Bayesian Analysis of Self-Tracked Health Data: A Technology Probe Study Examining Opportunities in Integration and Reflection

Shaan Chopra*
schopra7@cs.washington.edu
University of Washington
Seattle, Washington, USA

Alex Okeson*
amokeson@gmail.com
University of Washington
Seattle, Washington, USA

Yasaman S. Sefidgar
einsian@cs.washington.edu
University of Washington
Seattle, Washington, USA

James Fogarty
jfogarty@cs.washington.edu
University of Washington
Seattle, Washington, USA

Sean A. Munson
smunson@uw.edu
University of Washington
Seattle, Washington, USA

Abstract

People self-track for health, often to answer questions using their data. However, heterogeneity in tracking goals, data, and questions makes it challenging to support effective integration and reflection. Extending theoretical considerations around Bayesian modeling of trigger and symptom relationships in self-tracked health data, we empirically investigated self-tracker experiences using Bayesian analysis to examine their real-world data around a range of questions. We interviewed and observed 8 participants reflecting on their self-tracked data using a technology probe that applied Bayesian analysis. Participants successfully reflected on trigger and symptom relationships in their heterogeneous data, including answering specific and evolving questions. We also observed breakdowns due to incorrect or partial use of probe capabilities, misunderstanding of Bayesian analyses and conclusions, and misalignment with participant goals. We discuss considerations for designing Bayesian analysis to support exploration and reflection on relationships across a range of health goals, data, and questions.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Interactive systems and tools**.

Keywords

personal informatics; health tracking; bayesian analysis; technology probe; qualitative research

ACM Reference Format:
Shaan Chopra, Alex Okeson, Yasaman S. Sefidgar, James Fogarty, and Sean A. Munson. 2026. Bayesian Analysis of Self-Tracked Health Data: A Technology Probe Study Examining Opportunities in Integration and Reflection. In *Interactive Health Conference (IH '26), July 05–08, 2026, Porto, Portugal.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3786579.3804970

*These authors contributed equally to this work

IH '26, Porto, Portugal

ACM ISBN 979-8-4007-2422-0/2026/07
https://doi.org/10.1145/3786579.3804970

1 Introduction and Related Work

Many people seek to answer questions about their health by engaging in self-tracking, including for sleep [15, 26], activity [12, 32, 42], diet [2, 13, 14, 33], menstrual and reproductive health [8, 21], and chronic conditions [25, 41]. Personal informatics research provides frameworks for characterizing self-tracking journeys [22, 31, 34] and for supporting goals [36, 41]. Such frameworks emphasize reflection as a key need in understanding and acting on self-tracked data, but analyzing and interpreting collected data remains difficult [3, 17], due in part to the range of goals people bring to tracking, heterogeneity in data they collect, and different questions they want to ask over that data [3]. Such challenges mean that available tools often fail to support answering questions and providing insights toward health decisions and self-experimentation [7, 15, 25, 38].

Challenges of interpreting self-tracked health data have motivated consideration of Bayesian methods. For example, in contrast to likelihood-based methods, prior research argues Bayesian analysis (i) can utilize either population-level or individualized prior information [16], (ii) can work with less data collected [16], and (iii) might be more natural for clinicians to understand in decision making [29]. In response to such interest from the medical community, prior HCI research has investigated Bayesian methods for modeling relationships between potential triggers and symptoms in self-experimentation data [1, 38]. This included identifying common questions and providing design recommendations for Bayesian analyses that could support interpreting self-tracked data relative to these questions [38]. However, proposed approaches were theoretical and had not been examined with self-trackers, their real-world goals, and their real-world data. Addressing this gap, we extend prior work by (i) designing a technology probe [23] consisting of an integration framework and reflection interface supporting Bayesian analysis of self-tracked health data, focused on common questions about trigger and symptom relationships [38], and (ii) examining how participants use the technology probe for exploring and reflecting on relationships across a range of questions and forms of self-tracked data.

Informed by prior research [22, 31, 38, 39], our technology probe [23] combines an integration framework and a reflection interface applying Bayesian analysis to: (i) support a range of

self-tracking goals, data, and questions, (ii) support examining trigger and symptom relationships, (iii) be robust to many types of lapses in data collection, and (iv) scaffold reflection that can support and motivate action. We employed this probe in a two-stage study with 8 participants who self-track across 12 distinct health parameters and conditions, examining their experiences using Bayesian analysis. We first conducted semi-structured interviews to understand participant self-tracking goals, practices, and data. After integrating each participant's data and learning a Bayesian network from that data, we conducted a second set of participant sessions. Participants interacted with and reacted to Bayesian analysis in our technology probe, aiming to answer questions about trigger and symptom relationships within their self-tracked data.

Participants were able to successfully use Bayesian analysis via the technology probe: in examining potential trigger and symptom relationships and potential interactions among triggers, in support of reflection through evolution of their questions, and in motivating data-informed action. Participants also encountered challenges, caused by incorrect or partial use of probe capabilities, by misunderstanding of Bayesian capabilities and conclusions, and by misalignment of conclusions with participant goals. We discuss challenges and opportunities in Bayesian analyses supporting individuals in exploring and reflecting on relationships across a range of self-tracking goals, data, and questions. These include considerations for navigating reflection and action with support for working from goals to insights and for designing tools that ensure appropriate analysis.

2 Designing a Technology Probe for Bayesian Analysis

Informed by prior work in personal informatics [22, 31, 39] and on Bayesian analysis in self-tracking [38], we implemented Bayesian network analysis in a technology probe consisting of (i) an integration framework and (ii) a reflection interface. The integration framework focused on the integration stage of personal informatics [31], wherein we manually integrated participant-provided self-tracked data, translating prior research on needs and challenges into features, training settings, and data input for a Bayesian network. The integration framework then utilized the `bnlearn` R package [40] to learn a network from the provided data. The reflection interface next focused on the reflection stage of personal informatics [31], designed to support participants in exploring, interpreting, and reflecting on their data through the underlying Bayesian analysis. Figure 1 shows the two-tab interface designed to support reflection both in terms of in-depth analysis of aspects that answer a single question (a "Scenarios" tab) and more holistic exploration of a learned network (an "Overview" tab).

The scope of our technology probe is focused on support in examining self-tracking data around goals of identifying and understanding relationships between potential triggers and symptoms (i.e., not aiming to be comprehensive in support for Bayesian networks, Bayesian network learning, or other analyses of self-tracking data). A trigger is anything that might contribute to or result in a symptom. A symptom is any aspect of health or wellbeing and might be impacted by one or more triggers. Although our probe cannot guarantee a causal analysis (e.g., a causal variable may go untracked while merely a correlated variable is tracked), Bayesian analysis does evaluate beyond a simple correlation analysis, particularly in self-experimentation [38]. We focused on creating our technology probe [23] to facilitate interview and observation with people who self-track around a variety of health goals, data, and questions. Probe implementation details can be found in Appendix A.

3 Methods

We conducted an IRB-approved study with 8 participants, using the technology probe from Section 2. Our goal was to observe how people explore relationships between potential triggers and symptoms and how they answer questions using Bayesian analysis of their self-tracked data, as well as to examine potential challenges and opportunities around utilizing Bayesian network analysis.

3.1 Participants

Participants were recruited using purposive sampling [43] through a medical research participant pool associated with the University of Washington's Institute for Translational Health Sciences, Quantified Self message boards, social media groups, and our personal networks. Participants were required to have tracked data including potential triggers and symptoms related to a health condition. People with any health condition were eligible to participate, but study sessions focused on mental health tracking or questions were excluded under our IRB approval (i.e., because our study setting could not provide associated safety resources and protocols). After verifying participant eligibility, we asked participants to share their tracked data with the research team. Participants could withhold any data they did not want to share. Table 1 details participant demographics and tracking details.

3.2 Participant Sessions

The study consisted of two sessions with each participant, including interview and observation components (Figure 2). Session 1 was a background interview to learn about participant self-tracking goals and practices, followed by exploration of their existing self-tracked data. Between sessions, we performed necessary data transformations to map each participant's data into the integration framework (described in Section A.1). In Session 2, participants interacted with the technology probe, using it in Bayesian analysis of their self-tracked data. Further details and the complete study protocol are provided in Appendix C. Sessions each lasted 60 to 90 minutes, and participants were compensated a $30 gift card for each session. Informed consent, including permission to record, was obtained before each session. All sessions were conducted over Zoom with audio and video recording.

3.3 Data Analysis

Zoom-generated transcriptions were reviewed for correctness. We first used data from Session 1 interviews to prepare for Session 2. Specifically, we used what we learned about each participant's self-tracking goals, practices, and data to inform training of a Bayesian network for that participant, including decisions about how to integrate each participant's data and what questions to support.

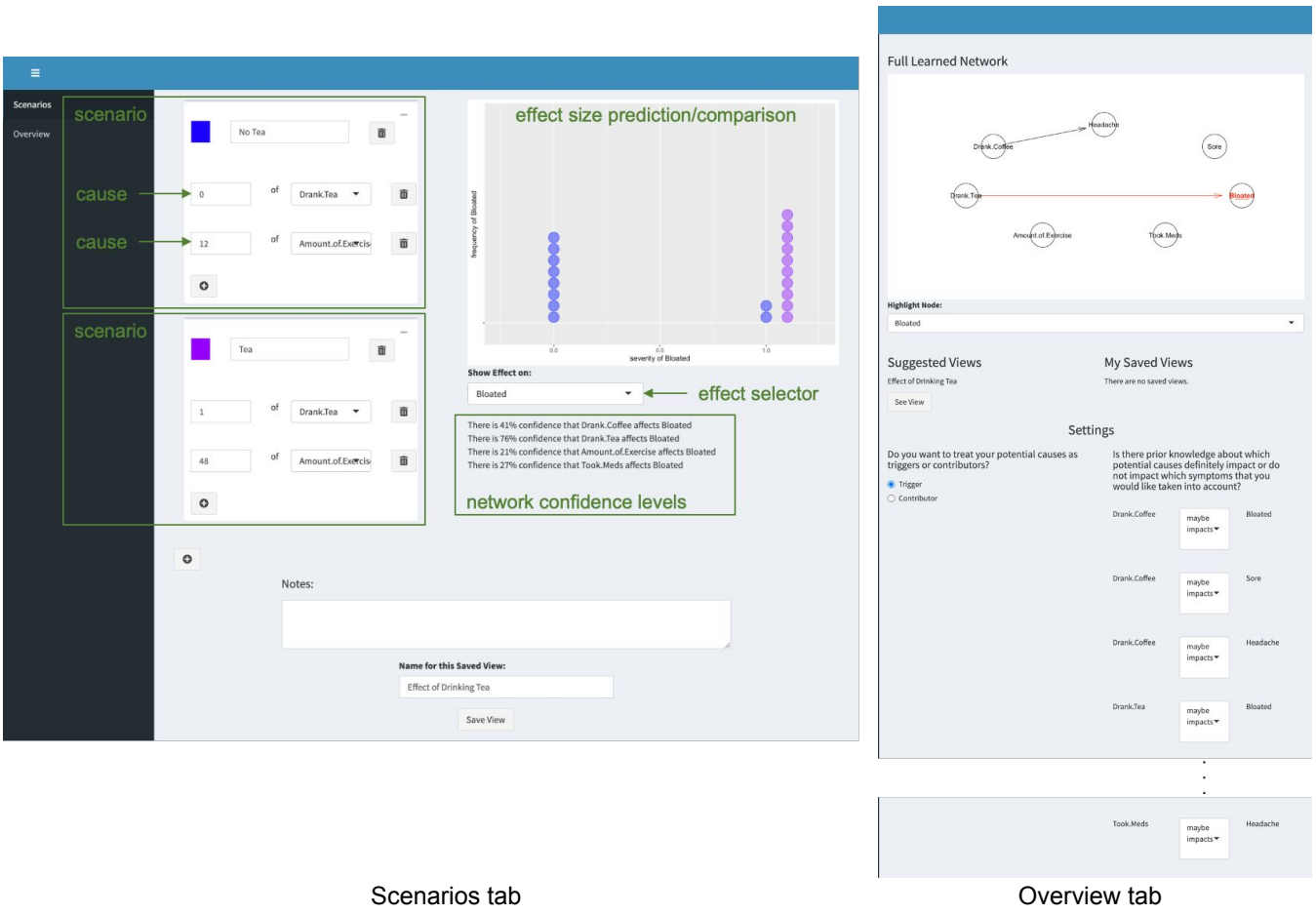

Figure 1: Elements of the technology probe's reflection interface, with data labels changed for anonymity. The Scenarios tab supports in exploring specific combinations of potential exposures to triggers and corresponding predicted probability distributions of symptoms. The Overview tab visualizes and supports exploration of the overall learned network.

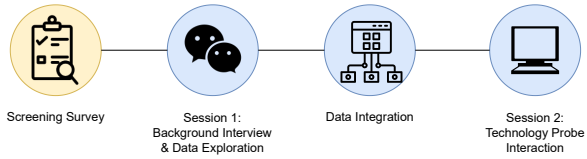

Figure 2: Study procedure. Potential participants filled out a screening survey (Appendix B) to determine their eligibility. Those eligible were invited to participate in the study. Session 1 consisted of a background interview and data exploration component to learn about participant health tracking data and questions. Between sessions, necessary data transformations were performed, mapping each participant's data into the integration framework of the technology probe. Session 2 consisted of participants interacting with and reacting to the technology probe that modeled their self-tracked data.

We then analyzed both sessions to understand participant experiences with the probe. Three researchers first individually open-coded both interviews of one participant. They discussed their coding, clustered codes based on patterns in the data, and created an initial outline of a codebook informed primarily by participant experiences with the probe. The three researchers then used the initial codebook to individually code both interviews of a different participant, again refining the codebook. One researcher then used the codebook to code for the remaining 6 participants (12 total sessions). Throughout this process, the researcher continued to open-code data relevant to our research questions, including what was not already reflected in the codebook. Researchers additionally categorized questions participants sought to answer according to Schroeder et al.'s question categories (e.g., any effect, interaction effect, varying effect) [38]. The researcher then used axial coding and memoing to construct preliminary themes. These were refined through further discussion with the full team.

Table 1: Self-reported demographic and tracking details of study participants. Two participants (P5, P11) dropped out. Two additional participants were not asked to complete the second interview because they did not have questions related to their data (P2) or had not tracked long enough or systematically enough (P6).

P#	Sex Age Race	Health Parameters and Tracking Data	Purpose of Tracking and Example Questions
P1	Female 18-33 White	POTS syndrome; using spreadsheet; 4 triggers, 3 symptoms; 1 month of data	Condition management, discovery of triggers; (i) "any effect" of stress or PT exercises on POTS flare-up, (iii) "interaction effect" between water & caffeine intake on dizziness, (iv) "temporal effect" of weekly weather patterns on dizziness, (v) "quantity/severity effect" of gaining weight on menstrual cycles
P3	Female 49-64 White	Workout recovery, weight; using spreadsheet, wearables; 23 triggers, 6 symptoms; 6 months of data	Self knowledge; (i) "any effect" of illness on high blood pressure, (ii) "noticeable effect" of carb, sodium, or water intake on glucose level
P4	Female 18-33 White	Chronic shoulder pain; using spreadsheet; 24 triggers, 10 symptoms; 13 months of data	Condition management, diagnosis; (i) "any effect" of exercises, side sleeping, or ice/heat on pain, (ii) "noticeable effect" of medication on pain thresholds, (iii) "interaction effect" between carrying groceries, exercise on pain, (v) "quantity/severity effect" for a specific pain on general pain
P7	Female 18-33 White	Heart palpations; using spreadsheet; 11 triggers, 5 symptoms; 2 months of data	Diagnosis; (i) "any effect" of anxiety level on heart rate, (ii) "noticeable effect" of sitting or standing on blood pressure, (v) "quantity/severity effect" of exercise on blood pressure
P8	Female 18-33 White	Menstrual cycles, migraines; using wearables, phone; 5 triggers, 4 symptoms; 3 months of data	Condition management, discovery of triggers; (i) "any effect" of caffeine intake or sleep on migraines, (ii) "noticeable effect" of period on caffeine craving, acne, or mood
P9	Female 34-48 White	Seizures (for son); using spreadsheet; 8 triggers, 6 symptoms; 26 months of data	Condition management, discovery of triggers; (i) "any effect" of moon phase, illness, or diet on seizures, (iii) "interaction effect" between moon phase, circadian rhythm on seizures, (iv) "temporal effect" of moon phase on seizures
P10	Male 18-33 White	Migraines, IBS, lower back pain; using paper journal; 17 triggers, 13 symptoms; 6 months of data	Condition management, discovery of triggers; (i) "any effect" of sugar, water, or alcohol intake on migraines or IBS, (iii) "interaction effect" between activity, sleep on migraines or IBS
P12	Male 34-48 White	Running performance, knee pain; using Google forms, wearables; 18 triggers, 6 symptoms; 9 months of data	Discovery of triggers, self knowledge; (i) "any effect" of mileage, rest, or hill running on knee pain, (iii) "interaction effect" between drinking water, hot weather on muscle cramping

4 Results

We observed participants using the technology probe and reacting to Bayesian analysis of their self-tracked data (Table 1). Participants experienced successes in reflecting within their data (Section 4.1), but also encountered breakdowns (Section 4.2) with the Bayesian network integration (Section 4.2.1) and reflection interface (Section 4.2.2), which could lead to misalignment and uncertainty about conclusions (Section 4.2.3).

4.1 Successes in Supporting Effective Interpretation & Reflection

All participants used the probe to successfully identify potential relationships ("any effect" questions [38], the most common type) and to identify interactions between two or more potential triggers ("interaction effect" questions [38], the second most common type). Even participants familiar with data science and statistics talked about how capabilities in the probe could support applying techniques they considered promising: *"there's one dad in one of the Facebook groups I'm in who who does stuff like this... he posts regularly to be like 'Oh, I found this correlation and like here's the*

very optimal dosing my son was like I've seen that and I've admired it, but this is not my forte.... it's like oh my gosh a tool where I could do that too, that's exciting" (P9, who works with machine learning).

4.1.1 Preferred Views and Visualizations for Different Needs. Participants described the probe's reflection interface as useful for exploring a broad range of variables (the Overview tab), before narrowing their focus in further questions (the Scenarios tab). Although participants used both, most found the Overview tab more *"intuitive"* (P3) for identifying contributing factors (P1), seeing connections (P4), and seeing *"direct correlations instead of having to like read through the percentages or figure out what the the dots on the graph mean"* (P8). P10 described the Overview tab's graph visualization as useful for *"putting cause and effect [relationships] into perspective"*, including for seeing the impact of multiple triggers together (i.e., interaction effects).

On the other hand, some liked diving into specific relationships using the Scenarios tab. P1 tracked a variety of data with the goal of discovering symptom triggers and managing their overall health. They used the probe to focus on specific triggers and their relationship to a symptom (e.g., *"remove effects of like exercise and only look at coffee"*). P1 described the ability to *"isolate"* the impact of individual potential triggers as more useful than their current reflection methods (i.e., creating *"simple"* line and bar graph visualizations), which they emphasized did not support their desired understanding of *"causation."* Others also praised the scores and confidence visualizations as readable (P4) and informative in identifying potential triggers (e.g., P9 said the probe *"explains more than my neurologist can"* in their goals for identifying potential triggers of their son's seizures).

4.1.2 Evolution of Participant Questions. We observed participants evolving their exploration and reflection beyond an initial question. P7 has self-tracked to understand potential triggers of heart palpation, and started by using the Overview tab to identify how their posture (sitting versus standing) impacted heart rate and blood pressure. After learning that postures did not impact blood pressure but somewhat impacted their heart rate, they evolved their question to examine whether there was an interaction effect [38] between a specific medication and posture. Seeing an impact on their heart rate, P7 further evolved their question to focus on the relationship between the medication and heart rate, examining how long after consumption the medication continued to impact their heart rate (i.e., temporal effect [38]). From the visualization, P7 concluded the medication seemed to impact their heart rate for more than 4 hours after consumption, an observation they were not expecting for a short acting medication. This led P7 to continue asking follow-up questions and investigating other potential relationships, such as the impact of the medication on sleep.

4.1.3 Potential for Motivating Action and Identifying Next Steps. Some participants identified and described actionable next steps based on conclusions reached through engaging their data with the probe. P1 wanted to try a quasi-experimental method to collect more varied data to confirm or disprove their conclusions. Although P1 had intermittently tracked water intake to see if more water reduced dizziness, P1 felt they did not yet have enough data to confirm the conclusion obtained with the probe. They were therefore motivated to consistently track and further self-experiment (e.g., *"I would encourage myself to drink 48 ounces of water more. I mean ideally I would rather try 32 ounces versus 64 ounces and see if that makes a difference [to my dizziness]"*). Other participants also appreciated insights gained through the probe that motivated action, such as *"entertaining the notion of [behavior] change"* (P4) and encouraging them to *"do better"* (P10): *"I really like this... I want to make a change... it kind of motivates me to want to do better for myself... I've been trying to do better with my new lifestyle, but I see these potential causes... and it motivates me personally to want to just continue to do even more"* (P10).

4.2 Breakdowns & Opportunities

4.2.1 Partial Use & Misunderstanding of Impacts. Our probe implemented Bayesian analyses for different question types identified in prior research [38], with a goal of supporting identification of relationships among variables in self-tracked data. However, some participants tried to explore and answer questions with subsets of the probe that did not provide a full answer. For example, although the Overview tab's graph visualization was useful for identifying presence of a relationship (i.e., indicated by an edge from a trigger node to a symptom node), it was not always possible to discern whether it was a positive or negative relationship, which could lead to errant conclusions. P4 used the graph visualization to examine what might be helping with their shoulder pain. Instead of considering nodes (e.g., new pillow, CBD, Tylenol) as pain "triggers" or "contributors", P4 interpreted a directed edge to mean reducing shoulder pain (a positive effect). However, because the direction was not specified in the graph, interpreting the graph visualization as P4 did, without checking that interpretation using the Scenarios tab, could lead to confusing a negative impact on their pain for a positive one.

4.2.2 Challenges of Interaction and Data Integration. Participants described challenges using the probe reflection interface. A key confusion was in the Scenarios tab, when inputting variables to create visualizations and in interpreting the quantile dot plot (described by participants as a *"bubble plot"* (P4)). When trying to investigate whether abdominal workouts can trigger dizziness, P1 accidentally added presence of a trigger (abdominal workout) three times. Although they later removed redundant triggers, they remained confused about what was encoded as a trigger versus as a symptom and where to input each on the Scenarios tab. Others also noted confusion around how variables were encoded (e.g., P7 was unsure why walking and exercise were assigned different data types). P9 encountered a breakdown while using the Scenarios tab to examine whether factors (e.g., changing light, keto ratio) affect the number of doses of Diastat (a rescue medicine) to stop a seizure. The visualization showed the most likely outcome was either 0.4 or 0.5 doses, which seemed nonsensical because Diastat is used as 0, 1, or 2 doses. In conversation, the researcher reflected that this problem resulted from modeling Diastat doses as a linear number, not as discrete events.

When encoding and integrating participant data, the research team made interpretations based on information participants provided in Session 1. Exposing more of the data integration work and rationale may have better supported participants in using the probe

and created opportunities to correct encoding issues. At the same time, participants (P4, P8, P9) anticipated challenges if they were to do data transformations and integration themselves, described not knowing how to integrate data in the probe, and wanted *"an instruction sheet on how to use it"* (P4).

4.2.3 Skepticism When Results Do Not Match Expectations. Some participants valued when results challenged existing expectations, but some also expressed curiosity or skepticism when results did not align with expectations (P7, P9, P12). For example, P7 explored how different triggers impacted her heart rate and contributed to heart palpation. She was confused why triggers such as *"exercise," "walking,"* and *"getting ready"* were not giving similar conclusions (all should impact her heart rate and blood pressure). She initially postulated this might be because they were encoded using different data types, but further reflected that insufficient data might be why she was not seeing expected impacts. In other situations, participants trusted that analysis represented available data, but did not investigate specific questions because of limited data pertaining to those questions. For example, P1 did not investigate migraine, as "*just because, that's not something that occurs frequently enough where I'm not sure that there would be enough data that it would be, you know, useful at all.*" Similarly, P12 believed there was an association between protein consumed and knee pain, but the model did not reflect this. He concluded *"Either I need to have more variables in the data I'm collecting or it doesn't have enough data with the variables I am currently putting in there."* In exploring their data, many participants first looked for support for expected relationships. P9 described she would first check if expected relationships are represented. If not, she would question either the data integration or the model.

Across these examples, we saw participants make judgments about whether to trust the learned network. Judgments relied on participant knowledge of data and statistics, discussions with the researcher, and intuition about what the model should show. This raises questions around challenges of confirmation bias in future systems, including whether people may mistrust and dismiss an accurate analysis that challenges their pre-existing beliefs.

4.2.4 Scaffolding Exploration of Questions in Scenarios View. Many participants used the Scenarios tab to explore specific questions, as it was designed, through quickly creating scenarios and specifying variables. However, this ability to quickly explore, absent scaffolding of those explorations, also led to challenges. For example, P12 initially configured comparisons between different variables (15 protein *versus* 240 minutes of sleep *versus* 60 minutes of activity) when intending to configure an additive question (what might happen if I eat 15 grams of protein *and* get 240 minutes of sleep *and* 60 minutes of activity). The researcher conducting the interview supported P12 in addressing this misunderstanding and reconfiguring the scenarios to match P12's question.

P9 encountered a barrier that could not be resolved during the interview session. They wanted to examine impacts of dosage of a seizure prevention medication. However, during integration the research team had noted there was insufficient data to model the question "what prevents seizures," and so this was not an option in the interface. This led P9 to comment *"it's not looking for correlations between the right things... it's not what I'm looking for."*

To some extent, difficulties participants encountered reflect anticipated usability limits of our technology probe. However, they also highlight participant needs in examining questions. P12 likely would have benefited from a summary of the question a scenario was configured to ask, so he could detect misalignment with his question. P9 may have benefited from being able to express her question, even though it could not be answered with the underlying model or data (e.g., if a tool had provided guidance about what tracking would support such a question).

5 Discussion

Our examination of participant experiences with a technology probe implementing Bayesian analysis highlights considerations for personal health informatics. These include considerations in navigating reflection and action from health goals to insights and in designing for appropriate trust in analysis.

5.1 Navigating Reflection and Action: From Goals to Insights

Participants self-tracked for a variety of health conditions, purposes, and goals (e.g., self-assessment of performance (P12), self-experimentation with treatments (P9), identifying potential triggers for a symptom (P1, P4), differentiating between triggers for different symptoms (P10)). Participants had previously tracked both potential triggers and symptoms, and we used Session 1 to examine each participant's data, goals, and questions before a researcher performed necessary data transformations (e.g., into relevant data types and aggregations). Other research on individualized self-tracking has successfully used similar approaches to provide tailored support for data collection and analysis aligned with individual goals [35, 41].

We expected that this individualized and intensive process would result in participants having the information and analysis needed to reflect on their data and answer questions using the probe. We also sought to support familiarization with using the probe by identifying questions (based on participant goals and questions from Session 1, refined to align with questions for which Bayesian analyses are well-suited [38]), and then creating suggested views (Section A.2.2) to guide participants through exploration of their data. However, some participants misunderstood the probe's capabilities, incorrectly used or only used parts of the probe, and derived conclusions misaligned with their health goals (Section 4.2).

These breakdowns highlight opportunities to more explicitly use an individual's specified questions in interactive guidance and tutorials that walk people through their data to demonstrate support for reflection. Techniques are needed to support data integration—including surfacing decisions made during integration and providing support if participants were to perform integration themselves (Section 4.2.2)—to direct people to appropriate views for their analyses (e.g., Section 4.1.1 notes different benefits to participants using the overview graph network versus scenarios aligned with specific questions), and to create suggested views and tutorials grounded in a person's own data. Barriers some participants encountered to effectively using the Scenarios tab (Section 4.2.4) also indicate opportunities for improving the experience of configuring and exploring scenarios, such as by asking what variables participants hold constant and which they wish to compare across scenarios.

To address participant challenges in reflection (Section 4.2), one opportunity would be to leverage generative AI to supplement Bayesian analysis. Recent research using LLMs for analyzing health data suggests their value for analyzing diverse self-tracking data and synthesizing multiple data streams, including generating natural language summaries [18]. Informed by our results, future research could examine potential for LLMs to help translate higher-level goals to concrete questions which can be examined using Bayesian analysis, to help streamline diverse self-tracking data before training a Bayesian network aligned with participant goals (e.g., removing data that might not be relevant for specific goals or questions), suggesting questions/relationships to examine according to collected data and higher-level goals, supporting natural language interactions to configure scenarios, and summarizing conclusions derived from Bayesian analysis.

Finally, upon reaching an insight, people may want to plan next steps [22, 31]. However, identifying appropriate action is a common challenge in self-tracking [31]. Although we did not observe participant action on insights, some participants noted actionable next steps and indicated an intent to act (Section 4.1.3). For example, P1 wanted to collect more varied data and self-experiment to test conclusions they reached with the probe. Future research could examine whether and how Bayesian analysis could help surface opportunities to support transition to action (e.g., through counter-factual analysis based on the network).

5.2 Designing for Appropriate Trust in Analysis

As described in Section 4.2.3, challenges of trust in analyses and conclusions were a barrier to reflection and potential action. We attribute these challenges to two primary causes: (i) misalignment between participant goals and what questions the resulting network could answer, and (ii) analysis results that challenged insights a participant had previously developed.

5.2.1 Misalignment between Goals and Questions the Bayesian Network Could Answer. One potential contributor was misalignment between participant goals and what questions the probe supported. In Session 1, we worked to translate participant goals into specific questions. Despite this hands-on, interactive opportunity to learn from participants about their goals, data, and questions, the later process of actually preparing and transforming the data for the Bayesian network still required the researcher to continue refining participant questions and making inferences about detailed participant goals.

This indicates an opportunity to design for integration that more tightly links processes of data transformation and question refinement. For example, some misalignments might be prevented by a design that indicates how certain data transformations will affect what questions can be answered. In this study, such a capability might have supported the researcher in getting participant feedback during the transformation process. More long-term, such designs might support people in preparing their own data. Additionally, P9's desire to ask a question the model could not support (Section 4.2.4) suggests opportunities for reflection interfaces that allow people to express questions they cannot answer. If a design were to provide guidance about a tracking plan that could answer the expressed question, this could provide actionable next steps.

5.2.2 Balancing Inclusion of Existing Insights with Mitigating Confirmation Bias. People engaging in self-tracking often seek both new insights and validation of beliefs about factors that affect their health [6]. People bring many prior beliefs into the tracking process, starting from the moment they decide to track. Both tracking tools and individual choices shape what data people do or do not collect, which then affects what is available in integration and reflection. For example, a person may strictly avoid a behavior they believe causes symptoms, thus never gathering data that would support testing that understanding. Although we discuss the risk of confirmation bias in Section 4.2.3, past work has also emphasized that people's experiential knowledge can be valuable in shaping and focusing their self-tracking and questions [10, 25].

A challenge for personal informatics continues to be how to best include experiential insights while mitigating biases and creating opportunities to challenge previous understanding. Prior research has suggested mixed-initiative approaches for identifying priors (e.g., a tool could suggest population or conditional priors, a person could adjust these based on their beliefs) [38]. This study chose not to encode priors based on any specific health parameters or conditions, allowing us to evaluate opportunities and challenges in reflection across a range of health conditions. Trying to develop priors requires extensive domain knowledge, and we recommend that tools which are intended for specific domains work with experts in those domains to establish appropriate priors. Participants could instead incorporate some prior knowledge through enforcing the presence or absence of learned edges. We also designed the probe to mitigate potential for confirmation bias, including through: (i) confidence scores in the Scenarios tab, and (ii) a graph of the full learned network in the Overview tab.

These choices meant participants encountered results that both matched their expectations and challenged previous understanding (Section 4.2.3). Some participants recognized the potential for confirmation bias, appreciated the components intended to mitigate it, and reached new conclusions during their study session (Section 4.1.1). When a mismatch occurred, further reflection was required to identify potential reasons. For example, P7 expected *"exercise"*, *"walking"*, and *"getting ready"* to impact her heart rate, but did not reach the same conclusion using the probe. Although P7 initially expressed skepticism about how her data was encoded for Bayesian analysis, she further engaged with the probe and then attributed the mismatch to having tracked insufficient data to support reliable conclusions. However, others encountering results that challenged their prior understanding expressed less trust in the technology probe. To support exploring data-driven conclusions, designers could build on prior research on visualizing personal informatics insights [4, 5] to design communicative tools for personal data reflection (e.g., visual annotations for highlighting insights, features that expose underlying knowledge structures that inform specific visualizations). Textual explanations and narratives alongside quantitative results can also enable deeper reflection and engagement with tools [42], and there is potential for LLM-based explanations to support understanding of health data and data-based insights [7], although LLM-based explanations also raise potential concerns (e.g., overreliance, confirmation bias) [28]. Self-trackers could thus benefit from designs that walk through

evidence and explain how conclusions are reached, further supporting self-experimentation and validation of understanding and knowledge. Designers may also develop support for preliminary analyses, integrating a person's lived experiences and experiential insights, then using those analyses to enlist health providers or other experts to help understand inconclusive results or plan for next steps when beliefs disagree with results [9]. Future work can thus explore how our examination of Bayesian formulations and analyses can complement and build upon prior examinations of patient-provider collaboration around identification of symptom triggers using patient self-tracking data [10, 37].

6 Conclusion

We conducted a technology probe study to examine participant interaction with Bayesian analysis of self-tracked health data across a range of health goals, data, and questions. Participants used the technology probe to successfully explore trigger and symptom relationships within their data, especially identifying presence of relationships and interaction among different triggers. They evolved their questions in reflection, obtained relevant conclusions, and identified actionable next steps on examining their data using the technology probe. We also observed breakdowns in participant interactions and reflection. We discussed our results in terms of considerations and opportunities for navigating reflection and action with support for working from goals to insights and designing tools that ensure appropriate trust in analysis.

Acknowledgments

This work was supported in part by the National Library of Medicine through award R01LM012810 and by the National Science Foundation through awards IIS-1813675 and IIS-1553167. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

References

[1] Farrokh Alemi, Shirley Moore, and Heibatollah Baghi. 2008. Self-experiments and analytical relapse prevention. *Quality Management in Healthcare* 17, 1 (2008), 53–65. doi:10.1097/01.QMH.0000308638.04850.48

[2] Eric P.S. Baumer, Sherri Jean Katz, Jill E. Freeman, Phil Adams, Amy L. Gonzales, John Pollak, Daniela Retelny, Jeff Niederdeppe, Christine M. Olson, and Geri K. Gay. 2012. Prescriptive persuasion and open-ended social awareness: expanding the design space of mobile health. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 475–484. doi:10.1145/2145204.2145279

[3] Janghee Cho, Tian Xu, Abigail Zimmermann-Niefield, and Stephen Voida. 2022. Reflection in Theory and Reflection in Practice: An Exploration of the Gaps in Reflection Support among Personal Informatics Apps. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 142, 23 pages. doi:10.1145/3491102.3501991

[4] Eun Kyoung Choe, Bongshin Lee, and M Schraefel. 2015. Characterizing Visualization Insights from Quantified Selfers' Personal Data Presentations. *IEEE computer graphics and applications* 35 (05 2015). doi:10.1109/MCG.2015.51

[5] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding self-reflection: how people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare* (Barcelona, Spain) *(PervasiveHealth '17)*. Association for Computing Machinery, New York, NY, USA, 173–182. doi:10.1145/3154862.3154881

[6] Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1143–1152. doi:10.1145/2556288.2557372

[7] Shaan Chopra, Katherine Juarez, James Fogarty, and Sean A. Munson. 2025. Engagements with Generative AI and Personal Health Informatics: Opportunities for Planning, Tracking, Reflecting, and Acting around Personal Health Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 75 (Sept. 2025), 33 pages. doi:10.1145/3749503

[8] Shaan Chopra, Rachael Zehrung, Tamil Arasu Shanmugam, and Eun Kyoung Choe. 2021. Living with uncertainty and stigma: self-experimentation and support-seeking around polycystic ovary syndrome. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18. doi:10.1145/3411764.3445706

[9] Chia-Fang Chung, Kristin Dew, Allison Cole, Jasmine Zia, James Fogarty, Julie A. Kientz, and Sean A. Munson. 2016. Boundary Negotiating Artifacts in Personal Informatics: Patient-Provider Collaboration with Patient-Generated Data. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 770–786. doi:10.1145/2818048.2819926

[10] Chia-Fang Chung, Qiaosi Wang, Jessica Schroeder, Allison Cole, Jasmine Zia, James Fogarty, and Sean A. Munson. 2019. Identifying and Planning for Individualized Change: Patient-Provider Collaboration Using Lightweight Food Diaries in Healthy Eating and Irritable Bowel Syndrome. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 7 (March 2019), 27 pages. doi:10.1145/3314394

[11] James Clawson, Jessica A. Pater, Andrew D. Miller, Elizabeth D. Mynatt, and Lena Mamykina. 2015. No longer wearing: investigating the abandonment of personal health-tracking technologies on craigslist. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 647–658. doi:10.1145/2750858.2807554

[12] Sunny Consolvo, David W. McDonald, Tammy Toscos, Mike Y. Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, Ian Smith, and James A. Landay. 2008. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1797–1806. doi:10.1145/1357054.1357335

[13] Felicia Cordeiro, Elizabeth Bales, Erin Cherry, and James Fogarty. 2015. Rethinking the Mobile Food Journal: Exploring Opportunities for Lightweight Photo-Based Capture. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3207–3216. doi:10.1145/2702123.2702154

[14] Felicia Cordeiro, Daniel A. Epstein, Edison Thomaz, Elizabeth Bales, Arvind K. Jagannathan, Gregory D. Abowd, and James Fogarty. 2015. Barriers and Negative Nudges: Exploring Challenges in Food Journaling. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1159–1162. doi:10.1145/2702123.2702155

[15] Nediyana Daskalova, Danaë Metaxa-Kakavouli, Adrienne Tran, Nicole Nugent, Julie Boergers, John McGeary, and Jeff Huang. 2016. SleepCoacher: A Personalized Automated Self-Experimentation System for Sleep Recommendations *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 347–358. doi:10.1145/2984511.2984534

[16] N Duan, I Eslick, N B Gabler, H C Kaplan, R L Kravitz, E B Larson, W D Pace, C H Schmid, I Sim, and S Vohra. 2014. Design and Implementation of N-of-1 Trials: A User's Guide. (2014).

[17] Elizabeth Victoria Eikey, Clara Marques Caldeira, Mayara Costa Figueiredo, Yunan Chen, Jessica L. Borelli, Melissa Mazmanian, and Kai Zheng. 2021. Beyond self-reflection: introducing the concept of rumination in personal informatics. *Personal Ubiquitous Comput.* 25, 3 (May 2021), 601–616. doi:10.1007/s00779-021-01573-w

[18] Zachary Englhardt, Chengqian Ma, Margaret E Morris, Chun-Cheng Chang, Xuhai" Orson" Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. 2024. From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–25.

[19] Daniel A. Epstein, Monica Caraway, Chuck Johnston, An Ping, James Fogarty, and Sean A. Munson. 2016. Beyond Abandonment to Next Steps: Understanding and Designing for Life after Personal Informatics Tool Use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1109–1113. doi:10.1145/2858036.2858045

[20] Daniel A. Epstein, Jennifer H. Kang, Laura R. Pina, James Fogarty, and Sean A. Munson. 2016. Reconsidering the device in the drawer: lapses as a design opportunity in personal informatics. In *Proceedings of the 2016 ACM International Joint*

Conference on Pervasive and Ubiquitous Computing (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 829–840. doi:10.1145/2971648.2971656

[21] Daniel A. Epstein, Nicole B. Lee, Jennifer H. Kang, Elena Agapie, Jessica Schroeder, Laura R. Pina, James Fogarty, Julie A. Kientz, and Sean Munson. 2017. Examining Menstrual Tracking to Inform the Design of Personal Informatics Tools. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 6876–6888. doi:10.1145/3025453.3025635

[22] Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. 2015. A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 731–742. doi:10.1145/2750858.2804250

[23] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 17–24. doi:10.1145/642611.642616

[24] Monique W.M. Jaspers, Thiemo Steen, Cor van den Bos, and Maud Geenen. 2004. The think aloud method: a guide to user interface design. *International Journal of Medical Informatics* 73, 11 (2004), 781–795. doi:10.1016/j.ijmedinf.2004.08.003

[25] Ravi Karkar, Jessica Schroeder, Daniel A. Epstein, Laura R. Pina, Jeffrey Scofield, James Fogarty, Julie A. Kientz, Sean A. Munson, Roger Vilardaga, and Jasmine Zia. 2017. TummyTrials: A Feasibility Study of Using Self-Experimentation to Detect Individualized Food Triggers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6850–6863. doi:10.1145/3025453.3025480

[26] Matthew Kay, Eun Kyoung Choe, Jesse Shepherd, Benjamin Greenstein, Nathaniel Watson, Sunny Consolvo, and Julie A. Kientz. 2012. Lullaby: a capture & access system for understanding the sleep environment. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (Pittsburgh, Pennsylvania) *(UbiComp '12)*. Association for Computing Machinery, New York, NY, USA, 226–234. doi:10.1145/2370216.2370253

[27] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5092–5103. doi:10.1145/2858036.2858558

[28] Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 420, 19 pages. doi:10.1145/3706598.3714020

[29] Richard L Kravitz, Naihua Duan, and Richard H White. 2008. N-of-1 trials of expensive biological therapies: a third way? *Archives of internal medicine* 168, 10 (2008), 1030–1033. doi:10.1001/archinte.168.10.1030

[30] Amanda Lazar, Christian Koehler, Theresa Jean Tanenbaum, and David H. Nguyen. 2015. Why we use and abandon smart devices. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 635–646. doi:10.1145/2750858.2804288

[31] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 557–566. doi:10.1145/1753326.1753409

[32] James J. Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B. Strub. 2006. Fish'n'Steps: Encouraging Physical Activity with an Interactive Computer Game. In *UbiComp 2006: Ubiquitous Computing*, Paul Dourish and Adrian Friday (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 261–278.

[33] Lena Mamykina, Elizabeth Mynatt, Patricia Davidson, and Daniel Greenblatt. 2008. MAHI: investigation of social scaffolding for reflective thinking in diabetes management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 477–486. doi:10.1145/1357054.1357131

[34] Lena Mamykina, Arlene M. Smaldone, and Suzanne R. Bakken. 2015. Adopting the sensemaking perspective for chronic disease self-management. *Journal of Biomedical Informatics* 56 (2015), 406–417. doi:10.1016/j.jbi.2015.06.006

[35] Jimmy Moore, Pascal Goffin, Jason Wiese, and Miriah Meyer. 2022. An Interview Method for Engaging Personal Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 173 (Dec. 2022), 28 pages. doi:10.1145/3494964

[36] Jessica Schroeder, Chia-Fang Chung, Daniel A. Epstein, Ravi Karkar, Adele Parsons, Natalia Murinova, James Fogarty, and Sean A. Munson. 2018. Examining Self-Tracking by People with Migraine: Goals, Needs, and Opportunities in a Chronic Health Condition. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 135–148. doi:10.1145/3196709.3196738

[37] Jessica Schroeder, Jane Hoffswell, Chia-Fang Chung, James Fogarty, Sean Munson, and Jasmine Zia. 2017. Supporting Patient-Provider Collaboration to Identify Individual Triggers using Food and Symptom Journals. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1726–1739. doi:10.1145/2998181.2998276

[38] Jessica Schroeder, Ravi Karkar, James Fogarty, Julie A Kientz, Sean A Munson, and Matthew Kay. 2019. A Patient-Centered Proposal for Bayesian Analysis of Self-Experiments for Health. *Journal of healthcare informatics research* 3 (2019), 124–155. doi:10.1007/s41666-018-0033-x

[39] Jessica Schroeder, Ravi Karkar, Natalia Murinova, James Fogarty, and Sean A. Munson. 2020. Examining Opportunities for Goal-Directed Self-Tracking to Support Chronic Condition Management. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4 (Sept. 2020), 151:1–151:26. doi:10.1145/3369809

[40] Marco Scutari. 2010. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software* 35, 3 (2010), 1–22. doi:10.18637/jss.v035.i03

[41] Yasaman S. Sefidgar, Carla L. Castillo, Shaan Chopra, Liwei Jiang, Tae Jones, Anant Mittal, Hyeyoung Ryu, Jessica Schroeder, Allison Cole, Natalia Murinova, Sean A. Munson, and James Fogarty. 2024. MigraineTracker: Examining Patient Experiences with Goal-Directed Self-Tracking for a Chronic Health Condition. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3613904.3642075

[42] Konstantin R. Strömel, Stanislas Henry, Tim Johansson, Jasmin Niess, and Paweł W. Woźniak. 2024. Narrating Fitness: Leveraging Large Language Models for Reflective Fitness Tracker Data Interpretation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3613904.3642032

[43] Maria Dolores C Tongco. 2007. Purposive sampling as a tool for informant selection. (2007).

[44] Ádamo L. de Santana, Carlos R. Francês, Cláudio A. Rocha, Solon V. Carvalho, Nandamudi L. Vijaykumar, Liviane P. Rego, and João C. Costa. 2007. Strategies for improving the modeling and interpretability of Bayesian networks. *Data & Knowledge Engineering* 63, 1 (2007), 91–107. doi:10.1016/j.datak.2006.10.005 Data Warehouse and Knowledge Discovery (DAWAK '05).

A Technology Probe Design and Implementation

A.1 Integration Framework

The integration framework applies Bayesian analyses and Bayesian network learning to support functionality across five self-tracking stages or aspects defined in prior work: collection stage [22], different underlying condition phenomena, different self tracker goals [38], lapsing and resuming stage [22], and tracking and acting stages [22]. Table 2 provides more details on each of these. We discuss the different stages and implementation of associated functionality, including how need for such functionality can arise across different self-tracking scenarios.

A.1.1 Collection Stage. The technology probe supports different overall approaches to self-tracking toward trigger and symptom relationship goals, from planned self-experimentation to observation. This is supported by abstracting out the type of tracking practice and modeling data in a trigger and symptom node framework. Different tracking methods can therefore be modeled within the same network, and the method of tracking will determine how much data is required for network confidence levels to converge. This provides flexibility to analyze data from settings where people are able and willing to systematically explore triggers and symptoms to answer questions faster (i.e., self-experimentation), where people experience symptoms but do not undertake rigorous self-experimentation (i.e., quasi-experiment), or where people are limited to observing

Table 2: Functionality implemented as a part of the integration framework of the technology probe, applying Bayesian network learning to explore support for examining trigger and symptom relationships in self-tracked health data. Further details for the data processing and network learning implementation can be found in the publicly available Github repo (github.com/aokeson/BayesNetsForSelfTracking) in the participant_specifics.R and network.R files, respectively. Further details for the Collection Stage are in Section A.1.1, the Reflection Interface in Section A.2, and the Integration Stage in Section C.2.

Self Tracking Stage	Functionality, Implementation Location	Bayesian Network Implementation
Collection Stage [22]	Different Data Types Data Processing Code Line 13	Boolean, categorical, ordinal, and continuous data are implemented as their associated R data types.
	Experiment Collection Stage	Experiment happens during participant data collection. All inputs are exactly the same for a specified period of time, while 1 potential trigger is varied and symptoms monitored to gain more confidence in exactly 1 potentially causal relationship.
	Quasi-Experiment Collection Stage	Quasi-experiment happens during participant data collection. Continues normal exposure to all other triggers, but varies 1 potential trigger systematically to gain diversity in observations of that trigger and associated symptoms.
	Observation Collection Stage	Observation happens during participant data collection. Tracks normal exposure, learned over sufficient data variation.
Different Underlying Condition Phenomena	Triggers Network Learning Code Line 35	Learn network edges from tracked potential triggers to tracked symptoms.
	Contributors Network Learning Code Line 35	Create "potential trigger" nodes for all possible combinations of tracked potential triggers, then learn network with them as individual network nodes.
Different Self Tracker Goals [38]	Any Effect Reflection Interface Overview Tab	Check for a presence of an arrow between the trigger and symptom of interest in the learned network.
	Noticeable Effect Reflection Interface Scenarios Tab	Compare predicted probability distribution of different amounts of triggers and see how large the difference is.
	Interaction Effect Reflection Interface Scenarios Tab	Directly compare predicted probability distribution of 2 triggers, alone and then together, to see if the combination results in significantly different symptom effect sizes.
	Temporal Effect Integration Stage	Compute and create new nodes for time elapsed, time of day, or aggregated data into a time scale of interest.
	Quantity/Severity Effect Reflection Interface Scenarios Tab	Compare the symptom effect size of different quantities of triggers.
	Confidence in Conclusion Reflection Interface Scenarios Tab	The strength metric of an edge indicates the confidence in the edge's existence and therefore if there is a relationship.
Lapsing and Resuming Stage [22]	Support Partial Lapse Integration Stage	If some nodes are sporadically not tracked, impute the learned impact based on other data.
	Support Complete Lapse and Resume Collection Stage	Individual tracking records are independent so resumption can begin at any time with no change to data processing.
Tracking and Acting Stages [22]	Prior Knowledge Network Learning Code Line 47, 144	Force an arc to exist or not exist based on existing knowledge of a relationship.
	New Trigger/Symptom Integration Stage	Add new node and relearn relationships by imputing past untracked values as 0 or based on prior observation data.
	Stop Tracking Trigger/Symptom Integration Stage	Impute values based on prior observation data or delete node.
	Stop Consuming/Doing Trigger Integration Stage	Impute 0s or delete node.

natural variation in triggers and symptoms in their day-to-day experiences (i.e., observation). Bayesian networks also support different data types for encoding triggers and symptoms, according to individual needs and self-tracking practices. In the technology probe, this is supported by assigning each "trigger" a numerical mapping (e.g., mapping Boolean data to 0 or 1, categorical data to an arbitrary number between 1 and the number of categories, ordinal data to its natural ordered value).

A.1.2 Different Underlying Condition Phenomena. The technology probe supports modeling underlying triggers associated with different health conditions as either triggers or contributors. Modeling as triggers can be appropriate when a single factor can be enough to induce a change in the symptom (e.g., in IBS, eating a single type of food or nutrient can cause symptom onset [25]). Modeling as contributors can also be appropriate when any one factor alone might not be enough to induce a change in the symptom, but the

sum of several factors may be (e.g. in migraine, multiple factors can compound as part of impacting symptoms [39]). In the technology probe, (i) modeling as triggers is supported by creating a single node for each self-tracked variable, and (ii) modeling as contributors is supported by creating additional nodes that represent possible combinations of different triggers, then adding the data values to get a sum of multiple triggers.

A.1.3 Different Self Tracker Goals. The technology probe provides support for exploring different questions based on the underlying Bayesian network through the reflection interface, further described in Section A.2. This includes support for key questions around trigger and symptom relationships among nodes in the Bayesian network, as highlighted in prior self-tracking literature [38]. It also includes a meta-goal of estimating confidence in answers.

A.1.4 Lapsing and Resuming Stage. People commonly lapse in their tracking [22], either partially or completely, so data about triggers and symptoms can often be incomplete or sparse. In a traditional experimental setup, such lapsing can violate assumptions of the experimental setup. Our technology probe addresses concerns of lapsing and resumption in self-tracking by treating all data as observational, so there is no time component to individual instances of tracking triggers and symptoms. It also integrates partial observations, where only some triggers and symptoms are not tracked, imputing the learned impact based on prior data values.

A.1.5 Tracking and Acting Stages. As people build up more understanding of their health condition through lived experience, self-tracking, and/or reflection, they often iterate upon or evolve their goals and questions [39]. While people sometimes want to test beliefs or knowledge for which they have limited evidence [25], tools prompting or requiring people to test or re-examine questions to which they believe they already know the answer can lead to frustrating experiences and tool abandonment [11, 19, 20, 30]. Our technology probe thus allows people to enforce prior knowledge (e.g., enforcing structure so a trigger is modeled as definitely or definitely not relating to a symptom), further supporting people in iterating upon and evolving their goals. People can also use our probe to begin tracking new triggers and symptoms over time, to end tracking of triggers and symptoms, or to note a permanent behavior change (e.g., ending an activity that is a potential trigger) by imputing data as needed in the data processing.

A.2 Reflection Interface

After a Bayesian network is learned using the R package `bnlearn` [40], it is important to facilitate review and interpretation of the network in the context of an individual's self-tracked data. Individuals bring additional knowledge and context around their data, self-tracking practices, and health that are essential for interpretation. However, Bayesian networks are known to be difficult to interpret without formal training [44]. We therefore designed a reflection interface as a part of our technology probe to support individuals in reviewing and reflecting on the results of Bayesian network analysis of their self-tracked data.

Shown in Figure 1, the desktop interface is implemented in R using the `Shiny` package. We created a two-tab interface to support reflection both in terms of in-depth analysis of aspects that answer a single question (Scenarios tab) and more holistic exploration of a learned network (Overview tab). The "Scenarios" tab (left in Figure 1) supports investigating individual questions via simulated predictions of symptoms experienced based on different exposures to triggers. The "Overview" tab (right in Figure 1) provides a graph of the entire learned network, suggestions of other scenarios to explore, and settings that support incorporating prior knowledge.

A.2.1 Scenarios Tab. This tab supports exploring specifics within a learned network without needing to parse through the full network. On the left of the scenarios tab, each white box represents a potential scenario consisting of different triggers an individual might experience. A person specifies which triggers, and how much of each, they experience in the scenario. Figure 1 shows Scenarios tab where a person is comparing a scenario specifying "0" of the "drank tea" trigger (i.e., they did not drink tea) and "12" amount of the "amount of exercise" trigger (i.e., 12 minutes of exercise) with another scenario specifying "1" of the "drank tea" trigger (i.e., they did drink tea) and "48" of the "amount of exercise" trigger (i.e., 48 minutes of exercise). The drop-down menu on the right side then supports selecting symptom nodes on which a person would like to see the impact of the scenarios. The network then predicts the probability distribution of symptom levels in each scenario, displaying it using quantile dotplots [27] in the graph at the top right. Prior research has shown that people without substantial statistical experience or training can understand and use quantile dotplots effectively to support decision making [27]. This graph then allows comparing effect size prediction among specified scenarios, shown on the same axis and differentiated by colors specified with each scenario. Below the symptom selector, a notation of the confidence level of a potential edge between each selected trigger and symptom provides an indication of how confident a person might be in the dotplot, based on the full learned network. Much less than 50% indicates network learning did not find a relationship between the trigger and a symptom; much greater than 50% indicates network learning found strong support for a relationship between the trigger and a symptom; near 50% indicates uncertainty. This tab also provides note-taking functionality and an ability to save the settings and view of the entire tab for later review via the Overview tab.

A.2.2 Overview Tab. Alongside answering specific questions, reflection should also support exploration of the overall learned network. This can be beneficial as a person's questions can overlap or be iterative and might be answerable upon further inspection of the overall network. A more holistic view of the network might also help people avoid confirmation bias from only examining and trusting conclusions that they already believe.

This tab shows the entire learned network, including all nodes corresponding to each column in the data. An edge between nodes is shown if there is greater than a 50% confidence that one node impacts another. The 50% edge threshold was chosen because it is the default threshold in the `bnlearn` package. The actual threshold percentage learned is also presented to the user in the network confidence levels section of Scenarios tab in Figure 1 to provide additional context to the user and allow them to assess confidence and potentially a different threshold for an edge in the network. Below the network diagram is an option to highlight a single node

and its incoming and outgoing edges (e.g., shown with the "Bloated" symptom node highlighted in red). Below the learned network are a set of suggested and saved views. Suggested views are configurations of the Scenarios tab that the research team created based on a participant's stated goals and their learned network. Saved views are configurations of the Scenarios tab that a participant saved for later viewing. The "See View" button associated with any of these opens the corresponding configuration in the Scenarios tab. Additional settings allow configuring functionality from Table 2. On the left is an option to model triggers either as triggers or as contributors. On the right is the option to include prior knowledge or hypotheses for each possible trigger and symptom relationship. The default is "maybe impacts", which allows the network learning to use provided data in learning the likelihood of an edge. Selecting "definitely impacts" enforces the presence of an edge, while selecting "definitely does not impact" enforces the absence of an edge. Changing any setting automatically triggers the `bnlearn` learning algorithm to run again with the updated parameters.

A.3 Study Focus and Ethical Considerations

Our technology probe is not a comprehensive tool or system for supporting all functionality of Bayesian networks, of Bayesian network learning, or of analysis of self-tracking data. We used our technology probe to facilitate participant sessions and examine how participants might use it to explore their self-tracking health data around goals of identifying and understanding relationships between potential triggers and symptoms. Although potentially useful for faster learning and more robust conclusions [38], we chose not to incorporate priors in the Bayesian network as part of supporting reflection across a diverse range of health parameters and conditions (i.e., relying on participant self-tracked data without encoding priors based on any specific health parameters or conditions). Participants could however specify in the interface settings if they already knew about the presence or absence of a relationship between any of the trigger and symptom nodes.

B Screening Survey

This survey has 3 pages and is expected to take 10-15 minutes to complete.

We are looking for participants for a study that explores possible improvement to collecting data about potential symptoms and triggers/causes of health conditions over time. We are looking for people who have kept a record of different symptoms and potential triggers/causes of these symptoms to try to understand more about their health and health conditions.

*We define **triggers** as: anything that may contribute to or cause symptoms or a change in symptom severity and that can be recorded.*

*We define **symptoms** as: anything that affects health or wellbeing that may be impacted by different **trigger/cause**.*

If you are selected to participate in this study, you will be asked to share some data that you have already collected about potential triggers and symptoms of health conditions. We will then ask to interview you about this data and about potential new ways of collecting and understanding your data.

Are you interested in participating in the study?

- *Yes*

*Examples of **symptoms***	*Examples of potential **triggers***
Migraines or headaches	Consuming caffeine, amount of sleep, *weather, lighting conditions*
Irritable Bowel Syndrome (constipation, gas, diarrhea)	*Consuming caffeine, consuming lactose, consuming gluten, consuming soy*
Juvenile arthritis flare ups	*Amount of physical activity, type of physical activity, eating gluten*
Menstrual cramps	*Amount of physical activity, consuming greasy foods, amount of sleep*

- *No -> exit survey*

Screening

- *Have you ever used web tools, apps, paper journals, spreadsheets or any other method of collecting data to understand more about potential triggers or causes of symptoms?*
  - *Yes*
  - *No -> exit survey*
- *Is the health condition you collected data about primarily a mental health condition?*
  - *No*
  - *Yes -> At this time, we are not focusing on collecting data about mental health conditions. We hope to extend this research in the future to include mental health conditions. If we do conduct a future study related to mental health data collection, may we contact you about participating in that study?*
    - *Yes -> collect email, exit survey*
    - *No -> exit survey*
- *Do you still have a record of the data you have collected? This might be spreadsheets, data inside tracking apps, handwritten records, or any other method of collecting data.*
  - *Yes*
  - *No -> Exit survey*

Details of what has been recorded

- *About what health conditions have you recorded data with the purpose of trying to understand potential triggers of symptoms?*
- *Please describe what potential triggers and symptoms of the health condition(s) you have recorded.*
- *Please describe anything else you have recorded related to the health condition(s).*
- *Please describe how you record this data. For example: did you record data in a spreadsheet, using a mobile phone or desktop app, or in a journal? How often did you record data?*
- *Please describe any understanding you have developed of potential triggers and symptoms of the health condition(s).*
- *Has your understanding of potential triggers and symptoms of the health condition(s) changed over time?*
  - *No*
  - *Yes -> Please describe how.*
- *Has your understanding of potential triggers and symptoms of the health condition(s) changed over time?*
  - *No*
  - *Yes -> Please describe how.*

Demographics/Contact Information

- *Age*
- *Gender*

- *Thank you for taking the time to complete this survey. If you are selected for this study, we will contact you about enrolling in the study. Please provide the best way for us to contact you if you are selected.*
  - *How would you like to be contacted?*
  - *What is the name of the person we should contact?*

C Session Details and Study Protocol

C.1 Session 1: Background Interview & Data Exploration

The goal of this semi-structured interview was: (1) to learn about participant experiences with and current practices of self-tracking for health, and (2) to review their self-tracked data and associated questions, in part because this would guide mapping of their data into the integration framework for Bayesian analysis. We asked participants questions about their self-tracking (e.g., why they tracked, what they tracked, what they learned from self-tracking, how they had learned through self-tracking, if/what they had used to reflect on their self-tracked data, what they still wanted to learn from self-tracking, if/how their tracking had changed over time). Next, the interviewer and participant looked through the participant's self-tracked data together, via screen share or synchronously on their own computers. Any unclear data entry was clarified, as well as which data fields corresponded to potential triggers and which corresponded to symptoms. Interview guide is as follows:

Hi _____. Thank you very much for taking the time to talk with me today. My name is _____ and I am a member of the research team who will be talking with you today. Today I expect the interview to take about 1 hour. I will be asking you about how you tracked the data you gave us, what your experience was while tracking, and what you learned from that data. As a reminder from the study consent sheet, we are recording this interview so that it may be referred back to by the research team. Do you have any questions before we begin?

- *Tell me about your experience with tracking data.*
- *Why did you begin tracking?*
- *Walk me through your data.*
- *For each of the variables you tracked,*
  - *Why did you track it?*
  - *Do you have any expectations for the relationship between that variable and your symptoms? What are they?*
  - *Have you gained any knowledge about how this variable affects your symptoms? Please elaborate.*
  - *Has this knowledge changed your tracking at all?*
- *Are there any other factors that you currently think may impact your symptoms?*
  - *Has this changed over time?*
- *Are there any other factors that you currently think do not impact your symptoms?*
  - *Has this changed over time?*
- *Did your reasons for tracking change over time?*
  - *How?*
  - *Did this affect how you tracked?*
- *Did what you tracked change over time?*
  - *How?*
  - *Why?*
  - *How did this affect your interpretation of the old versus new data?*
  - *How did this affect your evaluation?*
- *Did you reflect or look back on your tracking data after you collected it?*
  - *How? Any specific analysis techniques or specific visualizations?*
  - *Why did you choose your current interpretation method?*
  - *How confident do you feel in how you interpreted your data?*
  - *What questions were you able to answer?*
  - *Any questions you would like to answer but did not or were not able to? Why do you feel like you could not?*
- *Did you learn something during your tracking?*
  - *How did you learn this?*
  - *What did you learn?*
  - *Did what you learned shape your tracking after that? Why? Or why not?*
- *Is there anything you haven't learned that you would like to learn from tracking?*
- *Was there anything that made learning from your tracking easier?*
- *Was there anything that made learning from your tracking more difficult?*

C.2 Data Integration

After understanding participant self-tracking practices and data, researchers formatted participant data into an Excel spreadsheet to provide to the `bnlearn` Bayesian network learning algorithm and technology probe. A column was created for each trigger and symptom, recording each tracking entry (i.e., a single data point, typically one per day) in a row with associated trigger and symptom data in the relevant columns. Distinctions between trigger and symptom nodes were determined based on information about the participant's health condition and goals provided in the Session 1 interview. Data was also formatted into the correct data types to answer the questions participants identified. For example, P1 tracked how many ounces of tea they drank each day, but was only interested in whether drinking any tea affected their symptoms. The column for ounces of tea was therefore translated into a Boolean column containing a "yes" if they "did drink tea that day." If there was potential for temporal effects in the data, additional trigger columns were added to answer those questions or to better scaffold answers. For example, P3 tracked the amount of core nutrients they consumed every day and wanted to know any effect on their weight. Trigger columns were created to compute the average amount of each core nutrient consumed over the previous 7 days, aiming to account for the delays and fluctuations seen in weight. This spreadsheet was then read into R, run through the `bnlearn` Bayesian network learning package [40], and displayed in the technology probe's reflection interface. After this was complete, researchers explored each participant's results using the reflection interface and developed at least one suggested view for each participant based on their goals and results. This was meant to help participants familiarize themselves with the interface while also leaving room for participants to further explore their data using the reflection interface.

C.3 Session 2: Technology Probe Interaction

In this session, each participant was shown the technology probe, populated with their data, and was asked to explore and share their reactions. The interviewer shared their screen with the reflection interface, providing participants with a basic overview and explanation of the interface. The interviewer first explained the Scenarios tab, noted the suggested scenarios in the Overview tab, and presented the first suggested scenario. Next, the interviewer explained the Overview tab's full network diagram and settings. The interviewer then gave the participant control via the Zoom remote control option, asking them to explore their data using the reflection interface while thinking aloud [24]. The interviewer further probed whenever a participant seemed to discover something new, started exploring a new question or piece of data, or appeared confused. Further questioning included but was not limited to asking how participants arrived at a conclusion, why they decided to look at something, and what they were trying to accomplish. If participants had questions about how the reflection interface worked, they were first encouraged to try to find the answer themselves. If the participant was still confused or misunderstood the interface, the interviewer clarified so as to ensure minor points of interface confusion did not distract from the goal of the study. After exploration of their data using the reflection interface, participants were asked semi-structured interview questions soliciting their experience with using the technology probe, including what conclusions they derived, if they were going to change anything about their tracking or exposure to triggers based on their conclusions, if/how often they might use this probe on their own, what challenges they might expect if they were to use the probe on their own, and how this probe compared to previous methods they had used to make sense of their self-tracked data.

Now we will ask you to look at and explore a visualization of your tracking data. I'll give you a little tour, then I'll hand control of the screen over to you so you can click around and play with it. Take as long as you would like to explore this visualization. Please talk aloud as you look through the visualization and say what you see, what questions you have, what you are looking for, and anything you are thinking. Session protocol is as follows:

- *Explain*
  - *Scenarios*
    - *This is the "scenarios" tab. It allows you to explore different scenarios, that you specify on the left and see the predicted symptoms on the right. On the left, you can set multiple scenarios to compare how each one might impact your symptoms. Click the plus button to add a potential trigger or cause and use the color picker to specify the color on the graph on the right. You can use the up and down arrows here to specify the amount or you can manually type something in. If you use the arrows, it will only allow you to go up and down to the maximum and minimum value that was present in your data. If it shows up only as 1 or 0, that indicates "yes" or "no". 1 means yes and 0 means no. For example, [1 example for yes and no/up down arrows as present in the participant's data]. Press this plus to create and compare multiple scenarios. Once you have entered your scenarios, you can come to the right hand side and see the predicted symptom. Each dot represents a __ out of __ chance that this y-axis value of [symptom] will occur. Use the drop down menu to pick which symptom you want to look at on the y axis. Any questions so far?*
    - *Below is a list of confidence scores for every potential trigger and symptom pairing. So this says that it is __% confident that there is a relationship between ____ and ____. Here, 50% means unsure, 0% means it is very sure there is no relationship, and 100% means it is very sure there is a relationship.*
  - *Overall View*
    - *This shows all the relationships that the tool has found in your data. Each thing you tracked shows up as a bubble. An arrow between two bubbles represents that it is pretty sure one bubble impacts the other. The direction of the arrow indicates which bubble impacts which. So for example, it looks like it thinks ____ affects ____ in some way. What questions do you have?*
  - *Settings/Suggested Scenarios*
    - *Below the overall view are a few things, here you can see some suggested views that I've put together for you. If you click one, it will take you back to the scenarios tab and show you some pre-filled scenarios. You can also save your own scenarios if you want, and those will show up on the right. Below that are settings. On the right are different settings where you can specify a relationship between bubbles if you know there is or is not one there. So for example, if you know there is some kind of relationship between ____ and ____, you can click "definitely effects" here. Or if you know there is definitely not a relationship between ____ and ____, you can click "definitely does not effect" here. The default is "maybe effects" which means that the tool will try to figure out whether there is a relationship or not. When you make a selection, the tool will update and you will be able to immediately see the change in the overall view above and it will change what happens on the graph in the scenarios page. Then on the left hand side there is an option to select "trigger" or "contributor". A trigger means that any one potential bubble can cause one of your symptoms to occur. A contributor means that many things might have to happen before you would see a symptom occur. "Contributor" might run very slowly for you, just because there will be a lot of bubbles in the above graph. Any questions before I turn control over to you? You can feel free to ask more questions as you go.*
  - *Give time to play around with interface*
    - *Ask about immediate reactions (would you look at it, play with it on daily basis…)*
  - *Goals they have for data (have before)*
    - *Have them step through how to find answer, ask questions about how they did it, how they feel about it, how confident they are*
  - *Goals they had but couldn't answer before*
    - *Have them step through how to find answer, ask questions about how they did it, how they feel about it, how confident they are*
  - *Is there anything else you want to look into? Explore?*

    - *Have them explain what they are seeing and reactions as they go*
- *What did you look for in the tool?*
- *What about the tool was confusing or should be made clearer? Did anything confuse you about the tool?*
- *Did you see everything you expected to see in the data?*
- *Did you see anything you did not expect in the data?*
- *Do you see any of the same conclusions that you gained from your prior analyses?*
  - *Which ones?*
  - *Are you more or less confident in these conclusions now?*
- *Are there any conclusions that you had made in your prior analyses that you don't see in this visualization?*
  - *Which ones?*
  - *Are you more or less confident in these conclusions now?*
- *Are there any new conclusions that you see in this visualization that you didn't gain from your prior analysis?*
  - *Which ones?*
  - *Can you walk me through how you found this conclusion?*
- *Based on the [conclusions you saw / did not see] reflected in this data, do you plan to do anything different in your life? (e.g., change tracking, change behavior)*
- *Are there any new tracking questions that you have?*
  - *What brought about this question?*
  - *How would you try to answer these?*
  - *Would you like to try to answer these?*
- *Is there anyone you would like to share these results with?*
  - *Who?*
  - *For what purpose?*
  - *How would you show ____ the results?*
  - *Do you anticipate any problems with showing the results to ____?*
- *If you always had access to this tool, how often do you think you would use it?*
- *What challenges do you anticipate with using this tool on your own?*
- *Is there any additional information you wish the tool had included?*
- *How does the tool compare to previous methods for interpretation you have used/seen?*
- *What are 3 things you liked about the tool?*
- *What are 3 things you disliked about the tool*