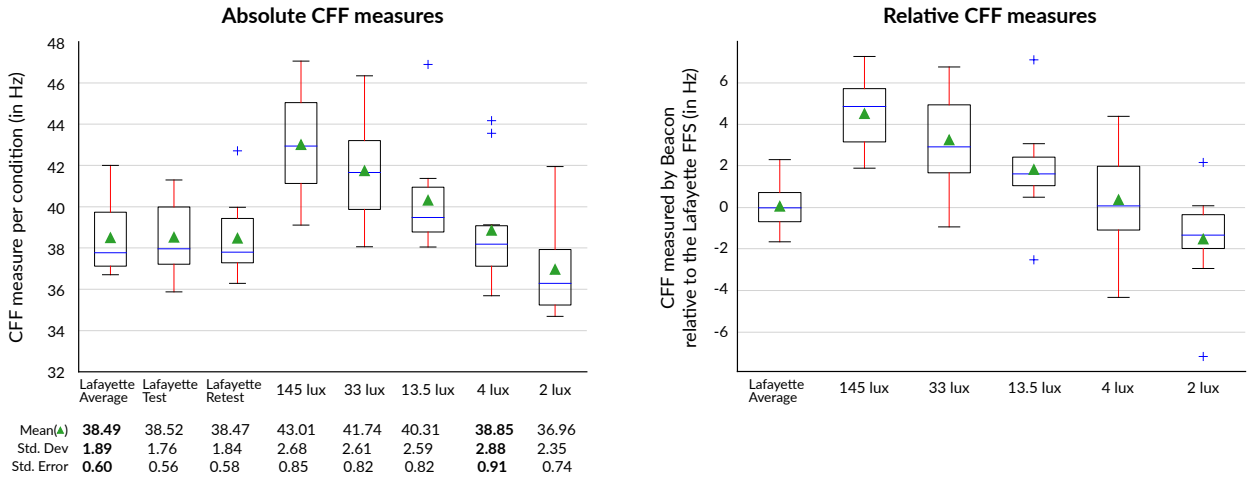


Figure 5.7: Analysis of formative study *Part 1* examining the impact of light source intensity on measured CFF. The **blue** lines in the plot represent the median; the **green** triangles, the mean. Left: The absolute values of the 7 conditions (5 light source intensities, Lafayette test, and Lafayette retest) alongside a new calculated measure called the "Lafayette average" obtained by combining the test and retest scores. The plot shows a trend that CFF value is directly proportional to light source intensity. The table underneath the plot shows the corresponding descriptive statistics. Right: The values of light source intensities relative to the Lafayette average (using that average as the baseline). Light source intensity value of 4 lux, which is the closest to Lafayette average, was chosen for the remaining studies.

a clinic. For perspective, 600 lux is similar to recommendations for office illumination while 45 lux is similar to recommended illumination for a relatively dark public area, such as a parking lot at night [136]. The light intensities were measured using a light meter sensor placed next to *Beacon* and facing the participant (i.e., facing the light sources). We used the same algorithm, range, and step rate as *Part 1* of the formative study. The light source intensity for *Beacon* was set at 4 lux as informed by the findings from *Part 1*.

Procedure. The study followed an identical procedure to that in *Part 1*, wherein participants were introduced to both the systems and completed practice runs before starting the study. The study consisted of seven conditions, using *Beacon* with the five varying intensities of ambient light of the room and using the Lafayette FFS device twice, with 12 measures per condition.

Recruitment. The 10 participants who participated in the previous study were recruited again for *Part 2*. Participants were provided with a $15 gift card as a compensation for participating in the study.

Data Cleaning. *Part 2* followed the same data cleaning process as *Part 1*.

Results. In this this part of the formative study, we examined the effect of ambient light intensity on the measured CFF and select an appropriate value for the comparative study. We observed that 600 lux (highest ambient light intensity) had the lowest mean CFF of 37.19Hz, while 13 lux (lowest ambient light intensity) had the highest mean CFF of 38.69 lux. The trend in Figure 5.8, shows that measured CFF is indirectly proportional to the ambient light intensity for *Beacon*, and is in alignment with prior research and our understanding [56]. The difference between the CFF measured using 600 lux and 13 lux intensities is 1.5Hz, or 3.85% of the Lafayette FFS's average mean (38.91Hz).

Examining the variation within the seven conditions, we observed low CFF variability between participants ($\sigma_{\bar{x}}$= std error = $\frac{\sigma}{\sqrt{n}}$: 0.58Hz to 0.73Hz). We observed similar variability for Lafayette FFS test, re-test, and average ($\sigma_{\bar{x}}$: 0.59Hz to 0.73Hz) as compared to *Beacon* with the five ambient intensities ($\sigma_{\bar{x}}$: 0.58Hz to 0.72Hz). Comparing differences between the highest and lowest intensities across the two parts of our formative studies, we observe that changing light source intensity has greater impact on resulting CFF measurement (a difference of 6.05Hz between the maximum and minimum, or 15.7% of mean) than changing ambient light intensity (a difference of 1.5Hz between the maximum and minimum, or 3.85% of mean).

For the comparative study (5.3.2), we desired a value of the light source intensity and ambient light intensity such that CFF measured by *Beacon* would be similar to the Lafayette FFS. Although ambient intensity of 13 lux was objectively the closest value to Lafayette_Avg (38.69Hz to 38.91Hz, difference of 0.22Hz), we decided to select **45 lux** for the comparative study because we deemed it was a much more reasonable level of ambient lighting for clinicians and patients to be able to achieve in their respective environments, and the impact on the CFF was small and not clinically meaningful (38.41Hz to 38.91Hz, a difference of 0.5Hz compared to a difference of 0.22Hz at 13

lux).

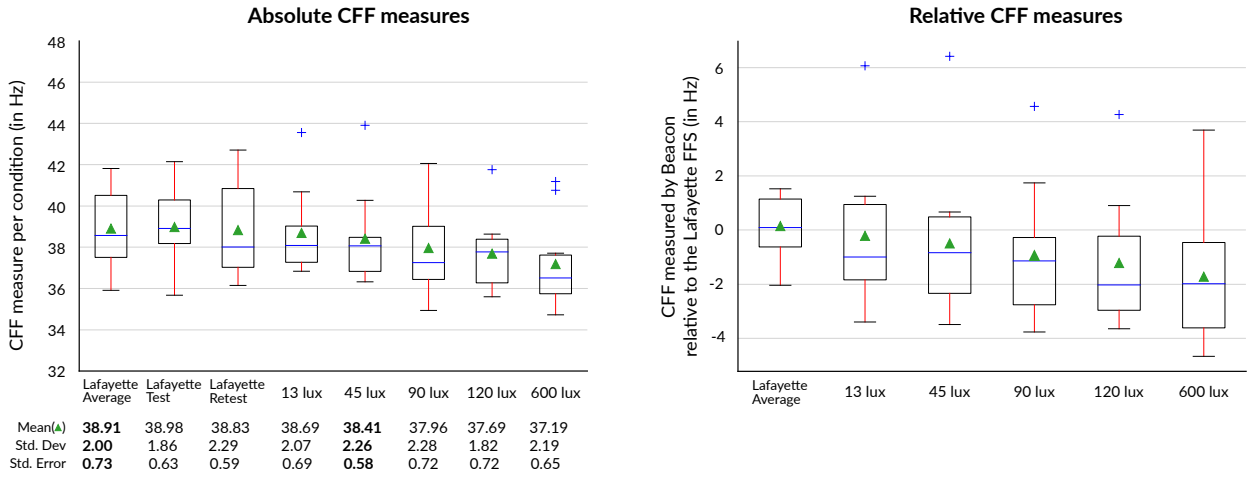


Figure 5.8: Analysis of formative study *Part 2* examining the impact of ambient light intensity on measured CFF. The **blue** lines in the plot represent the median; the **green** triangles, the mean. Left: This plot shows the absolute values of the 7 conditions (5 light source intensities, Lafayette test, and retest) alongside a new measure called Lafayette average which combines the test and retest score. The plot shows a trend that CFF value is indirectly proportional to ambient light intensity. Table underneath the plot shows the corresponding descriptive statistics. Right: This plot shows the values of ambient light intensities relative to Lafayette average by using it as the baseline. Ambient light intensity of 45 lux was chosen for the comparative study since it was deemed to be a more easily achievable ambient light setting in clinics and homes.

5.3.2 Comparative Study to Evaluate Beacon

The formative studies guided our decision of selecting appropriate light source and ambient light intensity parameters for *Beacon* to calibrate the measured CFF with the existing Lafayette FFS medical device. Using the selected parameters, we next conducted a comparative study to examine the CFF measured by *Beacon* relative to Lafayette FFS. Because our primary goal is to measure device performance, building upon prior research validating the potential for CFF measurement in diagnosis [159], we focus this comparison on healthy volunteer participants. Future research can then build upon this device validation to explore clinical usage with appropriate patients. Similarly, we do not collect psychometric tests such as PHES because we do not expect them to be informative

with healthy volunteer participants (i.e., there is no diagnosis to be made). Future clinical trials of *Beacon* could collect additional such measures as appropriate.

Study Design

We used the same algorithm, range, and step rate as the formative studies. Based on the previous studies, we set the light source intensity at 4 lux and ambient light intensity at 45 lux. We conducted a within subjects study with two conditions, using *Beacon* and using Lafayette FFS, counterbalanced across participants. For each of the **two conditions**, participants were asked to do eight ascending and eight descending runs of the method of limits, yielding **16 measures per condition**.

Procedure

To control the ambient light, the study was conducted in empty rooms with no windows. Similar to the formative studies, participant information was collected and they were introduced to both the systems and given a couple practice runs before starting the study.

Recruitment

43 healthy participants, 18 years or older, were recruited for the comparative study using mailing lists, fliers, and sign up sheet at a retirement community. 24 participants reported themselves as female and 19 as male. Participants were between the ages of 18 and 99 years with a mean of 41.7 years, median of 28 years, and standard deviation of 25.3 years.

Data Cleaning

Data cleaning for the comparative study involved: (1) removing two participants due to missing data, (2) using an adaptive algorithm to discard outliers, and (3) testing the normality of collected data to justify use of specific statistical test.

We removed data for two participants (P33 and P35) from the analysis because we could not obtain their CFF measures using *Beacon*, leaving us with 41 participants for analysis. P33 (87 y.o.) was unable to see flicker using *Beacon* at any frequency for unknown reasons, their CFF measure using the Lafayette FFS was 31Hz. P35 (88 y.o.) who self-reported as being legally blind in one

eye, had a very low CFF measure of 13Hz (using Lafayette FFS) which we were unable to capture without updating the *BeaconApp* during the session.

We collected 16 measures for each of the two conditions (i.e., *Beacon* and Lafayette FFS). Because of the potential for erroneous measurements (e.g., due to distractions), prior work using CFF has implemented a constraint where measures per participant should be collected repeatedly until the standard deviation is less than 3Hz [56]. We followed an adaptive algorithm for discarding high and low extremes based on a similar constraint:

- If max-sd $>$ 3Hz, discard two extreme measures — lowest and highest.
- If max-sd still $>$ 3Hz, repeat previous step.
- Terminate if number of measures is $<$ 8 per condition.

We tested the normality of the data using D'Agostino and Pearson's test that combines skew and kurtosis to produce an omnibus test of normality [47, 48]. The normality test has the null hypothesis the sample is not normal, which means a p $>>$ 0.05 implies that the sample is not significantly different from a normal distribution (i.e., sample is normal). The normality test for the lafayette CFF measures (coef=1.529, p=0.465) and *Beacon* CFF measures (coef=3.033, p=0.219) suggested that normality was a reasonable assumption.

Results

The goal of the comparative study was to measure the performance of *Beacon* as compared to Lafayette FFS. As discussed in the introduction, the effectiveness of CFF for diagnosing MHE is already proven [159]. Demonstrating that *Beacon* performs on par with Lafayette FFS is a first step toward our goal of providing an accessible clinical decision support tool for measuring CFF.

We first examine the performance of our adaptive algorithm and then present our analysis on the cleaned data. For us to consider the adaptive algorithm effective, it should (1) be less sensitive to outliers than original data, and (2) still give a reasonable central tendency. As we can see from Table 5.1, before discarding any extremes the maximum standard deviation for Lafayette FFS is 5.51Hz and for *Beacon* is 8.59Hz (i.e., before applying our algorithm). After discarding 6 extremes, we still have 10 measures per condition and find the maximum standard deviation is much more similar, at 2.78Hz for Lafayette and 2.93Hz for *Beacon*. During this process the mean CFF values

of Lafayette FFS changed from 37.60Hz to 37.58Hz and for *Beacon* they changed from 38.31Hz to 38.01Hz. Both of these difference are negligible, showing that the algorithm is removing outliers without affecting the mean. We also noted that while using the adaptive algorithm on Lafayette FFS, only 1 participant needed two extremes discarded, compared to *Beacon* where 6 participants had two extremes discarded, 5 had four discarded, and 1 had six discarded extremes. Although this did not impact the means, it suggests there is more internal variation in the measures taken by *Beacon* when compared to Lafayette FFS.

	Lafayette FFS			***Beacon***		
Without discarding extremes	Mean CFF: 37.60			Mean CFF: 38.31		
	Max SD	Min SD	Mean SD	Max SD	Min SD	Mean SD
	5.51	0.6743	1.70	**8.49**	0.7314	2.69
After discarding 2 extremes	Mean CFF: 37.59			Mean CFF: 38.11		
	Max SD	Min SD	Mean SD	Max SD	Min SD	Mean SD
	2.72	0.59	1.30	**7.56**	0.60	1.87
After discarding 4 extremes	Mean CFF: 37.58			Mean CFF: 37.97		
	Max SD	Min SD	Mean SD	Max SD	Min SD	Mean SD
	2.53	0.50	1.07	**6.02**	0.41	1.38
Adaptive until SD per condition $<$ 3Hz	Mean CFF: 37.58			Mean CFF: 38.01		
	Max SD	Min SD	Mean SD	Max SD	Min SD	Mean SD
	2.78	0.65	1.58	**2.93**	0.71	1.82

Table 5.1: Descriptive statistics for the Adaptive Algorithm used in the Comparative study data cleaning for each device (in Hz). The goal of the algorithm is to reduce maximum standard deviation while having minimal impact on the mean CFF. As seen here, using the adaptive algorithm *Beacon* achieved a maximum standard deviation of 2.93Hz compared to 2.78Hz achieved by Lafayette FFS. The mean CFF remains unaffected with a difference of only 0.29Hz in *Beacon* when not using and using the algorithm and 0.02Hz in Lafayette FFS.

We took the mean CFF measured per participant in the two conditions and used correlation analysis to evaluate the consistency in measured CFF values between the devices (Fig. 5.9). Despite the normality test results suggesting reasonable assumption of normality, we present results from both Pearson and Spearman correlation analyses for greater transparency (i.e., because Pearson assumes normality while Spearman is non-parametric). We observed a Pearson correlation

coefficient of **0.88** ($p < 0.001$) and Spearman correlation coefficient of **0.84** ($p < 0.001$), both of which correspond to strong correlations.

We used a Bland-Altman plot to better understand the agreement and the expected limits of difference between the CFF measurements taken by *Beacon* and Lafayette FFS. Also know as a difference plot, Bland-Altman plot is ideal for comparing two measurement techniques (or devices) that each produce some error in their measures and is extensively used to evaluate the agreement among two different instruments or measurement techniques [51, 56, 68]. Figure 5.9 shows the Bland-Altman plot on the right. The X-axis represents the mean of the CFF measurements for both devices and the Y-axis represents the absolute difference between the measurements taken by the two devices. The plot includes the line for the mean difference between the measurements (0.40Hz between Lafayette and *Beacon*) and the 2 lines showing the 2s (1.96 standard deviation) limits of differences between the measurements (also called 95% limits of agreement) which span from -3.27Hz to +4.07Hz for *Beacon* and Lafayette. *The limits of agreement tell us that the difference in CFF measured by Beacon and Lafayette will be at most $\pm$3.67Hz for 95% of the measurements*. A recent comparison of different CFF protocols using the same device reported limits of agreement from $\pm$3.02Hz to $\pm$6.74Hz [56]. Given that prior comparison of results from the same device using different algorithms, our results are promising when comparing two devices.

Our participant sample spanned a wide age distribution (18-99), which allowed us to examine the possible impact of age on the difference in CFF measurements by *Beacon* and Lafayette devices. Although we counterbalanced the order of measurement using each device, we also wanted to check if ordering had any systematic impact on differences in measurements. To examine both, we performed a regression analysis. We regressed the age and measurement order on the difference in CFF measurement between both devices. We found no significant impact of age ($\beta = -0.0043, SE = 0.013, p = 0.746$) or order ($\beta = -0.0171, SE = 0.619, p = 0.978$).

5.3.3 Focus Group with Clinicians to Understand Current Practices and Opportunities Surrounding Screening for MHE

The *Beacon* research team includes a gastroenterologist and a hepatologist, who provided medical expertise to support our development of *Beacon*. Our initial prototype and results demonstrated

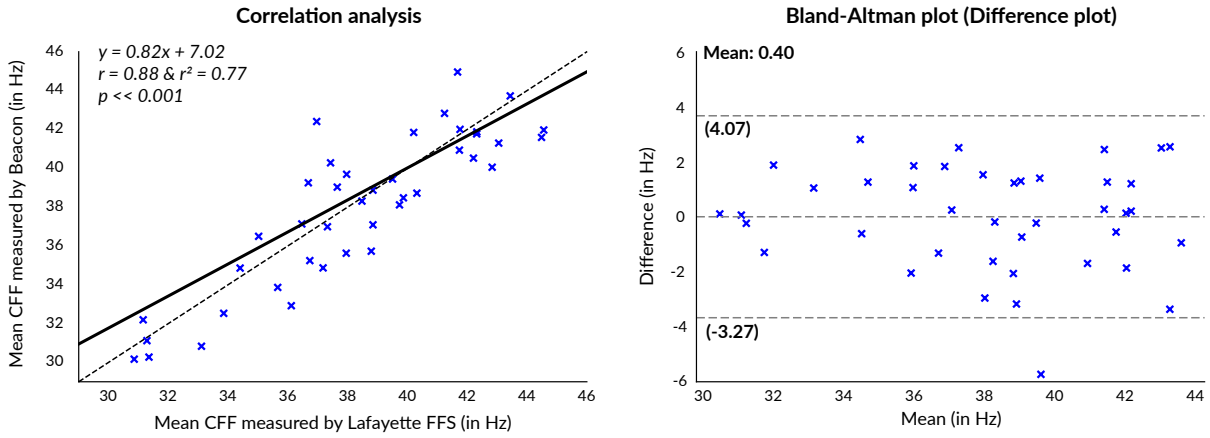


Figure 5.9: Analysis of comparative study evaluating the performance of *Beacon* when compared to the Lafayette FFS. Left: Regression analysis shows a strong correlation between the CFF measure by *Beacon* and Lafaytte FFS with a Pearson's R of 0.88. Right: The Bland-Altman plot shows the mean difference between *Beacon* and Lafayette FFS to be 0.4Hz with a maximum difference of at most $\pm$3.67Hz for 95% of the measurements.

technical feasibility of *Beacon*. To complement this, we gathered clinical perspectives on the acceptability and utility of such a device. Consultations with additional practitioners helped us better understand current diagnostic practices for HE, barriers to adoption of CFF, provider interest in a device like *Beacon*, and barriers they might envision to successful adoption. As the new form factor might allow new use cases for CFF, we were eager to explore provider views on more routine screening for MHE, such as patient use of *Beacon* for at-home self-monitoring of HE. Because HE primarily occurs among patients with end-stage liver diseases, treatment is often managed by specialists such as hepatologists. Consequently, we recruited five senior hepatologists at a medical school to participate in a focus group about challenges and opportunities associated with the adoption of a tool like *Beacon*.

Procedure

The focus group was conducted at a medical school for the convenience of the clinicians. Before starting the group discussion we collected background information about the clinicians, including title and years of experience treating patients with cirrhosis and HE. The focus group was structured

around three key areas of inquiry in order to understand the current state of:

- Screening practices for patients with HE (e.g., At what stage are patients screened? How are they screened? What is the clinician's satisfaction with the screening process?)
- Standard of chronic care for patients with HE (e.g., What is the advice given? What does routine care include? Do patients do any self-tracking of their condition?)
- Barriers to using CFF as a screening option for MHE (e.g., Why they do or do not use CFF? What benefits would they anticipate if CFF were an easily available measure? Why might they decide not to use CFF?)

Participants also completed a short paper-based questionnaire during the focus group to collect individual responses to specific questions.

We offered $30 compensation to the clinicians as a token of appreciation, but each clinician declined the compensation and indicated they would prefer it spent on further research and development.

Recruitment

The five hepatologists were recruited through the medical school mailing lists. All the clinicians had more than 10 years of experience of seeing patients with cirrhosis and hepatic encephalopathy.

Data Analysis

The focus group discussion's audio was recorded and later transcribed by the researchers along with the questionnaire responses. The transcript was used to collect and organize the responses and reactions to the specific themes outlined in the three key areas of inquiry mentioned above. For any result based on the questionnaire, we report how many of the five participants were in agreement with the particular response.

Results

The goal of the focus group was to better understand current practices surrounding diagnosing HE and exploring possible avenues of integrating *Beacon* in the screening and chronic care process. We

organize our findings from the focus group by the three key areas of inquiry outlined above. Quotes by specific clinicians are referred by C*X*.

Current Screening Practices & Diagnostic Workflow. All the participants reported that, as specialists (i.e., hepatologists), most of the patients with HE that they treat have been pre-diagnosed with potential HE and referred to them by either the primary care physician or a gastroenterologist. Each clinician reported using clinical evaluation (which can only diagnose grade 2 and above HE (Fig. 2.1)) as the primary method to screen patients for HE.

Clinicians agreed that many cases of MHE are missed or go undiagnosed until the condition has progressed to grade 1 or 2, where it is easily clinically diagnosable. While discussing MHE cases that go undiagnosed, clinicians mentioned time constraints during clinical visit as the primary reason for not using screening tests for MHE. However, clinicians also reported they would consider using a screening test for MHE if it were easy and quick enough to be administered by a medical assistant before a patient sees them, as highlighted by C4's comment *"if there was a tool that was accurate like the Stroop test that was available in the clinic which could be administered by a medical associate when they check their BP, heart rate, weight, test for encephalopathy.*"

Describing topics covered during a routine clinical visit for the patient with HE, the clinicians reported conducting a clinical evaluation, going over their lactulose (medication) use, compliance with medication, and checking if they have any financial issues with the medication.

Current Standard of Chronic Care. Clinicians reported that the first advice they give patients diagnosed with HE was to stop driving or operating heavy equipment. C4 explained *"it is pretty hard for [patients] to assess themselves that they have enough impairment that [them driving with HE] is probably worse than walking outside the bar and thinking it is safe to drive"*. As a part of their treatment of the chronic condition, all clinicians reported prescribing patients Lactulose to restrict further deterioration and control HE. Lactulose is a colonic acidifier that works by decreasing the amount of ammonia in the blood. Although this does not cure HE, it does help improve the mental status of the patient.

Discussing commonly prescribed at-home chronic-care, clinicians ask patients to self-titrate lactulose to achieve a goal of about 2-3 *smooth bowel movements per day*.

Explaining how they track whether a patient's MHE is worsening or stable, C5 reported using some measure of cognitive ability like the number connection test (Fig. 2.2): *"sometimes I give a stack of number connection test and then use that to track patients who are subclinical [MHE], or if people like to play tablet or phone games which require some cognitive activity like I have a patient who plays Sudoku or something and they notice that their score is going down then that is a sign."* Clinicians also reported asking caregivers, family or professional, to update them in case of any noticeable change in personality or behavior.

Barriers to using CFF as screening tool. None of the five clinicians currently use CFF measurement as a part of their screening process. However, everyone expressed their desire to use *Beacon*, and thereby CFF, *if* it were proven to be useful (i.e., clinically capable of screening for MHE) and could be administered quickly. C1 explained *"there are lots of people who don't even have encephalopathy clinically [diagnosed] yet, so we could use it [Beacon]."* When asked about potential uses of *Beacon* beyond a clinician's screening of patients with MHE, C2 responded that *Beacon* could be a useful tool that provides some objectivity in what usually is a subjective evaluation: *"We spend a lot of time with family members who are trying to (convince) the patient that they shouldn't be driving and the patient doesn't believe it. But, if you have something objective it can be used to help them realize the situation."*

Although clinicians were excited about the potential to use *Beacon* in the clinic, they were more skeptical of at-home use by patients. Discussing potential use of *Beacon* at home by the patients, the clinicians raised liability concerns. C1 said *"What if we get the data and they got in a car accident? Do we have medical or legal liability since we had data suggesting that they had abnormal flicker [CFF]. I don't want that data."* Clinicians also expressed concerns with data overload burdening their already strained workload, such as patients coming to their clinical visit with another piece of information that the clinician needs to interpret. When discussing about the value of *Beacon* for caregivers looking after patients with HE, C2 said *"I have sense that with the population at hand their families will pretty much appreciate the objective data so that they can save the patient / loved ones [by preventing them from driving]."*

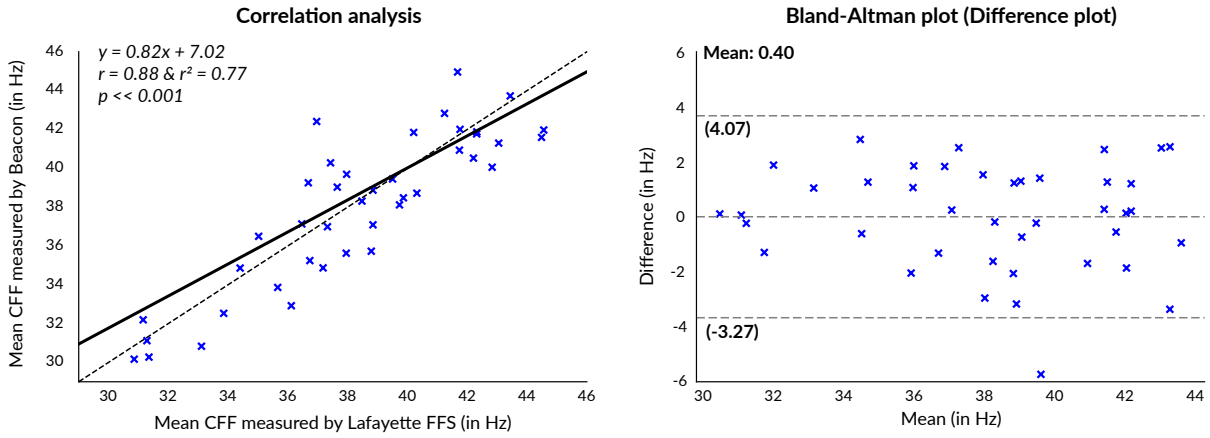


Figure 5.10: Results from the comparative study with 41 participants. **(left)** Regression analysis shows a strong correlation between the CFF measure by *Beacon* and Lafaytte FFS with a Pearson's R of 0.88. **(right)** The Bland-Altman plot shows the mean difference between *Beacon* and Lafayette FFS to be 0.4Hz with a maximum difference of at most $\pm$3.67Hz for 95% of the measurements.

5.4 Discussion

With *Beacon*, we seek to provide an accessible clinical decision support tool capable of measuring CFF. As the first step towards that goal, we conducted multiple studies to examine its feasibility for measuring CFF by comparing against an existing medical device. Although this is not the same as proving its diagnostic capabilities, it is a necessary precursor before testing the device with patients. We then collected impressions from experts to understand barriers to *Beacon*'s acceptance and provide feedback for future iterations. Our results indicate that *Beacon* has potential for use in screening and managing MHE, and that several barriers must still be overcome before it can be used as a screening tool in the clinic or at home. In addition to screening, development of a mobile, self-administered tool such as *Beacon* also enables new possibilities, such as observing CFF measurement trends over time to raise (i.e., observing a drop in CFF measurements over time) or reducing patient anxiety and concern (i.e., by observing stable or increasing CFF measurements over time). I finally conclude the discussion by providing an update on the status of the redesign of the device and the ongoing evaluation of Beacon among patient population.

5.4.1 Challenges & Opportunities in Designing Beacon

We designed *Beacon* to take a clinical measure which has been studied in research contexts and transition it in to clinical practice. Our comparative study demonstrated a strong correlation between the CFF measurements taken by *Beacon* and the gold standard Lafayette FFS. The reliability of *Beacon* makes new screening and monitoring possibilities available, such as providing patients with a device to measure CFF at home.

Through our studies, we also began refining the design in ways that will better support these opportunities. In our testing with participants at the retirement community, we discovered a design shortcoming in the *BeaconApp* wherein a person with limited dexterity or hand tremors might have unintended inputs due to accidentally pressing the screen in rapid succession. This resulted in two participants (P31 and P38) who had three and four measures respectively which were close to the range end-points of 25Hz and 55Hz (out of 16 measurements). We addressed this issue by introducing a 2-second delay between presses during which the screen remains inactive. This prevented any further misreports by other participants. Although adjustments like above can be easily implemented in a study, for a device like *Beacon* to be useful in practice it should be robust to adapt to a range of conditions. Although we observed that ambient light intensity has low impact on the measured CFF, one way to increase robustness for in-the-wild use would be to install components which automatically adjust the light source intensity by measuring the ambient light intensity (e.g., with a photocell).

During the comparative study, although the mean CFF reported by *Beacon* was close to that of the Lafayette FFS, we noticed slightly higher variance in the *Beacon* measures. The adaptive algorithm successfully removed the outliers and achieved the target of a $<$3Hz maximum standard deviation, affecting measures for 12 participants compared to just 1 for Lafayette FFS. We hypothesize that removal of the viewing chamber may have introduced this additional variance. We might explore increasing light source intensity and/or decreasing ambient light parameters to see if that reduces the variance.

Although greater measurement variability can be interpreted as an evidence of an unreliable measure, we believe this can be offset by the richness of data which could be collected by patients from their home. Daily or weekly CFF measures collected by a patient could provide an overview

of their HE that can be useful to bring in during a clinical visit, instead of being limited to a single data point collected on the day of the visit. Our work on *Beacon* suggests the design question of how to meaningfully present the data collected by patient in a succinct and actionable manner.

The focus group provided valuable insights regarding the challenges and opportunities that might be expected in clinical adoption of *Beacon*. Clinicians emphasized that for *Beacon* to help them screen MHE in a clinical context, it would need to be accurate as well as easy and quick to administer. We also unpacked some of the potential opportunities and challenges associated with using *Beacon* as an at-home self-tracking tool for HE patients. Although *Beacon* may provide a crucial objective measure where none exists now, there remain important ethical and legal questions surrounding the collected data and its implications for medical practice that will need to continue to be examined. Because our participant pool was recruited from a single location and does not represent the diverse practices or needs of clinics around the country and the world, future work to move *Beacon* or a similar device into clinical use should engage with clinicians, staff, administrators, and patients in a variety of contexts.

5.4.2 Potential Limitations of a Single-Point Threshold Based Assessment

Medical studies have put forward the notion that a CFF of 39Hz is the threshold for MHE (i.e., if a person has CFF below 39Hz they likely have MHE) [100]. Based on our work, we highlight problems with the notion of a single-point threshold based assessment that does not take other factors into consideration.

As we discovered during our focus group, medical professionals are overburdened and do not have the bandwidth to perform long screening tests during a clinical visit. Even if the CFF measurements are conducted in clinics, medical professionals will not take more than the minimum number of observations to get a CFF measurement. If, during those observations, the patient is distracted or misunderstands the process, their CFF measurement will be flawed. To base their treatment on such a flawed measure would be erroneous.

Contrary to the notion that a healthy individual's CFF score should be $>$39Hz [100], the comparative results showed otherwise [124]. CFF scores between 30Hz and 35Hz were common among our participants at the retirement community, across both the devices. Part of this can

be attributed to age-related degradation of sight reported by some participants (e.g., macular degeneration and astigmatism). Although the majority of the at-risk population might not be in that age group (65+ years old), it is crucial to understand that such thresholds or single-point assessments should be used with caution as like many other medical conditions, hepatic encephalopathy does not occur in a vacuum. Similarly, there can be other age or health related factors which can significantly reduce an otherwise healthy individual's CFF score.

We believe that by reducing the CFF measurement to a single-point assessment we are losing valuable insights like daily or weekly changes in condition, which can lead to better, more targeted health care interventions. *Beacon* offers a new potential approach to this limitation.

5.4.3 Monitoring CFF Trends as an Alternate to Single-Point Assessment

Prior to the development of *Beacon*, it was not practically feasible to collect CFF over time, and as such a single-point assessment was the only approach considered. To address the issues outlined in the above section, we propose reframing CFF from a clinical screening tool to a self-tracking measure collected at home.

As a preliminary investigation of the variability in CFF measurement over time, one member of the research team self-administered CFF measure using *Beacon* twice daily for one month: a morning measurement within 30 minutes of waking up and an evening measurement less than an hour before heading to bed. The measures were found to be relatively stable as seen in Figure 5.11. To measure the sensitivity of *Beacon*, after the first two weeks, the morning measurement was switched to be later (i.e., after breakfast). *Beacon* was able to detect that small shift in wakefulness as seen after the 14-day mark in Figure 5.11.

We envision a person with cirrhosis collecting their daily CFF measurements in relatively similar conditions (e.g., at a consistent point in their daily routine). Instead of comparing their data to a threshold value, we envision examining variation in a person's measurements over time. If a person's average CFF over a week drops by 1.5Hz or 2Hz, then they might reach out to their clinician or increase their dosage of lactulose. This shows an alternative possibility of thinking about the CFF measure. The frequency of CFF measurement in such a scenario might be agreed upon by the clinician and the patient. This practice can motivate further research in understanding

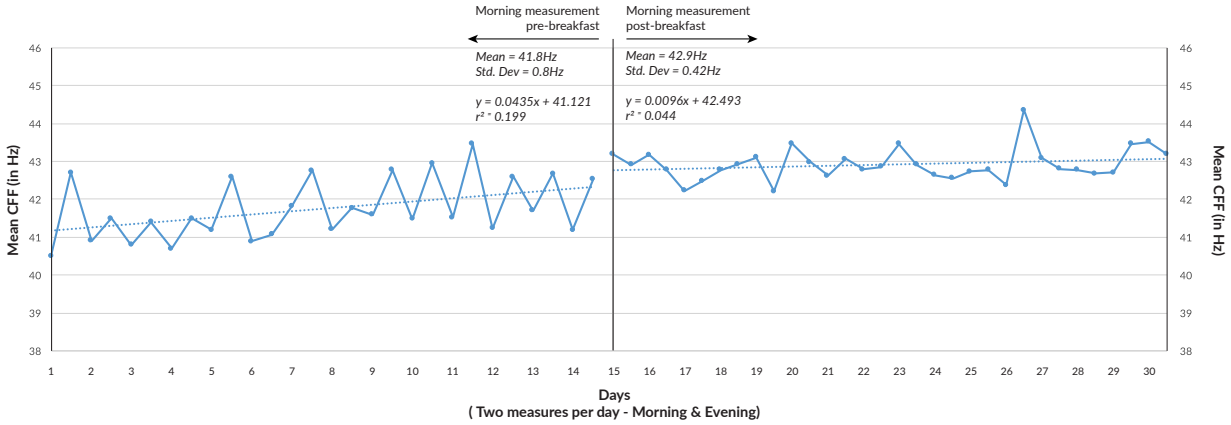


Figure 5.11: Month-long CFF measurement data collected by member of the research team using *Beacon*. Measures were taken twice daily—morning and evening. After 2 weeks, the morning measurement time was pushed back to see if the change in wakefulness would be picked up by *Beacon*. *Beacon* successfully picked up the slight variation in wakefulness, providing evidence of potential for long term CFF tracking.

the stability of CFF measurement for screening for and diagnosis of HE, and eventually treatment recommendations. Future work, enabled by a device like *Beacon*, can also explore what would be an ideal frequency of CFF measurements for people with different stages of HE, thus providing guidance to clinicians for tailoring treatment plans to a patient.

This new form of data and its representation raises interesting opportunities. It also raises important questions for patient-provider collaboration around such data in a patient's chronic self-management and about the ethics and legal implications of such data. Although we found support from the clinicians regarding using CFF in clinics and encouraging patients and caregivers to use CFF at home, there were still concerns regarding liability. This motivates further research investigating the longitudinal data collection of CFF measurements by patients, how to best communicate and present the measurements, and how to best enable collaboration around the data with clinicians.

Although our current focus with *Beacon* is to enable timely diagnosis of HE, CFF also has potential applications in other medical conditions like muliple sclerosis and Alzheimer's disease [146]. By making CFF measurement accessible, *Beacon* has the potential for impact across

a range of medical conditions.

5.4.4 Update on Evaluation Within Patient Population

The prior evaluation (section 5.3) demonstrated that Beacon performs on par with the Lafayette device in measuring CFF among healthy participants. As the next step, we moved towards establishing the performance among the cirrhotic population, which would be the eventual target demographic. Below I provide a brief update of the relevant update to the design of the device and the current status of the study.

Redesigning the Device

Although the design so far was sufficient for lab-based studies, we needed a design which would be robust enough for people to take home and use it over a few months. Additionally, given various stigmas associated with medical devices, we wanted to make it easier for people to use our device by making it look less like a medical device and more like a discreet household object.

I worked with an MHCI+D student Yue (Will) Wang to rethink the form factor of the device. Through several iterations, we decided to finalize a lamp-shaped prototype to be the design of Beacon moving forward. This design helped the device *blend into* a person's home instead of standing out as an obvious medical device. I then mentored a Mechanical Engineering captsone team (Drew Burack, Molly Foley, Neil Perrin, and Humza Talat) to build the physical device in a manner which is robust for long-term use and also was relatively easy for us to produce and assemble at a larger scale. Figure 5.12 depicts various iterations of the device including the current version on the right end of the image (also shown in Figure 5.13). Richard Li, a PhD student, is currently leading the effort to make final changes to the design in preparation for the upcoming at-home deployment.

Study Design

Medical literature examining the testing of HE often uses an array of tests including PHES, Stroop test, and CFF [147]. Given the lack of any true ground truth measure of HE classification, we decided to include these additional measures in our protocol to ensure we can leverage and build

Figure 5.12: Continuing evolution of the design of Beacon.

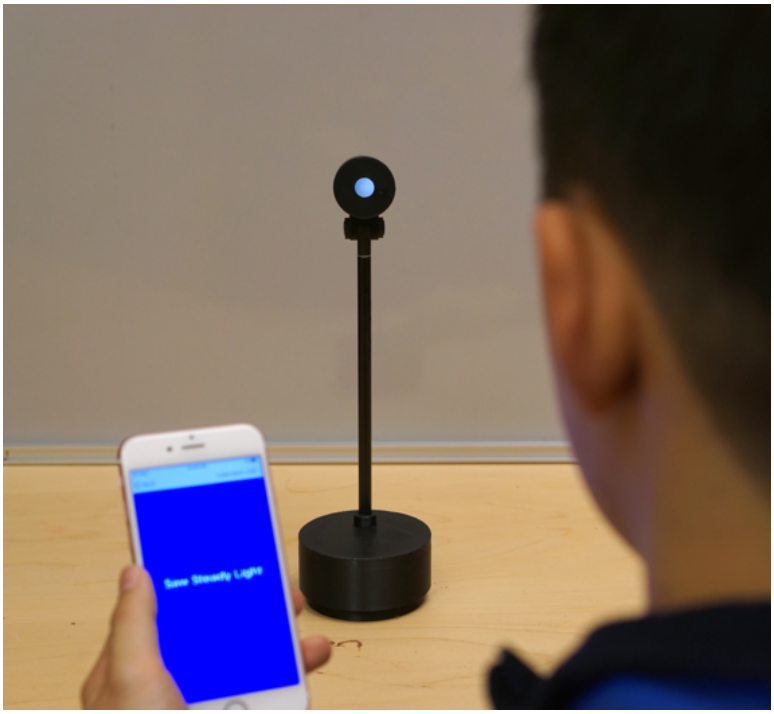


Figure 5.13: Current design of Beacon device.

on top of the results from prior literature. The study consists of participants with varying stages of cirrhosis being administered: (1) a CFF measure using Beacon, (2) a CFF measure using Lafayette

FFS, (3) Stroop test using EncephalApp [6], and (4) Number Connection Test (NCT) from the PHES test (2.2). The ordering is randomized and at the end of the stuyd participants fill out a 7-point ranking survey which states *[(test name) is easy to use]* on a scale of *strongly disagree* to *strongly agree* for Beacon, Lafayette, and EncephalApp. In addition, I also interview a subset of the participants to gather their perspectives on their current standard of care and the potential of Beacon assisting them in monitoring their HE.

Study Set Back

The study experienced an 8 month set back during its first iteration which ultimately lead to complete restart of the currently ongoing study in Autumn 2019. Given we had an updated design of the device we needed to ensure its performance had not been affected. I ran a pilot with 14 healthy participants and measured their CFF on both Beacon and Lafayette device. The numbers showed pretty strong alignment: average CFF of 14 participants using Lafayette was 39.507 Hz and using Beacon was 39.114 Hz. We were satisfied with this performance and decided to move to the testing in UW Medical Center.

For this study, I worked with Kara Walters to recruit and test participants when they were in the hospital for their check up. I managed to run six participants over a period of two months. While checking the data, I noticed that the patients were reporting a consistently lower CFF value on Beacon as compared to Lafayette. At this point, the study was halted and I retested the device on few of the team members and Beacon was indeed reporting a CFF off by 2-3 Hz consistently.

Given we were essentially restarting the study, we decided to recalibrate the Beacon device, but this time instead of calibrating for a dimly lit room, we calibrated it for a well-lit office room. This decision was taken with the longitudinal study in mind, since we will have little control over the quality of lighting in a participant's home, it would be easier to have them take the measures in a well-lit room than a poorly calibrated dim room.

Study Status

Our aim is to recruit 75 patients with cirrhosis at varying stages of HE and conduct interview with a subset of 12-15 of them. All the participants so far have been from the UW Medical Center, with

a plan to move the study to the Harborview center if and when we start experiencing a slow-down at this site.

We were able to run 37 participants through the study prior to the research closure owing to the ongoing pandemic in mid-March. We are awaiting instructions on when we can resume in-person research and are planning to pilot the at-home testing in a fully remote study in Autumn 2020.

5.5 Summary

My research aims to support people with chronic conditions get a personalized understanding of the unique nature of their condition. With Beacon, I am able to introduce a potential new measure of cognitive function for cirrhotic patients by making an existing measure, CFF, accessible and self-administrable. As I noted in the introduction to this chapter, Beacon aims to leverage population-level evidence that CFF threshold decreases as hepatic enceophalopathy worsens to provide a personalized understanding of the individual's current condition. Furthermore, Beacon enables a new data stream for future medical researchers and clinicians to better understand and characterize individual HE variations among the patients. At the same time, the data stream also raises new questions about the current models of care. For example, doctors do not have the bandwidth to continuously monitor data that patients are collecting at-home. Part of the concern is also that of liability; does the app provide a patient who tests low with guidance to avoid driving or to skip work and go to the hospital?

My experience with Beacon has also influenced how I think about the impact of personal health technology research which I discuss in section 7.3. It was also instrumental in securing an NIH R21 (#DK11743) to work towards initial clinical validation of the device among patient population. In addition to being useful for patients with cirrhosis, the CFF measure has applications in other conditions such as Alzheimer's, multiple sclerosis, and dementia [46, 149]. Over the years, I have received requests from researchers and medical professionals seeking Beacon to test on non-cirrhotic patient populations.

Chapter 6

OTHER RELEVANT PROJECTS

Although my dissertation's demonstration of my thesis is anchored in the previous chapters, I have also been involved in related projects which have helped shaped important aspects of my thinking. These projects are not part of my dissertation as they were led by other students and are not required for the core of my thesis statement. At the same time, they did inform my thinking and because the next chapter distills the generalizable design findings which were also influenced by these, I briefly summarize the three related projects below.

6.1 Categorizing the Range of Questions People Have about Their Health

This research was done in collaboration with Jessica Schroeder, James Fogarty, Julie A. Kientz, Sean A. Munson, and Matthew Kay and was published in JHIR in 2018 [154]. In this research we sought to understand the range of questions people want to answer using their health data.

To design effective tracking tools, designers need to be mindful of the range of questions people want to answer using such technologies. Using prior IBS interview data (from 25 patients and 10 providers) and a survey (78 respondents), we distilled the questions into nine templates based on the hypothesized relationship between dependent and independent variables. The nine questions were developed to cover a wide range of effects to be relevant for a wide variety of dependent variables (DV) and independent variables (IV). Q1 and Q2 are typical hypothesis testing types of questions. Q1 focuses on *any effect* between the IV and DV, and Q2 narrows that down to a *noticeable effect*. Q3 (*interaction effect*) focuses on scenarios in which there are *multiple IVs* affecting DV. Q4 focuses on a *temporal effect*. Q5 focuses on a *threshold for an effect* and Q6 focuses on a *varying effect*; both could be considered a precursor to a cost-benefit analysis, by helping a self-experimenter trade off how much of something they want against the symptoms they should expect. Q7-Q9 focus on *predictive variations* of the previous questions accounting for different relationships between the IV and DV (prediction under *avoidance* of the IV, *normal* exposure to the

IV, or *excess* exposure to the IV). The nine template questions we formulated (see Table 6.1) can serve as a starting point for future tool designers in explicitly deciding what kind of questions they can support or in expanding support for multiple question types.

Question	**Shorthand**
1. Does [independent variable] have any effect on my [independent variable]?	any effect
2. Does [independent variable] have a noticeable impact on my [independent variable]?	noticeable effect
3. Do different things in combination with [independent variable] affect the change in [independent variable]?	interaction effect
4. How does [independent variable] affect my [independent variable] differently depending on the time of day?	temporal effect
5. How much [independent variable] is needed to see an impact on my [independent variable]?	threshold for effect
6. By how much does my [independent variable] change with different amounts of [independent variable]?	varying effect
7. What will my [independent variable] be like in the future if I avoid [independent variable]?	avoidance prediction
8. What will my [independent variable] be like in the future after my normal amount of [independent variable]?	normal prediction
9. What will my [independent variable] be like in the future after more than my normal amount of [independent variable]?	excess prediction

Table 6.1: Types of health data related questions and their shorthand

In addition to providing the question templates, we also demonstrated that Bayesian methods can better answer the questions that people have about their self-experimentation data and can do so in a way that we believe is easier to understand than p-values and confidence intervals. Our findings aligned with recommendations from the Agency for Healthcare Research and Quality on analyzing n-of-1 data and endorse adoption of Bayesian analysis to better support a variety of question templates. Bayesian analysis provides a much needed alternative to frequentist statistics to better support flexible design and interpretation of self-experiments.

In Section 7.2, I discuss how this research has influenced my views on how to design tools that *empower the individual* though *understanding what questions* they want answered and *maintain data quality* through *robust* analytical methods.

6.2 Designing A Tool to Support People with Migraines

This research was led by Jessica Schroeder in collaboration with myself, Natalia Murinova, James Fogarty, and Sean A. Munson [151]. This research explored the design of tools to better support migraine patients and their providers better understand and manage the condition.

Migraine is characterized by unpredictable, intermittent, and poorly understood symptoms. Similar to IBS, providers often recommend their patients with migraine self-track to better understand and manage their migraines, but both again struggle to find value in the resulting data [154]. In this research, we investigated how to better support individualized migraine management. We first investigated challenges and pitfalls people currently face, characterizing distinct types of migraine tracking goals people would like to pursue [152]. Unpacking an overall management goal of reducing symptoms, we found four distinct categories of tracking goals that people often bring to migraine: (1) learning about their migraines; (2) predicting and preventing migraines; (3) monitoring migraines over time; and (4) fostering motivation and social recognition. Each goal category has different needs for data, analyses, and visualizations to support migraine-related tracking. We then developed and investigated goal-directed self-tracking, a new method that scaffolds the process of deciding what, when, and how to track toward a specific goal, and analyzes and visualizes the resulting data to support that goal [155]. The findings inform how tools can better support: (1) support people and their health providers in developing actionable goals; (2) ensure people are tracking exactly and only what they need to be tracking and when to achieve those goals; and (3) support people and their health providers in appropriately interpreting their data given their goals.

In section 7.2, I discuss how this research has influenced my views on the various aspects of how tools can be designed to *empower the individual*.

6.3 The Importance of Starting with Goals in N-of-1 Studies

This research is a culmination of various research projects including research presented in Chapters 3 & 4, Sections 6.1 & 6.2, and other projects [36, 37, 142, 152] through the collaborative efforts of Sean A. Munson, Jessica Schroeder, myself, Julie A. Kientz, Chia-Fang Chung, and James Fogarty and was published in Frontiers in Digital Health [133]. It presents a framing for the notion of

goal-directed tracking and its implication on the design of future tools.

N-of-1 tools offer the potential to support people in monitoring health and identifying individualized health management strategies. In this research, we argue that elicitation of individualized goals and customization of tracking to support those goals are a critical yet under-studied and under-supported aspect of self-tracking. We review examples of self-tracking from across a range of chronic conditions and self-tracking designs (e.g., self-monitoring, correlation analyses, self-experimentation). Together, these examples show how failure to elicit goals can lead to ineffective tracking routines, breakdowns in collaboration (e.g., between patients and providers, among families), increased burdens, and even designs that encourage behaviors counter to a person's goals. We found three ways in which tools contribute to these failures: (1) tools operationalize a broad goal in ways that are inconsistent with an individual's operationalization of that goal; (2) tools assume that a tracking goal implies other long-term goals; (3) data-first views that fail to scaffold use of that data to support individualized goals. We discuss potential techniques for eliciting and refining goals, scaffolding an appropriate tracking routine based on those goals, and presenting results in ways that advance individual goals while preserving individual agency. We then describe open challenges, including how to reconcile competing goals and support evolution of goals over time.

In section 7.2, I discuss how this research has influenced my views on how to design tools that *empower the individual* through better *understanding their questions and goals*.

Chapter 7

DISCUSSION

7.1 Review of Thesis Contributions

My dissertation focuses on human-centered approaches to building tools that support collecting, interacting with, and using health data toward improving human well-being through translating population-level evidence to personalized understanding.

I first described the framework for self-experimentation that I developed to motivate design of better tools aimed at assisting discovery of individual actionable variations among people with chronic conditions. I observed an unmet need among people with chronic conditions wherein they need support in answering questions about the individualized nature of their health. The framework provides guidelines on how to design tools that leverage existing population-level medical evidence to collect targeted personal health data and provide actionable personalized understanding. I also provided explicit guidelines on when the framework should be used by creating a list of required and desired criteria.

I then applied the framework to create TummyTrials, which is among the first apps aimed at supporting people with IBS conduct scientifically valid self-experiments to determine their individual food triggers. Within IBS, patients face challenges in finding effective tools which would help them determine individualized triggers for their symptoms. TummyTrials leveraged existing population-level evidence that food is the most common trigger for IBS and that majority people experience symptoms within a short window of consuming food. Combining those insights with a design that reduces collection burden and potential confounding factors, TummyTrials offered a guided self-experimentation process to enable people determine their individualized food triggers. Participants who used TummyTrials found that it scaffolded their journey by providing structure and support in a process where prior they had no alternatives or guidance. With TummyTrials, I was able to address the need where people with IBS were already collecting data and attempting to analyze it, but were unable to find value in the exercise owing to the lack of effective support from tools.

Next, I built Beacon, a platform aimed at supporting people with cirrhosis to better understand the impact of their liver failure on their cognitive abilities. People with cirrhosis are at high-risk of developing hepatic encephalopathy but, have no way to measure that. Beacon leveraged population-level evidence that CFF threshold decreases as a person's hepatic encephalopathy worsens and that a threshold of $<$39Hz is indicative of low-grade encephalopathy which is straightforward to treat. Beacon enabled self-measurement of hepatic encephalopathy through collection of CFF threshold data and laid the groundwork for future research exploring the individual interpretation of CFF variation. I have so far demonstrated the device's performance among *healthy participants* is on par with existing Lafayette research device. I am currently in the process of designing and deploying the at-home part of the study wherein patients will take the device home and provide longitudinal data to better understand and characterize individual variations among patients. This project began with a phenomenon which is well-understood in research but was difficult for an individual to track at-home and I was able to provide a platform capable of bridging that gap.

Across these projects, the common thread is to leverage existing *population-level evidence* to *collect and interpret* relevant data to provide *personal understanding*. I highlight these specific aspects across the three projects in the relevant chapters. The same human-centered research approach applied to vastly different contexts led to two distinct tools which help the people with chronic conditions better understand their individual variations. Together, this research demonstrates my thesis:

> *Novel personal health technologies can support collection and interpretation of personal data streams as part of translating available population-level evidence into personal understanding.*

7.2 Principles for Designing Personal Health Technologies

My experiences building personal health technologies that I presented in this dissertation have enabled me to formulate some general design recommendations for future tool designers. In explaining the principles, I use the word *designer* to mean researchers, designer, or someone in-charge of designing the tools. I now present these in the form of three design principles to

guide better design of personal health technologies.

7.2.1 Empower the Individual

A core aim of building personal health technologies is to support an individual in better understanding and managing their own health and well-being. To this end, the first principle of designing *better* tools is to provide value towards empowering an individual. I discuss four distinct approaches through which tools can accomplish this.

Provide Agency

Another approach towards empowering individuals is to promote agency. Instead of forcing a person down a pre-determined path of action, encourage the person choose what they would like to accomplish and provide necessary scaffolding. This approach to presenting *pathways* was discussed in section 4.5.3 and is central to the Migraine research I summarized in section 6.2. Instead of forcing a person to just track their migraines as most current apps do, the proposed design presents three pathways: (1) learn more about your migraines or related factors, (2) predict whether you are at risk for a migraine, and (3) monitor your migraine and related factors. A designer should make no assumption about the individual but it should simply present the alternatives along with relevant data to support informed decision making.

Scaffold Expertise

When building a tool to support a specific medical condition, it is up to the designer of the tool to scaffold the necessary domain expertise to ensure the end-user is successful in their pursuit. In the case of TummyTrials, the design of the app needed to scaffold the necessary medical, experimental design, and user experience expertise to ensure the individual will be able to successfully gather the self-experiment data (see section 4.2 for more details). TummyTrials was successful only through multi-disciplinary expertise that informed us of the temporal relationship between the trigger and symptoms; that it was a reasonable assumption to treat each day as an independent variable in experimental design because of daily bowel movement which *resets* the gut; and that we need daily reminders and other design elements to ensure maximum adherence. This was only

possible by having all the necessary experts as part of the team from the inception. The design meetings consisted of the of the team gastroenterologist, behavioral psychologist, human-centered design, and computer science experts providing their input on each aspect of the platform. Similar multi-disciplinary team setups have also contributed to the successful continuation of the Beacon and Migraine research.

Support Relevant Learning

Chronic conditions are complex and our understanding of them is still evolving. A designer should make no assumption of a person's understanding of their own condition and should instead provide venues for them to learn more if they desire. In TummyTrials, I accomplished this in two different ways: (1) providing in-context information buttons should a person need additional information about any task and (2) providing a centralized FAQ which combines many of these contextual pieces and provides additional information. The FAQ page (Figure 4.2) covered information about the food groups we supported experimenting with, definitions for the symptoms they could track, and a brief explanation of some elements of the self-experiment (e.g., how to choose a trigger to test, what is a p-value). Although not every participant engaged with it, those who did found value in it. As P12 said *"... when I read the frequently asked questions and this is the reason why I give you very high points for that. It's really one of the most comprehensive pages well, just in one page I'd say, that I've read a compendium of what I have researched on by myself, individually."*.

Understand the Individual's Questions and Goals

In order for a tool to effectively support an individual, the designer first needs to understand the motivation behind the individual's pursuit. This can in the form of a specific question they are seeking to answer or a specific goal they seek to achieve. One of the most common issue in self-tracking adherence is a mismatch between what the person desires and what the tool provides. For example, if a person seeks to find relationship between physical activity and sleep quality and the tool only supports tracking those two metric, but provides no analysis to support the person's task then it will likely lead to frustration and abandonment. There are a few different ways designers can reach shared a understating with the person: (1) At the most basic level, be

up front about what the tool is designed for and capable of. The research I discussed earlier about the types of health-related questions people want to answer can serve as a good starting point for this (section 6.1); (2) a level beyond that, the designer would be familiar with common goals people with a particular condition might have and present them explicitly as discussed in the *provide agency* paragraph above; and (3) finally, in an ideal scenario, the designer can create a tool capable of adapting to the current need and goal of an individual and then evolve as the individuals understanding of their own condition evolves.

7.2.2 Maintain Data Quality

Although empowering individuals is critical, they rely on these tools because they have a need for the data. Tool designs need to balance between how to ensure they fit the needs of an individual while still collecting data which provides them meaningful value. If the data is not reliable, then the tool cannot effectively support an individual in their journey of improving self-understanding. Below I highlight three approaches to improving and maintaining the necessary data quality to effectively support the individual.

Reduce Collection Burden

At the core of all the tools described in this dissertation is the practice of collecting only the *minimum viable data*. The personal informatics literature has provided us with plenty examples of people abandoning their tracking owing to burn-out from collecting too much data [59, 113]. Determining this *minimum viable data* takes the combined expertise of all the experts in the team. For example, during our initial conversation with the gastroenterologist about TummyTrials, they wanted the patients to fill out a 50 item questionnaire. Upon discussing further with them, we found out that this was essentially a replication of their current clinical practice in the digital form. Although not wrong, this was not what we intended for this tool. This is not an isolated incident, I have been involved in similar conversations where the medical expert wants to collect everything possible, instead of focusing on the key piece of data which supports our novel approach. A good way of thinking about what data the tool needs to collect is to think about the larger ecosystem within which the tool will be used. Instead of replacing every other source of information, the tool should be designed to compliment and enhance the existing infrastructure. It is the designer's duty

to cut through the perspectives of various experts and determine the *minimum viable data* needed to provide the necessary outcome to the individual.

Design for Robustness

Missing, corrupted, or sporadic data is a common concern in data collection which becomes significantly more challenging when dealing with *small data streams*. In addition to designing to improve adherence and reducing overall data collection burden, a common approach that I have adopted in my tools is to collect more data than what is needed. Although this may feel at odds with *reducing collection burden*, I view them as complementary to each other. By *reducing the collection burden* I emphasize figuring out the key data source and only collecting that, and by *robustness* I am emphasizing collection of addition data from that source to be more resilient to data concerns. In TummyTrials, this translated to collecting 12-days worth of data instead of the minimum 8-days and in Beacon, this translated to collecting 8 runs worth of data in the ongoing study instead of the minimum 4 runs. This additional data enables the statistical analysis methods to be more robust against certain forms of data corruptions such as outliers and inaccurate reports. Ultimately, it falls on the design team's expertise to determine where to draw the line in being robust while being low-burden enough for the task.

Understand and Support the Lived Experience

Any tool aimed at supporting individuals in their long-term health and well-being needs to be mindful of the contexts within which it might be used. This become particularly vital in case of self-experimentation, where any variation is a potential confounding factor. People may have different schedules on weekdays vs. weekends, in summer vs. winter, or when they are at home vs. travelling, or may have an unplanned emergency, all of which can impede reliable data collection. Understanding these differences exists and accounting for them through flexibility in the design of the tool can lead to not just a robust tool but better overall experience. Such flexibility can be introduced in different aspects of the tool such as: (1) in data collection where a trade-off can be made between collecting in-situ data vs. allowing post-hoc collection which might be less accurate; (2) in the experiment design by adopting a flexible design which permits pauses or skips in data

collection; and (3) in statistical analysis by using methods which can handle missing data points or can provide analysis during the experiment instead of only at the end [154].

7.2.3 Support Multiple Stakeholders

Designers need to be mindful of the different stakeholders who will be interacting with the tool or the data from the tool. For example, in IBS, both patients and providers currently struggle with an effective diagnosis technique. Given the lack of tools which meet their requirements, patients are asked to track detailed food journals without proper support mechanisms and providers who have no formal training in reviewing those journals are asked to draw treatment plans based on them. TummyTrials was designed with input from both patients and providers on how it could effectively support their individual challenges while keeping at eye on fostering effective collaboration.

If not managed properly personal health data can also contribute to information overload instead of providing value [35]. This was clearly highlighted in my focus group with the Hepatologists, wherein they were concerned about being able to monitor a constant stream of data and also about the inherent liability they have in such a situation (see section 5.3.3).

Additionally, the incentive structures need to be properly aligned to motivate the patients to collect data. If the incentive structure are not well-thought through or worse, misaligned, this can lead to unfavorable behaviors. For example, in Beacon, the treatment for early stage Hepatic Encephalopathy involves taking a medication for gut cleansing, which is a very unpleasant and possibly debilitating experience. If the protocol for using Beacon involved that every time a person tested low on their CFF, they had to take the medication, it could motivate a person to game the system to not get a low score. These concerns become more evident as these tools get closer to being integrated in healthcare ecosystem of the individual. Being cognizant of the overall ecosystem and how the data from the tool will impact the individual may allow designers create better incentive structures to minimize adversarial behaviors. The nature of these incentives may be perceived differently by different individuals and may even at times be at odds with the earlier approach of *providing agency*. There is no general solution to this conundrum and it will likely require expertise in human-centered design and medicine to ensure that all perspectives are well-understood and work together toward minimizing any incentive misalignment.

Overall, designers need to be aware of the different questions that arise when designing technologies at different point in the spectrum from supporting existing practices through a new tool (e.g., TummyTrials) to introducing a fundamentally new data stream for diagnosis (e.g., Beacon). Thinking through the various challenges and designing the tools to accommodate various stakeholders will enable better translation of the technology to eventual clinical use.

7.3 Future Directions

I am driven to empower and support individuals through building new technology and reducing barriers to collecting, understanding, and acting upon personal health data. Building on my track record of working as a part of multi-disciplinary teams, I plan to collaborate across engineering, design, and medicine to solve complex problems that researchers in this space will encounter in building the next generation of personal health technologies. My research agenda will continue to focus on using a human-centered approach to design, deploy, and evaluate novel health data experiences.

7.3.1 Next Generation of Personal Health Technologies

Building on my learnings from this dissertation research, I would like to continue my research on building personal health technologies. In my research with IBS and migraine populations, we identified the goals and needs of people using these technologies evolve over time. To address the ever increasing complexity of designing such dynamic technologies, I propose exploration of templates. These templates will be collections of relationships around (1) what questions people want to answer; (2) what data are necessary to answer those questions; (3) when and how to track such data; and (4) what analyses and visualizations are appropriate to answer a question with the data. Recognizing the changing goals of an individual, a platform can adapt to suit the person's current needs. For example, once caffeine is determined to be a likely trigger for abdominal pain, a next goal can be to check if there is a specific threshold of caffeine below which abdominal pain is manageable. Developing such templates necessitates a holistic approach requiring input from different stakeholders, balancing the trade-off between burden and value received, and evaluating for scientific rigor and practicality.

7.3.2 Interweaving Population and Personal Health Data

Much of modern medicine is based on large-scale RCTs, which provide population-level evidence regarding the effectiveness of an intervention. This top-down approach works well when the focus is on assisting a majority of a population, but RCTs have critical limitations in that they are not always feasible due to challenges of scale and associated economics, and individuals and providers who are unable to find the evidence they seek in existing trials are then left to fend for themselves. My work with n-of-1 experimental design aims to help fill this gap by adopting a bottom-up approach and focusing on the individual. Such studies provide outcomes that are actionable for an individual, and I believe they also can help build population-level knowledge, which can in turn inform future n-of-1 studies. I plan to bridge these two approaches to study design by collating n-of-1 data to form new population-level understanding and to use existing population-level understanding to guide better targeted n-of-1 interventions. Opportunities for research themes within this question can include better explainable machine learning models (owing to the rich data collected at an individual level) and actionable diagnostic experiments which can be run in a matter of days or weeks as opposed to months or years.

7.3.3 Translating Research to Practice

Many health innovations created in research do not make it into the hands of the individuals who need them, be that practitioners in clinics or individuals at home. In my experiences with early-stage commercialization attempts of TummyTrials and Beacon, I encountered systemic barriers that impede any effort to bridge these gaps. I aim to support other researchers and designers in bringing their tools to the market by identifying and scaffolding barriers pertaining to such translational efforts. As I continue to work on these translations and as I learn more through my interactions with various communities (e.g., HCI, global development, medicine, non-profits, investors), I plan to synthesize and disseminate recommendations for future researchers to follow and build upon in their own journeys.

7.4 Conclusion

Overall, in my dissertation I demonstrate how personal health technologies can leverage existing population-level medical evidence to provide personal understanding through collecting and interpreting targeted personal health data. Across this research, I have learned ways of building better tools in this space to help future designers and I shared those in the form of design principles above. Future work can leverage the insights of this dissertation to continue multi-disciplinary research on ways to better support people in their pursuit of individualized understanding of their health.

BIBLIOGRAPHY

[1] Actionable Bayesian Analysis for Evolving Health Goals.

[2] Amin Ahsan Ali, Syed Monowar Hossain, Karen Hovsepian, Md Mahbubur Rahman, Kurt Plarre, and Santosh Kumar. mpuff: automated detection of cigarette smoking puffs from respiration measurements. In *Proceedings of the 11th international conference on Information Processing in Sensor Networks*, pages 269–280. ACM, 2012.

[3] Javier Ampuero, Macarena Simón, Carmina Montoliú, Rodrigo Jover, Miguel Ángel Serra, Juan Córdoba, and Manuel Romero-Gómez. Minimal Hepatic Encephalopathy and Critical Flicker Frequency Are Associated with Survival of Patients with Cirrhosis. *Gastroenterology*, 149(6):1483–1489, 2015.

[4] Apple Health Kit.

[5] Amid Ayobi, Paul Marshall, Anna L Cox, and Yunan Chen. Quantifying The Body and Caring for The Mind: Self-Tracking in Multiple Sclerosis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 6889–6901. ACM, 2017.

[6] Jasmohan S Bajaj, Douglas M Heuman, Richard K Sterling, Arun J Sanyal, Muhammad Siddiqui, Scott Matherly, Velimir Luketic, R Todd Stravitz, Michael Fuchs, Leroy R Thacker, et al. Validation of encephalapp, smartphone-based stroop test, for the diagnosis of covert hepatic encephalopathy. *Clinical Gastroenterology and Hepatology*, 13(10):1828–1835, 2015.

[7] Jasmohan S Bajaj, Douglas M Heuman, James B Wade, Douglas P Gibson, Kia Saeian, Jacob A Wegelin, Muhammad Hafeezullah, Debulon E Bell, Richard K Sterling, R Todd Stravitz, Michael Fuchs, Velimir Luketic, and Arun J Sanyal. Rifaximin Improves Driving Simulator Performance in a Randomized Trial of Patients With Minimal Hepatic Encephalopathy. *Gastroenterology*, 140(2):478–487, 2011.

[8] Jasmohan Singh Bajaj, Ashkan Etemadian, Muhammad Hafeezullah, and Kia Saeian. Testing for Minimal Hepatic Encephalopathy in the United States: An AASLD Survey. *Hepatology*, 45(3):833–834, 2007.

[9] Jakob E Bardram, Mads Frost, Károly Szántó, Maria Faurholt-Jepsen, Maj Vinberg, and Lars Vedel Kessing. Designing mobile health technology for bipolar disorder: A field trial of the monarca system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2627–2636. ACM, 2013.

[10] David H Barlow and Steven C Hayes. Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of applied behavior analysis*, 12(2):199–210, 1979.

[11] David H Barlow and Michel Hersen. *Single case experimental designs: Strategies for studying behavior change*. Pergamon New York, NY, 1984.

[12] Colin Barr, Maria Marois, Ida Sim, Christopher H Schmid, Barth Wilsey, Deborah Ward, Naihua Duan, Ron D Hays, Joshua Selsky, Joseph Servadio, Marc Schwartz, Clyde Dsouza, Navjot Dhammi, Zachary Holt, Victor Baquero, Scott MacDonald, Anthony Jerant, Ron Sprinkle, and Richard L. Kravitz. The PREEMPT Study - Evaluating Smartphone-Assisted N-of-1 Trials in Patients with Chronic Pain: Study Protocol for a Randomized Controlled Trial. *Trials*, 16:67, 2015.

[13] Jared S Bauer, Sunny Consolvo, Benjamin Greenstein, Jonathan Schooler, Eric Wu, Nathaniel F Watson, and Julie Kientz. Shuteye: encouraging awareness of healthy sleep recommendations with a mobile, peripheral display. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1401–1410, 2012.

[14] Eric P.S. Baumer, Sherri Jean Katz, Jill E. Freeman, Phil Adams, Amy L. Gonzales, John Pollak, Daniela Retelny, Jeff Niederdeppe, Christine M. Olson, and Geri K. Gay. Prescriptive Persuasion and Open-Ended Social Awareness: Expanding the Design Space of Mobile Health. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 2012)*, pages 475–484, 2012.

[15] Eric PS Baumer, Sherri Jean Katz, Jill E Freeman, Phil Adams, Amy L Gonzales, John Pollak, Daniela Retelny, Jeff Niederdeppe, Christine M Olson, and Geri K Gay. Prescriptive persuasion and open-ended social awareness: expanding the design space of mobile health. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 475–484, 2012.

[16] Dror Ben-Zeev, Rachel Brian, Rui Wang, Weichen Wang, Andrew T Campbell, Min S H Aung, Michael Merrill, Vincent W S Tseng, Tanzeem Choudhury, Marta Hauser, John M Kane, and Emily A Scherer. Crosscheck: Integrating Self-report, Behavioral Sensing, and Smartphone Use to Identify Digital Indicators of Psychotic Relapse. *Psychiatric Rehabilitation Journal*, 40(3):266–275, 2017.

[17] Frank Bentley and Konrad Tollmar. The Power of Mobile Notifications to Increase Wellbeing Logging Behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2013)*, pages 1095–1098, New York, New York, USA, apr 2013.

[18] Hugh Beyer and Karen Holtzblatt. Contextual design. *interactions*, 6(1):32–42, 1999.

[19] Jessica R. Biesiekierski, Evan D. Newnham, Peter M. Irving, Jacqueline S. Barrett, Melissa Haines, James D. Doecke, Susan J. Shepherd, Jane G. Muir, and Peter R. Gibson. Gluten Causes Gastrointestinal Symptoms in Subjects without Celiac Disease: A Double-Blind Randomized Placebo-Controlled Trial. *The American Journal of Gastroenterology*, 106(3):508–514, 2011.

[20] Jessica R. Biesiekierski, Simone L. Peters, Evan D. Newnham, Ourania Rosella, Jane G. Muir, and Peter R. Gibson. No Effects of Gluten in Patients with Self-Reported Non-Celiac Gluten Sensitivity after Dietary Reduction of Fermentable, Poorly Absorbed, Short-Chain Carbohydrates. *Gastroenterology*, 145(2):320–328, 2013.

[21] Martin Blachier, Henri Leleu, Markus Peck-Radosavljevic, Dominique-Charles Valla, and Françoise Roudot-Thoraval. The Burden of Liver Disease in Europe: A Review of Available Epidemiological Data. *Journal of Hepatology*, 58(3):593–608, 2013.

[22] Jonathan B Bricker, Kristin E Mull, Julie A Kientz, Roger Vilardaga, Laina D Mercer, Katrina J Akioka, and Jaimee L Heffner. Randomized, controlled pilot trial of a smartphone app for smoking cessation using acceptance and commitment therapy. *Drug and alcohol dependence*, 143:87–94, 2014.

[23] John Brooke. SUS: A Quick and Dirty Usability Scale. *Usability Evaluation in Industry*, 189(194):4–7, 1996.

[24] Isis Bulté and Patrick Onghena. When the truth hits you between the eyes. *Methodology*, 2011.

[25] Christine Buttorff, Teague Ruder, and Melissa Bauman. *Multiple chronic conditions in the United States*. Rand Santa Monica, CA, 2017.

[26] Caroline Canavan, Joe West, and Timothy Card. The epidemiology of irritable bowel syndrome. *Clinical epidemiology*, 6:71, 2014.

[27] Summary Health Statistics: National Health Interview Survey 2015, 2015.

[28] M. Soledad Cepeda, Juan C. Acevedo, Hernando Alvarez, Nelcy Miranda, Catalina Cortes, and Daniel B. Carr. An N-of-1 trial as an aid to decision-making prior to implanting a permanent spinal cord stimulator. *Pain Medicine*, 9(2):235–239, 2008.

[29] Yu-Chen Chang, Jin-Ling Lo, Chao-Ju Huang, Nan-Yi Hsu, Hao-Hua Chu, Hsin-Yen Wang, Pei-Yu Chi, and Ya-Lin Hsieh. Playful toothbrush: ubicomp technology for teaching tooth brushing to kindergarten children. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 363–372, 2008.

[30] William D Chey, Monthira Maneerattaporn, and Richard Saad. Pharmacologic and complementary and alternative medicine therapies for irritable bowel syndrome. *Gut and liver*, 5(3):253, 2011.

[31] Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. Understanding Quantified-Selfers' Practices in Collecting and Exploring Personal Data. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2014)*, pages 1143–1152, New York, New York, USA, apr 2014.

[32] Chia-Fang Chung, Elena Agapie, Jessica Schroeder, Sonali Mishra, James Fogarty, and Sean A Munson. When personal tracking becomes social: Examining the use of instagram for healthy eating. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1674–1687, 2017.

[33] Chia-Fang Chung, Jonathan Cook, Elizabeth Bales, Jasmine Zia, and Sean A Munson. More than telemonitoring: Health provider use and nonuse of life-log data in irritable bowel syndrome and weight management. *Journal of Medical internet Research*, 17(8):E203, 2015.

[34] Chia-Fang Chung, Jonathan Cook, Elizabeth Bales, Jasmine K. Zia, and Sean A. Munson. More Than Telemonitoring: Health Provider Use and Nonuse of Life-Log Data in Irritable Bowel Syndrome and Weight Management. *Journal of Medical Internet Research*, 17(8):e203, 2015.

[35] Chia-Fang Chung, Kristin Dew, Allison Cole, Jasmine K Zia, James Fogarty, Julie A Kientz, and Sean A Munson. Boundary negotiating artifacts in personal informatics: Patient-provider collaboration with patient-generated data. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 2016.

[36] Chia-Fang Chung, Kristin Dew, Allison Cole, Jasmine K. Zia, James Fogarty, Julie A. Kientz, and Sean A. Munson. Boundary Negotiating Artifacts in Personal Informatics: Patient-Provider Collaboration with Patient-Generated Data. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 2016)*, pages 770—-786, 2016.

[37] Chia-Fang Chung, Qiaosi Wang, Jessica Schroeder, Allison Cole, Jasmine Zia, James Fogarty, and Sean A 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.

[38] C Clar, Katharine D Barnard, Ewen Cummins, Pamela Royle, and Norman Waugh. Self-monitoring of blood glucose in type 2 diabetes: Systematic review. *Health Technology Assessment*, 14(12):1–140, 2010.

[39] Sunny Consolvo, Predrag Klasnja, David W McDonald, and James A Landay. Designing for healthy lifestyles: Design considerations for mobile technologies to encourage consumer health and wellness. *Foundations and Trends in Human-Computer Interaction*, 6(3):167–315, 2014.

[40] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1797–1806, 2008.

[41] Sunny Consolvo, David W. McDonald, Tammy Toscos, Mike Y. Chen, Jon E. Froehlich, Beverly L. Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, Ian E. Smith, and James A. Landay. Activity Sensing in the Wild: A Field Trial of Ubifit Garden. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2008)*, pages 1797–1806, 2008.

[42] Thomas Coppetti, Andreas Brauchlin, Simon Müggler, Adrian Attinger-Toller, Christian Templin, Felix Schönrath, Jens Hellermann, Thomas F Lüscher, Patric Biaggi, and Christophe A Wyss. Accuracy of smartphone apps for heart rate measurement. *European journal of preventive cardiology*, 24(12):1287–1293, 2017.

[43] Felicia Cordeiro, Elizabeth Bales, Erin Cherry, and James Fogarty. Rethinking the Mobile Food Journal: Exploring Opportunities for Lightweight Photo-Based Capture. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2015)*, pages 3207–3216, 2015.

[44] Felicia Cordeiro, Elizabeth Bales, Erin Cherry, and James Fogarty. Rethinking the mobile food journal: Exploring opportunities for lightweight photo-based capture. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3207–3216, 2015.

[45] Felicia Cordeiro, Daniel A. Epstein, Edison Thomaz, Elizabeth Bales, Arvind K. Jagannathan, Gregory D. Abowd, and James Fogarty. Barriers and Negative Nudges: Exploring Challenges in Food Journaling. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2015)*, pages 1159–1162, 2015.

[46] Stephen Curran, Simon Wilson, Shabir Musa, and John Wattis. Critical flicker fusion threshold in patients with alzheimer's disease and vascular dementia. *International journal of geriatric psychiatry*, 19(6):575–581, 2004.

[47] Ralph D'Agostino and Egon S Pearson. Tests for departure from normality. empirical results for the distributions of b_2 and $\sqrt{b}_1$. *Biometrika*, 60(3):613–622, 1973.

[48] Ralph B d'Agostino. An omnibus test of normality for moderate and large size samples. *Biometrika*, 58(2):341–348, 1971.

[49] Jesse Dallery, Rachel N Cassidy, and Bethany R Raiff. Single-case experimental designs to evaluate novel technology-based health interventions. *Journal of medical Internet research*, 15(2):e22, 2013.

[50] Nediyana Daskalova, Danaë Metaxa-Kakavouli, Adrienne Tran, Nicole Nugent, Julie Boergers, John McGeary, and Jeff Huang. SleepCoacher: A Personalized Automated Self-Experimentation System for Sleep Recommendations. In *Proceedings of the Symposium on User Interface Software and Technology*, pages 347–358. ACM, 2016.

[51] Katy Dewitte, Colette Fierens, Dietmar Stöckl, and Linda M Thienpont. Application of the bland–altman plot for interpretation of method-comparison studies: a critical investigation of its practice. *Clinical chemistry*, 48(5):799–801, 2002.

[52] Steven Dow, T Scott Saponas, Yang Li, and James A Landay. External representations in ubiquitous computing design and the implications for design tools. In *Proceedings of the 6th conference on Designing Interactive systems*, pages 241–250, 2006.

[53] Dunne E. The unrealized potential of mhealth. *Journal of Mobile Technology in Medicine News*, 2014.

[54] Eugene Edgington and Patrick Onghena. *Randomization tests*. CRC Press, 2007.

[55] Eugene S. Edgington and Patrick Onghena. *Randomization Tests*. CRC Press, 2007.

[56] Auria Eisen-Enosh, Nairouz Farah, Zvia Burgansky-Eliash, Uri Polat, and Yossi Mandel. Evaluation of Critical Flicker-Fusion Frequency Measurement Methods for the Investigation of Visual Temporal Resolution. *Scientific Reports*, 7(1):2–10, 2017.

[57] Sigrid Elsenbruch. Abdominal Pain in Irritable Bowel Syndrome: A Review of Putative Psychological, Neural and Neuro-Immune Mechanisms. *Brain, Behavior, and Immunity*, 25(3):386–394, 2011.

[58] Daniel A Epstein, Felicia Cordeiro, James Fogarty, Gary Hsieh, and Sean A Munson. Crumbs: lightweight daily food challenges to promote engagement and mindfulness. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5632–5644, 2016.

[59] Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. A Lived Informatics Model of Personal Informatics. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2015)*, pages 731–742, 2015.

[60] Shanti Eswaran, Jan Tack, and William D. Chey. Food: The Forgotten Factor in the Irritable Bowel Syndrome. *Gastroenterology Clinics of North America*, 40(1):141–62, 2011.

[61] Aojg Farmer, Oj Gibson, L Tarassenko, and A Neil. A systematic review of telemedicine interventions to support blood glucose self-monitoring in diabetes. *Diabetic Medicine*, 22(10):1372–1378, 2005.

[62] Quentin Ferry, Julia Steinberg, Caleb Webber, David R FitzPatrick, Chris P Ponting, Andrew Zisserman, and Christoffer Nellåker. Diagnostically relevant facial gestalt information from ordinary photos. *Elife*, 3:e02020, 2014.

[63] Ronald Aylmer Fisher. Design of experiments. *Br Med J*, 1(3923):554–554, 1936.

[64] Fitbit.

[65] Carol Y. Francis, Julie Morris, and Peter J. Whorwell. The Irritable Bowel Severity Scoring System: A Simple Method of Monitoring Irritable Bowel Syndrome and its Progress. *Alimentary Pharmacology & Therapeutics*, 11(2):395–402, 1997.

[66] Robert H Friedman, Lewis E Kazis, Alan Jette, Mary Beth Smith, John Stollerman, Jeanne Torgerson, and Kathleen Carey. A telecommunications system for monitoring and counseling patients with hypertension: impact on medication adherence and blood pressure control. *American journal of hypertension*, 9(4):285–292, 1996.

[67] George A Gescheider. *Psychophysics: the fundamentals*. Psychology Press, 2013.

[68] Davide Giavarina. Understanding bland altman analysis. *Biochemia medica: Biochemia medica*, 25(2):141–151, 2015.

[69] Peter R. Gibson and Susan J. Shepherd. Food Choice as a Key Management Strategy for Functional Gastrointestinal Symptoms. *The American Journal of Gastroenterology*, 107(5):657–666, 2012.

[70] Google Fit.

[71] Erik Grönvall and Nervo Verdezoto. Understanding Challenges and Opportunities of Preventive Blood Pressure Self-Monitoring At Home. In *Proceedings of the European Conference on Cognitive Ergonomics*, page 1, 2013.

[72] Weixi Gu, Yuxun Zhou, Zimu Zhou, X I Liu, H A N Zou, P E I Zhang, Costas J Spanos, and L I N Zhang. SugarMate : Non-intrusive Blood Glucose Monitoring with Smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):54:1–27, 2017.

[73] Mieke Haesen, Jan Meskens, Kris Luyten, and Karin Coninx. Draw me a storyboard: incorporating principles & techniques of comics... *Proceedings of HCI 2010 24*, pages 133–142, 2010.

[74] Emma P. Halmos, Victoria A. Power, Susan J. Shepherd, Peter R. Gibson, and Jane G. Muir. A Diet Low in FODMAPs Reduces Symptoms of Irritable Bowel Syndrome. *Gastroenterology*, 146(1):67–75, 2014.

[75] Tian Hao, Guoliang Xing, and Gang Zhou. Isleep: Unobtrusive sleep quality monitoring using smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, SenSys '13, New York, NY, USA, 2013. Association for Computing Machinery.

[76] Lynsey R Harris and Lesley Roberts. Treatments for irritable bowel syndrome: patients' attitudes and acceptability. *BMC complementary and alternative medicine*, 8(1):65, 2008.

[77] Rex Hartson and Pardha S Pyla. *The UX Book: Process and guidelines for ensuring a quality user experience*. Elsevier, 2012.

[78] Steven C Hayes. Single case experimental design and empirical clinical practice. *Journal of consulting and clinical psychology*, 49(2):193, 1981.

[79] Margaret Heitkemper, Eric Carter, Vanessa Ameen, Kevin Olden, and Lin Cheng. Women with irritable bowel syndrome: differences in patients' and physicians' perceptions. *Gastroenterology nursing*, 25(5):192–200, 2002.

[80] Margaret M. Heitkemper, Monica E. Jarrett, Rona L. Levy, Kevin C. Cain, Robert L. Burr, Andrew Feld, Pamela Barney, and Pamela Weisman. Self-Management for Women with Irritable Bowel Syndrome. *Clinical Gastroenterology and Hepatology*, 2(7):585–96, 2004.

[81] Javier Hernandez, Daniel J McDuff, and Rosalind W Picard. Biophone: Physiology monitoring from peripheral smartphone motions. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7180–7183. IEEE, 2015.

[82] Mieke Heyvaert and Patrick Onghena. Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science*, 3(1):51–64, 2014.

[83] Jerry L Hintze and Ray D Nelson. Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998.

[84] Anne E. Jamieson, Paula C. Fletcher, and Margaret A. Schneider. Seeking Control Through the Determination of Diet: A Qualitative Investigation of Women with Irritable Bowel Syndrome and Inflammatory Bowel Disease. *Clinical Nurse Specialist*, 21(3):152–160, 2007.

[85] Susan L Janson, Kelly Wong McGrath, Jack K Covington, Su-Chun Cheng, and Homer A Boushey. Individualized Asthma Self-Management Improves Medication Adherence and Markers of Asthma Control. *Journal of Allergy and Clinical Immunology*, 123(4):840–846, 2009.

[86] Mark P Jensen, Wei Wang, Susan L Potts, and Errol M Gould. Reliability and validity of individual and composite recall pain measures in patients with cancer. *Pain Medicine*, 13(10):1284–1291, 2012.

[87] E Jonathan and Martin Leahy. Investigating a smartphone imaging unit for photoplethysmography. *Physiological measurement*, 31(11):N79, 2010.

[88] J. T. Jorgensen. New Era of Personalized Medicine: A 10-Year Anniversary. *The Oncologist*, 14(5):557–558, 2009.

[89] Sung Woo Kahng, Kyong-Mee Chung, Katharine Gutshall, Steven C Pitts, Joyce Kao, and Kelli Girolami. Consistent visual analyses of intrasubject data. *Journal of applied behavior analysis*, 43(1):35–45, 2010.

[90] Sunanda V Kane, Karen Sable, and Stephen B Hanauer. The menstrual cycle and its effect on inflammatory bowel disease and irritable bowel syndrome: a prevalence study. *The American journal of gastroenterology*, 93(10):1867–1872, 1998.

[91] Matthew R Kappus and Jasmohan S Bajaj. Assessment of Minimal He (With Emphasis on Computerized Psychometric Tests). *Clinical Liver Disease*, 16(1):43–55, 2012.

[92] Ravi Karkar, James Fogarty, Julie A Kientz, Sean A Munson, Roger Vilardaga, and Jasmine Zia. Opportunities and challenges for self-experimentation in self-tracking. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 991–996, 2015.

[93] Ravi Karkar, Rafal Kocielnik, Xiaoyi Zhang, James Fogarty, George N Ioannou, Sean A Munson, and Jasmine Zia. Toward a portable, self-administered critical flicker frequency test. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pages 1194–1199, 2016.

[94] Ravi Karkar, Rafal Kocielnik, Xiaoyi Zhang, James Fogarty, George N Ioannou, Sean A Munson, and Jasmine Zia. Beacon: Designing a Portable Device for Self-Administering a Measure of Critical Flicker Frequency. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3), 2018.

[95] Ravi Karkar, Jessica Schroeder, Daniel A. Epstein, Laura R. Pina, Jeffrey Scofield, James Fogarty, Julie A. Kientz, Sean A. Munson, Roger Vilardaga, and Jasmine Zia. TummyTrials: A Feasibility Study of Using Self-Experimentation to Detect Individualized Food Triggers. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2017)*, pages 6850–6863, 2017.

[96] Ravi Karkar, Jasmine K. Zia, Roger Vilardaga, Sonali R. Mishra, James Fogarty, Sean A. Munson, and Julie A. Kientz. A Framework for Self-Experimentation in Personalized Health. *Journal of the American Medical Informatics Association (JAMIA)*, 23(3):440–448, 2016.

[97] Matthew Kay, Eun Kyoung Choe, Jesse Shepherd, Benjamin Greenstein, Nathaniel F. Watson, Sunny Consolvo, and Julie A. Kientz. Lullaby: A Capture & Access System for Understanding the Sleep Environment. In *Proceedings of the ACM Conference on Ubiquitous Computing (UbiComp 2012)*, pages 226–234, 2012.

[98] Matthew Kay, Dan Morris, MC Schraefel, and Julie A. Kientz. There's No Such Thing as Gaining a Pound: Reconsidering the Bathroom Scale User Interface. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '13*, page 401, New York, New York, USA, 2013. ACM Press.

[99] Logan Kendall, Dan Morris, and Desney Tan. Blood Pressure Beyond the Clinic: Rethinking a Health Metric for Everyone Logan. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 1679–1688, 2015.

[100] Gerald Kircheis, Matthias Wettstein, Lars Timmermann, Alfons Schnitzler, and Dieter Häussinger. Critical Flicker Frequency for Quantification of Low-grade Hepatic Encephalopathy. *Hepatology*, 35(2):357–366, 2002.

[101] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. How to evaluate technologies for health behavior change in hci research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3063–3072. ACM, 2011.

[102] Predrag Klasnja and Wanda Pratt. Healthcare in the pocket: mapping the space of mobile-phone health interventions. *Journal of biomedical informatics*, 45(1):184–198, 2012.

[103] Kenneth D Kochanek, Jiaquan Xu, Sherry L Murphy, Arialdi M Minino, and Hsiang-Ching Kung. National Vital Statistics Reports Deaths : Final Data for 2009. Technical Report 3, Centers for Disease Control and Prevention, 2012.

[104] Thomas R Kratochwill, John H Hitchcock, Robert H Horner, Joel R Levin, Samuel L Odom, David M Rindskopf, and William R Shadish. Single-case intervention research design standards. *Remedial and Special Education*, 34(1):26–38, 2013.

[105] Uri Ladabaum, Erin Boyd, Wei K. Zhao, Ajitha Mannalithara, Annie Sharabidze, Gurkirpal Singh, Elaine Chung, and Theodore R. Levin. Diagnosis, Comorbidities, and Management of Irritable Bowel Syndrome in Patients in a Large Health Maintenance Organization. *Clinical Gastroenterology and Hepatology*, 10(1):37–45, 2012.

[106] Carney Landis. Something about Flicker-Fusion. *The Scientific Monthly*, 73(5):308–314, 1951.

[107] Carney Landis. Determinants of the Critical Flicker-Fusion Threshold. *Physiological Reviews*, 32(2):259–286, 1954.

[108] Larklife.

[109] Eric B. Larson. N-of-1 Clinical Trials: A Technique for Improving Medical Therapeutics. *The Western journal of medicine*, 152(1):52–56, 1990.

[110] Eric C. Larson, Mayank Goel, Gaetano Boriello, Sonya Heltshe, Margaret Rosenfeld, and Shwetak N. Patel. SpiroSmart: Using a Microphone to Measure Lung Function on a Mobile Phone. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*, pages 280–289, 2012.

[111] Eric C. Larson, TienJui Lee, Sean Liu, Margaret Rosenfeld, and Shwetak N. Patel. Accurate and privacy preserving cough sensing using a low-cost microphone. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, page 375–384, New York, NY, USA, 2011. Association for Computing Machinery.

[112] Jisoo Lee, Erin Walker, Winslow Burleson, Matthew Kay, Matthew Buman, and Eric B Hekler. Self-Experimentation for Behavior Change: Design and Formative Evaluation of Two Approaches. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 6837–6849. ACM, 2017.

[113] Ian Li, Anind K. Dey, and Jodi Forlizzi. A Stage-Based Model of Personal Informatics Systems. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2010).*, pages 557–566, New York, New York, USA, apr 2010.

[114] Elizabeth O Lillie, Bradley Patay, Joel Diamant, Brian Issell, Eric J Topol, Nicholas J Schork, Scripps Health, and La Jolla. NIH Public Access. *Future Medicine*, 8(2):161–173, 2012.

[115] James J. Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B. Strub. Fish'n'Steps: Encouraging Physical Activity with an Interactive Computer Game. *Ubiquitous Computing (UbiComp 2006)*, pages 261–278, 2006.

[116] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 351–360, 2012.

[117] Daniel M Maggin, Hariharan Swaminathan, Helen J Rogers, Breda V O'keeffe, George Sugai, and Robert H Horner. A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology*, 49(3):301–321, 2011.

[118] Martin Maguire. Methods to support human-centred design. *International journal of human-computer studies*, 55(4):587–634, 2001.

[119] Lena Mamykina, Elizabeth M. Heitkemper, Arlene M. Smaldone, Rita Kukafka, Heather J. Cole-Lewis, Patricia G. Davidson, Elizabeth D. Mynatt, Andrea Cassells, Jonathan N. Tobin, and George Hripcsak. Personal Discovery in Diabetes Self-management: Discovering Cause and Effect Using Self-monitoring Data. *Journal of Biomedical Informatics*, 76(June):1–8, 2017.

[120] Lena Mamykina, Elizabeth Mynatt, Patricia Davidson, and Daniel Greenblatt. MAHI: Investigation of Social Scaffolding for Reflective Thinking in Diabetes Management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2008)*, pages 477–486, 2008.

[121] Lena Mamykina, Elizabeth Mynatt, Patricia Davidson, and Daniel Greenblatt. Mahi: Investigation of social scaffolding for reflective thinking in diabetes management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 477–486, 2008.

[122] Lena Mamykina, Arlene M. Smaldone, and Suzanne R. Bakken. Adopting the Sensemaking Perspective for Chronic Disease Self-management. *Journal of Biomedical Informatics*, 56:406–417, 2015.

[123] Alex Mariakakis, Jacob Baudin, Eric Whitmire, Vardhman Mehta, Megan A Banks, Anthony Law, Lynn McGrath, and Shwetak N Patel. Pupilscreen: Using smartphones to assess traumatic brain injury. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):81, 2017.

[124] Ross A McFarland, A Bertrand Warren, and Charles Karis. Alterations in Critical Flicker Frequency As a Function of Age and Light: Dark Ratio. *Journal of Experimental Psychology*, 56(6):529–538, 1958.

[125] YA McKenzie, A Alder, W Anderson, A Wills, L Goddard, P Gulia, E Jankovich, P Mutch, LB Reeves, A Singer, et al. British dietetic association evidence-based guidelines for the dietary management of irritable bowel syndrome in adults. *Journal of Human Nutrition and Dietetics*, 25(3):260–274, 2012.

[126] Helena M Mentis, Anita Komlodi, Katrina Schrader, Michael Phipps, Ann Gruber-Baldini, Karen Yarbrough, and Lisa Shulman. Crafting a view of self-tracking data in the clinical visit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5800–5812. ACM, 2017.

[127] RD Mirza, S Punja, S Vohra, and G Guyatt. The history and development of n-of-1 trials. *Journal of the Royal Society of Medicine*, 110(8):330–340, 2017.

[128] Debanjali Mitra, Keith L. Davis, and Robert W. Baran. All-Cause Healthcare Charges among Managed Care Patients with Constipation and Comorbid Irritable Bowel Syndrome. *Postgraduate Medicine*, 123(3):122–132, 2011.

[129] Paul Moayyedi, Eamonn MM Quigley, Brian E Lacy, Anthony J Lembo, Yuri A Saito, Lawrence R Schiller, Edy E Soffer, Brennan MR Spiegel, and Alexander C Ford. The effect of dietary intervention on irritable bowel syndrome: a systematic review. *Clinical and translational gastroenterology*, 6(8):e107, 2015.

[130] Mariola Moeyaert, Maaike Ugille, John M Ferron, S Natasha Beretvas, and Wim Van den Noortgate. Three-level analysis of single-case experimental data: Empirical validation. *The Journal of Experimental Education*, 82(1):1–21, 2014.

[131] Kristina W. Monsbakken, Per Olav Vandvik, and Per G. Farup. Perceived Food Intolerance in Subjects with Irritable Bowel Syndrome – Etiology, Prevalence And Consequence. *European Journal of Clinical Nutrition*, 60(5):667–72, 2006.

[132] Margaret Morris and Farzin Guilak. Mobile heart health: project highlight. *IEEE Pervasive Computing*, 8(2):57–61, 2009.

[133] Sean A Munson, Jessica Schroeder, Ravi Karkar, Julie A Kientz, Chia-Fang Chung, and James Fogarty. The importance of starting with goals in n-of-1 studies. *Frontiers in Digital Health*, 2:3, 2020.

[134] Rajalakshmi Nandakumar, Shyamnath Gollakota, and Nathaniel F. Watson. Contactless Sleep Apnea Detection on Smartphones. In *Proceedings of the International Conference on Mobile Systems, Applications, and Services (MobiSys 2015)*, pages 45–57, 2015.

[135] Allen Neuringer. Self-Experimentation: A Call for Change. *Behaviorism*, 9(1):79–94, 1981.

[136] Recommended Light Levels (Illuminance) for Outdoor and Indoor Venues, 2015.

[137] Patrick Onghena and Eugene S Edgington. Customization of pain treatments: Single-case design and analysis. *The Clinical journal of pain*, 21(1):56–68, 2005.

[138] PACO: The Personal Analytics Companion.

[139] Olafur S Palsson, William E Whitehead, Miranda AL Van Tilburg, Lin Chang, William Chey, Michael D Crowell, Laurie Keefer, Anthony J Lembo, Henry P Parkman, Satish SC Rao, et al. Development and validation of the rome iv diagnostic questionnaire for adults. *Gastroenterology*, 150(6):1481–1491, 2016.

[140] Sun Young Park and Yunan Chen. Individual and Social Recognition: Challenges and Opportunities in Migraine Management. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1540–1551, New York, New York, USA, 2015. ACM Press.

[141] Richard I Parker, Kimberly J Vannest, and John L Davis. Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification*, 35(4):303–322, 2011.

[142] Laura R Pina, Sang-Wha Sien, Teresa Ward, Jason C Yip, Sean A Munson, James Fogarty, and Julie A Kientz. From personal informatics to family informatics: Understanding family practices around health monitoring. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2300–2315, 2017.

[143] Renard Xaviero Adhi Pramono, Syed Anas Imtiaz, and Esther Rodriguez-Villegas. A cough-based algorithm for automatic diagnosis of pertussis. *PloS one*, 11(9):e0162128, 2016.

[144] Srinivasa Prasad, Radha K. Dhiman, Ajay Duseja, Yogesh K. Chawla, Arpita Sharma, and Ritesh Agarwal. Lactulose Improves Cognitive Functions and Health-related Quality of Life in Patients With Cirrhosis Who Have Minimal Hepatic Encephalopathy. *Hepatology*, 45(3):549–559, 2007.

[145] William T Riley, Russell E Glasgow, Lynn Etheredge, and Amy P Abernethy. Rapid, Responsive, Relevant (R3) Research: A Call for a Rapid Learning Health Research Enterprise. *Clinical and Translational Medicine*, 2(1):10, 2013.

[146] Manuel Romero-Gómez. Critical Flicker Frequency: It Is Time to Break Down Barriers Surrounding Minimal Hepatic Encephalopathy. *Journal of Hepatology*, 47(1):10–11, 2007.

[147] Manuel Romero-Gómez, Juan Córdoba, Rodrigo Jover, Juan A. Del Olmo, Marta Ramírez, Ramón Rey, Enrique De Madaria, Carmina Montoliu, David Nuñez, Montse Flavia, Luis Compañy, José M. Rodrigo, and Vicente Felipo. Value of the Critical Flicker Frequency in Patients With Minimal Hepatic Encephalopathy. *Hepatology*, 45(4):879–885, 2007.

[148] John Rooksby, Mattias Rost, Alistair Morrison, and Matthew Chalmers Chalmers. Personal Tracking as Lived Informatics. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2014)*, pages 1163–1172, New York, New York, USA, apr 2014.

[149] Martin Sandry. Critical flicker frequency in multiple sclerosis. *Perceptual and motor skills*, 16(1):103–108, 1963.

[150] Jeff Sauro. Measuring usability with the system usability scale (sus), 2011.

[151] Jessica Schroeder. *Goal-Directed Self-Tracking in the Management of Chronic Health Conditions*. PhD thesis, University of Washington, 2020.

[152] Jessica Schroeder, Chia-Fang Chung, Daniel A Epstein, Ravi Karkar, Adele Parsons, Natalia Murinova, James Fogarty, and Sean A Munson. Examining self-tracking by people with migraine: Goals, needs, and opportunities in a chronic health condition. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018*, pages 135–148. ACM, 2018.

[153] Jessica Schroeder, Jane Hoffswell, Chia-Fang Chung, James Fogarty, Sean Munson, and Jasmine 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 - CSCW '17*, pages 1726–1739, 2017.

[154] Jessica Schroeder, Ravi Karkar, James Fogarty, Julie A. Kientz, Sean A. Munson, and Matthew Kay. A patient-centered proposal for bayesian analysis of self-experiments for health. *Journal of Health Informatics Research*, In press.

[155] Jessica Schroeder, Ravi Karkar, Natalia Murinova, James Fogarty, and Sean 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):1–26, 2019.

[156] SCRT Package.

[157] William R Shadish, Larry V Hedges, and James E Pustejovsky. An spss macro for a d-statistic for single-case designs. In *Workshop presented at the 2013 Spring Conference of the Society for Research on Educational Effectiveness*, 2013.

[158] Praveen Sharma and Barjesh Chander Sharma. Lactulose for Minimal Hepatic Encephalopathy in Patients With Extrahepatic Portal Vein Obstruction. *Saudi Journal of Gastroenterology*, 18(3):168, jun 2012.

[159] Praveen Sharma, Barjesh Chander Sharma, and Shiv Kumar Sarin. Critical Flicker Frequency for Diagnosis and Assessment of Recovery From Minimal Hepatic Encephalopathy in Patients With Cirrhosis. *Hepatobiliary and Pancreatic Diseases International*, 9(1):27–32, 2010.

[160] Debbie L. Shawcross, Arthur A. Dunk, Rajiv Jalan, Gerald Kircheis, Robert J. De Knegt, Wim Laleman, John K. Ramage, Heiner Wedemeyer, and Ian E.J. Morgan. How to Diagnose and Manage Hepatic Encephalopathy: A Consensus Statement on Roles and Responsibilities Beyond the Liver Specialist. *European Journal of Gastroenterology and Hepatology*, 28(2):146–152, 2016.

[161] Sandeep Singh Sidhu, Omesh Goyal, Bholeshwar Prashad Mishra, Ajit Sood, Rajoo Singh Chhina, and Ravinder Kumar Soni. Rifaximin Improves Psychometric Performance and Health-related Quality of Life in Patients With Minimal Hepatic Encephalopathy (the RIME Trial). *American Journal of Gastroenterology*, 106(2):307–316, 2011.

[162] Magnus Simrén, Agneta Månsson, Anna Maria Langkilde, Jan Svedlund, Hasse Abrahamsson, Ulf Bengtsson, and Einar S. Björnsson. Food-Related Gastrointestinal Symptoms in the Irritable Bowel Syndrome. *Digestion*, 63(2):108–15, 2001.

[163] Heidi M. Staudacher, Kevin Whelan, Peter M. Irving, and Miranda C. E. Lomer. Comparison Of Symptom Response Following Advice For A Diet Low In Fermentable Carbohydrates (FODMAPs) Versus Standard Dietary Advice In Patients With Irritable Bowel Syndrome. *Journal of Human Nutrition and DIetics*, 24(5):487–495, 2011.

[164] Hyewon Suh, Nina Shahriaree, Eric B. Hekler, and Julie A. Kientz. Developing and Validating the User Burden Scale: A Tool for Assessing User Burden in Computing Systems. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2016)*, pages 3988–3999, 2016.

[165] Xiao Sun, Zongqing Lu, Wenjie Hu, and Guohong Cao. SymDetector: Detecting Sound-Related Respiratory Symptoms Using Smartphones. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2015*, pages 97–108, 2015.

[166] Melanie Swan. Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. *International journal of environmental research and public health*, 6(2):492–525, 2009.

[167] F. J. Torlot, M. J W McPhail, and S. D. Taylor-Robinson. Meta-analysis: The Diagnostic Accuracy of Critical Flicker Frequency in Minimal Hepatic Encephalopathy. *Alimentary Pharmacology and Therapeutics*, 37(5):527–536, 2013.

[168] Khai N Truong, Gillian R Hayes, and Gregory D Abowd. Storyboarding: an empirical determination of best practices and effective guidelines. In *Proceedings of the 6th conference on Designing Interactive systems*, pages 12–21, 2006.

[169] Roger Vilardaga, Jonathan B Bricker, and Michael G McDonell. The promise of mobile technologies and single case designs for the study of individuals in their natural environment. *Journal of Contextual Behavioral Science*, 3(2):148–153, 2014.

[170] Hendrik Vilstrup, Piero Amodio, Jasmohan Bajaj, Juan Cordoba, Peter Ferenci, Kevin D. Mullen, Karin Weissenborn, and Philip Wong. Hepatic Encephalopathy in Chronic Liver Disease: 2014 Practice Guideline by the American Association for the Study of Liver Diseases and the European Association for the Study of the Liver. *Hepatology*, 60(2):715–735, 2014.

[171] Philip Vutien, Ravi Karkar, Richard Li, Kara Walter, Sean Munson, James Fogarty, and George Ioannou. Evaluating a novel, portable, self-administered device ("flicker-app") that measures critical flicker frequency as a test for hepatic encephalopathy in patients with cirrhosis. In *Hepatology*. Wiley 111 River St, Hoboken 07030-5774, NJ USA, 2020.

[172] Tarun Wadhawan, Ning Situ, Hu Rui, Keith Lancaster, Xiaojing Yuan, and George Zouridakis. Implementation of the 7-point checklist for melanoma detection on smart handheld devices. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3180–3183. IEEE, 2011.

[173] Kara Walter, Ravi Karkar, Sean Munson, James Fogarty, and George Ioannou. Developing a novel, portable, self-administered device (flicker-app) that measures critical flicker frequency as a test for minimal hepatic encephalopathy. In *Hepatology*, volume 70, pages 274A–274A. Wiley 111 River St, Hoboken 07030-5774, NJ USA, 2019.

[174] Edward Jay Wang, William Li, Doug Hawkins, Terry Gernsheimer, Colette Norby-Slycord, and Shwetak N Patel. Hemaapp: noninvasive blood screening of hemoglobin using smartphone cameras. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 593–604, 2016.

[175] Peter West, Richard Giordano, Max Van Kleek, and Nigel Shadbolt. The quantified patient in the doctor's office: Challenges & opportunities. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 3066–3078, New York, NY, USA, 2016. ACM.

[176] Jacob O. Wobbrock and Julie A. Kientz. Research contributions in human-computer interaction. *Interactions*, 23(3):38–44, April 2016.

[177] Kathryn Yorkston, Kurt Johnson, Estelle Klasner, Dagmar Amtmann, Carrie Kuehn, and Brian Dudgeon. Getting the Work Done: A Qualitative Study of Individuals with Multiple Sclerosis. *Disability and Rehabilitation*, 25(8):369–379, 2003.

[178] Emily Y Zeng, Roger Vilardaga, Jaimee L Heffner, Kristin E Mull, and Jonathan B Bricker. Predictors of utilization of a novel smoking cessation smartphone app. *Telemedicine and e-Health*, 21(12):998–1004, 2015.

[179] Haining Zhu, Joanna Colgan, Madhu Reddy, and Eun Kyoung Choe. Sharing patient-generated data in clinical practices: an interview study. In *AMIA Annual Symposium Proceedings*, volume 2016, page 1303. American Medical Informatics Association, 2016.