Journal of Healthcare Informatics Research (2019) 3:124–155
https://doi.org/10.1007/s41666-018-0033-x

RESEARCH ARTICLE

A Patient-Centered Proposal for Bayesian Analysis of Self-Experiments for Health

Jessica Schroeder[1] · Ravi Karkar[1] · James Fogarty[1] · Julie A. Kientz[1] · Sean A. Munson[1] · Matthew Kay[2]

Received: 2 December 2017 / Revised: 4 June 2018 / Accepted: 31 August 2018 /
Published online: 25 September 2018

Abstract

The rise of affordable sensors and apps has enabled people to monitor various health indicators via self-tracking. This trend encourages self-experimentation, a subset of self-tracking in which a person systematically explores potential causal relationships to try to answer questions about their health. Although recent research has investigated how to support the data collection necessary for self-experiments, less research has considered the best way to analyze data resulting from these self-experiments. Most tools default to using traditional frequentist methods. However, the US Agency for Healthcare Research and Quality recommends using Bayesian analysis for n-of-1 studies, arguing from a statistical perspective. To develop a complementary patient-centered perspective on the potential benefits of Bayesian analysis, this paper describes types of questions people want to answer via self-experimentation, as informed by (1) our experiences engaging with irritable bowel syndrome patients and their healthcare providers and (2) a survey investigating what questions individuals want to answer about their health and wellness. We provide examples of how those questions might be answered using (1) frequentist null hypothesis significance testing, (2) frequentist estimation, and (3) Bayesian estimation and prediction. We then provide design recommendations for analyses and visualizations that could help people answer and interpret such questions. We find the majority of the questions people want to answer with self-experimentation data are better answered with Bayesian methods than with frequentist methods. Our results therefore provide patient-centered support for the use of Bayesian analysis for n-of-1 studies.

Keywords Self-experiment · N-of-1 · Interface design · User-centered design · Self-tracking · Bayesian analysis

✉ Jessica Schroeder
jesscs@uw.edu

Extended author information available on the last page of the article

1 Introduction

Chronic diseases are the leading cause of sickness, disability, and death worldwide, contributing to 43% of the global burden of disease in 2014 [1]. Technology plays a vital role for patients with chronic conditions, as it allows them to monitor their health markers not just in clinics, but also at home, enabling ongoing self-management of health conditions (e.g., [2]). Another consequence of the rise of such health tracking technology is the newfound abundance of data in the hands of the patients, allowing them to investigate the impact of different factors on their condition (e.g., by how much eating a donut affects their blood sugar levels 2 hours later) [3, 4]. The goals of these self-experiments are often similar those of n-of-1 clinical studies conducted in medicine [5]. However, although self-trackers often want to perform self-experiments to answer questions they have about their health [6], they frequently struggle to design, follow, and interpret scientifically rigorous self-experiments. 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 [6].

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 self-experimentation. Recent work has examined how to help people avoid pitfalls concerning what they should track and when they should track it by introducing tools that support self-experiments in common self-tracking domains [7–9]. However, more work is needed to examine methods for effective analysis of self-experimentation data. Specifically, many self-experimentation tools rely on frequentist methods in their analyses without considering or supporting other kinds of analyses that may be better-suited to the data, easier for patients to interpret, or more relevant to the questions they are trying to answer.

Guidelines provided by the Agency for Healthcare Research and Quality (AHRQ) recommend Bayesian statistics as a more suitable approach for n-of-1 analysis than frequentist analysis [10], arguing largely from a statistical perspective (e.g., evidence from the population can be combined with evidence from a self-experiment to improve individual estimates using Bayesian methods). Others have argued that Bayesian estimation is more appropriate in small-sample settings where frequentist null hypothesis significance testing (NHST) tends to be unreliable [11–13]. As the dominant methodology in many fields, however, designers and researchers often naturally turn to NHST to analyze data.

This research aims to complement the prior statistical perspectives on the possible benefits of Bayesian analysis. We take a patient-centered approach to investigate how different statistical paradigms can support people in using self-experimentation to answer various types of health-related questions they may have. We examine the following research questions: (1) what questions do a lay audience want to answer with their self-experiments? and (2) given a self-experiment, rigorously designed for causal inference, how can designers determine and communicate experimental results to that audience to best answer those questions? Although these questions may have implications for communication of results of correlational analyses common in the broader class of self-tracking tools, we restrict our current scope to self-experimentation.

Drawing upon prior studies, we conducted with patients with irritable bowel syndrome (IBS) and their healthcare providers [9, 14], we identified questions patients wish to answer via self-experimentation. We then surveyed 78 people, asking them about the kinds of questions they would like to answer in a self-experiment. People want to answer a wide variety of questions, including some that can be answered by frequentist methods and many that cannot. Finally, we describe analyses of simulated self-experimentation data using frequentist NHST, frequentist estimation, and Bayesian estimation and prediction to compare their ability to answer several different types of questions that people might have about their self-experimentation data.

2 Background and Related Work

We first present related work in self-tracking, an active area of research in medicine, human-computer interaction, and other communities. We then present related work in self-experimentation, an increasingly popular subset of self-tracking, highlighting its specific nuances and additional requirements as compared to self-tracking more broadly. Finally, we introduce irritable bowel syndrome (IBS) to provide background on the domain in which we present our example analyses and in which our prior work in self-experimentation has been grounded.

2.1 Self-Tracking and Health

Self-tracking is an increasingly common practice of recording and reflecting on information about one's activities, context, and/or outcomes. Self-tracking tools, including self-monitoring applications, can help people understand their habits and behaviors [15]. Given the growing popularity of self-tracking, research has examined how to better design technologies to support these practices. For example, frameworks like the stage-based model for personal informatics [15] and the lived informatics model [16] describe how individuals engage with personal data, including the processes they follow and their motivations for self-tracking. Although both models emphasize the importance of reflecting on self-tracked data and acting upon it, they do not discuss how to analyze and interpret self-tracking data to support such reflection and action. Similarly, although the sense-making framework [17] takes into account the patient's perspective on self-tracking and provides insights into the motivations and needs of patients self-tracking for chronic care, it does not discuss how the collected data should be analyzed.

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) [18]. 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 [6]. Patients with various chronic conditions (e.g., diabetes, irritable bowel syndrome, migraine) often track related data (e.g., glucose, bowel movements, migraine symptoms) using various devices and apps [4, 19–22]. Self-tracking data is increasingly seen as an important contributor to improvements in both medical care and self-management [4].

Existing devices and apps in health-related domains, such as physical fitness (e.g., [23–28]), sleep (e.g., [24, 26, 29]), diet (e.g., [21, 30, 31]), smoking (e.g., [32]), and stress (e.g., [33]), 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, which are often insufficient to answer specific questions people have about relationships between variables.

2.2 Self-Experimentation

Self-experimentation is a subset of self-tracking that aims to help self-trackers more rigorously investigate the questions they have about their health. 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. 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* [34]) 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 [35], emphasizing the importance of personalized health.

Self-tracking methods that do not involve self-experimentation can identify 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 [36, 37]. 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 [5, 38–40]. Both mobile health (mHealth) and human–computer interaction researchers have recently noted the potential for technology to support individuals in conducting and analyzing self-experiments [8], independently or in collaboration with health providers [41].

The framework for self-experimentation in personalized health [8] provides a model for self-experimentation and provides guidelines for designing platforms to run valid self-experiments. Recently, researchers have developed various platforms to support self-experimentation across a range of domains. The personal analytics companion (PACO) helps people experiment with behavior change techniques [42], SleepCoacher identifies connections between potential sleep disruptors and sleep quality [7], Trialist helps patients and clinicians collaborate to find correct medication dosing for chronic pain [43], and TummyTrials helps IBS patients run n-of-1 experiments to determine if a specific food was a trigger for their symptoms [9]. However, nearly all these platforms use a different method for analyzing the self-experiment data, and most of those methods involve a frequentist analysis. Prior work examining n-of-1 studies has conjectured that Bayesian statistics may be beneficial in at least three ways. First, Bayesian statistics could reduce tracking fatigue by using multi-armed bandits, which

allow sampling from fewer treatments by dynamically taking more samples from those treatments which are most likely to be the best [44]. Second, Bayesian statistics may help people better understand the data and make decisions based upon it [44, 45]. Third, because corrections for multiple tests are usually unnecessary in Bayesian estimation,[1] outcomes can be measured in real time, rather than waiting for "enough" data [46, 47]. However, to date, research has not systematically investigated the questions people want to answer with self-tracking data and which types of analyses are best-suited to provide those answers.

2.3 Irritable Bowel Syndrome

We present examples of self-experimentation data analyses within the domain of irritable bowel syndrome (IBS). IBS is a chronic functional disorder characterized by episodic abdominal pain with diarrhea and/or constipation despite normal blood tests, X-rays, and colonoscopies. IBS affects 20% of the US population and is one of the top 10 reason people seek primary care [48, 49]. People with IBS report a lower quality of life and consume 50% more healthcare resources than non-IBS counterparts [50, 51]. IBS symptoms can be triggered by a range of potential factors, including certain foods, eating behaviors, stress, sleep disturbances, and menstruation, with foods as the most common trigger [52, 53]. However, specific food triggers vary across individuals [54–56]. Systematic identification of each patient's triggers traditionally involves an onerous, multiple-month elimination diet, in which a patient eliminates most foods and then slowly reintroduces them [57]. Instead of going through such a burdensome and unpleasant process, people with IBS often choose to track foods and symptoms to attempt to determine their specific triggers [58]. However, identifying triggers from food and symptom journals is difficult and error-prone [59], often lacking rigor in tracking design and analysis [60]. IBS is therefore a useful domain for understanding the potential for self-experimentation, as a 2-week self-experiment on a single possible food trigger is likely to be less burdensome than an elimination diet, but more rigorous than correlational self-tracking. In addition, because severe IBS symptoms usually occur within 4 hours of consuming a triggering nutrient [61], experiments can be designed to explicitly account for the dynamics of symptom occurrence after a given behavior (e.g., by having people avoid eating anything other than the experimental meal within the 4-h window and report symptoms after the window).

In this paper, we take a bottom-up approach to investigate the following: (1) what types of questions do people want to answer using self-experimentation data? and (2) given a self-experiment, rigorously designed for causal inference, how can designers and developers best analyze and communicate those results so that a lay audience to best answer those questions? In particular, we compare how effectively the questions identified in (1) can be answered by frequentist NHST, frequentist estimation, or Bayesian estimation and prediction. To provide a specific example, we use a self-experiment performed to examine an individual's IBS symptoms and triggers, but the designs we present could apply to many different types of questions people might want to answer about their health using self-experimentation data.

[1] So long as one makes use of informed priors (as we advocate here) and/or applies a hierarchical modeling approach.

3 Questions that Self-Experimenters Want to Answer

To determine what types of questions people want to answer through self-experimentation, we first re-examined qualitative data from two previous studies to identify different types of questions participants want to answer with data related to their condition. We then surveyed 78 people to more closely examine the types of questions a more general population would want to answer with data from a self-experiment.

3.1 Qualitative Analysis of Prior Studies

Two of our prior studies drew our attention to a disconnect between the questions that NHST *can* answer and the questions that people *want* to answer with their data. The first study involved 10 patients with IBS and 10 providers collaboratively interpreting interactive visualizations of the patient's food and symptom data to help generate hypotheses about which foods may trigger the patient's symptoms [14]. The second study consisted of 15 people with IBS rigorously testing their hypotheses with the assistance of a phone app that had been developed to support people through a low-burden self-experiment [9]. Both studies involved frequentist analyses performed to help people identify triggers. Only during the interviews did we discover a disconnect between the question those analyses could answer and the questions participants wanted to answer to improve their health and quality of life.

Specifically, participants in those studies wanted more detailed understandings of the relationship between potential triggers and symptoms than could be supported by the analyses we had conducted. For our current research, we re-analyzed interview data from these two studies to identify questions participants sought to answer. The interviews include both people with IBS and health providers. We refer to quotes from the first study as "s1-patient" and "s1-provider" and from the second study as "s2-patient."

3.1.1 By How Much Do My Symptoms Change When I Consume the Nutrient?

When informed that a nutrient was correlated with a change in symptoms, patients often wanted to know the quantity of that symptom change. As one patient put it, "When you say improving, how much improvement?" (s1-patient 9). People wanted to *quantify* the difference a nutrient made in their symptoms instead of just learning that a difference existed. Having a detailed understanding of how a nutrient affects symptoms could help people perform cost/benefit analyses to decide whether they want to consume the nutrient. One participant explained how she thought about that tradeoff: "is being more awake worth potentially having stomachache? Which matters more to me at this particular moment?" (s2-patient 8). Such a decision would be easier if people could quantify how a trigger affected a symptom.

Health providers were particularly interested in quantifying differences because it influenced the advice they might give. If the difference in symptoms between having and avoiding the nutrient was low, they did not feel the need to recommend the patient avoid the trigger, despite a significant p value. When examining a nutrient that had been found significant, one provider explained, "I bet it's correlated, but clinically, it doesn't make any difference. That's what my interpretation is" (s1-provider 4). Similarly, a second provider proclaimed, "I know there was significance statistically, but based on

this it's hard to say that fat has a strong association one way or the other. It seems like the symptoms and the amount of fat are all over the place," later commenting, "even if the data is saying it's significant, if it's not helpful in the real-world then what good is it?" (s1-provider 6). Despite the *statistical* significance, the low magnitude and high variance of the effect of the nutrient lead the provider to consider its *clinical* significance limited.

3.1.2 How Much of the Nutrient Is Associated with Increased or Decreased Symptoms?

Patients and providers also wanted to understand how much of the trigger the patient could eat before they saw an impact on their symptoms. For example, one patient wanted to know what specific amount of caffeine would help her avoid symptoms, saying she wished the analyses showed "not just that having caffeine improved it, but having a moderate amount of caffeine or one cup of it, however you convey that […] Like fiber, maybe having zero fiber is not as good as having five grams of fiber per meal. Maybe 'target five grams' is different than 'have more'" (s1-patient 8). Another participant described how she would like to know if she could have *any* of the nutrient without experiencing symptoms: "if there is a threshold, I would say, 'I would just stay beneath the threshold and not be the weird person who has to drink decaf'" (s2-patient 8). Providers also wanted to understand how much of the nutrient was associated with changed symptoms because they did not want to advise patients to eat more of a nutrient than was practical. Once provider said that "ideally, we would […] quantify it a little bit more to say moderate caffeine associated with no symptoms," because she believed "at higher levels of caffeine, [the patient is] just going to have diarrhea" (s1-provider 8). Simply indicating that a nutrient improved a symptom was not nuanced enough for her to give clinical advice with which she was comfortable.

3.2 Survey Method

Although our qualitative data gave us a sense of the types of questions people with IBS and their providers wanted to answer, we wanted to investigate whether people in a more general population would have similar types of questions about their own health. We therefore developed a list of nine question types we expected people might want to use self-experimentation to answer (Table 1).

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 survey asked participants to name an aspect of their life they wanted to examine (i.e., a dependent variable) and something they thought affected that aspect of their life

Table 1 Types of questions we expected a self-experimenter might wish to ask, with short-form names. The values of [independent variable] and [dependent variable] were populated dynamically in the survey for each respondent based on some aspect of their life they stated they wanted to examine

Question	Short-form name
Q1: Does [independent variable] have any effect on my [dependent variable]?	Any effect
Q2: Does [independent variable] have a noticeable impact on my [dependent variable]?	Noticeable effect
Q3: Do different things in combination with [independent variable] affect the change in [dependent variable]?	Interaction effect
Q4: How does [independent variable] affect my [dependent variable] differently depending on the time of day?	Temporal effect
Q5: How much [independent variable] is needed to see an impact on my [dependent variable]?	Threshold for effect
Q6: By how much does my [dependent variable] change with different amounts of [independent variable]?	Varying effect
Q7: What will my [dependent variable] be like in the future if I avoid [independent variable]?	Avoidance prediction
Q8: What will my [dependent variable] be like in the future after my normal amount of [independent variable]?	Normal prediction
Q9: What will my [dependent variable] be like in the future after more than my normal amount of [independent variable]?	Excess prediction

(i.e., an independent variable). To avoid biasing participants with our own question list, we asked participants to write what question they would be most interested in answering with those two variables before showing them our list. We then asked them to rate how useful they thought the answers to the nine questions we had developed would be for them on a 7-point Likert item from "very useless" to "very useful". For each question rated *somewhat useful*, *useful*, or *very useful*, we asked how long they would be willing to self-track to answer that question. We recruited through social media, university mailing lists, and posts to the Quantified Self Facebook group.

3.3 Data Analysis

We received 78 responses to the survey (49 female, 27 male, 2 preferred not to say). Participant ages ranged from 20 to 64 (mean = 34). Participants wanted to investigate a wide range of dependent and independent variables, including chronic illnesses, chronic pain, stress and anxiety, diet, sleep, and fitness and weight loss.

We used Bayesian regression to analyze *usefulness* ratings (Fig. 1a), the number of *days* people were willing to track to answer a question (Fig. 1b), the proportion of participants willing to *track indefinitely* (Fig. 2a), and the proportion of people saying they *already know* the answer to a question (Fig. 2b). For each outcome variable, we fit a generalized linear mixed model, with *participant* and *question* as random intercepts (i.e., we employed partial pooling to regularize estimates, which helps avoid overfitting and makes our estimates of differences between conditions conservative). The model for each outcome variable differed only in terms of the assumed response distribution.

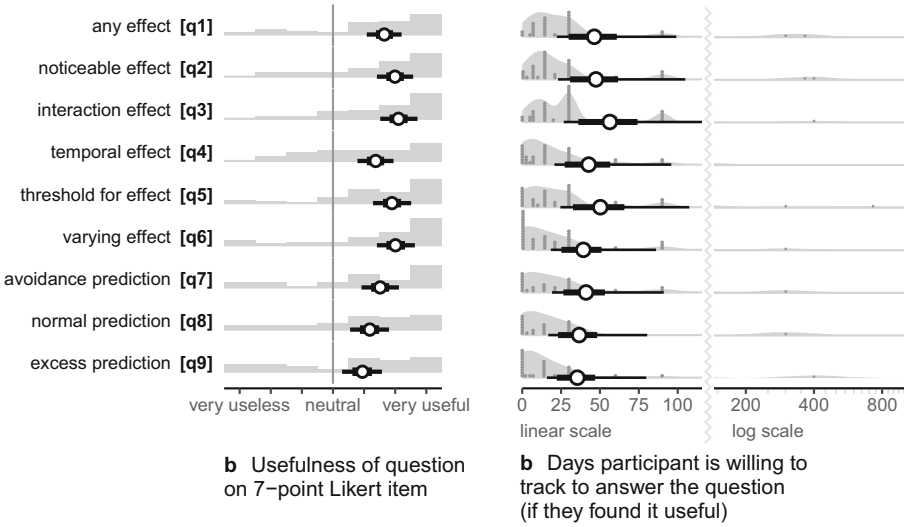

Fig. 1 Data (light gray) and population-level estimates (with 66 and 95% quantile credible intervals) from the survey. One data point is omitted from the plots (*tracking days* = 730 for Q5: *threshold for effect*) for space

For example, because we asked about *usefulness* on a 7-point Likert scale, we analyzed *usefulness* by modeling responses as coming from a latent conditional normal distribution rounded to the nearest value in {0,1,2,3,4,5,6} (using an interval-censored normal response model). We analyzed *days* using a negative binomial regression (appropriate for a count outcome variable), and *already knows* and *tracks indefinitely* using logistic regression (appropriate for a binary outcome variable). We marginalized out (i.e., averaged over) participant effects to report population-level estimates of the mean number of days (or proportions). Due to the use of random effects, these estimates will not exactly match the means or proportions of the survey data itself.

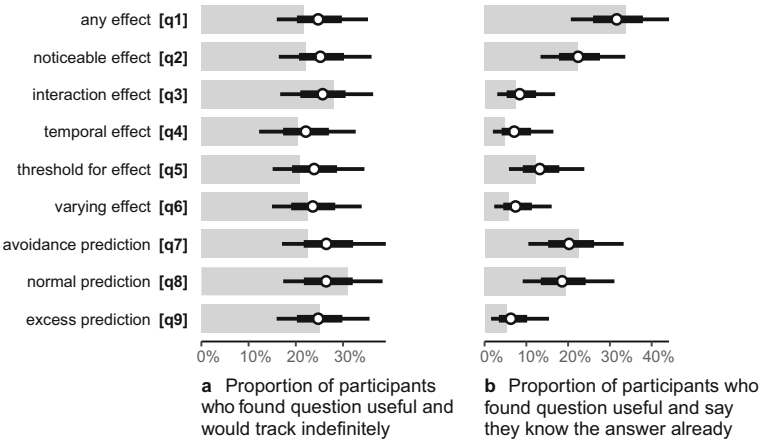

Fig. 2 Data (light gray) and posterior estimates (with 66 and 95% quantile credible intervals) from the survey

To analyze the qualitative results, two researchers coded the independent and dependent variables that participants said they wanted to investigate and the specific questions they said they were most interested in answering in a free-response text box at the start of the survey (Fig. 1). Any differences were resolved between the two researchers.

3.4 Survey Results

In general, participants rated the nine question types as approximately equally useful (Fig. 1a). They also indicated a willingness to spend approximately the same amount of time self-tracking to answer all nine questions (Fig. 1b).

About an equal number of people were also willing to track indefinitely for each question (Fig. 2a). However, the proportion of people who indicated that they already knew the answer to a question, given that they would find the answer useful, differs considerably between some questions (Fig. 2b).

In particular, Q1 (*any effect*) and Q2 (*noticeable effect*) were considerably more likely to be rated "already known" than any other question, and Q3 (*interaction effect*), Q4 (*temporal effect*), Q6 (*varying effect*), and Q9 (*excess prediction*) were least likely to be "already known" (Fig. 2b). Considering "least likely to be known" (low values on Fig. 2b) and "most likely to be useful" (high values on Fig. 1a) together, it would seem that Q3 (*interaction effect*) and Q6 (*varying effect*) might be particularly valuable questions to be able to answer.

When asked what question they would most like to answer with the variables they specified before viewing our question list, 41 participants specified questions that did not match the nine questions we developed. The majority of these miscellaneous questions (40/41) were not testable questions: they either expressed the need for more general information (e.g., P16 asked "What constitutes exercise?"), the need to form a hypothesis about the dependent variable, or the desire to self-track without performing an experiment. Of the questions that did fit our question types, those similar to Q2 (*noticeable effect*) and Q6 (*varying effect*) had the highest number of occurrences among these unprompted questions (Fig. 3).

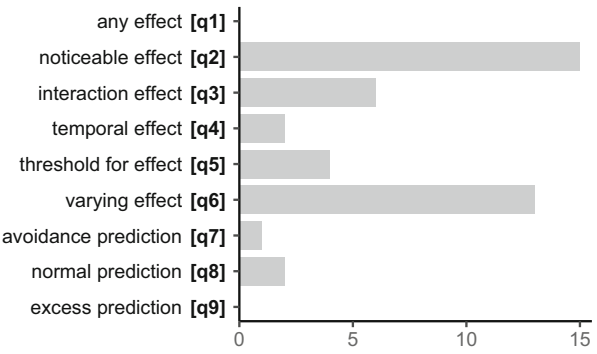

Fig. 3 Survey participants ($n = 78$) were asked to write a question to which they would be most interested in finding an answer. This graph shows the categorization of those questions into the nine survey template questions. Forty participants proposed questions which were not testable using an experiment (e.g., P43 wrote "what kinds of foods are safe to eat?"). One participant proposed a question that did not fall in any of our nine template questions

After we had determined the types of questions people would like to answer with self-experimentation data, we wanted to examine how different types of analyses could support answering those questions. Below, we describe frequentist and Bayesian analyses, then present example analyses that illustrate how those analyses address or fail to address the types of questions our survey participants want to ask.

4 Frequentist vs Bayesian Analyses of Self-Experimentation Data

In this section, we describe different statistical methods people might consider when analyzing data from a self-experiment to answer the questions they want to explore. To anchor our discussion on statistical methods, we consider a hypothetical self-experiment examining the effect of caffeine on IBS symptoms.

Scenario: Imagine a mobile app, such as that described by Karkar et al. [8, 9], which uses a single-subject fully randomized alternating treatment experimental design. The app randomly assigns a person to drink or not drink caffeine on each day then measures abdominal pain on a self-reported scale from 0 to 6. Thus, our predictor is caffeine consumption and our outcome is abdominal pain. Our hypothetical person might record 2 weeks of data, or 14 data points: 7 days on which they drank caffeine and 7 days on which they did not. We explore how such data may be analyzed to answer the nine different types of questions from our survey (Table 1) using frequentist NHST, frequentist estimation, and Bayesian analysis. For each approach, we review questions that can be answered with the approach, problems the approach introduces, and questions the approach cannot answer. Some question types described in the survey require more data than is collected in this scenario: the ability to help people understand interacting factors (i.e., Q3 (*interaction effect*)), differences based on time of day (i.e., Q4 (*temporal effect*)), or differences depending on the amount of the tested trigger (i.e., Q5 (*threshold for effect*)) all require more information than is collected in this proposed self-experiment design. We discuss these question types in Sect. 6.2.

4.1 Frequentist NHST

NHST performs binary inference: is there an effect of caffeine consumption on stomach pain? Is caffeine a *trigger*? These hypothesis testing questions best match Q1 (*any effect*) from our survey.

4.1.1 Answerable Questions with Frequentist NHST

Under frequentist NHST, we cannot directly produce evidence for caffeine being a trigger (e.g., $\boldsymbol{P}(\boldsymbol{caffeine\ is\ a\ trigger}|\ \boldsymbol{data})$). Instead, we must consider the *null hypothesis*: that caffeine is not a trigger. For example, assume we saw a mean increase of 2 points on the abdominal pain scale when drinking caffeine. We could ask: assuming caffeine is *not* a trigger, what is the probability we would see an increase in abdominal pain of 2 or more points when drinking caffeine (e.g., $\boldsymbol{P}(\boldsymbol{data}|\ \boldsymbol{caffeine\ is}$ **not** $\boldsymbol{a\ trigger})$)? The answer to this question is a p value. If this p value is low, the

reasoning goes, we should be skeptical that caffeine consumption does *not* have an effect (i.e., that the null hypothesis is true). Typically, "low" is defined by some cutoff, such as $p < 0.05$, below which we reject the null hypothesis and declare that caffeine is a trigger.[2] Frequentist NHST therefore can provide an answer to Q1 (*any effect*) from our survey.

4.1.2 Limitations with Frequentist NHST for n-of-1 Studies

Although the prospect of definitively answering this binary question using a short self-experiment is attractive, numerous practical issues arise. First and foremost, *binary inference is noisy in small samples* [11, 12, 62]. Unless caffeine is so strong a trigger that we can expect it to reliably increase pain every single time we drink it—and the measurement error in our pain scale is very low—our short experiment with only a few samples per condition has a high probability of failing to detect that caffeine is trigger. If caffeine consumption happens by chance to have no strong effect on abdominal pain on even 1 or 2 days out of the seven that the self-experimenter drinks it, we would be unlikely to detect the effect. If the experiment *does* detect that caffeine is trigger, the effect size (e.g., mean increase in pain) will likely be overestimated. Often in small studies an effect *must* be an overestimate to have been declared statistically significant; this is called a *magnitude error* [12]. With such a small sample, a *sign error* is also reasonably likely: the self-experimenter could have an unusually good set of days with caffeine, and on days without caffeine could have high abdominal pain due to a confounding factor (e.g., stress). In such a case, NHST might lead them to conclude that caffeine consumption *decreases* their symptoms even though it actually *increases* their symptoms. Studies with < 10% power (i.e., < 10% chance that the study will detect a present effect) have sign error rates from ~ 10 to 50% [10].

Traditional scientific fields solve this problem in one of two ways: (1) ensure studies have high power (a sample size large enough to reliably detect an effect of the size we care about) and (2) run many studies and combine their estimates in a meta-analysis. Neither approach seems feasible in self-experimentation. The first option is infeasible because extending the self-experiment defeats the goal of conducting a low-burden, short-term experiment. Similarly, because IBS triggers are personalized, having more participants collect data to build a population-wide understanding is also not a viable option.

In addition to being noisy in small samples, p values are not in the language of outcomes: as the American Statistical Association (ASA) explains, p values do not directly translate onto the magnitude of the effect [63]. p values therefore also *do not facilitate cost-benefit analysis* (e.g., Q6 (*varying effect*) in our survey). As the participants in our previous studies remarked (see Sect. 3.1), cost-benefit analysis can be critical to self-experimenters. Even knowing that caffeine is a trigger, the self-experimenter might want to decide if it is likely to be a large enough trigger for them to want to change their behavior. In other words, they might also want to consider the *effect size* and its *uncertainty*: perhaps the effect is present but mostly likely small, so

[2] The widespread use of this *null ritual* in scientific fields is not without criticism [78]. Most pointedly, Gigerenzer went so far as to declare it a symptom of "mindless statistics" [69]. We will describe why we believe it is not applicable to small self-experiments but leave aside the question of its broader applicability to science.

sometimes they might decide to drink caffeine anyway. To be empowered to make effective decisions, the self-experimenter needs a more nuanced understanding of caffeine's effects. The ASA notes, "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold." [63]. We believe the same is true for individual decisions about health.

Another limitation of frequentist NHST is that people often misunderstand the difference between failing to reject the null hypothesis, which is what happens if the *p* value is not below the specified threshold, and accepting the null hypothesis, which requires a different test [64]. For example, if someone performs a self-experiment to investigate whether caffeine causes their abdominal pain, analyzes the resulting data using frequentist NHST, and finds a *p* value above 0.5, they cannot conclude that caffeine *does not cause* their abdominal pain. They can only conclude that the experiment *failed to provide evidence* that caffeine causes their abdominal pain.

4.2 Frequentist Estimation

Other questions that people were interested in answering (e.g., questions involving quantifying relationships between the trigger and the symptom, such as Q5 (*threshold for effect*) and Q6 (*varying effect*) in our survey) might be addressed using frequentist estimation. For example, one could perform a regression to estimate the mean difference in self-reported pain when consuming caffeine compared to when avoiding caffeine.[3] However, that mean difference fails to convey any sense of uncertainty: although it gives an *average* effect size, it does not communicate how large or small the effect size *could* reasonably be. Confidence intervals are a frequentist procedure often used to try to address this question.

4.2.1 Answerable Questions with Frequentist Estimation

Frequentist confidence intervals can be defined in terms of a *confidence procedure*. A *confidence procedure* is a procedure that can be used to create a *confidence interval* for a parameter (e.g., a mean) in a given sample. An *X*% (e.g., 95%) *confidence procedure* is a procedure that, when used to construct confidence intervals in repeated samples, will generate a set of confidence intervals such that X% (e.g., 95%) of those intervals contain the true value of the parameter [65]. This idea is also called *coverage*: X% (e.g., 95%) of the confidence intervals *cover* the true parameter. Frequentist estimation is therefore well-suited to error control: if we have a problem in which we do not care about the particular value of the parameter in any one sample, but want to guarantee that over all of our samples we correctly estimate an interval containing the parameter X% (e.g., 95%) of the time, confidence intervals are well-suited to this task. This property makes confidence intervals applicable to industrial applications (e.g., where samples may be batches in a manufacturing process and it is desirable to ensure that estimates of properties of those batches are wrong only 1-X% (e.g., 5%) of the time).

[3] Readers familiar with standardized effect sizes (like Cohen's *d*) might ask why we do not use them here. Like Cummings [79], we believe that unstandardized effect sizes (e.g., mean differences) are easier to interpret, particularly for individual decision-making (a person should know what one point on a pain scale *that they have used* means to them; they are less likely to know what a difference of 1 standard deviation means).

4.2.2 Limitations with Frequentist Estimation

Although these long-run characteristics of confidence intervals (i.e., characteristics that describe *sets* of confidence intervals from *repeated* trials, rather than any one particular trial) make them useful in processes where many samples are observed over time, confidence intervals are less useful for making inferences about any particular sample. As Morey et al. commented [65], confidence intervals are more about pre-data error control than post-data inference about the sample at hand. A given X% confidence interval cannot be said to contain the true parameter with an X% "probability," "plausibility," or even "confidence"—such a statement would be a post-data inference about a particular interval and not a statement about the set of intervals generated by the procedure [65, 66].

We might dismiss this misinterpretation of confidence intervals as a statistician's quibble. However, even if we do dismiss it and interpret confidence intervals as statements of the likely location of the mean, confidence intervals are often very large in small studies, diminishing their usefulness. For example, a self-experimenter might estimate a 95% confidence interval that coffee increases their abdominal pain between 0 and 4 points on a 6-point scale, which is unlikely to help them make decisions. In the frequentist setting, to narrow that interval, the self-experimenter would have to perform a larger experiment—otherwise, more precise quantification of the relationships between the trigger and symptom is impossible with frequentist methods.

4.3 Bayesian Estimation and Prediction

The issue with confidence intervals described above stems from the fact they are derived from the probability of seeing the data at hand given a possible mean difference (in our example, this probability would be denoted as $\boldsymbol{P}(\boldsymbol{data}|\ \boldsymbol{mean\ increase\ in\ pain})$; this is called the *likelihood*) [46]. However, what we want to estimate and communicate (and how confidence intervals are often misinterpreted) is the probability the mean increase in pain is a particular value given the data we have observed (e.g., $\boldsymbol{P}(\boldsymbol{mean\ increase\ in\ pain}|\ \boldsymbol{data})$). For example, what is the chance that the mean increase in pain on days when a participant consumes coffee is 1 or more points on their pain scale? What is the chance that increase is 2 or more points?

4.3.1 Answerable Questions with Bayesian Estimation and Prediction

Bayesian analysis allows us to derive probability statements like that described above, as long as we also have a *prior* (i.e., a probability distribution that describes our knowledge of plausible values of the parameter before running the experiment). In this example, the prior would be $\boldsymbol{P}(\boldsymbol{mean\ increase\ in\ pain})$, a probability distribution describing plausible values of the mean increase in pain before running the experiment. We can then derive the probability distribution for the mean difference in pain after observing the data in the experiment, called the *posterior*, using Bayes' rule:

$$P(\textit{mean increase in pain}|\textit{data}) \propto P(\textit{data}|\textit{mean increase in pain}) \cdot P(\textit{mean increase in pain})$$

$$(\text{More generally}: \textit{posterior} \propto \textit{likelihood} \cdot \textit{prior})$$

Incorporating prior knowledge into an individual's self-experimentation model allows the model to start with a reasonable set of assumptions, rather than assuming each mean increase is equally probable a priori. In small-*n* experiments, doing so reduces estimation error [13]. Making use of prior knowledge is especially attractive in a domain like health, where population data (such as the proportion of IBS sufferers for whom caffeine is a trigger, or how strong a trigger it tends to be) might be available to form effective prior knowledge. Combining prior population-level estimates with data from a small self-experiment could help people make more effective use of their individual data, while *also* allowing people to interpret results as probabilistic statements.

For example, using the posterior, we can calculate a 95% *credible* interval, which unlike a 95% confidence interval, *is* an interval containing the parameter with 95% probability, conditional on the prior and the data [65]. We can also answer other probabilistic questions. For example, we can answer "what is the probability the mean pain increase is at least *X*?" and "What is the most likely value of the mean pain increase?" (i.e., Q6 (*varying effect*)). In addition, given a full Bayesian model, we can use posteriors to make probabilistic predictions: statements about new observations instead of just means. We do this by constructing a *posterior predictive* distribution [67], which is the predicted distribution of new responses, marginalizing over ("averaging out") the posterior distribution of the parameters in the model (e.g., averaging over the estimated mean and standard deviation of increase in pain). In this example, the posterior predictive distribution accounts for our uncertainty in the mean and standard deviation of the increase in pain, and for the uncertainty inherent in taking a new observation given the estimated mean and standard deviation, to give us ***P*(*increase in pain in one new observation*| *data***). We can then answer "what is the probability of experiencing abdominal pain of 4 points or more if I drink caffeine today?" (i.e., Q7 (*avoidance prediction*), Q8 (*normal prediction*), and Q9 (*excess prediction*) from our survey). Such predictions facilitate cost/benefit analyses because the self-experimenter can decide whether having caffeine is worth the risk of extra pain.

Finally, if the self-experimenter is looking to answer binary inference questions (i.e., Q1 (*any effect*) and Q2 (*noticeable effect*)), they can still do so. One could define a *trigger* as a food that causes a mean increase above some meaningful threshold of pain (e.g., 1 point on the scale—or whatever the *individual* considers personally significant) and then calculate:

$$P(\textit{mean increase in pain} > \textit{threshold} \mid \textit{data})$$

Based on that result, they could decide whether they consider the food a trigger (or even, whether they consider the probability that it is a trigger high enough to warrant cutting it from their diet—again, a personal cost/benefit analysis).

4.3.2 Limitations with Bayesian Estimation and Prediction in n-of-1 Studies

One problem with Bayesian methods is the requirement to find a prior that accurately summarizes the current state of knowledge. As the results of the analysis are

conditional on the prior, a poorly suited prior may result in inaccurate results. How to select priors for Bayesian analyses of self-experiments remains an open question. In Sect. 5.4.4, we discuss three possible prior types, but each has challenges. For example, designers could use a prior based on knowledge of the population, but such a prior requires existing population-level literature on the phenomenon. Alternatively, a designer might select a prior based on data from other people using a self-experimentation tool, but doing so requires such data to have been collected. A different prior may therefore be needed to bootstrap the analysis. A system could also elicit the prior from self-experimenter themselves, but personalized prior elicitation requires designing a process to help people quantify their personal beliefs for use in a prior—though recent work suggests this challenge is not insurmountable, particularly with the use of graphical elicitation methods [68]. We believe these challenges represent promising avenues for future work.

Another problem with Bayesian estimation and prediction is that people may be less familiar with those methods than with frequentist methods. Some people find p values provide a sense of scientific rigor [9]. p values also provide a better sense of an "answer"[4]; it provides a single number one compares to a threshold to interpret experimental results (despite the fact that the commonly defined threshold of 0.05 is arbitrary and under contention [69, 70]). Using Bayesian methods therefore may require more education about what the results mean for those familiar with frequentist methods.

5 Example Analyses of Self-Experimentation Data

Based on our self-experiment scenario investigating if caffeine consumption affects abdominal pain, we developed different analyses and representations that could help people interpret the results of their experiments. Analyses and representations were created using Stan [71] and R. We use these tools to propose possible designs that could help people interpret personal informatics data, envisioning the implementation of these designs in future self-tracking tools. For example, a future implementation of a self-experimentation tool such as TummyTrials [9] could include these designs, rather than its current approach of providing a p value. Such an integration would be relatively simple; as the quantity of data in the proposed self-experiments is small, few computational resources are required to perform the analyses, and the algorithms for doing so have already been developed. In case of higher-than-expected computation requirements, the analysis could be done in the cloud. To determine what analyses and representations might be helpful to answer the questions people have about their health, we examine how frequentist and Bayesian approaches can be used to analyze the data.

[4] We do not discuss the use of Bayes factors—one approach to Bayesian hypothesis testing—in this paper, as the sensitivity of Bayes factors to irrelevant details of the prior make them difficult even for experienced analysts to use in practice [80]. Instead, if hypothesis testing is desired, we prefer estimation-based approaches, such as regions of practical equivalence, which we believe are also easier to interpret. Regions of practical equivalence answer questions like "how likely is the effect to be 0 (or close enough to 0 that I will not care)?" [46, 80].

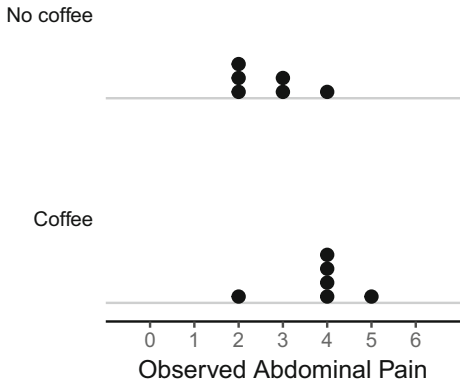

Fig. 4 Simulated data from a self-experiment used for our example analyses

5.1 Sample Dataset

Based on our work on self-experimentation in IBS, we generated a simulated dataset relating caffeine consumption to abdominal pain. We chose this data from several simulated datasets generated by a member of the research team to resemble data collected in a previous study. The particular dataset was chosen for our example because it did not show a 'clear' effect based on visually inspecting the data, a common approach for analyzing self-experiment data [8]. Caffeine consumption was binary (yes or no) and abdominal pain was rated on a 7-point Likert item from 0 (no pain) to 6 (extreme pain). The self-experiment was 12 days long, yielding 6 days or data points with caffeine and 6 without. The dataset we use in the following analyses is illustrated in Fig. 4.

5.2 Frequentist NHST

Because the scale we used for abdominal pain is limited to {0,1,2,3,4,5,6}, we again used an interval-censored normal regression model to analyze our simulated data. With this analysis, people could use frequentist NHST to estimate a two-sided p value against the null hypothesis that the mean difference in pain is 0 (yielding $p = 0.08$). This p value fails to reject the null at the customary $\alpha = 0.05$. A charitable reading of NHST would say that a person gets a kind of an answer for Q1 (*any effect*) from our survey: that the data failed to provide evidence for the effect, and is thus *inconclusive*, given our 0.05 threshold. An important note is that the logic of NHST does *not* allow us to conclude that the effect is 0 (i.e., we cannot use this test to *accept* the null hypothesis; we can only fail to reject it).

5.3 Frequentist Estimation

Despite being unable to find a statistically significant increase in pain, one might use frequentist estimation[5] to produce confidence intervals to estimate the mean pain in each

[5] We used a variant of our Bayesian regression model with flat priors (i.e., priors in which all possible outcomes are equally likely, which is the implicit assumption a frequentist analysis makes) on the parameters to simulate the frequentist regression.

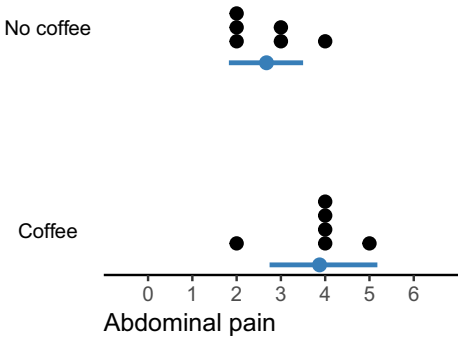

Fig. 5 Ninety-five percent confidence intervals of the means of each condition (blue line) given the raw data (black dots)

condition (Fig. 5). One could also generate a confidence interval for the mean difference between conditions (Fig. 6).

That the 95% confidence interval for this difference overlaps 0 reflects the finding in the previous section that p is greater than 0.05. From this confidence interval, a person might conclude there is a mean difference in pain of between a little less than 0 and almost 3 or estimate their mean pain level when they drink coffee as between about 3 and 5—*if* they interpret the confidence interval as a statement of probability. As noted above, confidence intervals are *not* probable regions containing the mean. From these results, the person therefore cannot determine a probability distribution of the mean difference or make probabilistic predictions of future stomach pain.

5.4 Bayesian Estimation

In addition to analyzing the data via the frequentist approach, we performed Bayesian regression (using the same type of rounded-normal response model), with three different priors. These regressions illustrate the effect a prior belief can have when analyzing self-experimentation data. In *large* samples, the statistician's quibble that frequentist confidence intervals cannot be interpreted as Bayesian credible intervals is arguably a distinction without a difference—the large amount of data often outweighs most reasonable priors, so the intervals overlap almost perfectly. There, interpreting a confidence interval as an approximation to a Bayesian interval is more defensible. However, in *small* samples like these, what we believe prior to the experiment has a much greater effect on what our posterior intervals look like. Indeed, with frequentist results we have had the experience that people will disagree with evidence from small self-experiments, citing their prior

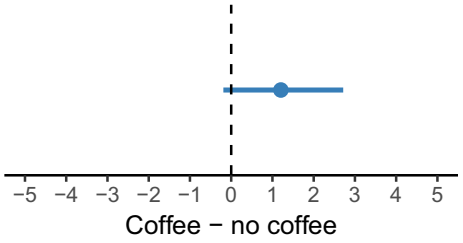

Fig. 6 Ninety-five percent confidence interval of the mean difference in abdominal pain the self-experimenter experiences between consuming and avoiding caffeine

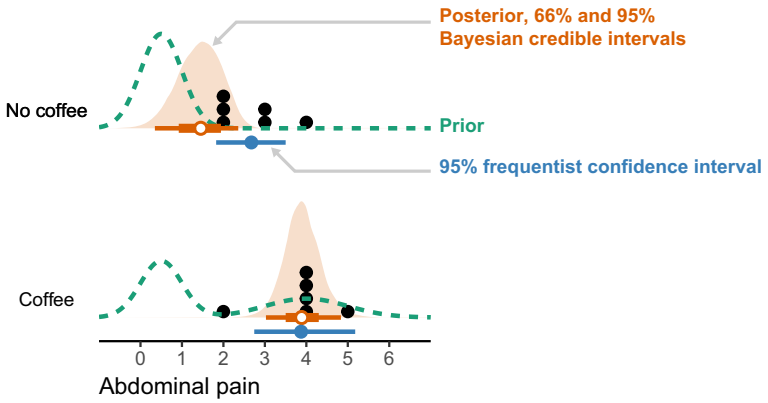

Fig. 7 Bayesian analysis of the raw data (black dots). The informed population prior assumes no abdominal pain when coffee is avoided (green dotted line, top), and abdominal pain for some people when coffee is consumed (green dotted line, bottom). The posterior distributions (orange filled area), and corresponding orange 95 and 66% Bayesian credible intervals, show the likely means for the individual in the two conditions

beliefs and the small amount of data—perhaps rightly so [9, 14]. With a Bayesian approach, we can incorporate those prior beliefs into the analysis to produce more reasonable and believable probability intervals. We fit the model using three different priors to demonstrate the impact prior knowledge can have on conclusions from small self-experiments.

5.4.1 Informed Population Prior

The first prior we used is an informed population prior: one in which without coffee, most people experience low abdominal pain, and where with coffee, a small portion of the population experiences increased pain. This bimodal prior represents one possible prior for this analysis. The Bayesian analysis combines this prior with the data collected from the participant to form a posterior distribution describing the likely means for *coffee* and *no coffee* (prior is dotted green, posterior is filled orange, frequentist confidence interval is shown for reference; Fig. 7). As with the frequentist analyses, one can also calculate intervals for the estimated mean difference in abdominal pain between consuming and avoiding caffeine (Fig. 8).

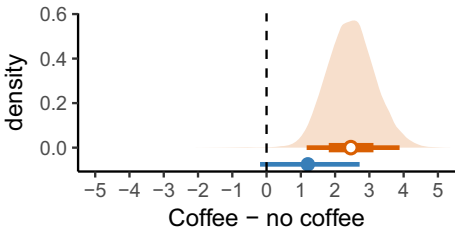

Fig. 8 Estimated mean difference in abdominal pain between consuming and avoiding coffee, using 95% frequentist confidence intervals (blue line) and 95% and 66% Bayesian credible intervals (orange line) with an informed population prior

In contrast to frequentist *confidence* intervals, however, these Bayesian *credible* intervals can be interpreted as describing the probable location of the mean (or mean difference).

Because the population prior assigns low probability to high baseline pain (in *no coffee*), the posterior for *no coffee* is shrunk towards zero—the Bayesian credible interval in the *no coffee* condition is closer to the prior (and therefore zero) than the frequentist interval. This bias towards zero is because this study has only a small number of observations: the model and prior suggest we would need more evidence to be convinced that this person's pain in the *no coffee* condition is normally as high as was observed. These results therefore suggest that, conditional on believing pain in *no coffee* is low a priori, we should conclude that the observations of higher pain in *no coffee* during the study most likely happened due to chance. Similarly, where in the frequentist model observing a handful of high values under the *coffee* condition results in a 95% confidence interval that covers about half the scale (providing little precision), the Bayesian model combines these results with our prior knowledge of the bimodal nature of the population to infer that this person is likely among the second "hump" that corresponds to people with a higher pain response to caffeine. Thus, the Bayesian intervals for the *coffee* condition are narrower than the frequentist one (Fig. 7).

Bayesian credible intervals can help people answer Q2 (*noticeable effect*) and Q6 (*varying effect*) from our survey by showing how much their abdominal pain changes between consuming and avoiding caffeine. People could also use the information in the more informative posterior distributions (and posterior predictive distributions) in cost/benefit analyses between consuming and avoiding caffeine. For example, we can use the Bayesian model to create posterior predictive distributions of *future* abdominal pain by randomly sampling from the response distribution, conditional on the posterior mean and standard deviation in each condition. We calculated predictive distributions for drinking and avoiding caffeine and discretized the predictions into a 20-dot quantile dotplot (Fig. 9). A quantile dotplot is a continuous analog to an icon array (commonly used in medical risk communication [72]), and can be thought of as depicting n (here 20) equally likely predicted outcomes. Evidence suggests quantile dotplots may yield better estimates and decisions in lay populations than visualizations of continuous distributions, such as density plots [73, 74].

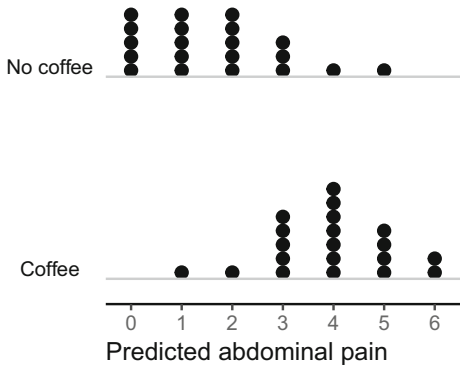

Fig. 9 Quantile dotplots of predictive distributions of *future* abdominal pain if the self-experimenter avoided (top) or consumed (bottom) coffee, calculated using Bayesian posterior prediction with an informed population prior. Each plot shows 20 approximately equally likely predicted outcomes

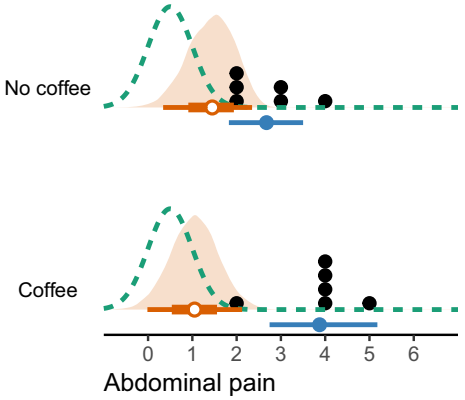

Fig. 10 Bayesian analysis of the raw data (black dots). The skeptical prior assumes no abdominal pain, regardless of whether coffee is avoided or consumed (green dotted lines). The posterior distributions (orange filled area), and corresponding orange 95% and 66% credible intervals, show likely means for the individual in the two conditions

For example, Fig. 9 shows that abdominal pain at level 4 or more is predicted to occur 2 times out of 20 when caffeine is avoided (or 10% of the time), but the same level of pain is predicted to occur 13 times out of 20 when caffeine is consumed (65% of the time). With this dotplot, people could therefore predict their abdominal pain when they consumed or avoided caffeine, thus allowing them to answer Q7 (*avoidance prediction*), Q8 (*normal prediction*), and Q9 (*excess prediction*) from our survey.

5.4.2 Skeptical Prior

Instead of using an informed population prior, a person that doubts they are sensitive to coffee before running this experiment may want to use a skeptical prior. Such skepticism might be expressed by placing priors near 0 pain in both the *coffee* and *no coffee* conditions. We can re-run the same Bayesian model with those priors (Fig. 10) and again calculate intervals for the estimated mean difference in abdominal pain between consuming and avoiding caffeine (Fig. 11). Now the results suggest no large difference between *coffee* and *no coffee*—there being few observations in both cases, the effect of the prior pulls both estimated means towards 0, causing the estimate of the mean difference also to pull towards 0. We can say that, *given* the patient's skepticism that there is an effect prior to their self-experiment, the evidence in the experiment is not enough that they should abandon this belief. This prior therefore yields an opposing conclusion to the informed population prior, despite using identical data.

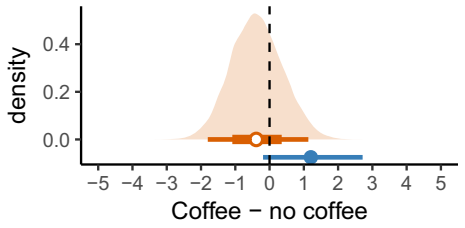

Fig. 11 Estimated mean difference in abdominal pain between consuming and avoiding coffee, using 95% frequentist confidence intervals (blue line) and 95 and 66% Bayesian credible intervals (orange line) with a skeptical prior

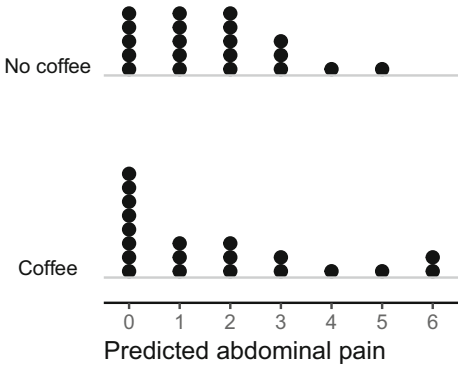

Fig. 12 Quantile dotplots of predictive distributions of *future* abdominal pain if the self-experimenter avoided (top) or consumed (bottom) coffee, calculated using Bayesian posterior prediction with a skeptical prior. Each plot shows 20 approximately equally likely predicted outcomes

We can also again plot a quantile dotplot of predictions to view predicted future abdominal pain if the self-experimenter avoided or consumed coffee (Fig. 12). We see relatively similar predictions of future pain in both cases, with perhaps a slightly higher chance of very high pain if coffee is consumed.

5.4.3 Optimistic Prior

Finally, we imagine a third scenario, where our prior belief (perhaps based on the patient's own self-tracking data) is that the patient has higher baseline pain than 0 in *no coffee* (say around 2), and also is likely to have some increased pain in the *coffee* condition (say around 5). Given such a prior, the results are illustrated in (Fig. 13). In this case, the priors have the effect of focusing our posterior beliefs into regions that are more credible a priori: the Bayesian credible intervals overlap the frequentist intervals, but exclude regions deemed less probable according to the prior. This narrowing occurs in both the estimates of the mean in each condition (Fig. 13) and the estimates of the mean difference (Fig. 14).

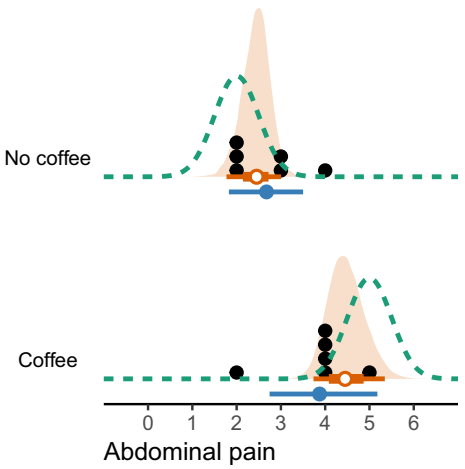

Fig. 13 Bayesian analysis of the raw data (black dots). The optimistic prior assumes minimal abdominal pain when coffee is avoided (green dotted line, top), and high abdominal pain when coffee is consumed (green dotted line, bottom). The posterior distributions (orange filled area), and corresponding orange 95% and 66% credible intervals, show the likely means for the individual in the two conditions

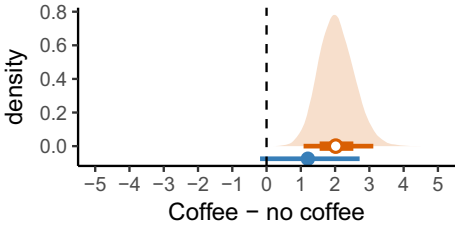

Fig. 14 Estimated mean difference in abdominal pain between consuming and avoiding coffee, using 95% frequentist confidence intervals (blue line) and 95% and 66% Bayesian credible intervals (orange line) with an optimistic prior

This effective combination of prior and data leads to much more precise estimates, as our posterior predictions reflect (Fig. 15). As in the informed population prior, we find that abdominal pain at level 4 or more is predicted to occur 2 times out of 20 when caffeine is avoided (or 10% of the time), but pain at level 4 or more is now predicted to occur 16 times out of 20 when caffeine is consumed (80% of the time).

We can now examine the results of the Bayesian analyses using our three different priors together to facilitate further comparison (Fig. 16).

5.4.4 Choosing an Appropriate Prior

As we have illustrated, one of the main advantages of using Bayesian models is the ability to specify informed priors (e.g., based on prior knowledge of the population or knowledge about the specific individual). Bayesian analyses based on uninformed priors are often equivalent to the analogous frequentist analyses, with the same limitations. However, exactly what information to use to derive an informed prior for a self-experiment may be difficult to determine. We describe three possible sources for defining a prior and discuss how they relate to the example priors we explored (Fig. 16).

Prior Based on Global Population The use of informed priors based on a larger population is an attractive prospect for medical self-experimentation, where existing population studies could provide excellent priors. These priors may even deviate from a unimodal distribution, as in our example: a reasonable hypothesis is that the true distribution of abdominal pain given caffeine consumption is *bimodal*, as the prior

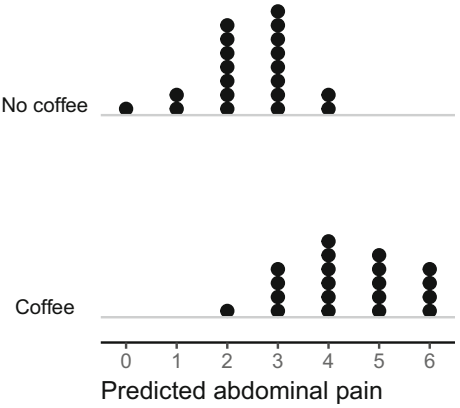

Fig. 15 Quantile dotplots of predictive distributions of *future* abdominal pain if the self-experimenter avoided (top) or consumed (bottom) coffee, calculated using Bayesian posterior prediction with an optimistic prior. Each plot shows 20 approximately equally likely predicted outcomes

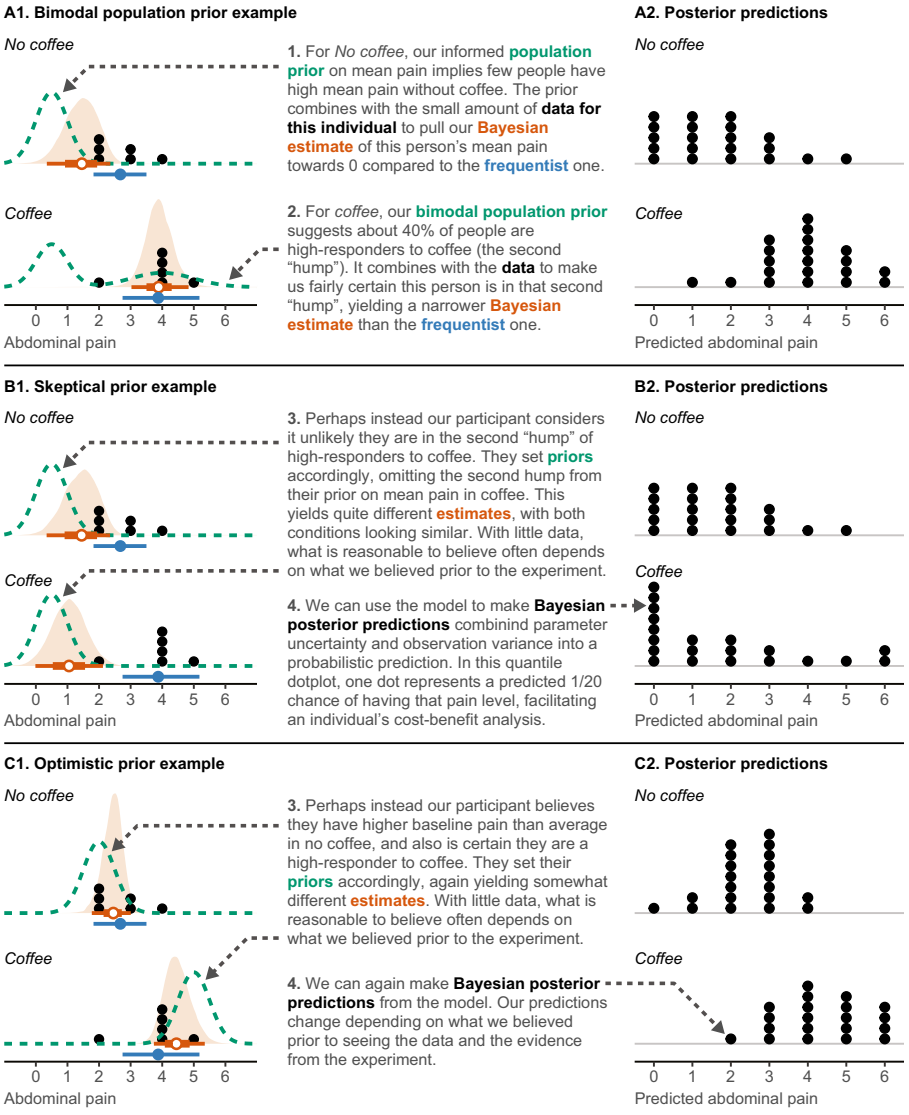

Fig. 16 Comparison of the analysis results when using the three different priors described above

we chose for the "informed population prior" (Fig. 16a). For a large percentage of the population, abdominal pain after caffeine consumption might be a distribution with low variance and a mean around zero, and for a small percentage of the population (i.e., the percentage with IBS wherein they get abdominal pain after consuming caffeine), abdominal pain after consuming caffeine is a distribution with higher variance and a mean higher than zero.

Conditional Prior Based on Tool Use, Demographics, and/or Medical History Another reasonable theory is that people who want to perform certain self-experiments are systematically different from the global population. Although a global population prior

could be drawn from the literature, population data could also be aggregated into priors by the tool itself, based on data from people who use the tool. Conditional priors (e.g., based on demographics or medical history) could also to improve estimates. The tool might derive a slightly different prior for each new person automatically, depending on the literature, a person's demographics, and data already collected on that phenomenon by people who use the tool. Such a prior may still be bimodal, like the one we chose for the "informed population prior" (Fig. 16a), but it would likely have slightly different distributions than that of the global population.

Patient-Specified Prior A third option is for the tool not to make assumptions for every new patient, but to instead let the patients themselves have a say in the prior. Someone who doubts that caffeine is causing their abdominal pain, but wants to formally test the relationship, could indicate their belief that they belong to the population that does not experience abdominal pain after consuming caffeine, as illustrated in our skeptical prior example (Fig. 16b). On the other hand, someone who is fairly sure that caffeine is a trigger may want to incorporate that belief as a prior, which matches our optimistic prior (Fig. 16c). This elicitation process could be mixed initiative, with the tool suggesting relevant population or conditional priors, and the person adjusting the priors to match their beliefs. A patient's priors might also be informed by correlational analyses they have conducted using self-tracking tools. Not only would such a prior take advantage of information people already know about themselves, but it could also mitigate confirmation bias by giving people a principled way to *update* their prior belief based on the data, rather than presenting a result independent of any prior beliefs they have (as in a frequentist approach).

6 Discussion

We discuss design implications of using Bayesian methods to analyze self-experimentation data, including the difficulties around eliciting specific, measurable questions people have about their health; the design of self-experiments to provide the data necessary to answer those questions; the necessity of supporting actionability for self-experiments for health; and the perceived credibility and interpretability of Bayesian analyses.

6.1 Eliciting Specific, Measurable Questions People Have About Their Health

Our survey revealed that people are not always good at stating a testable (or estimable) question they want to answer about their health. For example, P43 wanted to know what kinds of foods were safe to eat. Neither frequentist nor Bayesian analysis can provide analysis for such a vague question. Answering that question in a self-experiment would first require forming a definition of "safe" and "unsafe," hypothesizing what foods *might* be causing their symptoms (i.e., are therefore "unsafe"), and then designing a self-experiment to evaluate that hypothesis (i.e., to estimate the effects of those foods). Because our survey suggests that people tend to start with a high-level

question that cannot be directly translated into a self-experiment, tools aimed at supporting self-experimentation must guide people through expressing what general questions they have about their health and then helping them state those questions in a specific, measurable format. This guidance may be in the form of a step-by-step wizard which assists in eliciting the necessary independent and dependent variables to design a self-experiment based on the abstract question a person might have [9]. In some cases, a recommendation system could help by suggesting questions that people with similar condition have asked in the past (e.g., suggesting someone with IBS ask the same questions that others with IBS have asked). However, as self-experiments for health are inherently personal, different individuals will want to investigate different aspects of their health, and some may even want to purposefully remain ignorant about the answers to some possible questions [9]. Systems created to support self-experiments for health will need to have functionality to support the generation of measurable questions, either through community outsourcing, collaboration with health providers, automated methods, or some combination of methods.

6.2 Designing Self-Experiments Based on the Questions People Have About Their Health

Designing an appropriate self-experiment with which people could collect and analyze the necessary data to answer that question can be difficult, even when starting from a concrete, measurable question someone wants to answer. For example, although severe IBS symptoms often occur within a 4 hour window of exposure to a dietary trigger [61], some people might want to investigate other conditions that do not have such a clear relation between a trigger and symptom (e.g., the effect of exercise on sleep quality). Such an investigation would likely require domain expertise (e.g., guidance from a health provider, guidance directly built into the self-experimentation app as in [9]).

In addition, although the combination of the experimental design we used as an example in this paper and Bayesian estimation can address the majority of the types of questions people in our survey had about their health, the example experimental design does not address question types 3–5. These question types require different data than is collected by our example self-experiment. To find an interaction effect between multiple factors (i.e., Q3 (*interaction effect*)), people would need to systematically experiment with and track their exposure to all of those factors. To help people understand differences based on time of day (i.e., Q4 (*temporal effect*)), people would have to further restrict their meals so they could experiment throughout the day, rather than just experimenting with breakfast. Finally, to determine differences depending on the amount of the tested trigger (i.e., Q5 (*threshold for effect*)), people would need to note the quantity of the possible trigger they were exposed to, rather than just indicating whether or not they were exposed. Although these kinds of self-experiments are possible, they would be more complicated, longer, and more burdensome than those described by Karkar et al. [8, 9] and used as an example in this paper. However, supposing the existence of an app that could support those kinds of self-experiments, the necessary analyses are straightforward extensions of the

Bayesian regression framework. For example, Bayesian regression can be naturally extended to handle effects that vary at different time scales (e.g., days, weeks, months) by using techniques like Gaussian process regression [67] or Bayesian structural time series [75]. However complicated the model, the Bayesian posterior predictive distribution remains well-defined, allowing the display of probabilistic predictions of future outcomes regardless of the specific model used.

6.3 Acting upon Self-Experiments for Health

Integrating our proposed analyses and visualizations into self-tracking tools could help people interpret their data and better answer questions about their health. Such analyses of a self-experiment may provide a concrete answer to a person's question. However, a person's health goals often include not just learning the answers to their questions, but also making changes based on those answers [8]. The next step in an individual's overall process may therefore be changing their behavior. Additional opportunities remain in considering how a person uses the information they gain from a self-experiment to improve their health or quality of life. Going back to our example scenario, if caffeine is a trigger for a person's abdominal pain, a first recommendation may be to give up caffeine. However, as we discovered in our interviews with IBS patients [9], many factors could prevent someone from taking this step (e.g., they may need it to stay awake during their job). In such a scenario, it is helpful to think of the "result" of the self-experiment as empowering someone to make an informed decision. If one cannot give up caffeine completely, is there a certain threshold up to which they are willing to trade off their need for caffeine and the severity of their abdominal pain? Are there times when the caffeine and risk of resulting pain are worth it to them and times when it is not? To make these sorts of decisions, people need results that help them anticipate the range of possible outcomes for the different choices they can make.

6.4 Perceived Credibility of Bayesian Analysis

We have demonstrated that Bayesian estimation could be advantageous in the analysis of self-experimentation data, both from a statistical and a patient-centered viewpoint. However, some open questions must be answered to determine how to conduct such analyses in practice. For example, for people to trust the results of an experiment, people must believe the experiment and analysis are sufficiently rigorous. In a prior examination of self-experimentation, some people felt p values served as an indicator of scientific rigor and thus the credibility of the results [9]. Although this opinion is a misconception, it nevertheless raises an important question: do common presentations of Bayesian analyses project similar rigor to people? If not, how can they be adapted to similarly communicate rigor and credibility to people who hope to act on their results? The result of a self-experiment could influence behavior changes (e.g., someone with IBS might stop drinking coffee if caffeine is triggering symptoms), but it is unlikely to do so if the self-experimenter does not trust the results.

6.5 Interpretability of Bayesian Analyses

A related issue is that some people may have trouble interpreting posterior distributions from a Bayesian analysis [68]. To ensure people can understand and act upon results, analyses need to be presented in ways that that people without a strong background in statistics can correctly interpret. We believe that framing results in terms of predicting future outcomes—rather than as posterior distributions or credible intervals—may help. Doing so would make the display of self-experimentation results amenable to discrete outcome presentations, such as the quantile dotplots [73] we used (e.g., in Fig. 8) or icon arrays that have been found to be effective in medical risk communication [76]. When evaluated in other domains, representations such as these (e.g., those which communicate a range of possible outcomes) can help people make better decisions and increase their trust in the system [74, 77].

7 Conclusion

Prior research has argued that Bayesian analyses are well suited for the small sample sizes people are likely to generate via self-experimentation, in part because Bayesian estimates tend to have lower error than traditional frequentist analyses [10]. In addition to those recommendations, we contribute a patient-centered examination of Bayesian analyses, finding 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. Through probabilistic predictions of future outcomes, Bayesian analysis offers a potential to enable patients to conduct personal, actionable cost-benefit analyses. By contrast, frequentist NHST can only answer *whether* an independent variable has an effect on a dependent variable, and frequentist estimation results in confidence intervals that are often less informative and harder to interpret than the analogous Bayesian credible intervals. For myriad statistical and patient-centered reasons, Bayesian estimation is therefore a superior method for analyzing self-experiments compared to frequentist NHST and should be investigated further for use in self-experimentation tools.

Acknowledgements We thank Eric B. Heckler and Roger Vilardaga for conversations that informed this research.

Funding information This research was funded in part by a University of Washington Innovation Research Award, the National Science Foundation under awards IIS-1553167 and SCH-1344613, and the Agency for Healthcare Research Quality under award 1R21HS023654.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

References

1. Global Status Report on Noncommunicable Diseases. Geneva: World Health Organization; 2014
2. Mamykina L, Mynatt ED, Kaufman DR (2006) Investigating health management practices of individuals with diabetes. Proc SIGCHI Conf Hum Factors Comput Syst - CHI '06. :927
3. Riggare S, Unruh KT, Sturr J, Domingos J (2017) Patient-driven n-of-1 in Parkinson's disease. 123–8
4. Mamykina L, Heitkemper EM, Smaldone AM, Kukafka R, Cole-Lewis HJ, Davidson PG, Mynatt ED, Cassells A, Tobin JN, Hripcsak G (2017) Personal discovery in diabetes self-management: discovering cause and effect using self-monitoring data. J Biomed Inform. 76(June):1–8
5. Cepeda MS, Acevedo JC, Hernando A, Miranda N, Cortes C, Carr DB (2008) An n-of-1 trial as an aid to decision-making prior to implanting a permanent spinal cord stimulator. Pain Med (United States) 9(2): 235–239
6. Choe EK, Lee NB, Lee B, Pratt W, Kientz JA (2014) Understanding quantified-selfers' practices in collecting and exploring personal data. In: Proc ACM Conf Hum Factors Comput Syst (CHI 2014). New York, New York, USA; p. 1143–52
7. Nediyana D, Metaxa-Kakavouli D, Tran A, Nugent N, Boergers J, McGeary J, Huang J (2016) SleepCoacher: a personalized automated self-experimentation system for sleep recommendations. In: Proc ACM Symp User Interface Softw Technol (UIST 2016). p. 347–58
8. Karkar R, Zia JK, Vilardaga R, Mishra SR, Fogarty J, Munson SA, Kientz JA (2016) A framework for self-experimentation in personalized health. J Am Med Informatics Assoc. 23(3):440–448
9. Karkar R, Schroeder J, Epstein DA, Pina LR, Scofield J, Fogarty J, Kientz JA, Munson SA, Vilardaga R, Zia JK (2017) TummyTrials: a feasibility study of using self-experimentation to detect individualized food triggers. In: Proc ACM Conf Hum Factors Comput Syst (CHI 2017). p. 6850–63
10. Kravitz RL, Duan MSPHN (2014) Panel De.M.C.N.-1 G. Design and implementation of n-of-1 trials: a user's guide. Agency Healthc Res Qual 13(14):1–88
11. Gelman A, Weakliem D (2008) Of beauty, sex, and power: statistical challenges in estimating small effects. Am Sci 97(4):310–316
12. Gelman A, Carlin J (2014) Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspect Psychol Sci. 9(6):641–651
13. Kay M, Nelson GL, Hekler EB (2016) Researcher-centered design of statistics: why Bayesian statistics better fit the culture and incentives of HCI. Proc 2016 CHI Conf Hum Factors Comput Syst. :4521–32
14. Schroeder J, Hoffswell J, Chung C-F, Fogarty J, Munson S, Zia JK (2017) Supporting patient-provider collaboration to identify individual triggers using food and symptom journals. Proc 2017 ACM Conf Comput Support Coop Work Soc Comput - CSCW '17. :1726–39
15. Li I, Dey AK, Forlizzi J (2010) A stage-based model of personal informatics systems. In: Proc ACM Conf Hum Factors Comput Syst (CHI 2010). New York, New York, USA; p. 557–66
16. Epstein DA, Ping A, Fogarty J, Munson SA (2015) A lived informatics model of personal informatics. In: Proc ACM Int Jt Conf Pervasive Ubiquitous Comput (UbiComp 2015). p. 731–42
17. Mamykina L, Smaldone AM, Bakken SR (2015) Adopting the sensemaking perspective for chronic disease self-management. J Biomed Inform. 56:406–417
18. Rooksby J, Rost M, Morrison A, Chalmers MC (2014) Personal tracking as lived informatics. In: Proc ACM Conf Hum Factors Comput Syst (CHI 2014). New York, New York, USA; p. 1163–72
19. Chung C-F, Cook J, Bales E, Zia JK, Munson SA (2015) More than telemonitoring: health provider use and nonuse of life-log data in irritable bowel syndrome and weight management. J Med Internet Res 17(8):e203
20. Park SY, Chen Y (2015) Individual and social recognition: challenges and opportunities in migraine management. In: Proc ACM Conf Comput Support Coop Work Soc Comput. ACM Press, New York, USA, pp 1540–1551
21. Mamykina L, Mynatt E, Davidson P, Greenblatt D (2008) MAHI: investigation of social scaffolding for reflective thinking in diabetes management. In: Proc SIGCHI Conf Hum Factors Comput Syst (CHI 2008). p. 477–86
22. Schroeder J, Chung C-F, Epstein DA, Karkar R, Parsons A, Murinova N, Fogarty J, Munson SA (2018) Examining self-tracking by people with migraine: goals, needs, and opportunities in a chronic health condition. In: Proc ACM Conf Des Interact Syst (DIS 2018) To Appear. https://doi.org/10.1145/3196709.3196738
23. Consolvo S, McDonald DW, Toscos T, Chen MY, Froehlich JE, Harrison BL, Klasnja P, La Marca A, Le Grand L, Libby R, Smith IE, Landay JA (2008) Activity sensing in the wild: a field trial of Ubifit Garden. In: Proc ACM Conf Hum Factors Comput Syst (CHI 2008). p. 1797–806

24. Fitbit [Internet]
25. Jawbone UpBand [Internet]
26. Larklife [Internet]
27. Lin J.J., Mamykina L, Lindtner S, Delajoux G, Strub HB (2006) Fish'n'Steps: encouraging physical activity with an interactive computer game. Ubiquitous Comput (UbiComp 2006). 261–78
28. Nike Fuelband [Internet]
29. Kay M, Choe EK, Shepherd J, Greenstein B, Watson NF, Consolvo S, Kientz JA (2012) Lullaby: a capture & access system for understanding the sleep environment. In: Proc ACM Conf Ubiquitous Comput (UbiComp 2012). p. 226–34
30. Baumer EPS, Katz SJ, Freeman JE, Adams P, Gonzales AL, Pollak J, Retelny D, Niederdeppe J, Olson CM, Gay GK (2012) Prescriptive persuasion and open-ended social awareness: expanding the design space of mobile health. In: Proc ACM Conf Comput Support Coop Work (CSCW 2012). p. 475–84
31. Cordeiro F, Bales E, Cherry E, Fogarty J (2015) Rethinking the mobile food journal: exploring opportunities for lightweight photo-based capture. In: Proc ACM Conf Hum Factors Comput Syst (CHI 2015). p. 3207–16
32. Ali AA, Hossain SM, Hovsepian K, Plarre K, Kumar S (2012) mPuff: automated detection of cigarette smoking puffs from respiration measurements. In: Proc Conf Inf Process Sens Networks (ISPN 2012). p. 269–80
33. Morris M, Guilak F (2009) Mobile heart health: project highlight. IEEE Pervasive Comput. 8(2):57–61
34. Jorgensen JT (2009) New era of personalized medicine: a 10-year anniversary. Oncologist. 14(5):557–558
35. Swan M (2009) Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. Int J Environ Res Public Health. 6(2):492–525
36. Lillie EO, Patay B, Diamant J, Issell B, Topol EJ, Schork NJ (2011) The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? Per Med 8(2):161–173
37. Riley WT, Glasgow RE, Etheredge L, Abernethy AP (2013) Rapid, responsive, relevant (r3) research: a call for a rapid learning health research enterprise. Clin Transl Med 2(1):10
38. Barlow DH, Hayes SC (1979) Alternating treatments design: one strategy for comparing the effects of two treatments in a single subject. J Appl Behav Anal. 12(2):199–210
39. Larson EB (1990) N-of-1 clinical trials: a technique for improving medical therapeutics. West J Med 152(1):52–56
40. Barlow DH, Nock MK, Hersen M (2008) Single case experimental designs: strategies for studying behavior change. Third. Pearson; 416
41. Barr C, Marois M, Sim I, Schmid CH, Wilsey B, Ward D, Duan N, Hays RD, Selsky J, Servadio J, Schwartz M, Dsouza C, Dhammi N, Holt Z, Baquero V, MacDonald S, Jerant A, Sprinkle R, Kravitz RL (2015) 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
42. PACO: The Personal Analytics Companion [Internet]
43. Tiralist - ohmage [Internet]
44. Daskalova N, Desingh K, Kim JY, Zhang L, Papoutsaki A, Huang J (2017) Lessons learned from two cohorts of personal informatics self-experiments. In: Proc ACM Conf Ubiquitous Comput. p. 46
45. Lee J, Walker E, Burleson W, Kay M, Buman M, Hekler EB (2017) Self-experimentation for behavior change: design and formative evaluation of two approaches. In: Proc SIGCHI Conf Hum Factors Comput Syst. p. 6837–49
46. Kruschke JK, Liddell TM (2017) The Bayesian new statistics : hypothesis testing, estimation, meta-analysis, and planning from a Bayesian perspective. Psychon Bull Rev. :1–29
47. Gelman A, Hill J, Yajima M (2012) Why we (usually) don't have to worry about multiple comparisons. J Res Educ Eff 5(2):189–211. https://doi.org/10.1080/19345747.2011.618213
48. Elsenbruch S (2011) Abdominal pain in irritable bowel syndrome: a review of putative psychological, neural and neuro-immune mechanisms. Brain Behav Immun. 25(3):386–394
49. Lovell RM, Ford AC ((2012)) Effect of gender on prevalence of irritable bowel syndrome in the community: systematic review and meta-analysis. Am J Gastroenterol. 107:991–1000
50. Ladabaum U, Boyd E, Zhao WK, Mannalithara A, Sharabidze A, Singh G, Chung E, Levin TR (2012) Diagnosis, comorbidities, and management of irritable bowel syndrome in patients in a large health maintenance organization. Clin Gastroenterol Hepatol. 10(1):37–45

51. Mitra D, Davis KL, Baran RW (2011) All-cause healthcare charges among managed care patients with constipation and comorbid irritable bowel syndrome. Postgrad Med. 123(3):122–132
52. Harris LR, Roberts L (2008) Treatments for irritable bowel syndrome: patients' attitudes and acceptability. BMC Complement Altern Med. 8:65
53. Heitkemper M, Carter E, Ameen V, Olden K, Cheng L (2002) Women with irritable bowel syndrome: differences in patients' and physicians' perceptions. Gastroenterol Nurs 25(5):192–200
54. Monsbakken K, Vandvik P, Farup P (2006) Perceived food intolerance in subjects with irritable bowel syndrome—etiology, prevalence and consequences. Eur J Clin Nutr 60(5):667–672
55. Simrén M, Månsson A, Langkilde AM, Svedlund J, Abrahamsson H, Bengtsson U, Björnsson ES (2001) Food-related gastrointestinal symptoms in the irritable bowel syndrome. Digestion 63(2):108–115
56. Zia JK, Barney P, Cain KC, Jarrett ME, Heitkemper MM (2016) A comprehensive self-management irritable bowel syndrome program produces sustainable changes in behavior after 1 year. Clin Gastroenterol Hepatol 14(2):212–219
57. Parker TJ, Naylor SJ, Riordan AM, Hunter JO (1995) Management of patients with food intolerance in irritable bowel syndrome: the development and use of an exclusion diet. J Hum Nutr Diet 8(3):159–166
58. American Gastroenterological Association. American Gastroenterological Association Medical Position Statement: Irritable Bowel Syndrome. Vol. 123, Gastroenterology. American Gastroenterology Association; p. 2105–72002
59. Zia JK, Chung C-F, Xu K, Dong Y, Cain KC, Munson SA, Heitkemper MM Inter-rater reliability of healthcare provider interpretations of food and gastrointestinal symptom paper diaries of patients with irritable bowel syndrome. :In Preparation
60. Choe EK, Duarte ME, Kientz JA (2010) Understanding and designing computing technologies that convey concerning health news. In: Proc Int Conf Des Emot (D&E 2010). p. 1–12
61. Eswaran S, Tack J, Chey WD (2011) Food: the forgotten factor in the irritable bowel syndrome. Gastroenterol Clin N Am 40(1):141–162
62. Loken E, Gelman A (2017) Measurement error and the replication crisis. Science (80-). 355(6325):584–585
63. Wasserstein RL, Lazar NA (2016) The ASA's statement on p -values: context, process, and purpose. Am Stat. 70(2):129–133
64. Walker E, Nowacki AS (2011) Understanding equivalence and noninferiority testing. J Gen Intern Med 26(2):192–196. https://doi.org/10.1007/s11606-010-1513-8
65. Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers E-J (2016) The fallacy of placing confidence in confidence intervals. Psychon Bull Rev 23(1):103–123
66. Hoekstra R, Morey RD, Rouder JN, Wagenmakers E-J (2014) Robust misinterpretation of confidence intervals. Psychon Bull Rev. 21(5):1157–1164
67. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis. Third Edit. Chapman and Hall/CRC; 675 p
68. Goldstein DG, Rothschild D (2014) Lay understanding of probability distributions. J Soc Judgm Decis Mak 9(1):1–14
69. Gigerenzer G (2004) Mindless statistics. J Socio Econ. 33(5):587–606
70. Benjamin D.J., Berger J.O., Johannesson M., Nosek B.A., Wagenmakers E.-J., Berk R., Bollen K.A., Brembs B., Johnson V.E., et al. (2017) Redefine statistical significance. Nat Hum Behav.
71. Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker MA, Li P, Riddell A (2016) Stan: a probabilistic programming language. J Stat Softw. 76(1)
72. Ancker JS, Senathrajah Y, Kukafka R, Starren JB (2006) Design features of graphs in health risk communication : a systematic review. J Am Med Informatics Assoc 13(6):608–619. https://doi.org/10.1197/jamia.M2115.Introduction
73. Kay M, Kola T, Hullman JR, Munson SA (2016) When(ish) is my bus?: user-centered visualizations of uncertainty in everyday, mobile predictive systems. Proc ACM Conf Hum Factors Comput Syst (CHI 2016). 5092–103
74. Fernandes M, Walls L, Munson S, Hullman J, Kay M (2018) Uncertainty displays using quantile dotplots or CDFs improve transit decision-making. In: Proc ACM Conf Hum Factors Comput Syst (CHI 2018). p. To Appear
75. Scott SL, Varian HR (2014) Predicting the present with Bayesian structural time series. Int J Math Model Numer Optim. 5(1/2). doi:https://doi.org/10.1504/IJMMNO.2014.059942
76. Garcia-Retamero R, Cokely ET (2013) Communicating health risks with visual aids. Curr Dir Psychol Sci. 22(5):392–399

77. Jung MF, Sirkin D, Gür TM, Steinert M (2015) Displayed uncertainty improves driving experience and behavior. Proc 33rd Annu ACM Conf Hum Factors Comput Syst - CHI '15. (April):2201–10
78. McShane BB, Gal D, Gelman A, Robert C, Tackett JL (2017) Abandon statistical significance. 1–12
79. Cummings P (2011) Arguments for and against standardized mean differences (effect sizes). Arch Pediatr Adolesc Med. 165(7):592–596
80. Betancourt M (2018) Calibrating model-based inferences and decisions. 1–35

Affiliations

Jessica Schroeder[1] · Ravi Karkar[1] · James Fogarty[1] · Julie A. Kientz[1] · Sean A. Munson[1] · Matthew Kay[2]

Ravi Karkar
rkarkar@uw.edu

James Fogarty
jaf1978@uw.edu

Julie A. Kientz
jkientz@uw.edu

Sean A. Munson
smunson@uw.edu

Matthew Kay
mjskay@umich.edu

[1] University of Washington, Seattle, WA, USA

[2] University of Michigan, Ann Arbor, MI, USA