DeepStress: Supporting Stressful Context Sensemaking in Personal Informatics Systems Using a Quasi-experimental Approach

Gyuwon Jung
KAIST
School of Computing
Daejeon, South Korea
gwjung@kaist.ac.kr

Sangjun Park
KAIST
School of Computing
Daejeon, South Korea
sangjun@kaist.ac.kr

Uichin Lee*
KAIST
School of Computing
Daejeon, South Korea
uclee@kaist.ac.kr

ABSTRACT

Personal informatics (PI) systems are widely used in various domains such as mental health to provide insights from self-tracking data for behavior change. Users are highly interested in examining relationships from the self-tracking data, but identifying causality is still considered challenging. In this study, we design DeepStress, a PI system that helps users analyze contextual factors causally related to stress. DeepStress leverages a quasi-experimental approach to address potential biases related to confounding factors. To explore the user experience of DeepStress, we conducted a user study and a follow-up diary study using participants' own self-tracking data collected for 6 weeks. Our results show that DeepStress helps users consider multiple contexts when investigating causalities and use the results to manage their stress in everyday life. We discuss design implications for causality support in PI systems.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**; **User studies**.

KEYWORDS

Personal Informatics, Mental Health, Quasi-experimental Approach, Causal Relationship

ACM Reference Format:
Gyuwon Jung, Sangjun Park, and Uichin Lee. 2024. DeepStress: Supporting Stressful Context Sensemaking in Personal Informatics Systems Using a Quasi-experimental Approach. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3613904.3642766

1 INTRODUCTION

As digital devices such as smartphones make it easier for users to self-track diverse data in their daily lives, their interest in gaining insights from their own data is increasing. Personal informatics (PI) systems [46] are designed to meet this demand by helping users reflect on their self-tracking data and plan actions to improve their future selves [2, 67]. PI users want to understand the impact of their behaviors on health goals [11]. Thus, they generate hypotheses about influencing factors, evaluate them using self-tracking data, and revise their strategies to improve health.

PI systems target diverse domains of health and well-being, including physical activity, chronic condition management, sleep, diet, and weight [22]. This work focuses on stress management in mental health. Stress is associated with contextual factors such as activity, place, social setting, or time. For instance, some people become stressed around others, regardless of the activity or location. As context could affect one's stress, a holistic view of contextual factors is needed to understand the causes of stress. Moreover, users should identify causally related contextual factors for effective long-term stress management, given their routine and repetitive occurrence.

Yet, prior PI systems are limited in effectively supporting users in investigating contexts causally linked to stress. Existing studies have focused on presenting a correlation between factors [3, 8, 49], but analyses overlooking the complex relationships within the data may lead to different conclusions. For example, when identifying the effects of studying on stress, it is necessary to compare stress levels during studying and non-studying. However, comparing studying in a cafe to not studying in a dormitory may not be reasonable, as the difference in place could also influence the causal relationship. Instead, all other factors besides the factor of interest should be identical (e.g., comparing stress levels while studying in the cafe to those while not studying in the cafe) to ensure that the difference in outcome comes only from the factor of interest. This shift from an 'apples-to-oranges comparison' to an 'apples-to-apples comparison' minimizes fundamental differences in comparison targets.

PI users unaware of this can be susceptible to errors if causality is identified solely based on correlation, without considering external factors introducing bias in comparison groups. Alternatively, some studies have employed self-experimentation to investigate causal relationships [18, 19, 35]. This approach is rigorous and scientifically valid, but its utility is limited because the random assignment of conditions is challenging in daily life. Asking users to always adhere to required experimental settings when investigating causal relationships between context and stress is difficult.

To address such practical issues, we suggest a quasi-experimental approach with observational data in PI system design. Unlike the experimental approach, this method uses observational data obtained without random allocation of samples to the treatment and

*Corresponding author

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642766

control groups. Before analyzing the causality of a specific context, it balances the distribution of confounding factors (i.e., contexts other than that under analysis) between the two groups to minimize their effect on stress. If a significant difference in stress levels exists after balancing, we may conclude that the context has a causal relationship with stress. We opted to employ 'matching' among the various balancing methods, as it is intuitive and widely used [28, 33, 45, 79, 80]. The quasi-experimental approach is less rigorous than experimentation in controlling for confounders, but beneficial since it allows users to mimic the experimental setting in data analysis even when experimentation is not applicable or practical. Human-computer interaction (HCI) studies have used this approach to understand the relationships between smartphone usage and contexts or emotional states [54, 82]. Notably, there is a lack of research that leverages this method in PI systems to help individuals analyze causality.

In this study, we designed DeepStress, a PI system aiding users in exploring stressful contexts (i.e., activities, places, social settings, and times) using a quasi-experimental approach. As a PI system, it visualizes self-tracked data such as stress changes over time, and provides summarized stress information for each context. In addition, it offers a list of stressful contexts (i.e., contexts demonstrating causality) with relevant contextual information.

The primary goal of this study is to explore user experiences in exploring causal relationships through DeepStress. We specifically examined (1) how DeepStress supports users in investigating stressful contexts, (2) how users understand and interpret causality results, (3) which challenges occur while using DeepStress, and (4) how users utilize insights from DeepStress in their daily lives. DeepStress facilitated data-driven self-reflection, enabling users to pinpoint stressful contexts and understand relationships between various contexts. We revealed a sensemaking process about how users reflect on their stress and interpret the causal relationships. A follow-up diary study indicated that causal insights helped participants manage their daily routines and stressful contexts.

Overall, our contribution can be summarized as follows:

- We demonstrated a use case for a quasi-experimental approach in PI systems for exploring unbiased causal relationships between contexts and stress.
- We investigated how users explored and interpreted causal relationships by leveraging sensemaking frameworks.
- We discussed challenges and design considerations for supporting causal analysis in PI systems.

2 RELATED WORK

2.1 Data-driven Relationship Analyses in Personal Informatics Systems

As daily life tracking with smart devices has become routine, people leverage these devices to track physical, behavioral, and contextual information [81]. Personal informatics (PI) systems help users reflect on and understand themselves based on the gathered data [46]. Previous studies examined PI system usage behavior and intentions, proposing models to identify barriers and better assist users [23, 46, 47]. PI systems are typically utilized for health management, enhancing quality of life, and seeking new experiences [11], encouraging users to shape a better future self [2, 67].

PI systems are useful in the affective computing domain, interacting with users to sense, analyze, and respond to their emotions [61]. They infer emotional states from data, including physiological signals, facial expressions, and body gestures [26, 27, 83]. Smartphones and wearables also detect user states via passive sensors (e.g., accelerometer, GPS, etc.) and interactions (e.g., app usage) or questionnaires [55, 63]. The systems then provide users with information about mental states, facilitating their self-reflection. For example, they encourage users to (1) review past events in connection with their emotions or mood [32, 43], (2) explore triggers influencing them [78], or (3) examine affective states or stress levels along with contextual information [42, 53]. Such systems employ visual elements to help users' understanding [6, 84] and provide interventions to address negative emotions at opportune moments [29, 73].

PI systems are expected to provide diverse data-driven insights [9, 10]. Particularly, PI users are interested in linking different types of factors in their data [69] such as exploring correlation and causality [66]. Fleck and Fitzpatrick [25] called this 'dialogic reflection,' where individuals try to determine relationships between their experiences, hypothesize why they happened, and generate explanations. HCI studies have proposed PI systems to help users understand relationships between diverse factors. Bentley et al. [3] designed a system that pinpoints correlations between contexts and well-being indicators. PI systems also assessed sleep quality based on the contributing factors or visually displayed correlated contexts [8, 49]. Furthermore, they predicted users' stress using contextual factors and delivered useful results [38].

Still, most existing PI studies rely on correlation rather than causality when analyzing contextual factors' impact, limiting the rigor of the results. For causal analysis, studies have allowed users to conduct self-experimentation, such as identifying problematic food triggers [35], activities for improving sleep [18, 19], or causal relationships between various activities and conditions [17]. Self-experimentation is practical when manipulating the condition is readily available following the system's random assignment. However, controlling contextual factors may not be easy; for example, users are not always able to change their place immediately to examine the causality between places and stress levels. Thus, we chose an alternative way, a quasi-experimental approach, to address this limitation and enable causal analysis through PI systems.

2.2 Investigating Causal Relationships with Quasi-experimental Approaches

When examining causal relationships between factors, the Randomized Controlled Trial (RCT) is considered the gold standard [60, 76]. This method randomly assigns each subject to either the control or treated group and evaluates whether there is a statistically significant difference in outcome between the two groups. Random allocation ensures that groups are identical except for the treatment condition, implying any outcome differences are solely due to the treatment. This process minimizes bias arising from external factors like confounding variables, distinguishing between correlation and causality (as the well-known phrase "correlation does not imply causation"). Moreover, the concept of RCT can be extended to an individual level called 'single-case designs' (or n-of-1 trials) [15, 50],

using a certain participant as a control for themselves to examine the effect of the treatment on their outcome.

However, RCTs are not always applicable for practical reasons, such as ethical concerns, costs related to sample size, and the spillover effect where treatment in one group affects another [70, 74]. In such situations, a quasi-experimental design [13] can be used to verify causality, similar to RCTs addressing counterfactual (i.e., what-if) questions. However, this approach is considered less rigorous due to the absence of random allocation of subjects. Thus, controlling for bias before comparing groups is necessary to ensure that the difference in outcomes originates from the treatment.

One method is to compare subjects having similar combinations of confounding variables, known as 'matching' [28]. In matching, it is essential to pair the most similar subjects to mimic the random assignment. Several methods have been proposed to define the distance (similarity) for matching samples. For instance, researchers can directly compare the similarity of two subjects using the Euclidean or Mahalanobis distance [20] of confounding variables, or calculate a scalar value such as propensity score (i.e., the probability of being allocated to control or treated group given confounding variables) [71] as a distance metric. Another approach is coarsened exact matching (CEM) [31], which coarsens each confounding variable into discrete categories (or 'bins') and matches subjects if they are in the same bin. When matching subjects, additional options can be considered, including the matching ratio and the algorithm for minimizing the total distance [79].

Previous studies have analyzed causality in human behavior by implementing the quasi-experimental approach in a single-case design setting. Tsapeli and Musolesi [82] investigated the causality between contextual factors and stress levels, while Mehrotra et al. [54] explored the relationship between emotions and mobile phone interactions. In both studies, analysis was conducted for each individual user, acknowledging variations in lifestyles, smartphone usage, and emotions. Moreover, the quasi-experimental approach was preferable, as participants could not manipulate their emotions to evaluate their effect on smartphone usage. Referring to these studies, we also applied the quasi-experimental approach in investigating causality between contextual factors and stress levels.

2.3 Making Sense of Self-tracking Data and Visualized Information

Sensemaking is the process of retrieving, organizing, and utilizing information. Russell et al. [72] illustrated this concept as a process of generating and modifying a representation to process a task, iteratively reducing the associated costs. Pirolli and Card [62] extended the sensemaking process by introducing two main loops: (1) a foraging loop where the process between retrieving relevant information and creating schema happens, and (2) a sensemaking loop where the generation of hypothetical mental models and their assessment occur to find the optimal one. The sensemaking concept was further detailed in Klein et al.'s work [41], where they introduced a data-frame theory of sensemaking. They distinguished the concept of data and frame as instance and structure and described the sensemaking process as creating and updating the frame using data. In their framework, individuals may reinforce the frame or seek another frame that better explains the data, depending on whether a gap exists between the data and the current frame.

Sensemaking was further elaborated in self-tracking. Mamykina et al. [52] proposed a sensemaking framework to explain the process involved in managing chronic diseases. Their framework comprised two modes depending on the gaps in understanding new information: (1) a sensemaking mode where individuals actively investigate data and generate explanations to determine actions, and (2) a habitual mode where they passively engage with their data and use pre-existing knowledge to continue routine actions. During the construction of a new model in the sensemaking mode, individuals form hypotheses (or explanations) from the data and test whether the hypotheses adequately describe the collected data [51].

Sensemaking actively occurs when exploring and understanding data in a visualized form. For example, when visualizing stress changes over time with contextual factors such as activities and places, users inspect the relationships among them and design personalized, just-in-time interventions [75]. In addition, PI system users could follow the sensemaking process with their health data, for instance, by creating the initial frame of explaining what affects their symptoms, supporting or revising the frame based on the newly found information from the data, and selecting the most reasonable frame for them [64]. However, such information visualization should be carefully designed while considering the sensemaking process, as users may not be familiar with integrating different types of factors, rely heavily on their pre-existing knowledge, and result in a biased interpretation if they inspect the data only using their eyes [9, 24, 57, 66]. Our study explored how the sensemaking process occurs when users explore the causal relationships in their data through self-tracking data visualization.

3 USER STUDY DESIGN

As illustrated in Figure 1, this study was conducted following several steps. We recruited participants from a university and collected self-reported data about their contexts and stress levels over six weeks. One week after the data collection began, we conducted a preliminary interview to investigate user needs for our system. Then, we developed DeepStress based on insights gathered from the interview and the literature review. After the remaining five-week data collection, we conducted a user study in a lab setting to allow participants to explore their stressful contexts using DeepStress. Following this, we asked them to use DeepStress freely in their daily lives for one week and record a diary about their experiences. Note that each participant followed all these steps.

In Section 3, we provide information about the participants (Section 3.1), the collected data (Section 3.2), and the procedures of the preliminary interview and user study (Section 3.3). Next, we describe the design process of DeepStress based on the preliminary interview (Section 4) and present the user study results (Section 5).

3.1 Participants Recruitment

For this study, participants were recruited through an online community at a large university. Initial applicants were screened using the Perceived Stress Scale (PSS) survey [12], and recruitment occurred only if the score exceeded 13 (i.e., moderate to high perceived

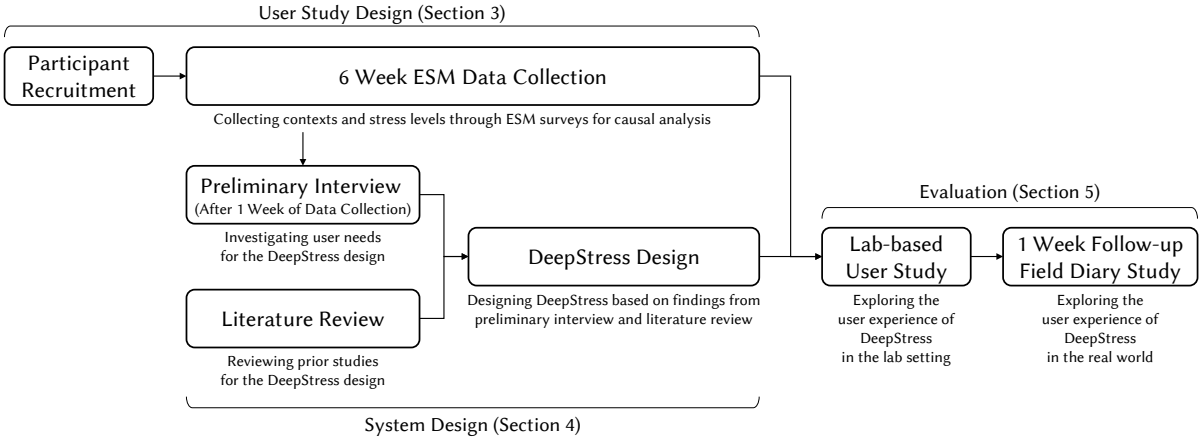

Figure 1: An overview of the study procedure, illustrating each step along with corresponding sections.

stress). This criterion was set for two main reasons. Firstly, participants with low perceived stress might be less motivated to engage in a stress management study, making them less suitable as our target users. Secondly, individuals consistently experiencing low stress levels may yield biased data with small variance, resulting in fewer meaningful causal insights into stressful contexts. As a result, 24 participants were recruited (9 women, 15 men; age: M = 21.3, SD = 2.1), and their average PSS scores were 23.3 (SD: 5.0).

The participants in the study had diverse academic majors, including natural sciences, engineering, industrial design, and business management. Nevertheless, participants had the opportunity to grasp the concept of correlation through data analysis in their mandatory experimental courses, such as physics experiments, during their freshman year. Thus, they might have developed a certain level of understanding of experimental design (e.g., setting up control and treatment groups), implying that the notion of confounding variables may not have been entirely unfamiliar to them.

3.2 Data Collection

The participants were instructed to collect data on their stress and contexts for six weeks, as described in Figure 1. Since we designed and evaluated a PI system, it was more reasonable to evaluate the system using participants' own data rather than those from others.

We developed a mobile application that uses the Experience Sampling Method (ESM) [44] to collect data. The participants reported their stress levels and contexts when responding. Each ESM session consisted of four questions: one regarding the participant's stress level and three others related to context information. The participants rated their stress levels on a 5-point Likert scale, ranging from 1 (not stressed at all) to 5 (very stressed). Also, they provided three types of contextual information containing where they had been (place), what they had been doing (activity), and who they had been with (social setting) until they responded to the survey.

To reduce the data collection burden on participants, we provided a list of contexts (Table 1) referring to similar ESM studies conducted with college students [37, 77]. The participants chose eight contexts for each context type from a given list before collecting data. They made their choices based on the frequency of occurrence, and additional contexts could also be reported through an open-ended question. Consequently, each ESM response consisted of stress level, place, activity, and social setting. Note that time was not reported, as it was logged automatically when responding to the ESM survey. In addition, if participants encountered multiple contexts within the same type when reporting, we requested them to report the one most relevant or impactful to them.

We also configured the ESM survey frequency in the data collection application. A minimum of 30-minute gap between samples was set to avoid excessive repetition of ESM surveys. Consequently, after reporting one ESM survey, the app remained disabled for 30 minutes. Moreover, to prevent insufficient data, a reminder was sent if no response was received within one hour after the system was enabled again. The notification was sent every hour until a response was received, and the system was disabled for 30 minutes again as soon as it received any response.

Participants could customize the start/end time of receiving notifications to avoid receiving them late at night, and they set the duration for receiving ESM notifications to an average of 15.7 hours (SD: 2.9). By setting the notification permission period to a total of N hours, participants could potentially receive up to N-1 ESM notifications daily (in the case where they do not respond to any ESM at all). However, the actual ESM notifications received per day averaged

Context Type	Contexts Provided By the ESM Survey
Place	Home, Classroom, Dormitory, Library, Restaurant, Cafe, Pub, Club room, Laboratory, Place for exercise, Place for leisure, Outdoor, Place for part-time job, Public transportation
Activity	Class, Studying, Research, Resting, Meeting, Eating, Drinking, Part-time work, Club activity, Socializing, Leisure activity, Exercise, Moving
Social Setting	Alone, Family, Boyfriend/Girlfriend, Roommate, Friend, Colleague, Professor

Table 1: The list of contexts provided by the ESM survey application, referring to Kim et al. [37]. These contexts are known to be common in the daily lives of university students.

4.9 times (SD: 1.6), suggesting that participants reported their states frequently, even without notifications. Existing studies suggested that the notification frequency we set was reasonable [34].

As a result, the participants reported 566.9 ESM surveys on average (SD: 156.8, max:867, min: 258). Following the data collection, we asked participants to provide three to four bins of contexts for each context type, organized based on their similar characteristics. For example, the 'activity' type could be divided into three bins: (1) academic and work (class, study, research, meeting), (2) hobbies (social activities, leisure activities), and (3) health (rests, meals). These bins were employed to match similar samples using CEM, aiming to balance confounding factors for causal analysis. Further details are provided in Section 4.3.

3.3 Study Procedure and Data Analysis

Before designing a PI system that analyzes stressful contexts using a quasi-experimental approach, we established key design considerations through a literature review and preliminary interviews. We focused on previous studies that examine how PI systems support users in reflecting on and analyzing self-tracking data. Given that we aim to offer causal results, we investigated how previous systems helped users identify meaningful relationships between factors. Moreover, we examined challenges in understanding and interpreting relationships in data through existing PI systems.

A brief preliminary interview was conducted to investigate the participants' needs for the PI system supporting causal analysis. This interview, held one week after data collection began, uncovered analysis needs arising from participants' actual experiences of self-tracking. We asked two main questions: (1) What information should our PI system provide? and (2) How would you analyze the data to identify the 'cause' contexts of stress? These questions delved into the information to establish connections between contexts and to assess participants' understanding of deriving causality from data. DeepStress was then designed based on the key considerations obtained from interviews and the literature review.

In the lab-based user study, we examined the user experience of exploring stressful contexts using DeepStress in a lab setting. We first introduced our study and provided a brief overview of the main functions of DeepStress. Then, DeepStress was installed on each participant's smartphone, and they were given up to 30 minutes to freely use the system to explore their stressful contexts. Participants could view their stress and context information, along with DeepStress' identified stressful contexts and relevant information.

After using DeepStress, we measured its usability with the System Usability Scale (SUS) [5]. This allowed quantitative evaluation of our design's appropriateness for supporting PI users' understanding of causality from their data. We also conducted semi-structured interviews to investigate how people make sense of DeepStress. All interview sessions were recorded and transcribed to capture participants' responses in detail. We analyzed the interviews by repeatedly reading the transcripts, generated initial codes, and organized the codes and data into relevant themes [4]. The themes and codes were reviewed and revised iteratively by conducting affinity diagraming until all the researchers agreed on the final themes.

In the follow-up diary study, we examined participants' experiences of using DeepStress in the real world, focusing on how they applied the findings from DeepStress in their daily lives. Participants revisited the collected data and analysis results without additional data collection. We requested them to run DeepStress at least once a day for a week and write a diary describing their experience of using it. To guide them in providing detailed experiences, we gave sample questions, such as what they explored, whether they utilized the information, and whether any information influenced their thoughts or actions. The participants submitted their diaries after the one-week study period, and we analyzed the diary contents following a process similar to the lab-based user study.

The participants were paid 10 USD as compensation for each preliminary interview, user study, and follow-up diary study. For the ESM survey, they were given 120 USD (20 USD per week) as a baseline, considering the repetitive tasks over a long period. This study was approved by the Institutional Review Board (IRB) of a university and obtained written consent from all participants.

4 SYSTEM DESIGN

4.1 Design Rationale

Based on the literature review and the preliminary interview, we summarized our findings into three key design considerations.

4.1.1 The System Allows Users to Navigate Past Stress History with Contextual Information. One basic requirement of PI systems is allowing users to explore the data to recall the past. As illustrated in previous studies [9, 10], PI users typically navigate through collected data to interpret specific situations or answer their own questions. Cho et al. [7] noted that commercial PI systems commonly present data by visualizing it or adding further explanations for users to review. In presenting data, we may refer to Sharmin et al.'s approaches [75], offering stress levels over time or for each context to establish connections between stress and contexts.

Our interview also revealed similar needs to explore stressful contexts through PI systems. Most participants wanted to monitor stress changes over time. In addition, they were interested in how stress levels varied across different contexts, requesting summaries of stress levels for each context. "*If I recorded 'library' as a place, I hope to see the number of records and the overall stress level in that place*" (P16). They suggested that the average stress level for each context may be useful in comparing the contexts. Moreover, the participants wanted to see their records while considering the relationship between various factors. "*I think my stress levels may vary if I do the same activity in different places. I'd like to see those relationships*" (P08). P17 also mentioned, "*Showing what I've done frequently in that place could be useful. Maybe, I can think about how they relate to my stress.*" Thus, we chose to provide summaries and details of data and to help users diversely revisit their data.

4.1.2 The System Analyzes and Delivers the Causal Relationship. Prior research reported that PI users face challenges when analyzing data, investigating relationships, and interpreting findings correctly [57, 66]. Similarly, PI users may experience difficulties when examining causal relationships between factors. The effect of confounding factors should be minimized, but this was not well supported in existing PI systems. For instance, PI systems leveraging correlations were limited in addressing confounding factors [49], leading to misunderstandings of correlation as causality without

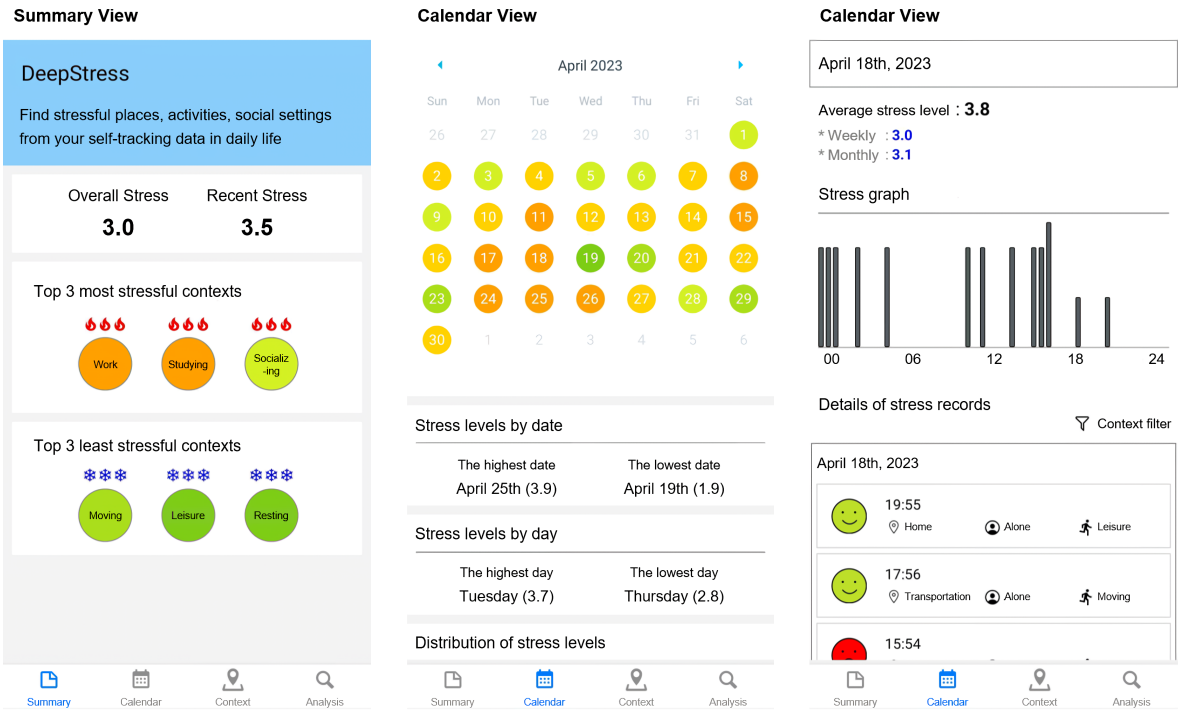

Figure 2: The summary view and calendar view. These features overview stress levels and important contexts (a). They allow users to explore stress trends over time at a summarized level with a calendar (b) and in detail by reviewing the timeline (c).

considering dependencies between contextual factors. In this regard, previous studies highlighted the importance of designing PI systems to prevent users from obtaining biased results [9].

In the interview, participants had difficulties in identifying causal relationships from self-tracking data. In the simplest way, many participants suggested using the average stress level of each context, as P12 stated, "*I would calculate the average stress level of all places and then compare it with that in a particular place.*" They also thought that comparing stress levels between contexts would inform them of relative stress and help them evaluate causal relationships. However, some participants noticed that it may not be clear which of the contexts resulted in the high or low stress. "*Since I can perform various activities in a specific place, it may be difficult to pinpoint whether my stress is due to the place or activity*" (P01). We found it challenging for participants to independently identify causal relationships, prompting the PI system to analyze the causalities and deliver the results to them.

4.1.3 The System Provides Contextual Explorations Based on Causality to Manage Stress. One of the major purposes of using PI systems is to change behaviors and achieve goals such as improving health [46, 47]. Choe et al. [11] illustrated that PI users are interested in monitoring their current conditions, identifying factors that affect their health, and making data-driven health decisions.

Our participants also showed interest in utilizing the analysis results to manage their stress in their daily lives. "*If I can check where I get stressed a lot, maybe I can plan my day in a way to avoid such contexts as possible*" (P08). Some participants mentioned that they could leverage the analysis when they experience mental health issues. P02 noted, "*If I undergo burnout or depression, I can get some information from the system to investigate the cause.*" In addition, they wanted to identify ways to reduce stress within various contextual relationships. Based on these responses, we determined to present how much each context affects stress as a whole and within a certain context.

4.2 DeepStress

Based on the key design considerations derived, we designed DeepStress, a PI system that helps users understand their stressful contexts in everyday life. Unlike previous systems, it directly showed (1) stressful contexts determined by causal analysis and (2) information about other relevant contexts. This was intended to avoid false conclusions from simple investigations (e.g., exploring similar trends between two factors only with eyeballing) and allow users to consider the relationships between multiple contexts when examining stressful (i.e., causally related) contexts. The system consists of four main views for describing stressful contexts: (1) Summary, (2) Calendar, (3) Context, and (4) Analysis.

4.2.1 Summary View. The summary view provides a brief overview of the user's stress level and stressful contexts as a landing page, letting users quickly grasp the overall stress status (Figure 2 (a)).

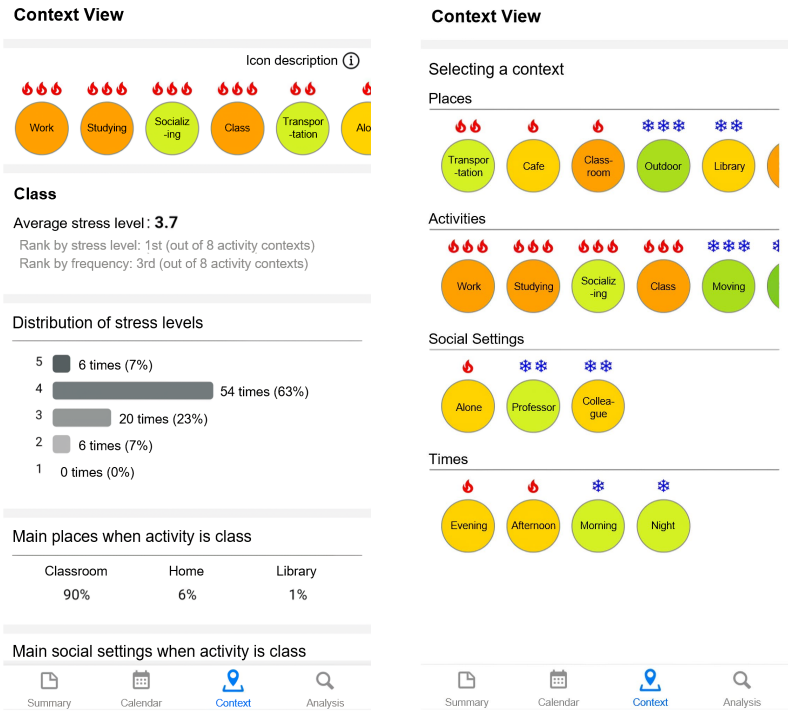

Figure 3: The context view. This feature enables users to explore each context, about stress levels and co-occurring contexts (a). When selecting contexts, it overviews the causal relationships of each context, represented by fire and ice icons (b).

This view shows the average stress level for the entire record and the most recent data. It also highlights the three most stressful contexts (i.e., increasing stress) and least stressful (i.e., decreasing stress). Users can then navigate to the calendar view for stress by date or the context view for details about each context.

4.2.2 Calendar View. The calendar view displays summary and detailed records together, allowing users to reflect on previous stress and contexts they experienced. The default stress information is provided monthly (Figure 2 (b)); this monthly view displays the average stress level for each day using colored circles in the calendar. The view marks days with high stress levels (near level 5) in red and those with low stress levels (near level 1) in green. It also summarizes the highest and lowest stress levels by date and day, respectively. The distribution of stress levels recorded over a month is also presented to assist users in understanding their overall stress.

When selecting a specific date from the calendar, the system provides detailed information about that date (Figure 2 (c)). It shows the average stress level of that date, along with the weekly and monthly averages. Below that, two components are placed; a graph describing changes in stress levels over time and a timeline of detailed user-reported records. For each record in the timeline, stress levels are displayed using facial expressions (e.g., smile, anger, etc.), and contextual factors composed of time, place, social setting, and activity are provided. Moreover, a context filter is implemented to extract and show only the records with that context.

4.2.3 Context View. We described the stressful context using two separate views that emphasize different aspects. The context view (Figure 3) primarily describes what happened in a given context, including the overall stress level within the context and its relationship with other contexts. Conversely, the analysis view (Figure 4) addresses the causal relationship between context and stress, taking the effects of other contexts into account.

To help users quickly review and explore the contexts, we organized a list of contexts at the top of both the context and analysis view (the upper section of Figure 3 and Figure 4). For each context, the relationship with stress is represented using a circle with a color inside and a set of fire or ice icons above the circle. The circle's color denotes the average stress level for a given context, using the same color spectrum as in the calendar view. The icons represent the causal relationship: fire and ice icons indicate the direction of the causality (i.e., increasing or decreasing the stress) and their quantity denotes statistical significance. These icons are not displayed if there is no causality between the given context and stress. The circle's color and the fire/ice icon have different meanings since the latter considers other relevant contexts (i.e., confounding factors) in calculation whereas the former does not. We intended this design to inform users that high-stress contexts are not necessarily causal.

The context view (Figure 3 (a)) first shows the average stress level of the context and its ranking against other contexts of the same

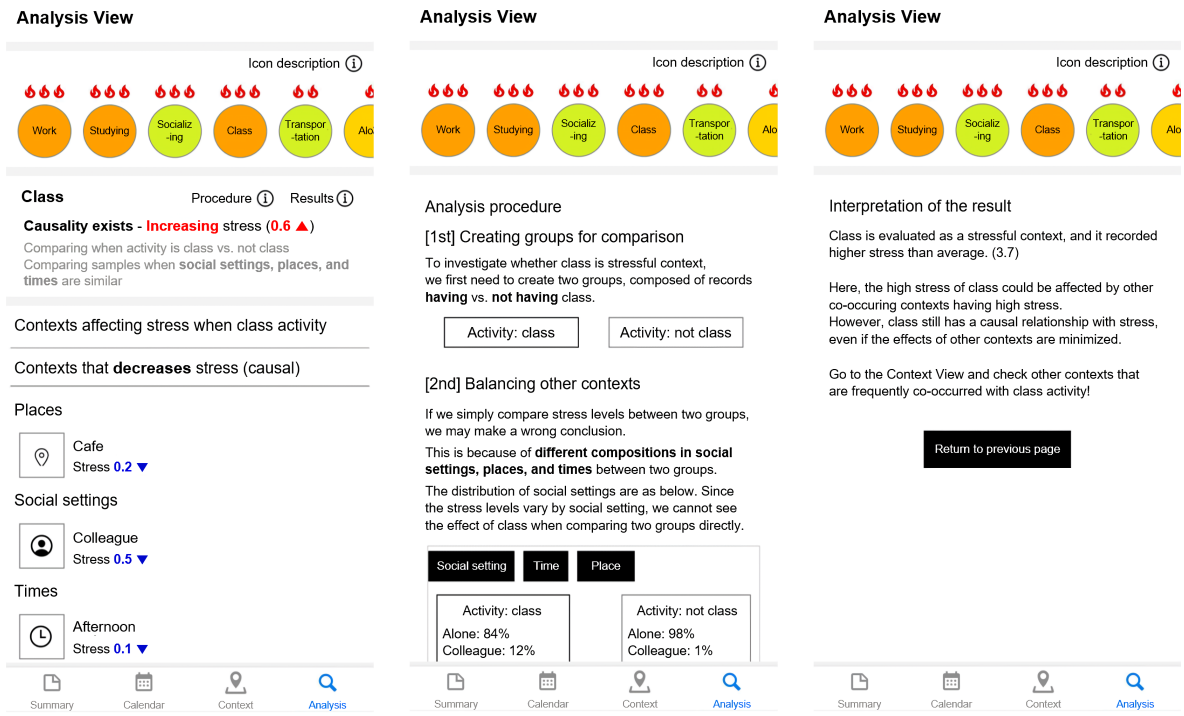

Figure 4: The analysis view. This feature offers the results of causal analysis at two levels: determining whether the context is causally linked to stress levels and identifying other factors affecting stress within the given context (a). It illustrates the process of causal analysis (b) and provides a brief interpretation of the results (c).

type. For instance, if we choose a 'class' context, the average stress level in this context and its ranking compared to other 'activity' contexts are displayed. The system shows the context's frequency ranking among the same type, denoting its relative frequency. The users can check the distribution of stress levels for the given context. We also displayed three of the most co-occurring contexts from each context type to illustrate their relationship. For example, the activity 'class' may occur frequently in certain places (e.g., classroom, library, dormitory), social settings (e.g., friends, colleagues, alone), and during the day (e.g., afternoon, evening, morning). Furthermore, the system shows the frequency of co-occurring (in ratio), informing how closely they are related to the context being explored. Figure 3 (b) shows that users can move to other contexts while reviewing information of the whole context by type.

4.2.4 Analysis View. The analysis view (Figure 4 (a)) illustrates the causal relationship between the chosen context and stress. It first describes the existence of the causal relationship (e.g., taking a class is stressful) and its effect on stress levels (e.g., 0.6 points increase). For users' understanding of the result, we added an explanation in a short sentence about (1) the samples compared in the analysis and (2) the other context types balanced to minimize their effects on the stress level. For more details about the quasi-experimental approach, we added separate pages explaining the analysis procedure (Figure 4 (b)) and the interpretation of the result (Figure 4 (c)).

Moreover, for a given context, we listed the contexts of the remaining types that may increase or decrease the stress level. We extracted samples containing the chosen context and conducted the quasi-experimental approach to identify the causal relationship between the context of other types and stress. For example, if the chosen context was 'class' (activity type), the system will only use samples with 'class' to conduct causal analysis for places, social settings, and time. As a consequence, users were allowed to explore which of the contexts of the remaining types would affect their stress within a given context.

4.3 Quasi-experimental Approach

The essence of the quasi-experimental design lies in mimicking a controlled experiment by minimizing the differences in confounding factors between the control and treated groups. To achieve this goal, we utilized CEM, a matching method widely employed to identify similar pairs of samples using predefined bins. Ideally, achieving control over confounding variables involves matching identical samples except for the treated state; however, this is challenging to achieve in practice. CEM considers samples sufficiently similar (i.e., comparable) if each of their confounding variables falls in the same coarsened bin [31]. Previous studies have demonstrated the advantages of this method in controlling confounders [40]. Unlike other methods employing scalar values as distance metrics

Place bins	Social setting bins	Time bins
P1: Home, Dormitory	S1: Alone, Family	T1: Morning
P2: Public transportation, Outdoor	S2: Professor, Colleague	T2: Afternoon
P3: Classroom, Library, Laboratory	S3: Boyfriend/Girlfriend, Friend	T3: Evening

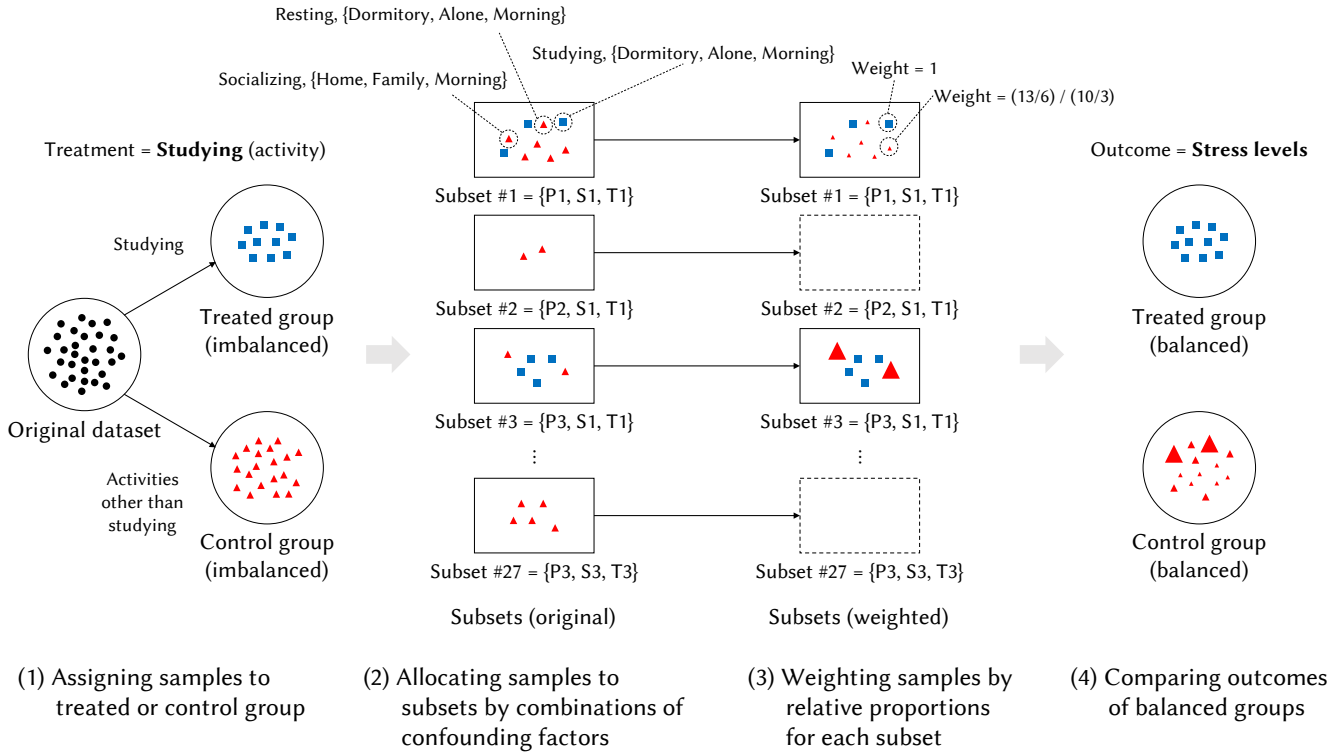

Figure 5: The process of conducting CEM as a quasi-experimental approach in four steps. The table at the top right illustrates the coarsened bins assumed in this example case, including three bins for each context (i.e., place, social setting, and time) that should be balanced during matching. The blue squares and red triangles denote samples from the treated and control groups, respectively, and their changed size represents the weights assigned to them after matching.

(e.g., propensity score matching), the balance of one factor does not impact the balance of another.

While binning is typically applied to categorize continuous variables, we extended this concept to match samples based on the similarity of their contexts. As shown in Section 3.2, samples with 'rests' or 'meals' can be considered similar in activity type context as they belong to the same 'health' bin. Matching these samples is feasible when other context types (place, social setting, and time) are also similar, meaning they were included in the same bin for each context type. In the following explanation, we name the combination of bins for each context type a "subset" to avoid confusion.

Our quasi-experimental approach using CEM is shown in Figure 5, illustrating a sample case investigating whether "studying" causes an increase in stress. (1) When a user chooses a context for investigation, like "studying," samples from the original dataset are allocated to either the treated group (with studying activity) or the control group (with an activity other than studying). (2) Next, samples from both groups are assigned to the same subset if the combinations of coarsened bins for their confounding factors are identical (e.g., Subset #1 = P1, S1, T1). By comparing coarsened bins instead of the raw context, this approach allows matching sufficiently similar samples, even when not exactly the same (e.g., considering home and dormitory similar, both included in bin P1). (3) After assigning samples to subsets, the samples in each subset are weighted based on the ratio of treated and control samples. Samples within subsets exclusively consisting of treated or control are discarded as they are considered unmatched (Subset #2 and #27 in Figure 5). Throughout this process, the treated and control groups achieve balance, minimizing bias introduced by confounding factors. (4) Finally, the average treatment effect on treated units (ATT) [80] is calculated using linear regression to estimate the effect of the treatment context on stress levels. In our example, we may conclude that "studying causes more stress" if there is a statistically significant difference in the stress levels of the two balanced groups.

Note that the balanced dataset after matching can differ from the original dataset because unmatched samples are discarded and matched ones are weighted by their relative ratio. This may lead to different conclusions between simple correlation and causality. For

instance, there could be a significant correlation between a specific context and stress levels, but no causal relationship. In addition, this approach is conducted on one user's data, employing the same individual's data for both treated and control conditions. As the comparison is within one user's data, interpersonal differences like varying lifestyles, sensitivity to stress, or self-evaluation of stress do not affect the inferred results. Moreover, causal relationships can vary among users, and even for the same context, stress can be increased, decreased, or not changed.

Further details regarding the quasi-experimental approach and each participant's causal analysis results are provided in the Supplementary Material.

5 EVALUATION

5.1 RQ1: How Does DeepStress Support Users in Exploring Their Stressful Contexts?

We analyzed how DeepStress supported the user's exploration of stressful contexts. The SUS survey yielded a mean score of 74.1 (SD: 15.8) from 24 participants, indicating good usability [1] for supporting PI users in exploring causality. Our interview further revealed the various ways that DeepStress supported users.

5.1.1 Enabling Participants to Recall Past Context and Stress States Readily. As in the existing PI systems, DeepStress supported participants to easily recall what happened in the past by presenting their previous records. Most participants started reflecting on their stress by reviewing their historical data through the calendar view and context view. They tried to make connections between their life patterns and stress levels, as P20 mentioned. "*I could find my stress level increasing as the midterm comes and decreasing after the exams.*" Also, DeepStress helped them recall the details of previous contexts and stress.

The colored circles in the calendar view facilitated participants in recognizing deviations in their stress levels and recalling what was behind them. The participants were particularly interested in days with higher stress levels. "*I took a closer look at the days with distinct red colors. With the timeline, I could remember why my stress was high*" (P19). In addition, participants could explore trends in stress changes and think about recurring events that could affect stress. "*I found relatively high stress on Monday, Wednesday, and Friday. This pattern may be related to my part-time job*" (P23).

In the context view, the participants could check how they evaluated a particular context in terms of stress level, which was a novel experience for them. "*It was impressive because I had no chance to analyze my stress by place and time*" (P17). In particular, they focused on the distribution of stress levels, as P03 stated. "*After checking my stress level was mostly 4 or 5, I noticed that I was truly stressed out when taking a class.*" Some participants realized the importance of the recorded data when their recalls were incorrect. "*I thought I was stressed out when I was with my girlfriend, but it was not. I was mistaken as bad memories often remain longer*" (P15).

5.1.2 Allowing Participants to Identify Stressful Contexts while Considering Confounders. DeepStress presented information on stressful contexts determined through a quasi-experimental approach, assisting participants in rigorously identifying contexts with causality. Most of those stressful contexts were already recognized by the participants, and in that case, DeepStress served to reaffirm the results. However, participants also discovered new stressful contexts while using the system. P07 mentioned, "*It was meaningful to identify stressful contexts that I didn't even think of.*" Consequently, this process may enhance their self-knowledge.

The participants could understand the difference in the quasi-experimental approach when identifying stressful contexts, particularly in terms of confounding factors. From the summary of contexts using colors and icons, they noticed that a context with a high stress score does not necessarily have causality. P23 noted, "*The average does not consider other contexts' effects on stress, so I think it cannot indicate causality*" Through the explanations about procedure and analysis result, they learned how to control confounding factors when identifying the stressful one. "*After all, we have to match the other three contexts before the analysis to minimize their effect on stress*" (P19). For some participants, this concept was familiar, as P17 mentioned. "*This was similar to when planning an experiment. We should change only one condition to see the impact of it.*" After understanding the matching process, some participants reported that the results became more reliable.

Participants also found it helpful for DeepStress to provide causal results directly to them. "*DeepStress automatically analyzed and delivered in which context I was stressed, so I could check the stressful contexts easily and accurately*" (P09). Other participants mentioned the advantages of DeepStress, comparing it to other apps; e.g., allowing them to think deeply about stressful contexts (P10), providing accurate results based on the statistical testing (P01), and consequently making the self-tracking data more meaningful (P03).

5.1.3 Letting Participants Consider the Relationship Between Contexts. DeepStress not only informed the stressful contexts but also enabled participants to explore the relationship between contexts. They could explore how stress levels in one context can be affected by others, by investigating stress level distribution and co-occurring contexts in the context view. P23 mentioned, "*When I checked 'dormitory' in the context view, there was much more 'studying' than other activities. So I thought if I got stressed due to studying, the dormitory would show relatively high stress as well.*" Through this process, the participants reviewed one context and naturally moved to another that was highly related (i.e., frequently co-occurring).

The information explaining which other contexts would increase or decrease the stress level in a given context also assisted participants in understanding the relationships between contexts. "*I found I was more stressed when I studied at home, with colleagues, and in the evening. It was interesting to see that the combinations varied the stress level, and I learned that the relationships among contexts also mattered*" (P08). Based on this information, participants could also set up their plans to manage their stress, as P02 stated. "*From the analysis, I realized places also affect my stress when studying. I should study at the cafe instead of at home.*"

5.2 RQ2: How Do Users Interpret and Conceptualize the Causality Results Provided by DeepStress?

5.2.1 Reconfirming Stressful Contexts That Are Consistent with Prior Self-knowledge. Since DeepStress directly provided causal analysis

results, the participants first questioned whether those results made sense. In most cases, they could confirm the stressful contexts from DeepStress were consistent with their prior self-knowledge and simply reconfirmed the facts through data analysis. "*Most of the results were as expected. I thought the club activity would be the most stressful, and the result told me the same*" (P01). They interpreted these results based on the characteristics of the contexts. "*The fact that 'studying' is a stressful activity for me is natural because I usually have no choice but to study*" (P04). They also found the results understandable considering their lifestyle. "*I hate taking the bus or train since they are too crowded, so I expected that the transportation is a stressful place*" (P19).

These interpretations mostly supported that the analyzed results were correct, and the participants did not express doubt about the results as they found them acceptable. While some results showed gaps between DeepStress's analysis and participants' self-knowledge, not everyone questioned the reasons behind these discrepancies. They accepted these results by assuming the algorithm worked correctly (P07, P11, P19), considering potential issues with data collection (P17), or simply skipping the details (P24).

5.2.2 Hypothesizing about the Reason for Unexpected Causal Analysis Results. However, when the participants faced unexpected results, they questioned the outcomes and formed new hypotheses to explain them. They proposed alternative explanations, such as suggesting that another context might be truly stressful or recognizing limitations in the data contributing to the unexpected conclusion.

For instance, they assumed that there could be other unrecorded reasons in that context. P04, who enjoyed drinking, explained why the pub was determined to be a stressful place as follows: "*The atmosphere of the pub might increase stress though it was not recorded. The pub was somewhat noisy and crowded, and accidents may happen more frequently.*" Some participants supposed that the results might be influenced by other co-occurring contexts. P19 mentioned, "*DeepStress told me that my colleagues got me stressed more, but I suppose this would be because of class-taking activities that happened often with them.*" Similarly, P17 said, "*The dormitory was evaluated as a stressful place, but it seems this result was affected by the fact that I often study in the dormitory.*"

They also hypothesized that those inconsistent results might be due to the recorded contexts not being specific enough. "*The stress really depends on the class. The classes I took in the afternoon or evening were easy, and this might lead to the unexpected causal result that taking a class lowers my stress*" (P04). The number of samples and temporal precedence of the context were suggested as potential reasons for the unexpected results. P06 said, "*I don't think I had a meeting that much, so the low frequency of the activity may make the result biased.*" Additionally, P07 mentioned, "*I usually take a walk outside when I get stressed. Maybe my stress lasted while I was walking and that is why the place 'outside' was determined as a stressful context.*"

5.2.3 Evaluating Alternative Explanations Using Self-knowledge and Self-tracking Data. The participants then evaluated whether the hypotheses made sense by leveraging their prior self-knowledge and the DeepStress data. The former usually happened when they could not find any supportive evidence from DeepStress, such as hypotheses related to unrecorded contexts, details of the context, or temporal precedence. In this situation, the participants simply recalled past events and reflected on whether the hypothesized situation happened in their lives.

Meanwhile, the latter occurred if the participants could explore the data supporting their hypotheses. They re-examined the detailed records about the context and the frequency of other co-occurring contexts. P19 stated, "*Upon revisiting the context view, one of the frequent co-occurring activities with colleagues was taking a class. So, I thought my assumption seemed to be correct.*" Moreover, they utilized the results in the analysis view; another context that further increases stress in a given context. P15, whose 'lab' was a stressful context said, "*I went over which contexts increased my stress when I was in the lab. 'Studying' and 'Afternoon' were there, so I realized that doing homework in the lab got me stressed.*" The participants also checked how many times such contexts were recorded from the context view and evaluated whether their hypotheses were correct. "*There aren't many records in this view that include meeting activities, so the result may be biased. Perhaps, I happened to experience less stress at those times*" (P06).

If the hypotheses seemed reasonable, participants accepted them as a new explanation for interpreting the causal results from DeepStress. Otherwise, they revisited the hypotheses formulation stage and considered other alternative explanations. This process continued until the participants could generate their own interpretation of the results; in other words, it continued until the gap between the participants' self-knowledge and the quasi-experimental analysis became minimized. After that, they moved to another context of interest to see if the causal result was understandable.

5.3 RQ3: What Are the Key Challenges in Identifying Stressful Contexts in Practice?

5.3.1 Evaluating Stress Levels Using Scores. The participants found it challenging to report their stress on a 5-point scale during data collection, expressing difficulty in making clear distinctions between scores. As P12 mentioned, "*I reported stress level 1 when I felt nothing, but the criteria were not clear to me since it was subjective.*" P09 tried to refer to previous records, saying "*I made a relative evaluation when reporting my stress. It was difficult at the beginning as there was no reference, but got better as the data accumulated.*" They thought their results might become unexpected because of their ways of scoring the stress level. "*Maybe the results would have become clearer if I reported the stress scores with big differences*" (P01).

5.3.2 Recording the Context in a Fine-grained Way. In addition, the participants noted that how fine-grained the context was may also affect the analysis results. For example, P09, a part-time tutor of several students, said, "*My stress varies depending on the tutee, but it seemed the analysis results didn't cover it properly.*" Also, P15 said, "*If I could report contexts more in detail, DeepStress would identify the stressful context better.*"

However, they noticed that there would be a trade-off when reporting the context in detail. "*Recording contexts by given categories would make the data collection process simple. However, I was concerned whether the category would represent my context in detail and have an accurate analysis*" (P22). They suggested having more context options for users (P08) or allowing them to customize the set of

context categories based on their own lifestyles (P16). Another option could be leaving an annotation on the record. "*Since I couldn't fully record my situation, I wished to provide more information about myself to DeepStress using a short diary*" (P05).

5.3.3 Handling the Results Derived from Small Data. As shown in RQ2, insufficient samples (i.e., not enough data) posed a challenge for both researchers and participants in identifying stressful contexts. Some contexts were recorded less frequently than others, yet they occasionally proved to be stressful, which could be difficult for participants to understand. P16 stated, "*Social activities were recorded only a few times for me, so I thought the results would be affected by outliers like cases with extremely high stress.*"

Participants understood this situation, acknowledging that not all contexts occur at the same frequency. However, they mentioned the need for careful handling of the analysis results for infrequent contexts to avoid potentially undermining the trust in other results. They proposed several solutions, for instance, showing analysis results only for contexts recorded above a certain threshold (P07) or prioritizing the presentation of stressful contexts with more samples to establish trust in the results (P24).

5.3.4 Differences in Understanding by Context Type. While exploring the stressful contexts, the participants showed differences in understanding the result depending on the context type. They understood the result clearly only when the context type was activity. "*For me, what I did could directly affect my stress rather than when or where it was*" (P17). P18 also mentioned, "*If we remember being stressed because of the assignment, we often recall the activity as a source of stress, not who we did it with.*" Additionally, they thought the place was closely related to the activity, which made it unfamiliar to view a place itself as a stressful context. "*Since the classroom was mainly designed for study and lecture, it would be easier to think that I was stressed because of those activities*" (P04).

5.4 RQ4: How Do Users Utilize the Information about Stressful Contexts in Everyday Life?

5.4.1 Understanding Their Own Stress by Revisiting the Deepstress Data. In the main study, most participants were interested in the analysis results and wanted to go over the details even after the interview. During the diary study, they revisited their previous records to explore their overall stress levels and check if they were experiencing high stress. They tried to figure out patterns in stress changes, for instance, focusing on stress on specific days to see if there was a difference between the days of the week (P08). The participants also tried to compare their current stress levels with records in DeepStress. "*I looked at the past records, imagining what it would be like if I recorded stress today. April was a tough month for me*" (P19). After reviewing their previous records, P06 stated, "I'd like to collect more data to make meaningful comparisons."

5.4.2 Planning Their Every Day Towards Lowering Their Stress Levels. In the diary, participants reported that they utilized insights from DeepStress when planning a day. They revisited the stress information for each context to simulate potential stress levels. P01 stated, "*I have a dinner meeting today, and I try to see how leisure time affects my stress before I join there.*" They utilized the information when planning their schedule. "*I checked the 'studying' context and decided to go to a cafe where my stress gets lower*" (P24). P04 reflected on his day to plan his tomorrow, "*I compared what happened today and the stressful contexts from DeepStress and decided to live with less stress tomorrow by changing contexts.*"

5.4.3 Performing Causality-driven Coping Actions When Stress Management Is Required. The information in DeepStress was frequently used when the participants were in stressful situations. They identified whether their context was determined to be stressful in DeepStress if they got stressed. P24 reported, "*I checked the information about the 'classroom' because I was getting stressed in that place.*" P03 also mentioned, "*I went over the stressful contexts to identify what made me get stressed right now.*" In those situations, they utilized contexts that lowered their stress. "*I tried to move around more, have time with my friends, and do exercises regularly to lower my stress*" (P05). They also leveraged diverse contexts to lower their stress in a given context, as P23 mentioned, "*I went to the cafe instead of other places since DeepStress said my stress gets lower if I study in the cafe.*" However, in some cases, they could not do anything even if they noticed the stressful context. P22 said, "*I was still questioning whether I could control the stressful context. Knowing what is stressful and handling it is different for me.*" Nevertheless, they could better understand themselves, as P21 noted, "*Although I couldn't change where to take a class, at least I could see my stress more objectively.*"

5.4.4 Conducting Re-evaluation and Detailed Analysis of Stressful Contexts. The participants also wondered whether the analyses from DeepStress were correct in practice. In the main study, they generated their own explanations for the results based on their prior self-knowledge or DeepStress data. Conversely, they attempted to validate the results in situ during the diary study, especially when experiencing contexts determined as stressful. In doing so, some participants could specify the true stressful contexts. P15's result showed that 'morning' is a stressful time for him, which he was uncertain about during the main study. However, he reported in the diary, "*I found my bad sleep quality affected my morning time. I will secure my sleep time more and see how my stress changes.*" P02 found that the stress of studying can vary depending on the course taken and began thinking about how to plan his schedule. "*As I deeply looked into my stress while I was studying, I realized that my stress levels varied by my preference for the course. I have to re-plan my study schedule to handle the non-preferred courses.*" As such, the re-evaluation process with DeepStress in the wild helped them gain a more specific understanding of themselves and plan their lives to better manage stress.

6 DISCUSSION

6.1 Leveraging the Quasi-experimental Approach in PI Systems

DeepStress employed a quasi-experimental methodology to help users investigate which contexts were causally related to stress using their self-tracking data. While previous PI systems have also explored relationships between various factors (e.g., correlation analysis and self-experimentation), there are methodological differences from this study's approach. Additionally, this method would be beneficial for exploring causal relationships that (1) are unknown

or less intuitive, (2) can be influenced by multiple, complex external factors, or (3) are challenging to test individually.

Compared to studies conducting correlation analyses, the quasi-experimental approach enables users to assess the effects of specific contexts on stress levels with unbiased data, by balancing the distribution of confounding factors. While our approach shares a fundamental similarity with correlation analysis in eventually comparing stress levels between cases with and without the target context, the key difference lies in whether the data were balanced before the comparison, which potentially leads to different conclusions. In this study, we observed cases where the results of correlation (i.e., without balancing) and causality (i.e., with balancing) were inconsistent. Some showed significant causality but no correlation or presented opposite directions of correlation and causality (details in Supplementary Material). Therefore, we may consider delivering results from the quasi-experimental approach together in PI systems so that users do not miss important relationships from a rigorous analysis. Given that users may utilize PI systems for critical purposes such as health management, these inconsistencies should be carefully considered to avoid potential type 2 errors, particularly when numerous confounding factors complicate the estimation of the unbiased effect of a target context.

The quasi-experimental approach is less rigorous than experimentation but has the advantage of investigating causalities from collected data without conducting controlled experiments for each context. Testing all contexts with n-of-1 trials can be challenging in terms of time and cost, particularly as the number of contexts and confounding factors increases. When experimentation is limited, our approach could serve as an alternative for inferring causal relationships, enabling unbiased evaluations. Furthermore, self-experimentation and quasi-experimental approaches can be employed together to complement each other's limitations. For example, PI system users can first explore causal relationships using quasi-experimental approaches and proceed with follow-up self-experimentation if needed. This combined approach helps narrow down the scope of experimentation, allowing users to analyze causality rigorously and address practical concerns such as costs.

The quasi-experimental approach aims to minimize bias in collected data, similar to randomized trials, before comparing groups. Compared to correlation analysis without balancing data, this approach may require more data to find proper counterfactual samples (i.e., pairs of treated and control samples). The necessary sample size for causal inference varies depending on factors such as data dimension. As the dimension increases, finding sufficiently similar samples for all confounding factors may become challenging, potentially requiring further data collection. The required sample size also depends on the specific quasi-experimental method or the users' determination of statistical power and significance level [21]. Collecting as much data as possible is advantageous for analysis; however, the burden on users associated with data collection must be considered carefully (further discussed in Section 6.3).

In matching similar samples, we employed the intuitive CEM method. Unlike other methods like propensity score matching, CEM did not require additional abstraction for balancing confounding factors. It allowed users to create bins for each confounding factor based on similarity, making it easily understandable that this method enables an apples-to-apples comparison. CEM offers flexibility in adjusting bin size (or the number of bins), letting users set the maximum imbalance bound for each confounding factor based on their knowledge of the data [30]. However, this approach involves a trade-off: smaller bins (i.e., more fine-grained) increase sample similarity but decrease the likelihood of having both treated and control samples within the same bin, reducing matched sample size. For user-defined coarsened bins, there are no specific guidelines such as deciding the appropriate number of bins or their sizes. King et al. [39] proposed the "matching frontier" that reaches the jointly optimal case for the trade-off, but they primarily focused on pruning samples for given coarsened bins. Moreover, as explained in Iacus et al. [31], different numbers of bins can be generated based on the target domain and the importance of each variable.

One suggestion is to test numerous binning cases and choose the one that best balances the confounding factors between groups [28]. This can be supported by an automated procedure [31], where the algorithm progressively reduces the number of bins until reaching the minimum allowable number. As shown in their example using the Lalonde dataset, this procedure improves the matched cases and reveals which confounding factors have the largest impact on the imbalance level. However, it is crucial to avoid bins that drop treated samples, if possible, because the analysis aims to measure the effect of the treated cases on the outcome. In our study, determining the similarity between contexts posed a challenge, even when participants provided reasons for clustering them in the same bins. This challenge may arise partly because contexts are categorical variables, and measuring their closeness can be ambiguous. To improve PI systems, we propose allowing users to input for binning criteria (e.g., "This activity requires concentration a lot.") and score each context accordingly (e.g., "Strongly agree"). By leveraging these responses, the system can generate bins by considering the closeness of contexts (e.g., k-nearest neighbors), with bin sizes iteratively revised to find the optimal coarsened bins.

Determining appropriate rigor in PI systems is an open question. One suggestion is that systems should support different rigor levels by users' questions and analysis scope [35]. We may improve DeepStress to deliver analyzed results with varying rigor levels incrementally, starting with correlations [3] and progressing to unbiased causal analysis using the quasi-experimental approach. If users require more rigorous evidence (e.g., testing serious health issues), systems may guide them on self-experimentations or provide randomized study designs. However, PI systems should inform users if the inconsistency in analyzed results between biased and unbiased data occurs, enabling more careful examinations. Throughout this process, users may learn rigorous data analysis and experiment design from PI systems and answer their questions.

6.2 Delivering Causal Analysis Results to PI Users

In this study, DeepStress directly delivered the results of causal relationships to users. Previous studies [9] reported that allowing users to identify causality from visualized data may lead to wrong conclusions. Therefore, we chose to provide users with the analyzed stressful contexts along with explanations and relevant information, rather than letting them investigate causality on their own.

In the user study, participants understood the concept of causality and explored various relationships between contexts while identifying stressful situations. When encountering a context with high stress that was not identified as a cause, participants closely examined its relationship with other contexts and recognized the potential influences on stress levels. Prior studies [24, 57, 66] highlighted users' reliance on prior knowledge due to the complexity of investigating confounding factors. However, DeepStress played a crucial role in raising awareness of these potential confounders, encouraging a more balanced perspective when evaluating causality.

The overall process of exploring and understanding stressful contexts using DeepStress can be interpreted through the Mamykina et al.'s sensemaking framework [52]. When there was no gap between the inferred causal relationship and users' existing self-knowledge, they entered the 'habitual mode,' accepting the causal results without further analysis. In contrast, in the presence of inconsistency, the 'sensemaking mode' was initiated, where users actively explored the data and generated interpretations for the given results. Users' analytic thinking was activated with unexpected causal relationships (perception of the gap), leading them to explain how such results emerged by reflecting on past records (construction of new inferences). Also, they could maintain or update their mental models of causal relationships and apply the newly acquired knowledge in the real world (explicit actions).

In particular, activities occurring in the inference stage aligned with the personal discovery framework [51]. For instance, users might hypothesize that unexpected causal relationships were influenced by other co-occurring contexts. They evaluated the hypothesis by examining the co-occurring ratio and checked the details in the calendar view to assess their hypothesis.

We found some behaviors fell between the two modes. Users occasionally integrated analyzed results with their existing knowledge or past experiences, instead of revisiting the data or collecting additional information. This heuristic integration process could be interpreted as fluid contextual reasoning [36], where individuals quickly connect existing models without elaborate sensemaking. As this mode does not require effortful thinking for sensemaking, users apply their existing knowledge along with the analysis from DeepStress when managing stress entangled with diverse contexts.

The use of DeepStress can also be interpreted through another sensemaking perspective, namely the data-frame theory of sensemaking [41]. In this framework, samples collected from self-tracking typically served as 'data,' employed to construct new 'frames' illustrating relationships between various factors. However, as DeepStress directly informed the stressful context analyzed by the system, there were some variations in the sensemaking process compared to existing studies on PI systems.

Our work showed that the data-frame relationship could extend to two different levels. Initially, users' self-knowledge of how each context affects stress can be seen as a frame, while the causal result for a certain context inferred from DeepStress can be regarded as data. Depending on the gap between the existing frame and the data in determining stressful contexts, users may either elaborate on the frame or question and adjust it using the data. When the gap is observed, another round of sensemaking begins with a new data-frame relationship, wherein the DeepStress's causal result acts as a frame and the raw self-tracking samples serve as data. During this process, users may re-analyze the data, potentially generating an alternative frame or understanding and accepting the given frame. If the alternative makes sense to them, the frame is integrated into the prior self-knowledge, thereby extending the users' own frame. Therefore, the system-driven analysis results in the PI system may serve as both data and frame, effectively connecting self-tracking data and self-knowledge in the sensemaking process. Moreover, acting as an anchor in the analysis, they could assist users in quickly understanding the overall causalities while exploring unfamiliar cases where new knowledge could be discovered.

6.3 Actions after Reflection on Causality

As with prior healthcare studies highlighting the significance of contextual factors [48, 64, 65], this study empowered participants to understand how context influences stress and implement data-driven strategies. Participants utilized information on the presence of causality and its intensity in each context for effective stress management. This information could be integrated into existing just-in-time interventions, guiding users on stress management in stressful contexts. By leveraging the causal relationships, existing PI systems can be extended to predict stress levels in specific contexts and proactively inform users to help plan their day.

The information from DeepStress could be used to (1) check whether a certain context was stressful and (2) identify other contexts that influenced stress within a given situation. This novel feature, employing a quasi-experimental approach in two stages, offered users flexibility in developing coping strategies for managing their stress. Previous PI systems typically focused on the relationship between two factors (e.g., stress levels and a specific context), leaving users with the only option of 'avoiding the stressful context.' However, DeepStress additionally analyzed the causalities within a given context and identified which other contexts have causal relationships with stress. This can be practical for developing an alternative plan for stress management in situations where controlling the problematic context directly is challenging, such as attending a class in the dormitory instead of simply skipping it. Mamykina et al. [51] also demonstrated that providing flexible options for addressing (health) issues would make users more actively engaged in self-management. Therefore, the two-stage causal analysis in DeepStress would empower PI system users with more alternatives, potentially motivating users to explore causal relationships from various angles.

In our follow-up diary study, DeepStress motivated participants to revisit the reflection and data collection steps in the stage-based model. They iterated reflecting on whether each context's causal relationship was accurate while experiencing that context in their daily lives, which can be considered as 'reflection-in-action' [58]. This differed from the main study's 'reflection-on-action,' where they relied on analysis results or memory. From the data-sensemaking perspective, users utilized this active reflection to find cases supporting hypotheses generated during the 'reflection-on-action' process. This process may help users construct a rationale for analysis results and expand their self-knowledge. In addition, the participants recognized the need for further data collection, questioning whether the results could be reproduced with the new data. Therefore, supporting continuous data collection is necessary to enable

users to examine stressful contexts over any period of interest and to investigate how the relationships change over time.

Given the recurring process after reflecting on causal relationships, our system could be more beneficial when the value of the quasi-experimental approach exceeds the users' burden. Previous studies [66, 68] also emphasized balancing the benefits and burdens of self-tracking to encourage users to continue data collection while gaining new insights.

First, DeepStress could be used with passively collected data, reducing the data collection burden. For instance, the system could assist users with diabetes in identifying behavioral causes of blood glucose spikes by utilizing sensor data, including accelerometers (physical activity), GPS (place), ambient light (sleep), and heart rate (stress), along with photos of food intake [14, 56]. By combining them with continuously monitored glucose levels, PI systems could estimate causal relationships and alert users to potential risk factors without requesting users' manual data collection. Moreover, it would be valuable for laborious cases that purely rely on self-experimentations, such as identifying triggers of gastrointestinal discomfort. Our approach reduces the complicated process of setting experiments and testing each case individually, as users can derive causality from data they naturally collect in their daily lives.

Alternatively, we may reduce user burden by narrowing down the scope of data collection. Given that users formulate hypotheses and assess them using self-tracking data [59], PI systems could offer specific guidelines for data collection based on the hypotheses rather than simply requesting extensive data collection. The causal analysis process can be improved by referring to suggestions in [16], letting users explore testable variables, conduct a preliminary analysis, and perform a main evaluation focusing on specific hypotheses. Furthermore, PI systems may empower users to start with a narrow, specific question, such as determining whether context X increases their stress, as opposed to a more general inquiry about which contexts elevate stress levels. This approach would enable users to collect sufficient samples for causal inference within a shorter period because the treatment condition is predefined. This facilitates the acquisition of more treated samples, which are often sparser than controlled ones. Considering the iterative nature of the reflection-and-action cycle in PI systems [23], users can identify relevant factors (to the initial context) during their exploration and subsequently conduct the same analysis on those factors. These approaches would encourage continued self-tracking for users to explore and test various aspects of daily life.

6.4 Limitations and Future Work

One crucial challenge is determining the level of detail for user-recorded context. While we offered preset contexts to ease data collection, it had limitations in capturing detailed context information. In addition, there was a trade-off between facilitating the analysis (i.e., requiring more samples) and providing results for very specific contexts (i.e., having fewer samples). To address these issues, PI systems could allow users to record more specific situations like a diary. These records could serve as additional contextual information for causal inference. Furthermore, the system could provide information on the feasibility of causal inference based on the amount of collected data. This would encourage users to collect more samples for the context or create another coarsened bin including that context to enable the causal analysis.

To explore the feasibility of a PI system supporting causal inference, we provided analysis results only after completing a certain data collection period. In future work, systems could gradually deliver analyzable insights based on the collected data, motivating users for data collection and supporting reflection-in-action.

For the generalizability of the study, considering alternative quasi-experimental approaches like propensity score matching is crucial for analytical robustness. Further investigation should involve individuals beyond university students, especially those with limited data analysis knowledge, to assess how they utilize the system. Also, experiments comparing DeepStress with other PI systems help reveal the relative benefits of investigating causality.

7 CONCLUSION

We introduced 'DeepStress,' a PI system that supports users in investigating stressful contexts determined by a quasi-experimental approach. Our goal was to identify user experiences in exploring causal relationships through DeepStress. We showed that DeepStress helped users conduct data-driven self-reflection where they could pinpoint both stressful contexts and relationships between contexts. We also found that the sensemaking process occurred while interpreting the causal analysis results, which may be continued to the application of findings in daily life to manage stress.

Our study showed the feasibility of a PI system in exploring causal relationships using observational, self-tracking data. We expect our system to be extended to address diverse health challenges by leveraging causalities and to be utilized complementary with other methods such as self-experimentations in delivering meaningful causal insights.

ACKNOWLEDGMENTS

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. 2021M3A9E4080780) and by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2022-0-00064).

REFERENCES

[1] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies* 4, 3 (2009), 114–123. https://doi.org/10.5555/2835587.2835589

[2] Eric PS Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing Reflection: On the Use of Reflection in Interactive System Design. In *Proceedings of the 2014 Conference on Designing Interactive Systems*. Association for Computing Machinery, New York, NY, USA, 93–102. https://doi.org/10.1145/2598510.2598598

[3] Frank Bentley, Konrad Tollmar, Peter Stephenson, Laura Levy, Brian Jones, Scott Robertson, Ed Price, Richard Catrambone, and Jeff Wilson. 2013. Health Mashups: Presenting Statistical Patterns between Wellbeing Data and Context in Natural Language to Promote Behavior Change. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20, 5 (2013), 1–27. https://doi.org/10.1145/2503823

[4] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706QP063OA

[5] John Brooke. 1996. SUS: A 'Quick and Dirty' Usability. *Usability Evaluation in Industry* 189, 3 (1996), 189–194.

[6] Clara Caldeira, Yu Chen, Lesley Chan, Vivian Pham, Yunan Chen, and Kai Zheng. 2017. Mobile Apps for Mood Tracking: An Analysis of Features and User Reviews.

In *AMIA Annual Symposium Proceedings*, Vol. 2017. American Medical Informatics Association, Washington DC, USA, 495.

[7] 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*. Association for Computing Machinery, New York, NY, USA, 1–23. https://doi.org/10.1145/3491102.3501991

[8] Eun Kyoung Choe, Bongshin Lee, Matthew Kay, Wanda Pratt, and Julie A Kientz. 2015. SleepTight: Low-Burden, Self-Monitoring Technology for Capturing and Reflecting on Sleep Behaviors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery, New York, NY, USA, 121–132. https://doi.org/10.1145/2750858.2804266

[9] Eun Kyoung Choe, Bongshin Lee, and M.C. Schraefel. 2015. Characterizing Visualization Insights from Quantified Selfers' Personal Data Presentations. *IEEE Computer Graphics and Applications* 35, 4 (2015), 28–37. https://doi.org/10.1109/MCG.2015.51

[10] 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*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/3154862.3154881

[11] Eun Kyoung Choe, Nicole B Lee, Bongshin Lee, Wanda Pratt, and Julie A Kientz. 2014. Understanding Quantified-Selfers' Practices in Collecting and Exploring Personal Data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1143–1152. https://doi.org/10.1145/2556288.2557372

[12] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1994. Perceived Stress Scale. *Measuring Stress: A Guide for Health and Social Scientists* 10, 2 (1994), 1–2.

[13] Thomas D Cook, Donald Thomas Campbell, and William Shadish. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Vol. 1195. Houghton Mifflin, Boston, MA, USA.

[14] 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*. Association for Computing Machinery, New York, NY, USA, 3207–3216. https://doi.org/10.1145/2702123.2702154

[15] Jesse Dallery, Rachel N Cassidy, and Bethany R Raiff. 2013. Single-Case Experimental Designs to Evaluate Novel Technology-Based Health Interventions. *Journal of Medical Internet Research* 15, 2 (2013), e22. https://doi.org/10.2196/jmir.2227

[16] Nediyana Daskalova, Karthik Desingh, Alexandra Papoutsaki, Diane Schulze, Han Sha, and Jeff Huang. 2017. Lessons Learned From Two Cohorts of Personal Informatics Self-Experiments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–22. https://doi.org/10.1145/3130911

[17] Nediyana Daskalova, Eindra Kyi, Kevin Ouyang, Arthur Borem, Sally Chen, Sung Hyun Park, Nicole Nugent, and Jeff Huang. 2021. Self-E: Smartphone-Supported Guidance for Customizable Self-Experimentation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445100

[18] 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. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 347–358. https://doi.org/10.1145/2984511.2984534

[19] Nediyana Daskalova, Jina Yoon, Yibing Wang, Cintia Araujo, Guillermo Beltran Jr, Nicole Nugent, John McGeary, Joseph Jay Williams, and Jeff Huang. 2020. SleepBandits: Guided Flexible Self-Experiments for Sleep. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376584

[20] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. 2000. The Mahalanobis Distance. *Chemometrics and Intelligent Laboratory Systems* 50, 1 (2000), 1–18. https://doi.org/10.1016/S0169-7439(99)00047-7

[21] Nianbo Dong and Rebecca Maynard. 2013. PowerUp!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies. *Journal of Research on Educational Effectiveness* 6, 1 (2013), 24–67. https://doi.org/10.1080/19345747.2012.673143

[22] Daniel A. Epstein, Clara Caldeira, Mayara Costa Figueiredo, Xi Lu, Lucas M. Silva, Lucretia Williams, Jong Ho Lee, Qingyang Li, Simran Ahuja, Qiuer Chen, Payam Dowlatyari, Craig Hilby, Sazeda Sultana, Elizabeth V. Eikey, and Yunan Chen. 2020. Mapping and Taking Stock of the Personal Informatics Literature. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–38. https://doi.org/10.1145/3432231

[23] 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*. Association for Computing Machinery, New York, NY, USA, 731–742. https://doi.org/10.1145/2750858.2804250

[24] Sarah Faisal, Ann Blandford, and Henry WW Potts. 2013. Making Sense of Personal Health Information: Challenges for Information Visualization. *Health Informatics Journal* 19, 3 (2013), 198–217. https://doi.org/10.1177/1460458212465213

[25] Rowanne Fleck and Geraldine Fitzpatrick. 2010. Reflecting on Reflection: Framing a Design Landscape. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*. Association for Computing Machinery, New York, NY, USA, 216–223. https://doi.org/10.1145/1952222.1952269

[26] Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. 2022. Review on Psychological Stress Detection Using Biosignals. *IEEE Transactions on Affective Computing* 13, 1 (2022), 440–460. https://doi.org/10.1109/TAFFC.2019.2927337

[27] Shalom Greene, Himanshu Thapliyal, and Allison Caban-Holt. 2016. A Survey of Affective Computing for Stress Detection: Evaluating Technologies in Stress Detection for Better Health. *IEEE Consumer Electronics Magazine* 5, 4 (2016), 44–56. https://doi.org/10.1109/MCE.2016.2590178

[28] Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. 2007. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* 15, 3 (2007), 199–236. https://doi.org/10.1093/pan/mpl013

[29] Esther Howe, Jina Suh, Mehrab Bin Morshed, Daniel McDuff, Kael Rowan, Javier Hernandez, Marah Ihab Abdin, Gonzalo Ramos, Tracy Tran, and Mary P Czerwinski. 2022. Design of Digital Workplace Stress-Reduction Intervention Systems: Effects of Intervention Type and Timing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3491102.3502027

[30] Stefano M Iacus, Gary King, and Giuseppe Porro. 2011. Multivariate Matching Methods That Are Monotonic Imbalance Bounding. *J. Amer. Statist. Assoc.* 106, 493 (2011), 345–361. https://doi.org/10.1198/jasa.2011.tm09599

[31] Stefano M Iacus, Gary King, and Giuseppe Porro. 2012. Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis* 20, 1 (2012), 1–24. https://doi.org/10.1093/pan/mpr013

[32] Ellen Isaacs, Artie Konrad, Alan Walendowski, Thomas Lennig, Victoria Hollis, and Steve Whittaker. 2013. Echoes From the Past: How Technology Mediated Reflection Improves Well-Being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1071–1080. https://doi.org/10.1145/2470654.2466137

[33] Gyuwon Jung, Sangjun Park, Eun-Yeol Ma, Heeyoung Kim, and Uichin Lee. 2024. A Tutorial on Matching-based Causal Analysis of Human Behaviors Using Smartphone Sensor Data. *Comput. Surveys* (2024), 1–33. https://doi.org/10.1145/3648356 Just Accepted.

[34] Soowon Kang, Cheul Young Park, Auk Kim, Narae Cha, and Uichin Lee. 2022. Understanding Emotion Changes in Mobile Experience Sampling. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3491102.3501944

[35] 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*. Association for Computing Machinery, New York, NY, USA, 6850–6863. https://doi.org/10.1145/3025453.3025480

[36] Dmitri S Katz, Blaine A Price, Simon Holland, and Nicholas Sheep Dalton. 2018. Designing for Diabetes Decision Support Systems With Fluid Contextual Reasoning. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174199

[37] Inyeop Kim, Hwarang Goh, Nematjon Narziev, Youngtae Noh, and Uichin Lee. 2020. Understanding User Contexts and Coping Strategies for Context-aware Phone Distraction Management System Design. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–33. https://doi.org/10.1145/3432213

[38] Taewan Kim, Haesoo Kim, Ha Yeon Lee, Hwarang Goh, Shakhboz Abdigapporov, Mingon Jeong, Hyunsung Cho, Kyungsik Han, Youngtae Noh, Sung-Ju Lee, and Hwajung Hong. 2022. Prediction for Retrospection: Integrating Algorithmic Stress Prediction Into Personal Informatics Systems for College Students' Mental Health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3491102.3517701

[39] Gary King, Christopher Lucas, and Richard A Nielsen. 2017. The Balance-Sample Size Frontier in Matching Methods for Causal Inference. *American Journal of Political Science* 61, 2 (2017), 473–489. https://doi.org/10.1111/ajps.12272

[40] Gary King and Richard Nielsen. 2019. Why Propensity Scores Should Not Be Used for Matching. *Political Analysis* 27, 4 (2019), 435–454. https://doi.org/10.1017/pan.2019.11

[41] Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. 2007. *A Data-Frame Theory of Sensemaking*. Psychology Press, New York, NY, USA, 113–155.

[42] Rafal Kocielnik and Natalia Sidorova. 2015. Personalized Stress Management: Enabling Stress Monitoring With LifelogExplorer. *KI-Künstliche Intelligenz* 29 (2015), 115–122. https://doi.org/10.1007/s13218-015-0348-1

[43] Artie Konrad, Simon Tucker, John Crane, and Steve Whittaker. 2016. Technology and Reflection: Mood and Memory Mechanisms for Well-Being. *Psychology of Well-Being* 6 (2016), 1–24. https://doi.org/10.1186/s13612-016-0045-3

[44] Reed Larson and Mihaly Csikszentmihalyi. 2014. *The Experience Sampling Method*. Springer Netherlands, Dordrecht, 21–34. https://doi.org/10.1007/978-94-017-9088-8_2

[45] Uichin Lee, Gyuwon Jung, Eun-Yeol Ma, Jin San Kim, Heepyung Kim, Jumabek Alikhanov, Youngtae Noh, and Heeyoung Kim. 2023. Toward Data-Driven Digital Therapeutics Analytics: Literature Review and Research Directions. *IEEE/CAA Journal of Automatica Sinica* 10, 1 (2023), 42–66. https://doi.org/10.1109/JAS.2023.123015

[46] 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*. Association for Computing Machinery, New York, NY, USA, 557–566. https://doi.org/10.1145/1753326.1753409

[47] Ian Li, Anind K Dey, and Jodi Forlizzi. 2011. Understanding My Data, Myself: Supporting Self-Reflection With Ubicomp Technologies. In *Proceedings of the 13th International Conference on Ubiquitous Computing*. Association for Computing Machinery, New York, NY, USA, 405–414. https://doi.org/10.1145/2030112.2030166

[48] Ian Li, Anind K Dey, and Jodi Forlizzi. 2012. Using Context to Reveal Factors That Affect Physical Activity. *ACM Transactions on Computer-Human Interaction (TOCHI)* 19, 1 (2012), 1–21. https://doi.org/10.1145/2147783.2147790

[49] Zilu Liang, Bernd Ploderer, Wanyu Liu, Yukiko Nagata, James Bailey, Lars Kulik, and Yuxuan Li. 2016. SleepExplorer: A Visualization Tool to Make Sense of Correlations Between Personal Sleep Data and Contextual Factors. *Personal and Ubiquitous Computing* 20 (2016), 985–1000. https://doi.org/10.1007/s00779-016-0960-6

[50] Elizabeth O Lillie, Bradley Patay, Joel Diamant, Brian Issell, Eric J Topol, and Nicholas J Schork. 2011. The N-of-1 Clinical Trial: The Ultimate Strategy for Individualizing Medicine? *Personalized Medicine* 8, 2 (2011), 161–173. https://doi.org/10.2217/pme.11.7

[51] Lena Mamykina, Elizabeth M Heitkemper, Arlene M Smaldone, Rita Kukafka, Heather J Cole-Lewis, Patricia G Davidson, Elizabeth D Mynatt, Andrea Cassells, Jonathan N Tobin, and George Hripcsak. 2017. Personal Discovery in Diabetes Self-Management: Discovering Cause and Effect Using Self-Monitoring Data. *Journal of Biomedical Informatics* 76 (2017), 1–8. https://doi.org/10.1016/j.jbi.2017.09.013

[52] 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. https://doi.org/10.1016/j.jbi.2015.06.006

[53] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. AffectAura: An Intelligent System for Emotional Memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 849–858. https://doi.org/10.1145/2207676.2208525

[54] Abhinav Mehrotra, Fani Tsapeli, Robert Hendley, and Mirco Musolesi. 2017. MyTraces: Investigating Correlation and Causation Between Users' Emotional States and Mobile Phone Interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–21. https://doi.org/10.1145/3130948

[55] Varun Mishra, Tian Hao, Si Sun, Kimberly N Walter, Marion J Ball, Ching-Hua Chen, and Xinxin Zhu. 2018. Investigating the Role of Context in Perceived Stress Detection in the Wild. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. Association for Computing Machinery, New York, NY, USA, 1708–1716. https://doi.org/10.1145/3267305.3267537

[56] David C Mohr, Mi Zhang, and Stephen M Schueller. 2017. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual Review of Clinical Psychology* 13 (2017), 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

[57] Jimmy Moore, Pascal Goffin, Jason Wiese, and Miriah Meyer. 2021. Exploring the Personal Informatics Analysis Gap: "There's a Lot of Bacon". *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 96–106. https://doi.org/10.1109/TVCG.2021.3114798

[58] Hugh Munby. 1989. Reflection-in-Action and Reflection-on-Action. *Current Issues in Education* 9, 1 (1989), 31–42. https://doi.org/10.1353/eac.1989.a592219

[59] Sean A Munson, Jessica Schroeder, Ravi Karkar, Julie A Kientz, Chia-Fang Chung, and James Fogarty. 2020. The Importance of Starting With Goals in N-of-1 Studies. *Frontiers in Digital Health* 2 (2020), 3. https://doi.org/10.3389/fdgth.2020.00003

[60] World Health Organization. 2016. Monitoring and Evaluating Digital Health Interventions: A Practical Guide to Conducting Research and Assessment.

[61] Rosalind W Picard and Jonathan Klein. 2002. Computers That Recognise and Respond to User Emotion: Theoretical and Practical Implications. *Interacting with Computers* 14, 2 (2002), 141–169. https://doi.org/10.1016/S0953-5438(01)00055-8

[62] Peter Pirolli and Stuart Card. 2005. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. In *Proceedings of International Conference on Intelligence Analysis (IA '05)*, Vol. 5. International Conference on Intelligence Analysis, McLean, VA, USA, 2–4.

[63] Eugenia Politou, Efthimios Alepis, and Constantinos Patsakis. 2017. A Survey on Mobile Affective Computing. *Computer Science Review* 25 (2017), 79–100. https://doi.org/10.1016/j.cosrev.2017.07.002

[64] Shriti Raj, Joyce M Lee, Ashley Garrity, and Mark W Newman. 2019. Clinical Data in Context: Towards Sensemaking Tools for Interpreting Personal Health Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–20. https://doi.org/10.1145/3314409

[65] Shriti Raj, Kelsey Toporski, Ashley Garrity, Joyce M Lee, and Mark W Newman. 2019. "My Blood Sugar Is Higher on the Weekends" Finding a Role for Context and Context-Awareness in the Design of Health Self-Management Technology. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300349

[66] Amon Rapp and Federica Cena. 2016. Personal Informatics for Everyday Life: How Users Without Prior Self-Tracking Experience Engage With Personal Data. *International Journal of Human-Computer Studies* 94 (2016), 1–17. https://doi.org/10.1016/j.ijhcs.2016.05.006

[67] Amon Rapp and Maurizio Tirassa. 2017. Know Thyself: A Theory of the Self for Personal Informatics. *Human–Computer Interaction* 32, 5-6 (2017), 335–380. https://doi.org/10.1080/07370024.2017.1285704

[68] Sara Riggare, Therese Scott Duncan, Helena Hvitfeldt, and Maria Hägglund. 2019. "You Have to Know Why You're Doing This": A Mixed Methods Study of the Benefits and Burdens of Self-Tracking in Parkinson's Disease. *BMC Medical Informatics and Decision Making* 19 (2019), 1–16. https://doi.org/10.1186/s12911-019-0896-7

[69] John Rooksby, Mattias Rost, Alistair Morrison, and Matthew Chalmers. 2014. Personal Tracking as Lived Informatics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1163–1172. https://doi.org/10.1145/2556288.2557039

[70] Paul R Rosenbaum. 2005. *Observational Study*. John Wiley & Sons, Ltd, Chichester, UK, 1451–1462. https://doi.org/10.1002/0470013192.bsa454

[71] Paul R Rosenbaum and Donald B Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70, 1 (1983), 41–55. https://doi.org/10.1093/biomet/70.1.41

[72] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The Cost Structure of Sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 269–276. https://doi.org/10.1145/169059.169209

[73] Akane Sano, Paul Johns, and Mary Czerwinski. 2017. Designing Opportune Stress Intervention Delivery Timing Using Multi-Modal Data. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE Computer Society, Los Alamitos, CA, USA, 346–353. https://doi.org/10.1109/ACII.2017.8273623

[74] Robert William Sanson-Fisher, Billie Bonevski, Lawrence W Green, and Cate D'Este. 2007. Limitations of the Randomized Controlled Trial in Evaluating Population-Based Health Interventions. *American Journal of Preventive Medicine* 33, 2 (2007), 155–161. https://doi.org/10.1016/j.amepre.2007.04.007

[75] Moushumi Sharmin, Andrew Raij, David Epstien, Inbal Nahum-Shani, J Gayle Beck, Sudip Vhaduri, Kenzie Preston, and Santosh Kumar. 2015. Visualization of Time-Series Sensor Data to Inform the Design of Just-in-Time Adaptive Stress Interventions. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery, New York, NY, USA, 505–516. https://doi.org/10.1145/2750858.2807537

[76] Bonnie Sibbald and Martin Roland. 1998. Understanding Controlled Trials. Why Are Randomised Controlled Trials Important? *BMJ: British Medical Journal* 316, 7126 (1998), 201. https://doi.org/10.1136/bmj.316.7126.201

[77] Elena Smets, Emmanuel Rios Velazquez, Giuseppina Schiavone, Imen Chakroun, Ellie D'Hondt, Walter De Raedt, Jan Cornelis, Olivier Janssens, Sofie Van Hoecke, Stephan Claes, Ilse Van Diest, and Chris Van Hoof. 2018. Large-Scale Wearable Data Reveal Digital Phenotypes for Daily-Life Stress Detection. *npj Digital Medicine* 1, 1 (2018), 67. https://doi.org/10.1038/s41746-018-0074-9

[78] Aaron Springer, Victoria Hollis, and Steve Whittaker. 2018. Mood Modeling: Accuracy Depends on Active Logging and Reflection. *Personal and Ubiquitous Computing* 22 (2018), 723–737. https://doi.org/10.1007/s00779-018-1123-8

[79] Elizabeth A Stuart. 2010. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25, 1 (2010), 1. https://doi.org/10.1214/09-STS313

[80] Elizabeth A Stuart, Gary King, Kosuke Imai, and Daniel Ho. 2011. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* 42, 8 (2011), 1–28. https://doi.org/10.18637/jss.v042.i08

[81] Melanie Swan. 2013. The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. *Big Data* 1, 2 (2013), 85–99. https://doi.org/10.1089/big.2012.0002

[82] Fani Tsapeli and Mirco Musolesi. 2015. Investigating Causality in Human Behavior From Smartphone Sensor Data: A Quasi-Experimental Approach. *EPJ Data Science* 4, 1 (2015), 24. https://doi.org/10.1140/epjds/s13688-015-0061-1

[83] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, and Wenqiang Zhang. 2022. A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances. *Information Fusion* 83 (2022), 19–52. https://doi.org/10.1016/j.inffus.2022.03.009

[84] Mengru Xue, Rong-Hao Liang, Bin Yu, Mathias Funk, Jun Hu, and Loe Feijs. 2019. AffectiveWall: Designing Collective Stress-Related Physiological Data Visualization for Reflection. *IEEE Access* 7 (2019), 131289–131303. https://doi.org/10.1109/ACCESS.2019.2940866