TummyTrials: A Feasibility Study of Using Self-Experimentation to Detect Individualized Food Triggers

Ravi Karkar[1], Jessica Schroeder[1], Daniel A. Epstein[1], Laura R. Pina[1,2], Jeffrey Scofield[1], James Fogarty[1], Julie A. Kientz[2], Sean A. Munson[2], Roger Vilardaga[3], Jasmine Zia[4]

[1]Computer Science & Engineering, [2]Human Centered Design & Engineering, [4]Division of Gastroenterology
DUB Group, University of Washington, Seattle, WA, United States
[3]Center for Addiction Science and Technology, Duke University, Durham, NC, United States
{rkarkar, jesscs, depstein, lpina, jeffsco, jfogarty}@cs.washington.edu, {jkientz, smunson}@uw.edu
roger.vilardaga@duke.edu, jzia@medicine.washington.edu

ABSTRACT

Diagnostic self-tracking, the recording of personal information to diagnose or manage a health condition, is a common practice, especially for people with chronic conditions. Unfortunately, many who attempt diagnostic self-tracking have trouble accomplishing their goals. People often lack knowledge and skills needed to design and conduct scientifically rigorous experiments, and current tools provide little support. To address these shortcomings and explore opportunities for diagnostic self-tracking, we designed, developed, and evaluated a mobile app that applies a self-experimentation framework to support patients suffering from irritable bowel syndrome (IBS) in identifying their personal food triggers. TummyTrials aids a person in designing, executing, and analyzing self-experiments to evaluate whether a specific food triggers their symptoms. We examined the feasibility of this approach in a field study with 15 IBS patients, finding that participants could use the tool to reliably undergo a self-experiment. However, we also discovered an underlying tension between scientific validity and the lived experience of self-experimentation. We discuss challenges of applying clinical research methods in everyday life, motivating a need for the design of self-experimentation systems to balance rigor with the uncertainties of everyday life.

Author Keywords

Self-Tracking; Self-Experimentation; Irritable Bowel Syndrome; Personal Informatics; Symptom Triggers; Food.

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g., HCI): User Interfaces; J.3. Life and Medical Sciences: Health.

INTRODUCTION

Many people have an interest in tracking aspects of their health, with Pew reporting 69% of all U.S. adults track at least one health indicator (e.g., weight, diet, exercise routine, symptoms) [19]. People with chronic conditions are even more likely to track, since they need meaningful and actionable information from their tracking. Many technology-based tools currently exist for people to use to improve their health (e.g., Fitbit, MyFitnessPal, RunKeeper, Weight Watchers).

Current tools are generally designed for data collection toward goals predetermined by the tool's designer (e.g., staying physically active, losing weight, eating healthy). Because self-tracking is an increasingly common and everyday consumer practice [17,63], goals supported by tools are often generic and intended to appeal to the broadest set of people. However, many people have specific and personal questions about their health, such as "Does caffeine impact my sleep?" Rooksby et al. define self-tracking with the goal of answering such specific questions as *diagnostic self-tracking* [63].

Widely-available tools do not yet support a systematic approach to answering such diagnostic questions. Self-tracked data may suggest a relationship between sleep quality and caffeine, but determining if caffeine is actually *causing* poor sleep quality is not well supported (i.e., the difference between correlation versus causation). A person could also be consuming more caffeine because they are tired due to a lack of sleep that has other causes (e.g., stress). Such uncertainty leaves people hesitant to make lifestyle changes that could improve health outcomes (e.g., eliminating caffeine) based on self-tracked data.

Choe et al.'s examination of a community of expert self-trackers identified three major pitfalls people encounter in diagnostic self-tracking, even when they have the knowledge and skills to build their own tools [11]: (1) tracking too many things at once, (2) not tracking triggers and context, and (3) lacking scientific rigor in experimental design and analysis. Although some individuals eventually succeeded in their efforts to build custom solutions, the process was often long and frustrating. Many people also lack the knowledge, skills, and motivation to succeed in diagnostic self-tracking without better support.

We examine this need for better tool support for personal diagnostic self-tracking through the domain of people who experience gastrointestinal symptoms they believe may be triggered or worsened by certain foods. Successfully identifying such personal triggers could help people reduce their symptoms and improve their quality of life. Although symptom trackers exist (e.g., [54,71]), no tools currently

CHI 2017, May 06 - 11, 2017, Denver, CO, USA

ACM 978-1-4503-4655-9/17/05...$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025480

exist to help people navigate the process of systematically collecting and analyzing the data required to confidently identify foods that trigger their personal symptoms.

To help bridge this gap, we have applied a framework for self-experimentation [34] in the design and development of a mobile app called TummyTrials. The app uses single-case experimental designs (SCD), also called n-of-1 trials, to help people suffering from irritable bowel syndrome (IBS) determine whether certain foods worsen their gastrointestinal symptoms. TummyTrials includes support for people designing, executing, and analyzing a scientifically valid self-experiment to determine whether a food is impacting their symptoms. Specifically, TummyTrials supports this process through an interface that allows choosing an independent variable (i.e., a food type) and one or more dependent variables (i.e., symptoms). It then generates a randomized study protocol of days on which to consume or avoid a potential trigger, provides daily reminders and prompts to record symptoms, and presents analyses and visualizations to help a person understand and interpret their personal self-experiment. The TummyTrials design was motivated and informed by scientifically robust approaches to single-case experimental design and analysis, as well as prior formative research with people suffering from IBS [34].

We conducted an evaluation of TummyTrials to examine the feasibility of conducting self-experiments as a form of diagnostic self-tracking. In a study with 15 IBS patients, participants were guided in configuring a 12-day self-experiment to determine if a particular food was affecting their symptoms, then asked to undergo the self-experiment. For the duration of the self-experiment, participants reported their compliance with the self-experiment as well as their daily symptom levels. At the end of the experiment, participants received analyses and visualizations of results showing the evidence and impact of their experimental trigger food on each of their reported symptoms. We used questionnaires and a semi-structured interview to gather feedback after the self-trials. We found that: (1) participants were able to conduct a 12-day self-experiment with the support of instructions and notifications to scaffold the self-experimentation process, and (2) there are new challenges of applying clinical research methods to everyday life that motivate a need for the design of self-experimentation systems that balance introduction of necessary rigor with the uncertainties of everyday life.

BACKGROUND AND RELATED WORK

Diagnostic Self-Tracking

Personal informatics tools, including self-monitoring applications, help people understand their habits and behaviors [41]. One common motivation for self-monitoring is diagnostic self-tracking: tracking to answer a specific question [63]. People often use diagnostic self-tracking to manage a condition, find triggers, or identify relationships pertaining to their health or other aspects of life [11].

Existing devices and apps often focus on tracking physical fitness (e.g., [12,18,29,40,43,56]), sleep (e.g., [18,35,40]), diet (e.g., [2,13,46]), smoking [1], and stress [53]. Their primary focus is to support a high-level health goal, such as staying healthy or 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 may not provide any feedback or give only correlational results, which are often insufficient to answer diagnostic questions. These apps also tend to be burdensome, particularly those for supporting a healthy diet, in part because they often require accurately logging every meal to be valuable to the person tracking [13,14].

Recent research examines support for self-experimentation. PACO helps people experiment with behavior change techniques [57]. SleepCoacher identifies connections between potential sleep disruptors and sleep quality [55]. Trialist helps patients and clinicians collaborate to find correct medication dosing for chronic pain [70]. TummyTrials builds on this work by focusing on helping end-users design their own single-case experiment for identifying causal relationships between food triggers and symptoms.

Single-Case Designs

To assist IBS patients in determining their individualized food triggers, TummyTrials uses a single-case experimental design. This design contrasts group randomized controlled trials (RCTs), where participants are randomly assigned to a treatment or control condition [48]. RCTs are considered the highest level of clinical efficacy evaluation for an intervention, but the population-based estimates they produce do not inform how a specific individual and their symptoms will respond [60]. SCDs, or n-of-1 experiments, can be used to understand how an individual responds to a certain intervention [42,62]. In these experiments, the individual serves as their own control, testing the individual's specific response to an intervention rather than a group's average. SCDs are therefore more sensitive to individual differences than RCTs, which makes them ideal in our use case.

Within SCDs, a number of experimental design alternatives are available. AB and ABAB phase designs are among the more common, where phase A is a baseline measurement and phase B is the intervention. To address our specific use case, we used a variation of Alternating Treatment Design (ATD) that applies A and B phases completely at random [15,27]. Random assignment of treatment phases helps to overcome common criticisms regarding the internal validity of SCDs [24,32,37,45,51,59,66] and allows the use of a statistical method for SCDs called randomization tests [15,27]. Our prior work summarizes the design rationale behind the SCD framework in more depth [34].

Irritable Bowel Syndrome and Food Intolerances

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 reason people seek primary care [16,44]. People with IBS report a lower quality of life and consume

50% more healthcare resources than non-IBS counterparts [39,49]. Potential triggers for IBS symptom flare-ups include certain foods, eating behaviors, stress, sleep disturbances, and menstruation, with foods as the most common trigger [23,26]. Traditional IBS medications have only marginal therapeutic gains of 7-15% over placebo [10]. 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 [21,22]. The elimination diet process can last up to six months [47,50]. Patients find elimination diets frustrating because they are high burden, are unintuitive, and lack sufficient instructions to successfully undergo [5,6,22,68]. Fortunately, total elimination of all possible trigger foods is excessive for most people. Individual responses to specific foods vary, with a given food triggering bowel symptoms in some people but not others [28]. 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.

During an elimination diet, people are often asked to journal their food and IBS symptoms. However, journals are typically handwritten, incomplete, disorganized, and unreliable [25,31]. Information such as meal time, food ingredients, and symptom severity are often missing because journaling is complex and high burden [14]. Also, clinicians do not receive formal training on how to review journals, and such interpretations result in a high degree of inter-observer variability [38]. Not surprisingly, most people with IBS are dissatisfied with the journal feedback they receive from health providers [28].

IBS is a useful domain for understanding the potential for self-experimentation because patients struggle to manage their condition, particular triggers are highly individualized [52,67], symptoms tend to be experienced within a short time window of consuming the trigger food [34,61,67], and the current identification process is lengthy, tedious, and frustrating [28]. We aim to improve both process and outcome, aiding IBS patients in effectively determining whether a particular food is a trigger while minimizing impact on their daily life.

TUMMYTRIALS DESIGN

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 [34]. The goal of TummyTrials is to provide an effective and low-burden approach for people suffering from IBS to systematically test potential food-based 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.

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 1A), (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 interviews we conducted with IBS patients suggest these are the most common symptoms and trigger foods for patients with IBS [34].

Patients have reported that the onset of symptoms generally occurs within a short duration after consuming a trigger food, typically under three hours [34]. 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 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 or bread or English muffin or toast, cereal, muffin, waffle or 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

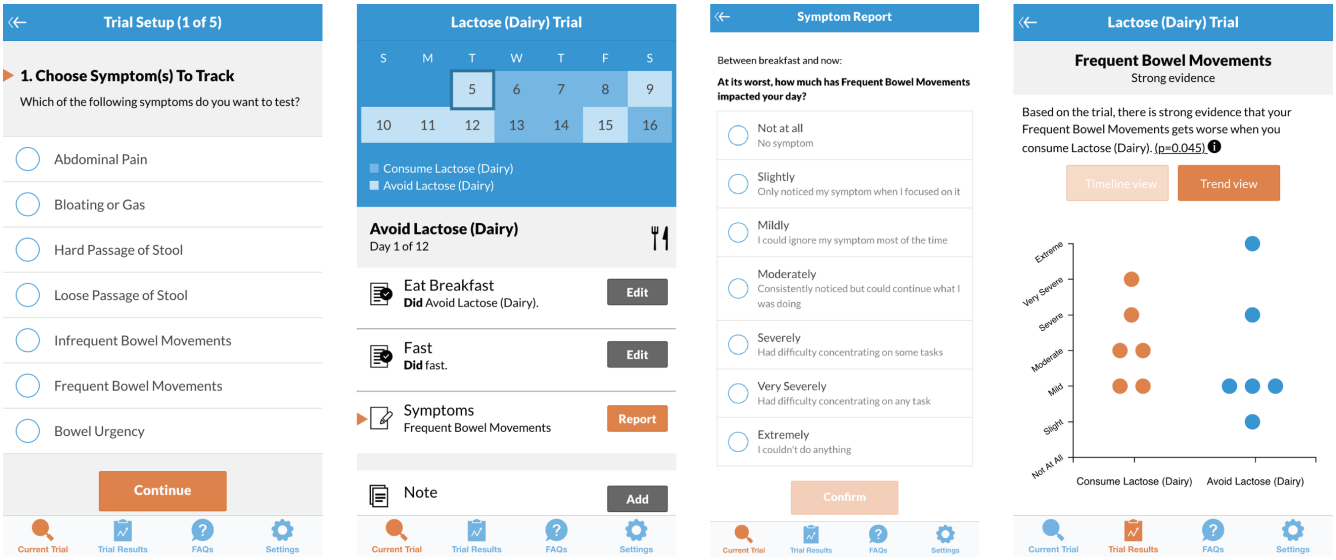

Figure 1: TummyTrials supports scientifically valid self-experimentation for identifying individualized food triggers, including: A) self-experiment configuration, B) daily prompts and reminders, C) compliance and symptom tracking, and D) analysis of results.

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 [15,27]. 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 [33], 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 [3], 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.

Self-Experiment Execution and Data Collection

We designed TummyTrials to be low burden relative to current standards of care: elimination diets and food and 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 1C). TummyTrials provides an optional notes section to record any additional information a person wants to add (Figure 1B).

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 1B), giving an overview of the entire self-experiment and highlighting the "avoid" (control) or "consume" (experimental) condition for each day. The person can review reports for prior days and the food plan for the current day. 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.

Self-Experiment Results Review

Upon completing a self-experiment, TummyTrials generates a results page for each symptom a person tracked (Figure 1D). Visual analysis is the traditional approach to analyze

single-case designs [8]. 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 [34]. We therefore do not plot trend lines, instead illustrating the data in a timeline plot and a trend plot (Figure 2), as proposed in [34]. 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 [65]. 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 1D is an example of strong evidence (p=0.045) while Figure 2 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.

METHOD

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 [36]. 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.

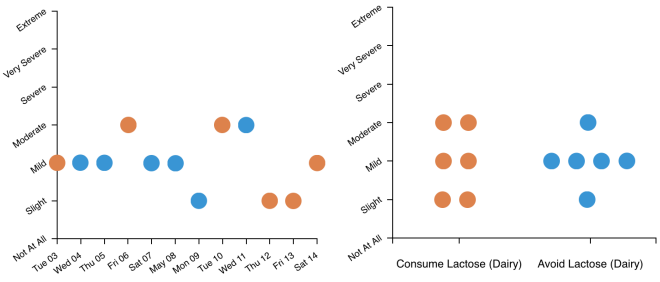

Figure 2: TummyTrials visualizes self-experimentation both as a timeline (left) and by trend in experimental condition (right).

Patients were asked to avoid testing known (diagnosed or strongly suspect) 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).

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 [58], 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 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 [9].

Procedures

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) [20], 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) [7], the User Burden Scale (UBS) [69], 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 [4], iterating on themes with the rest of the research team through discussion.

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.

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 were instructed to avoid testing known triggers, 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*").

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 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.

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 [64].

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, median, and standard deviation within each subscale of the scale was as follows: difficulty of use ($\bar{x}$=0.73, M=0, $\sigma_{\bar{x}}$=1.1, Grade=C), physical ($\bar{x}$=0.2, M=0, $\sigma_{\bar{x}}$=0.56, Grade=C), time and social ($\bar{x}$=0.5, M=0, $\sigma_{\bar{x}}$=1.35, Grade=B), mental and emotional ($\bar{x}$=0.47, M=0, $\sigma_{\bar{x}}$=0.9, Grade=B), privacy ($\bar{x}$=0.8, M=0, $\sigma_{\bar{x}}$=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 >50 points is considered clinically significant according to the IBS-SSS scale [20]. 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 accountability 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 it 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. 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.

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 [50,67]. 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 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.

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 changing the wording of the existing scale to adding different measures for tracking. 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 three days in a row with no caffeine, maybe that helped my stomach settle.*

Similarly, having a condition mostly on weekdays or weekends could cofound 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.

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 can keep a check on their progress.

DISCUSSION

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 valid for three days [30]. 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).

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. 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.

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. Toward this, 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.

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 [34]. 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 [34]. 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.

CONCLUSION

We designed and examined TummyTrials, an app that applies a framework for self-experimentation in personalized health to help IBS patients conduct self-experiments to identify their individualized food triggers. In a field study in which 15 IBS patients completed 12-day self-experiments using the app, we found TummyTrials effectively supports self-experimentation. However, interviews with participants revealed a tension between scientific rigor and uncertainties of lived experience. This research therefore motivates further development of self-experimentation as an approach, together with additional explorations of how to support the realities of everyday life.

ACKNOWLEDGEMENTS

We acknowledge Andrea Martin, Jonathan Cook, and Kai-Ting Huang, and our study participants for their assistance in the research. This research has been funded in part by the University of Washington Innovation Research Award, the Intel Science and Technology Center for Pervasive Computing, Nokia Research, the National Science Foundation under awards IIS-1553167 and SCH-1344613, the Agency for Healthcare Research Quality under award 1R21HS023654, and the National Institute on Drug Abuse under award 1K99DA037276-01.

REFERENCES

1. Amin Ahsan Ali, Syed Monowar Hossain, Karen Hovsepian, Kurt Plarre, and Santosh Kumar. (2012). mPuff: Automated Detection of Cigarette Smoking Puffs from Respiration Measurements. *Proceedings of the Conference on Information Processing in Sensor Networks (ISPN 2012)*, 269–280. http://doi.org/bwtk
2. 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. (2012). Prescriptive Persuasion and Open-Ended Social Awareness: Expanding the Design Space of Mobile Health. *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 2012)*, 475–484. http://doi.org/bbkm
3. Frank Bentley and Konrad Tollmar. (2013). The Power of Mobile Notifications to Increase Wellbeing Logging Behavior. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2013)*, 1095–1098. http://doi.org/f2zs59
4. Hugh Beyer and Karen Holtzblatt. (1999). Contextual Design.
5. 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. (2011). 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. http://doi.org/bwtm
6. Jessica R. Biesiekierski, Simone L. Peters, Evan D. Newnham, Ourania Rosella, Jane G. Muir, and Peter R. Gibson. (2013). 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. http://doi.org/f2kqcx
7. John Brooke. (1996). SUS: A Quick and Dirty Usability Scale. *Usability Evaluation in Industry*, *189*(194), 4–7.
8. Isis Bulté and Patrick Onghena. (2012). When the truth hits you between the eyes: A software tool for the visual analysis of single-case experimental data. *Methodology*, *8*(3), 104–114. http://doi.org/bgx86g
9. Caroline Canavan, Joe West, and Timothy Card. (2014). The Epidemiology of Irritable Bowel Syndrome. *Clinical epidemiology*, *6*, 71–80.
10. William D Chey, Monthira Maneerattaporn, and Richard Saad. (2011). Pharmacologic and Complementary and Alternative Medicine Therapies for Irritable Bowel Syndrome. *Gut and Liver*, *5*(3), 253–266. http://doi.org/cxgfcq
11. Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. (2014). Understanding Quantified-Selfers' Practices in Collecting and Exploring Personal Data. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2014)*, 1143–1152. http://doi.org/bbpd
12. 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. (2008). Activity Sensing in the Wild: A Field Trial of Ubifit Garden. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2008)*, 1797–1806. http://doi.org/fj37wd
13. Felicia Cordeiro, Elizabeth Bales, Erin Cherry, and James Fogarty. (2015). Rethinking the Mobile Food Journal: Exploring Opportunities for Lightweight Photo-Based Capture. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2015)*, 3207–3216. http://doi.org/bbkc
14. Felicia Cordeiro, Daniel A. Epstein, Edison Thomaz, Elizabeth Bales, Arvind K. Jagannathan, Gregory D. Abowd, and James Fogarty. (2015). Barriers and Negative Nudges: Exploring Challenges in Food Journaling. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2015)*, 1159–1162. http://doi.org/bbdt
15. Eugene S. Edgington and Patrick Onghena. (2007). *Randomization Tests*. CRC Press.
16. Sigrid Elsenbruch. (2011). Abdominal Pain in Irritable Bowel Syndrome: A Review of Putative Psychological, Neural and Neuro-Immune Mechanisms. *Brain, Behavior, and Immunity*. http://doi.org/bmn48f
17. Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. (2015). A Lived Informatics Model of Personal Informatics. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2015)*, 731–742. http://doi.org/bdsr
18. Fitbit. https://www.fitbit.com/
19. Susannah Fox and Maeve Duggan. (2013). Tracking for Health. *Pew Research Center's Internet & American Life Project*, (October), 1–40.

20. Carol Y. Francis, Julie Morris, and Peter J. Whorwell. (1997). The Irritable Bowel Severity Scoring System: A Simple Method of Monitoring Irritable Bowel Syndrome and its Progress. *Alimentary Pharmacology & Therapeutics*, *11*(2), 395–402. http://doi.org/chgr63
21. Peter R. Gibson and Susan J. Shepherd. (2012). Food Choice as a Key Management Strategy for Functional Gastrointestinal Symptoms. *The American Journal of Gastroenterology*, *107*(5), 657–666. http://doi.org/bwtn
22. Emma P. Halmos, Victoria A. Power, Susan J. Shepherd, Peter R. Gibson, and Jane G. Muir. (2014). A Diet Low in FODMAPs Reduces Symptoms of Irritable Bowel Syndrome. *Gastroenterology*, *146*(1), 67–75. http://doi.org/f2n9rr
23. Lynsey R. Harris and Lesley Roberts. (2008). Treatments for Irritable Bowel Syndrome: Patients' Attitudes and Acceptability. *BMC Complementary and Alternative Medicine*, *8*, 65. http://doi.org/d2p6vg
24. Steven C. Hayes. (1981). Single Case Experimental Design and Empirical Clinical Practice. *Journal of Consulting and Clinical Psychology*, *49*(2), 193–211. http://doi.org/bqzkhf
25. Reetta Heinonen, Riitta Luoto, Pirjo Lindfors, and Clas-Håkan Nygård. (2012). Usability and Feasibility of Mobile Phone Diaries in an Experimental Physical Exercise Study. *Telemedicine and e-Health*, *18*(2), 115–119. http://doi.org/fx8dgq
26. Margaret Heitkemper, Eric Carter, Vanessa Ameen, Kevin Olden, and Lin Cheng. (2002). Women with Irritable Bowel Syndrome: Differences in Patients' and Physicians' Perceptions. *Gastroenterology Nursing*, *25*(5), 192–200.
27. Mieke Heyvaert and Patrick Onghena. (2014). 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. http://doi.org/bwtp
28. Anne E. Jamieson, Paula C. Fletcher, and Margaret A. Schneider. (2007). 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. http://doi.org/c9nvp3
29. Jawbone UpBand. https://jawbone.com/up
30. Mark P. Jensen, Wei Wang, Susan L. Potts, and Errol M. Gould. (2012). Reliability and Validity of Individual and Composite Recall Pain Measures in Patients with Cancer. *Pain Medicine (United States)*, *13*(10), 1284–1291. http://doi.org/bwtr
31. Mehul Jhaveri and Elizabeth Lee. (2007). Performance of Electronic Diaries in Diabetes Clinical Trials Measured through Overall Satisfaction of Site Coordinators. *Journal of Diabetes Science and Technology*, *1*(4), 522–30. http://doi.org/bwts
32. SungWoo Kahng, Kyong-Mee Chung, Katharine Gutshall, Steven C. Pitts, Joyce Kao, and Kelli Girolami. (2010). Consistent Visual Analyses of Intrasubject Data. *Journal of Applied Behavior Analysis*, *43*(1), 35–45. http://doi.org/dswgqx
33. Sunanda V. Kane, Karen Sable, and Stephen B. Hanauer. (1998). 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. http://doi.org/dmg8sx
34. Ravi Karkar, Jasmine Zia, Roger Vilardaga, Sonali R. Mishra, James Fogarty, Sean A. Munson, and Julie A. Kientz. (2015). A Framework for Self-Experimentation in Personalized Health. *Journal of the American Medical Informatics Association (JAMIA)*, 1–9. http://doi.org/bwtt
35. Matthew Kay, Eun Kyoung Choe, Jesse Shepherd, Benjamin Greenstein, Nathaniel F. Watson, Sunny Consolvo, and Julie A. Kientz. (2012). Lullaby: A Capture & Access System for Understanding the Sleep Environment. *Proceedings of the ACM Conference on Ubiquitous Computing (UbiComp 2012)*, 226–234. http://doi.org/bwtv
36. Predrag Klasnja, Sunny Consolvo, Wanda Pratt, Health Informatics, and Intel Labs Seattle. (2011). How to Evaluate Technologies for Health Behavior Change in HCI Research. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2011)*, 3063–3072. http://doi.org/dzgdm6
37. Thomas R. Kratochwill, John H. Hitchcock, Robert H. Horner, Joel R. Levin, Samuel L. Odom, David M. Rindskopf, and William R. Shadish. (2013). Single-Case Intervention Research Design Standards. *Remedial and Special Education*, *34*(1), 26–38. http://doi.org/bwtw

38. Theodore Kueper, Dean Martinelli, Wayne Konetzki, Ralph W. Stamerjohn, and Jeanne B. Magill. (1995). Identification Of Problem Foods Using Food And Symptom Diaries. *Otolaryngology - Head and Neck Surgery*, *112*(3), 415–420. http://doi.org/cps5rf
39. Uri Ladabaum, Erin Boyd, Wei K. Zhao, Ajitha Mannalithara, Annie Sharabidze, Gurkirpal Singh, Elaine Chung, and Theodore R. Levin. (2012). Diagnosis, Comorbidities, and Management of Irritable Bowel Syndrome in Patients in a Large Health Maintenance Organization. *Clinical Gastroenterology and Hepatology*, *10*(1), 37–45. http://doi.org/c2rcf9
40. Larklife. http://lark.com/products/larklife/experience
41. Ian Li, Anind K. Dey, and Jodi Forlizzi. (2010). A Stage-Based Model of Personal Informatics Systems. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2010).*, 557–566. http://doi.org/bh8zsb
42. Elizabeth O. Lillie, Bradley Patay, Joel Diamant, Brian Issell, Eric J. Topol, and Nicholas J. Schork. (2011). The N-of-1 Clinical Trial: The Ultimate Strategy For Individualizing Medicine? *Personalized Medicine*, *8*(2), 161–173. http://doi.org/b6mzkp
43. James J. Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B. Strub. (2006). Fish'n'Steps: Encouraging Physical Activity with an Interactive Computer Game. *Ubiquitous Computing (UbiComp 2006)*, 261–278. http://doi.org/crcvd9
44. Rebecca M. Lovell and Alexander C. Ford. (2012). Effect of Gender on Prevalence of Irritable Bowel Syndrome in the Community: Systematic Review and Meta-Analysis. *The American Journal of Gastroenterology*, *107*, 991–1000. http://doi.org/bwtx
45. Daniel M. Maggin, Hariharan Swaminathan, Helen J. Rogers, Breda V. O'keeffe, George Sugai, and Robert H. Horner. (2011). A Generalized Least Squares Regression Approach For Computing Effect Sizes In Single-case Research: Application Examples. *Journal of School Psychology*, *49*, 301–321. http://doi.org/d4nc3m
46. Lena Mamykina, Elizabeth Mynatt, Patricia Davidson, and Daniel Greenblatt. (2008). MAHI: Investigation of Social Scaffolding for Reflective Thinking in Diabetes Management. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2008)*, 477–486. http://doi.org/dck23q
47. Y. A. McKenzie, A. Alder, W. Anderson, A. Wills, L. Goddard, P. Gulia, E. Jankovich, P. Mutch, L. B. Reeves, A. Singer, and M. C. E. Lomer. (2012). British Dietetic Association Evidence-Based Guidelines for the Dietary Management of Irritable Bowel Syndrome in Adults. *Journal of Human Nutrition and Dietetics 25*, 260–274. http://doi.org/bwtz
48. Marcia L. Meldrum. (2000). A Brief History Of The Randomized Controlled Trial: From Oranges and Lemons to Gold Standard. *Hematology/Oncology Clinics of North America*, *14*(4), 745–760. http://doi.org/fc9mpg
49. Debanjali Mitra, Keith L. Davis, and Robert W. Baran. (2011). All-Cause Health Care Charges among Managed Care Patients with Constipation and Comorbid Irritable Bowel Syndrome. *Postgraduate Medicine*, *123*(3), 122–132. http://doi.org/fwcctz
50. Paul Moayyedi, Eamonn M. M. Quigley, Brian E. Lacy, Anthony J. Lembo, Yuri A. Saito, Lawrence R. Schiller, Edy E. Soffer, Brennan M. R. Spiegel, and Alexander C. Ford. (2015). The Effect of Dietary Intervention on Irritable Bowel Syndrome: A Systematic Review. *Clinical and Translational Gastroenterology*, *6*, e107. http://doi.org/bwt2
51. Mariola Moeyaert, Maaike Ugille, John M. Ferron, S. Natasha Beretvas, and Wim Van den Noortgate. (2014). Three-level Analysis of Single-Case Experimental Data: Empirical Validation. *The Journal of Experimental Education*, *82*(1), 1–21. http://doi.org/bwt3
52. Kristina W. Monsbakken, Per Olav Vandvik, and Per G. Farup. (2006). Perceived Food Intolerance in Subjects with Irritable Bowel Syndrome – Etiology, Prevalence And Consequence. *European Journal of Clinical Nutrition*, *60*(5), 667–72. http://doi.org/fqd4d9
53. Margaret Morris and Farzin Guilak. (2009). Mobile Heart Health: Project Highlight. *IEEE Pervasive Computing*, *8*(2), 57–61. http://doi.org/fcwrgd
54. mySymptoms. http://skygazerlabs.com/wp/
55. Daskalova Nediyana, Danaë Metaxa-Kakavouli, Adrienne Tran, Nicole Nugent, Julie Boergers, John McGeary, and Jeff Huang. (2016). SleepCoacher : A Personalized Automated Self-Experimentation System for Sleep Recommendations. *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST 2016)*, 347–358. http://doi.org/bwt4
56. Nike Fuelband. http://www.nike.com/us/en_us/c/nikeplus/nikefuel

57. PACO: The Personal Analytics Companion. https://www.pacoapp.com/
58. Olafur S. Palsson, William E. Whitehead, Miranda A. L. van Tilburg, Lin Chang, William Chey, Michael D. Crowell, Laurie Keefer, Anthony J. Lembo, Henry P. Parkman, Satish Sc Rao, Ami Sperber, Brennan Spiegel, Jan Tack, Stephen Vanner, Lynn S. Walker, Peter Whorwell, and Yunsheng Yang. (2016). Rome IV Diagnostic Questionnaires and Tables for Investigators and Clinicians. *Gastroenterology*, *150*(6), 1481–1491. http://doi.org/bwt5
59. Richard I. Parker, Kimberly J. Vannest, and John L. Davis. (2011). Effect Size in Single-Case Research: A Review of Nine Nonoverlap Techniques. *Behavior Modification*, *35*(4), 303–22. http://doi.org/dsdfs4
60. Elazar J. Pedhazur and Liora Pedhazur Schmelkin. (1991). *Measurement, Design, And Analysis: An Integrated Approach*. Erlbaum, New Jersey.
61. Iris Posserud, Hans Strid, Stine Störsrud, Hans Törnblom, Ulla Svensson, Jan Tack, Lukas Van Oudenhove, and Magnus Simrén. (2013). Symptom Pattern Following a Meal Challenge Test in Patients with Irritable Bowel Syndrome and Healthy Controls. *United European Gastroenterology Journal*, *1*(5), 358–67. http://doi.org/f25c23
62. Shireen L. Rizvi and Matthew K. Nock. (2008). Single-Case Experimental Designs For The Evaluation Of Treatments For Self-Injurious And Suicidal Behaviors. *Suicide & life-threatening behavior*, *38*(5), 498–510. http://doi.org/dpcc9h
63. John Rooksby, Mattias Rost, Alistair Morrison, and Matthew Chalmers Chalmers. (2014). Personal Tracking as Lived Informatics. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2014)*, 1163–1172. http://doi.org/bbdz
64. Jeff Sauro. (2011). Measuring Usability with the System Usability Scale (SUS). http://www.measuringu.com/sus.php
65. SCRT Package - CRAN. http://cran.r-project.org/web/packages/glmnet/index.html
66. William R. Shadish, Larry V. Hedges, and James E. Pustejovsky. (2013). An SPSS Macro for a d-Statistic for Single-Case Designs.
67. Magnus Simrén, Agneta Månsson, Anna Maria Langkilde, Jan Svedlund, Hasse Abrahamsson, Ulf Bengtsson, and Einar S. Björnsson. (2001). Food-Related Gastrointestinal Symptoms in the Irritable Bowel Syndrome. *Digestion*, *63*(2), 108–15. http://www.ncbi.nlm.nih.gov/pubmed/11244249
68. Heidi M. Staudacher, Kevin Whelan, Peter M. Irving, and Miranda C. E. Lomer. (2011). 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 Dietetics*, *24*(5), 487–495. http://doi.org/dhpg8h
69. Hyewon Suh, Nina Shahriaree, Eric B. Hekler, and Julie A. Kientz. (2016). Developing and Validating the User Burden Scale: A Tool for Assessing User Burden in Computing Systems. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2016)*, 3988–3999. http://doi.org/bwt6
70. Trialist - ohmage. https://github.com/ohmage/trialist-front-end
71. TummyTrends. https://itunes.apple.com/us/app/tummy-trends-constipation/id513358882?mt=8