UNIVERSITY OF CALIFORNIA SAN DIEGO

SAN DIEGO STATE UNIVERSITY

Optimizing Just-In-Time Adaptive Interventions: Incorporating Idiographic, Dynamic Predictions to Support Physical Activity

A dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Public Health (Health Behavior)

by

Junghwan Park

Committee in charge:

University of California San Diego

Professor Eric Hekler, Chair
Professor Job Gideon Godino
Professor Gregory J Norman

San Diego State University

Professor Jonathan Helm
Professor Melody Karen Schiaffino

2024

The Dissertation of Junghwan Park is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

_____

_____

_____

_____

_____

Chair

University of California San Diego

San Diego State University

2024

DEDICATION

This dissertation is dedicated to my beloved wife, Nayeon Park, whose unwavering support and dedication have been my anchor throughout this challenging journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

API	Application Programming Interface
ASHA	Async Successive Halving
CONSORT	Consolidated Standards of Reporting Trials
EMA	Ecological Momentary Assessment
GPS	Global Positioning System
INUS conditions	Insufficient but Necessary parts of Unnecessary but Sufficient conditions
JDP	Joint Doctoral Program
JIT	Just-In-Time
JIT states	Just-In-Time states
JITAI	Just-In-Time Adaptive Intervention
MAP	Maximum *a posteriori* Point
MCC	Mathew Correlation Coefficient
mHealth	Mobile Health
MLP	MultiLayered Perceptron
N	Need
N+O	Need and Opportunity
N+O+R	Need, Opportunity, and Receptivity
N+R	Need and Receptivity
O	Opportunity
PA	Physical Activity
PRBS	PseudoRandom Binary Sequence

R	Receptivity
SCT	Social Cognitive Theory
SDSU	San Diego State University
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis
UCSD	University of California San Diego

ACKNOWLEDGEMENTS

I owe my deepest gratitude to Professor Eric Hekler, who has been the guiding light throughout the arduous journey of my doctoral studies. His steadfast academic support and insightful guidance have been invaluable to my development as a scholar. Professor Hekler's dedication to my research and his persistent encouragement have enabled me to pursue my research with rigor and passion.

I am also profoundly thankful to the members of my dissertation committee: Drs. Job Godino, Melody K Schiaffino, Gregory Norman, and Jonathan Helm. Their expert advice and invaluable support throughout this process have greatly contributed to the substance and quality of my work.

This research would not have been possible without the support and collaboration of several key individuals and institutions. I am particularly thankful to Professor Predrag Klasnja at the University of Michigan and Professor Daniel E. Rivera at Arizona State University, whose support and guidance were critical to the success of this project. My academic partner, Mohamed El Mistiri, PhD, has been a constant source of inspiration and shared excitement in every new discovery. I am also grateful for the collegial and supportive interactions with Meelim Kim, PhD, which have been immensely comforting during challenging times. My thanks also extend to Michael Higgins, Shadia J. Assi, Sarasij Banerjee, Olga Perski, Steven De La Torre, and Manas Bedmutha. Furthermore, I appreciate the foundational support from the University of California San Diego, National Library of Medicine, Fitbit Inc., Fitabase Inc., and most importantly, the invaluable participants of this study.

On a personal note, my heartfelt thanks go to my beloved wife, Nayeon Park, and our daughter/co-researcher, Seohee, who have been my pillars of strength and solace throughout this

demanding period. Their love and unwavering support have been my greatest motivation. Additionally, I am forever grateful for the continuous encouragement from my parents, parent-in-laws, my sister Jihyang Park, and all the other supportive families and relatives back in Korea.

Lastly, I extend my gratitude to the Korean government, which has funded my doctoral studies as a national scholar, and to the Ministry of Health and Welfare for their faith in my professional growth. I am deeply appreciative of the support from the people of Republic of Korea, whose belief in my potential has been a great honor to uphold.

Chapter 3, in full, is a reprint of the material as it appears in Park, Junghwan, Meelim Kim, Mohamed El Mistiri, Rachael Kha, Sarasij Banerjee, Lisa Gotzian, Guillaume Chevance, Daniel E. Rivera, Predrag Klasnja, and Eric Hekler. 2023. "Advancing Understanding of Just-in-Time States for Supporting Physical Activity (Project JustWalk JITAI): Protocol for a System ID Study of Just-in-Time Adaptive Interventions." *JMIR Research Protocols* 12 (September): e52161. The dissertation author was the primary investigator and author of this paper. The study described in this chapter was funded by the National Library of Medicine (R01LM013107).

Appendix 1, in full, is a reprint of the material as it appears in Park, Junghwan, Gregory J. Norman, Predrag Klasnja, Daniel E. Rivera, and Eric Hekler. 2023. "Development and Validation of Multivariable Prediction Algorithms to Estimate Future Walking Behavior in Adults: Retrospective Cohort Study." *JMIR mHealth and uHealth* 11 (January): e44296. The National Library of Medicine (R01LM013107) funded JP's stipend.

Chapter 1, 2, 4, 5, 6, and 7 of this thesis, in part, are currently being prepared for submission for publication of the material. Park, Junghwan; Kim, Meelim; El Mistiri,

Mohamed; Kha, Rachael; Banerjee, Sarasij; Gotzian, Lisa; Chevance, Guillaume; Rivera, Daniel E.; Klasnja, Predrag; Hekler, Eric. The dissertation author was the primary researcher and author of this material.

VITA

2011. Bachelor of Science in Bioinformatics, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

2011 – present. Ministry of Health and Welfare, Sejong, Republic of Korea

2024. Doctor of Philosophy in Public Health (Health Behavior), University of California San Diego, San Diego State University, San Diego, California, USA

PUBLICATIONS

Park, Junghwan, Minho Lee, and Bhak Jong. 2005. "HExDB: Human EXon DataBase for Alternative Splicing Pattern Analysis." *Genomics & Informatics* 3 (3): 80–85.

Park, Junghwan, Gregory J. Norman, Predrag Klasnja, Daniel E. Rivera, and Eric Hekler. 2023. "Development and Validation of Multivariable Prediction Algorithms to Estimate Future Walking Behavior in Adults: Retrospective Cohort Study." *JMIR mHealth and uHealth* 11 (January): e44296.

Park, Junghwan, Meelim Kim, Mohamed El Mistiri, Rachael Kha, Sarasij Banerjee, Lisa Gotzian, Guillaume Chevance, Daniel E. Rivera, Predrag Klasnja, and Eric Hekler. 2023. "Advancing Understanding of Just-in-Time States for Supporting Physical Activity (Project JustWalk JITAI): Protocol for a System ID Study of Just-in-Time Adaptive Interventions." *JMIR Research Protocols* 12 (September): e52161.

ABSTRACT OF THE DISSERTATION

Optimizing Just-In-Time Adaptive Interventions: Incorporating Idiographic, Dynamic Predictions to Support Physical Activity

by

Junghwan Park

Public Health (Health Behavior)

University of California San Diego, 2024
San Diego State University, 2024

Professor Eric Hekler, Chair

Background.

Physical Activity (PA) plays a crucial role as a protective factor against many diseases. Even though it is widely known that an appropriate level of PA contributes to overall physical and mental health, a significant portion of the population fails to achieve the recommended PA levels.

Methods.

We conducted an optimization trial, called a system identification experiment, meant to guide the development of a future digital health just-in-time adaptive intervention to increase PA. The system identification experiment was conducted to test the assumption that we could identify "just-in-time" states whereby individuals would reliably increase steps taken when support is offered (relative to no support offered in the same state). We specifically predicted that these patterns would not be easily detectable using classic population-based (also known as nomothetic) statistical approaches and, instead, would require idiographic Bayesian modeling. Two articles, one on the operationalization of just-in-time states and the second about the trial protocol have been published. A series of analyses, including Mixed Effects Models, Bayesian Regression, Machine Learning Models, and exploratory analysis, were conducted to rigorously and experimentally study the nomothetic, idiographic and dynamic nature of people's response to PA intervention within each person and across people.

Results

We found that it is feasible to identify individualized states whereby people would reliably increase steps/3 hours post support (compared to no support given in the same state) for 91% (40/44) of participants with sufficient data (83% using an intent to treat approach, 40/48).

Conclusion

This study demonstrates the capacity of our approach for identifying individualized states whereby each person could benefit from receiving support for most of our target sample. These results provide strong justification for the next step in this systematic line of research whereby we would integrate this system identification optimization trial into a control optimization trial (COT) that enables these insights to be used in real-time and at scale to support increases in physical activity.

OVERVIEW

The purpose of this study was to empirically understand how people respond to suggestions to walk depending on the dynamic states of individuals, to determine if there are different patterns for each individual that are also predictable and, thus, could be used within a future individualized intervention that uses walking suggestions as one intervention component. To accomplish this, we designed a unique research study and corresponding technology platform that allowed us to study the dual possibility both that context matters, things change, and people are different AND that the influence of context, time, and individual differences occurs in a predictable fashion within individuals across time. If our hypothesis is true, it creates a pathway for using that predictive knowledge to develop highly personalized interventions. Doing this work required a significant amount of technical work, including creating technical systems and algorithms, providing this technology to participants and supporting them in using it, and the development and implementation of a robust data analytic approach that could both honor that context, time, and individual differences are likely while also predictably so. This dissertation thesis is organized as a scientific record of the process and results, documenting the role of the PhD student in developing research questions, study design, technology, and analytic approach. It culminates in the final results, and stand as the key product of the dissertation.

Organization of the Thesis

The Introduction outlines the organization of the thesis and an overview of the overall study. Chapter 1 covers concepts essential to the study's design, conduct, and analysis, as well as to interpret our work in a larger context, and lists necessary references. At the end of Chapter 1, we outline our research questions and hypotheses. Chapter 2 focuses on the details rather than the concepts already covered in Chapter 1, listing design choices that were materialized or

operationalized in designing or conducting this study or prior research or considerations that are directly relevant to this study. For Chapter 2, we minimized the discussion of concepts (as these were covered in Chapter 1) and focused on our actual design choices and our rationale for them.

Chapter 3 is a complete reprint of a previously published protocol paper written as the lead author, addressing the specifics of this trial. In describing the trial's details, overlap with Chapters 1 and 2 naturally occurs, but the already published manuscript is included without modification. However, where some of the manuscript's content (e.g., analysis methods) overlaps with the content of this dissertation thesis and may be somewhat confusing or misleading, this is noted this in a footnote.

Chapter 4 briefly discusses the analytical methods specific to this dissertation thesis. As Chapter 1 covered the conceptual work, Chapter 4 only describes the technical choices made in the analysis process and their rationale, just as Chapter 2 did for the clinical trial design. Chapter 5 presents the results of the analysis. Rather than a lengthy result report full of numbers, the results are summarized visually so that the reader can succinctly grasp the key takeaways from the study. The code that produced them will be made available using open science practices. Chapter 6 is a discussion section that outlines limitations, strengths, and future research. Chapter 7 is a brief conclusion.

Chapter 8 is an appendix, which includes a reprint of an essential related study, how the algorithm to define "Opportunity" was operationalized. While it was critical to enable the implementation of the final study, it is not directly related to the primary logical flow of this dissertation thesis. That said, given the roles as the first and corresponding author and the importance of this core piece that supported the final study, it is included in the interest of producing an archival document of the work completed during PhD program. Other

supplementary research materials are included in the appendices including the detailed results of the data analyses.

Study Overview

This study aims to improve mobile health (mHealth) interventions to increase physical activity. Specifically, it seeks to determine if a system identification approach can be used to learn how to individually optimize interventions to help different individuals engage in more physical activity. To this end, among other possible strategies, the study focused on both recognizing the 1) dynamic (time-varying), 2) context-dependent (depending on the individual's condition and external context), and 3) idiosyncratic (differences between individuals) of an individual's response to an intervention and the possibility that there would be predictable patterns, detectable using within-participant (also called idiographic) modeling approaches. This line of thinking is entirely consistent with the theorized notion of a Just-In-Time Adaptive Intervention [1], which aims to provide the necessary support when needed for each person.

The point we wanted to explore in this study is enhanced personalization, which goes beyond considering the context and time of day (e.g., morning/afternoon, weekday/weekend) of each individual and instead factors in some dynamic constructs that can be dynamically estimated from the individual. If individualized yet predictable patterns can be gleaned from a majority or more of participants, this would provide the evidence needed to justify the next step in this systematic line of research; namely the design of a control optimization trial (COT) [2], which is an idiographic study design that can be embedded within a digital health intervention to produce the evidence needed to develop personalized algorithms, called a controller, which is the specific type of personalization algorithm that is the ultimate target and approach for operationalizing a personalized intervention.

The type of experiment used both in this dissertation and COT trials to develop personalized predictions is an open-loop system identification experiment [3,4]. Several intervention strategies based on a priori hypotheses are presented to all participants for them to experience, and they are alternately exposed on a temporal axis in a method similar to a micro-randomized trials (MRT) [5]. It is called "open-loop" because individuals' responses to the intervention do not depend on their responses to the strategy.

This complex experiment, real-time data sharing, and algorithmic automated decision-making to implement the study required creating a complex computer system. The idea was to receive real-time data from Fitbit to estimate each participant's dynamic state, make different decisions for each individual based on the data, and tailor the intervention accordingly. This required a dedicated mobile app for the study, delivered to users via smartphones.

Finally, on the theoretical and statistical side, we utilized Bayesian regression and multilayer perceptron algorithms to look at causal models and nonlinear internal structure specifically within idiographic models. We chose Bayesian models in particular based on their more common use within idiographic modeling, particularly in N-of-1 cross-over designs [6], which is similar, experimentally, to our system identification case in that interventions are experimentally varied across time, within individuals.

Summary of Key Findings and Implications

This study aimed to assess whether timely, personalized notifications and daily step goals could significantly increase individual step counts compared to scenarios without such interventions. By analyzing participants' responses under specific time conditions and decision policies during a trial, we were able to predict the effectiveness of the intervention for a majority of participants. The trial successfully identified "Just-In-Time" (JIT) states for nearly all

participants (91% if using a priori approaches to account for need, opportunity and receptivity and up to 98% of participants if need, opportunity, and receptivity could be independently considered), demonstrating the potential to empirically identify individualized JIT states that could be utilized in future control-system-driven JITAI (Just-In-Time Adaptive Interventions).

The results underscore the effectiveness of Identifying Individualized states that can benefit from real-time support, covering most of the target sample. This paves the way for integrating these findings into a larger control optimization trial (COT), aiming to implement these insights on a larger scale to enhance physical activity. The positive outcomes from this study provide substantial justification for advancing this research line, focusing on real-time and scalable applications to boost physical activity through personalized interventions.

Chapter 1 INTRODUCTION

In this chapter, the rationale of this study, theoretical and practical concepts relevant to this study are enumerated, then the relationships between them are explained, and how they can be operationalized is discussed.

The Study of Physical Activity

Nature of Physical Activity

Physical Activity (PA) plays a crucial role as a protective factor against many diseases. There is convincing evidence indicating that PA is valuable for reducing the risk of bladder, breast, colon, endometrial, esophageal adenocarcinoma, renal, and gastric cancers [7] and cardiovascular disease [8,9] and improving glycemic control [10,11]. With an aging population, step interventions could help prevent chronic diseases, reduce healthcare costs, and improve functional life years and quality of life [12–17].

Moderate-to-vigorous physical activity is the most important and frequently mentioned type. According to US guidelines, adults can achieve clinically significant benefits from at least 150 minutes of MVPA per week [18]. In the past, the guidelines have been somewhat restrictive. For example, the guidelines stipulated that 10 minutes of MVPA must be done consecutively to be beneficial [19] . However, subsequent research has shown that shorter bouts of exercise also have positive cumulative effects [20], which has been relaxed in recent guidelines.

It is also a recent view that the benefits of physical activity do not necessarily come from MPVA alone. Even lower-intensity exercise, such as walking [21] or light jogging (Light Physical Activity) [22], has been shown to have health benefits. The practical advice of walking at least 8000 steps [23,24] or 10000 steps per day [19,25] has shifted in guidelines to describe the

value of physical activity in terms of MVPA [18], but the value of walking is still recommended. Studies suggest that the general recommendation to "walk more" has clinical benefits [13,23,26].

In summary, there is profound evidence that any form or intensity of physical activity benefits inactive adults, and even modest improvements can have proportionate clinical benefits. For adults, the recommended level is 150 minutes of moderate-intensity PA per week, but physical activity is beneficial even at lower levels.

On the other hand, prevalent sedentary and inactive lifestyles have been repeatedly reported [27–37]. These low physical activity levels result in significant financial and quality of life losses [14,38]. There is an urgent need to develop accessible interventions to improve low physical activity levels as a strategy for reducing risk of a range of chronic diseases.

Theoretical Model of the Study: Social Cognitive Theory (SCT)

This section introduces the theoretical model used in this study. This section relies heavily on the textbook Health Behavior: Theory, Research, and Practice (2015) [39] by Glanz et al. and will be cited repeatedly.

SCT is a highly used interpersonal health behavior theory [39,40]. According to SCT, an individual's health behaviors, cognitive functioning, and environmental factors interact, called *reciprocal determinism*. Important psychological constructs include a sense of control (called *agency*), vital in regulating behavior. The factors that explain the relationship between cognitive function and behavior include self-efficacy, outcome expectancy, and knowledge. Social and environmental factors surrounding the individual include observational learning, normative beliefs, social support, opportunities, and barriers. Behavioral skills, intentions, reinforcement, and punishment are considered to support behavior.

SCT also played a central role in the design of this study. Although not used in the central analysis of this thesis, Ecological Momentary Assessment (EMA), which participants

were asked to respond to daily, included daily measures of key constructs (See Supplemental Table 3.1 of Chapter 3). In addition, the mobile app provided to participants incorporated elements of the Behavior Change Technique (BCT) Taxonomy [41] linked to the agency, such as setting exercise goals and entering exercise history.

Relevant Psychological Constructs

This section introduces the relevant psychological constructs and relevant framework that support this study. This section relies heavily on the Nahum-Shani et al. 2015 [1] and will be cited repeatedly.

Opportunity: In health behavior change, an opportunity, also known as teachable moment [42], is a temporary state in which an individual has a high probability of internalizing information, taking action, and actively embracing positive change [1]. We recognized the need to consider past behavior to define this. More specific to this and in line with Nahum-Shani, we define opportunity as that next 3-h time window was predicted to have an 50-80% likelihood that someone could take steps.

Receptivity: We define receptivity as a person is deemed to be receptive when they have received ≤6 messages in the last 72 hours and have responded (i.e., walked in the following 3 h) after ≥50% of the walking notifications sent in that period.

Understanding whether an individual is receptive to support is important for ensuring that support is not wasted and delivered in the environment where it is needed [1]. Receptivity is expected to be dependent on context, including the media used to deliver it, its content, and frequency, and to change dynamically [1].

We add one more construct to this, called Need: We define this as accounting for if a particular type of support is required at a given time for a person. We include this to prevent offering support to individuals that may have the opportunity to respond favorably and receptive

but who otherwise would not benefit from support to reduce the likelihood of sending unnecessary support. We believe it is ethical not to provide support, even if it is valid, if we it is unlikely that a person needs it, particularly as a strategy to reduce the likelihood (and indeed prevalence) of notification fatigue [43,44].

Idiosyncrasy, Context Dependency, and Dynamics

Idiosyncrasy: Individuals May Have Different Patterns

It has long been discussed that the effectiveness of PA interventions varies significantly between individuals [45]. This stream of research has been gradually strengthened theoretically and methodologically [46–51] and empirically [49,52–54]. This trend has also been linked to research looking for behavioral patterns observed through digital signatures (i.e., digital phenotyping) [55–59], genetic backgrounds in patterns of behavior or response to interventions [60–62] or psychological factors, including personality [52,63].

We assume that the moderating effect of facets of JIT states mentioned above on the effectiveness of the intervention may vary across individuals. Since we consider three facets of JIT states (Need, Opportunity, and Receptivity), the pattern can be quite complex.

Context Dependency: People May Behave and Respond Differently Based on Context

Context dependency refers to people's behavior, psychological constructs, or behavioral response to the intervention varies over contextual environment where the person is put. This dynamic psychological state can naturally trigger or inhibit an individual's behavior [39]. For example, our prior research has shown that some people walk more when they are stressed, whereas others walk less [3]. While this study highlights inter-individual differences in these tendencies, we have found various cases where the fast-changing and dynamic concept of stress has short-term effects on walking behavior [3].

It can also be further assumed that this psychological state may Influence the Individual's receptivity to the intervention, or the short-term response to the intervention (in other words, the proximal outcome), and the long-term effect (in other words, the distal outcome) [1,5,64–66]. To extend the example above, we may hypothesize that some people respond more favorably to the walking suggestions when stressed, whereas others may do so when they are less stressed. This process is illustrated in Figure 1.1 (b). In other words, the psychological construct is assumed to act as a time-varying moderator in the causal relation between intervention and behavior. Figure 1.1 (b) illustrates what we hope to uncover in this thesis. We want to examine the time-varying moderation effect of the three psychological constructs mentioned above (i.e., Need, Opportunity, and Receptivity), and the intervention is in-app notification, with these time-varying moderations experimentally varied in the form of different decision policy algorithms (defined in chapter 2 on page 53).

Dynamics: Constructs Vary Over Time

The word dynamic refers to a variety of concepts, depending on the contexts [67]. In this thesis, dynamics refers to the property that the constructs of interest have varied values *over time* within a defined time range that is measurable and studied within-person/idiographic modeling. In particular, the change in these values frequently defies simple predictability, diverging markedly from straightforward patterns such as constant levels, sinusoidal forms like sine waves, highly periodic cycles, or linear trends of increase or decrease. One of the most prominent examples is dynamics observed in the behavior of interest, steps/min. In this study, the number of steps per minute measured by the Fitbit wearable is one indicator of the focal behavior of physical activity. The number of steps per minute ranges from 0 steps/min when not moving or not wearing it to 81-138 steps/min at a normal pace [23,68], and in some extreme cases, up to 200 steps/min. Most of the other constructs, such as psychological constructs, weather, and

schedule, may show large fluctuations and are not easily predictable. Thus, in planning our experiment and analysis, we must account for the variable and shifting dynamics of all components, particularly our primary outcome of steps/min (and its various aggregations) and JIT states.

Figure 1.1 (a) shows an illustrative directional acyclic graph for the dynamic behavior. In each diagram of Figure 1.1, each circle denotes a dynamic construct (i.e., variable) that can be temporarily measured.

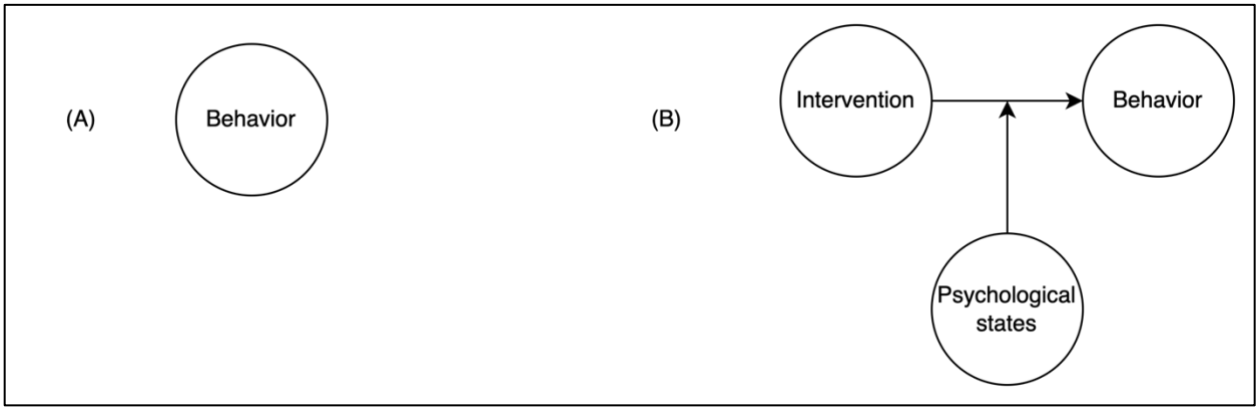


Figure 1.1 Illustrative Directional Acyclic Graph (DAG) for the causal relationships between dynamic behavior, psychological states, and intervention.

In this study, we assumed that, among other things, need, opportunity, and receptivity can vary across individuals (idiosyncrasy), across context (context dependency), and dynamically [1]. These three focal target psychological constructs will be referred to in this thesis as "JIT states" because they support and enrich a special type of intervention called a Just-In-Time Adaptive Intervention (JITAI, see page 23 for detailed concepts).

From Context Dependency to Causality

Advances in study designs from psychology and medicine

In most areas of medicine, between-person randomized controlled trials (RCTs) are the type of study accepted as the most valid for supporting causal inferences. However, traditional

RCTs are not well suited to the focus of this research for two reasons. First, RCTs have difficulty capturing individual differences, particularly causally. The strength of RCTs is to use randomization to seek to eliminate the impact of confounders by averaging out their effects across a large enough group of participants randomized to an intervention vs. control condition. With this strategy, individual-level deviations are treated as "variance" that needs to be "averaged out". These results are highly valuable for supporting decision-making at the level of populations. For example, this type of evidence provides guidance on selecting populations to focus on who may be experiencing disparities compared to other populations or the selection of interventions that are "generally" useful as the first line treatment provided by a healthcare system.

With that said, there is mounting evidence that human behavior is a non-ergodic phenomenon [69]. Ergodicity is an assumption related to the alignment of between-person variance observed – the focus of classic population-focused statistics, also called nomothetic statistics – and within-person variance observed – the focus of this dissertation, which involves analyses conducted on individual times-series data, also called idiographic modeling. The key assumption of ergodicity is that the average outcome observed from a group is the same as the average outcome of an individual over time. An ergodic system is flipping a coin. If 100 people flip a coin or 1 person flips a coin 100 times, the average outcome would be the same. In non-ergodic systems, the individual does not experience the average outcome from the group over time. Based on this, a simple way to figure out if a phenomenon is ergodic or non-ergodic is to compare predictions/projections of an individual about a phenomenon across time or look at a phenomenon at a single time point with multiple individuals. If we get the same results, then the phenomenon under study is ergodic. If it is not the same results, it is a non-ergodic phenomenon.

Based on this definition of ergodicity, it is highly unlikely that human behavior is ergodic, though, of course, this is not a clear distinction between ergodic vs. non-ergodic. Instead, it is more likely exists on a continuum between the two and is contingent upon the exact focal phenomena of interest.

This has potentially large implications for how we conduct our work to develop interventions that can help people improve behavior and helps to flag the need for BOTH population-focused (nomothetic) and individual-across-time-focused (idiographic) statistical approaches. The traditional nomothetic models are valuable for supporting decisions made about populations, such as identifying general issues to focus on, such as physical inactivity, that is prevalent across a population. Another good use would be to support decision-making about what first-line treatment should be provided to a population, such as vaccinations, or selection of the type of behavioral intervention to offer a population. If the goal is to support decisions made to support individuals' changing behaviors over time, given non-ergodicity, it requires the use of idiographic models that can identify the patterns and dynamics that occur with each person.

With this, we return to our core hypothesis, which is that people are different, context matters, and things change in a predictable fashion when studied using idiographic modeling. The RCT strategy of "averaging out variance" literally "averages out" the core focus of our hypothesis and, unless human behavior is ergodic, provides limited insights on what the appropriate decisions should be to support people in behavior change, save only the very first decision offered.

This problem becomes worse when it is hypothesized that the causal effects will be context-dependent and multicausal (for more details on this, see discussion on INUS conditions below). Traditional RCTs assign a group to an individual once at the beginning of the trial (i.e.,

randomization) and do not change it thereafter. With this, any context-dependent and time-variant causal effects, such as the facets of JIT states we hypothesize, would not be experimentally varied and thus, could not be meaningfully studied with traditional nomothetic methods.

To address these challenges in studying dynamic and context-dependent causal relationships, numerous study design variations have been proposed [70]. As each has a different focus and functional goal, we describe them in parallel in this section, to help further situate our work within the broader context of new experimental approaches for guiding intervention optimization, which have been advanced in the context of the multiphase optimization strategy (MOST).

MOST is a framework for guiding the optimization of biobehavioral and behavioral interventions [71]. MOST is a framework, not a single methodology, and can be used in conjunction with various other tools and study designs, which are collectively called "optimization trials." MOST was explicitly inspired by engineering [70,71]. The most common design used in MOST is a screening experiment. Within it, multiple intervention components options are identified and, a series of steps are taken to screen out components that do not significantly affect efficacy. The final "optimized" intervention is then tested in a standard RCT to validate its effectiveness. Within our system identification trial, we technically included three experimentally varied factors, goals, notifications, and our variations in decision policies. Technically, our trial could be considered a factorial screening experiment, that utilized a within-person randomization strategy instead of between persons. With this, the results of our nomothetic tests could be thought of as akin to a classic MOST factorial screening experiment.

Another common trial is the Sequential Multiple Assignment Randomized Trial (SMART). SMART is an experimental approach in which an intervention is multilayered with additional randomizations for some participants to support deeper exploration of the effects of individual intervention components that are difficult to capture in a single time point, single-stage randomization in a typical RCT [72,73]. This randomization can also be applied to intervention components that will be tried later in the intervention.

This randomization can be fused with a scheme that directs participants to different interventions via a participant classification tree separated by a series of if-else logic [73]. The branching logic of some of the nodes in this classification tree could include whether the participants responded to the initial intervention, the severity of their symptoms, whether they are getting better, or whether they are conscientiously participating.

Getting closer to the design we used is the micro-randomized trial (MRT). The MRT is a method that allows testing of how the same participant responds to different environments, types of interventions, and contexts by repeatedly randomizing them multiple times [5]. In essence, although its original concept focuses on the nomothetic nature of the populations, because each randomization may expose each individual to diverse contexts, it is potentially useful to build a detailed model of each individual, even for research hypotheses that view each individual as an independent system. With this, it is a logical and valuable tool to explore the facets of human behavior that may be ergodic vs. non-ergodic.

In MRT, a decision point (i.e., the point at which each randomization occurs) can semantically mean any point on the time axis. However, depending on the study design, it can also mean something procedurally specific. The decision does not necessarily have to be a coin flip; it can also involve slightly modulating some element of the intervention, as was considered

in SMART, or tweaking how the intervention is operationalized. As long as the randomization is done systematically and carefully, it is part of a microrandomized trial.

MRTs and N-of-1 crossover designs are the closest study designs to our design that has its lineage from RCTs. With this, nomothetic estimates can be produced that can account for timing, thus, partially addressing potential challenges of non-ergodicity by providing far more targeted predictions in repeatable states. With that said, if the phenomenon is highly non-ergodic, then these study designs, particularly the MRT or producing nomothetic predictions polled from a series of N-of-1 cross-over designs, would still produce relatively limited value. Our study design, called a ***System Identification study***, comes from the disciplinary lineage of control systems engineering, but is akin to a mixture of a an MRT and N-of-1 cross-over design.

MRTs shift the randomization from something that occurs at a single time point between people to, instead, randomization that occurs repeatedly within-person across time and at pre-determined decision points. For example, an MRT could use a decision point of 8a each morning to randomize sending a notification to walk or not, or a decision point could be defined algorithmically, such as when a wearable detects that a person seems to be stressed. This experimental structure supports efforts to explore causal relationships operating in complex contexts. A system identification experiment largely follows this same pattern, but with incorporation of pseudo-random non-linear deterministic signals and with an explicit focus on running idiographic statistician analyses. Pseudo-random signals produce the valuable properties of randomness to support counterfactual causal inference but, critically, are repeatable across time. They are also designed with an explicit use of accounting for orthogonality across the frequency domain. This means that the temporal effects of experimental manipulations can be

disambiguated across time within persons. This provides a structure to support counterfactual causal testing within individuals.

Drawing from N-of-1 cross-over designs, it is common to use Bayesian statistical analyses instead of frequentist statistics to support causal inferences for each targeted individual; hence the choice in this dissertation to use Bayesian statistics instead of frequentist statistics for all idiographic modeling. Finally, and unique to our generic hypothesis that context matters, things change, and people are different but predictably so, our study involved, to the best of our knowledge a unique shift in what was experimentally manipulated. Traditional MRTs and N-of-1 cross over trials experimentally manipulate the intervention strategy. In our system ID experiment, we experimentally manipulated decision policies, which are the algorithms that we designed to operationally define JIT states. We included four variations (defined in depth below). This manipulation of decision policies was based on the philosophical causal notion of an INUS condition, described next.

Control Systems Engineering

Control systems engineering refers to the process of the design, analysis, and improvement of dynamic systems [2]. These systems consist of devices that manage the operation of other devices or systems and may be mechanical, electronic, or both. The aim of control systems engineering is to create equipment and systems that perform as expected within controlled settings.

While system identification guides us to attain the internal structure of unknown dynamic systems, control systems engineering requires the accurate internal structures of the system to control it or peripheral systems. Thus, two engineering concepts, system identification and control systems engineering, are tightly connected. Particularly, if we target idiographic control systems engineering to optimize the intervention to improve the individual's behavior response,

we may naturally need to use system identification methodology for each individual, as they began to be actively used in human behavior research [2,3,74,75].

Using the methods described in the previous sections, we want to detect the predictable vulnerable states, for each person, to improve PA. If we can find them, this will be strong evidence to move on to multi-timescale COT that incorporate this approach for learning individualized models and, with that, enact a robust JITAI for supporting PA. This process requires the team to develop an innovative experimental design that could enable causal inferences that compare different JIT States, operationalized via guidance from INUS condition causal logic, described next.

INUS Condition

This study utilized INUS conditions [76] as a conceptual model for causality. INUS conditions stands for Insufficient but Necessary parts of Unnecessary but Sufficient conditions and was first proposed by philosopher J. L. Mackie in 1965 [76]. An INUS condition philosophy was not the causal model that was used to design the experiment. As just mentioned, we are using causal inference insights, particularly the notion of counter-factual logic, to guide our study design, just as is the case with MRT and N-of-1 cross over designs. With that said, INUS condition causal logic was used to guide the development of the decision policies that operationally defined and varied different JIT state operationalizations. INUS Condition causal logic was used as it provides guidance on how to theorize about moment-to-moment state-shift-focused causal phenomena. Beyond this, while not a traditional part of INUS condition causal logic, we found that did lend itself well, conceptually, towards using deterministic signals to produce probabilistic predictions that could be useful in our targeted future personalized digital health intervention.

In INUS causal logic, we do not attribute a single cause for a focal event we are interested in happening. Instead, multiple ***factors*** are theorized to form a ***condition***, and when this condition is satisfied, in classical INUS causal logic, the event will happen. For example, providing an intervention to a participant via an in-app notification does not automatically cause the participant to walk. Instead, the intervention, along with some other factors (e.g., the participant is feeling the need for a walking behavior right now, and there is an opportunistic setting for walking, such as a gap in their schedule), is theorized to cause the participant to go for a walk when they all occur at the same moment. Using our concepts of an intervention, need, and opportunity, each can be thought of as a single ***INUS factor***. When these three factors occur in a given moment (e.g., "get the intervention when you need it and have the opportunity"), that moment meets the hypothesized ***INUS condition*** that can produce the targeted effect.

While there is great value in this logic, there are limitations to its classical use. Most critically, Mackie formulated INUS Conditions in a deterministic way. If the INUS condition is met, the effect will occur. If not, it will not occur. Human behavior is not merely non-ergodic (and thus, requires study across time), but, particularly when looking towards causal assertions towards repeating states, it is likely much more like the weather; meaning inherently non-linear, non-predictive, and influenced by stochasticity (or, as statement more colloquially, human behavior is likely chaotic, as used within chaos theory). Returning to our focus, there are many other potential triggers for the target behavior besides the "receive intervention when needed and available" condition. For example, "getting interventions when I am busy and stressed" or "getting interventions on a weekend morning when I have had a good night's sleep but am still tired" might both also be INUS conditions that can produce a walk. With this, we needed to

develop an empirically replicable approach for studying INUS condition causal hypotheses while still recognize the likely inherent chaotic nature of human behavior.

Specifically, in the empirical use of INUS conditions within the context of a phenomenon influenced by non-linear, non-predictive, and stochastic influences, one important starting point is to probabilistically reinterpret the meaning of each logical concept dealing with INUS causal logic, including Sufficiency [77,78]. The four terms in the name of INUS are commonly used to describe deterministic causal drivers. For example, when we say that A is a sufficient condition for B, we mean that if A is True, then B will necessarily be True. However, from a probabilistic view, we mean that if A is true, the probability that B is true is significantly higher than when A is false. Suppose we further adopt a probabilistic view of the probability of B being true, meaning that the true probability of B being true has a probability distribution, such as a normal distribution, rather than a fixed value. In that case, the probability distribution is shifted to the right (i.e., larger values) when A is true compared to when A is false. (See Figure 1.2) It also means that even when A is false, the probability of an event happening is not completely zero, but can have a small value.

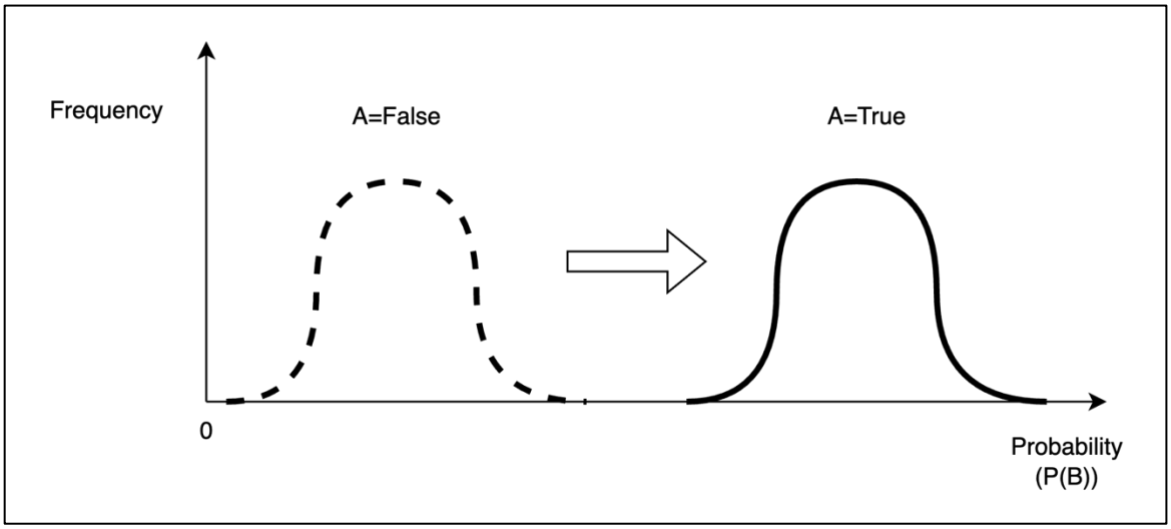


Figure 1.2 An illustrated concept of probabilistic view on the INUS conditions.

Bayesian statistics is suitable for dealing with these mathematical concepts, because it allows for more flexible assumptions about the probabilistic distribution of the estimate [79,80]. As described earlier as well, it is also the more common statistical toolset used for supporting causal inference within N-of-1 trials [6]. More details are described on page 37.With this, we explicitly incorporated Bayesian notions into our work to honor and recognize the likely inherent chaotic nature of human behavior within our experiment while still allowing us to achieve our aims of seeking to detect INUS Condition patterns that were reliably predictive ideographically.

As the names of the INUS conditions hint, INUS conditions are formed by combinations of INUS factors. Just as "intervention" is an INUS factor, factors such as "need," "opportunity," "weekend," "morning," "stress," "busy," and "tired" can all optionally but combinatorially intertwined to form a specific INUS condition. In this study, we focused on five INUS factors: "need," "opportunity," "receptivity," "weekday/weekend," and "morning/afternoon."

We also note that most of these INUS factors are not capable of causing an event on their own. The first letter of INUS, I (Insufficient), refers to this. However, when these INUS factors meet with other INUS factors to form a condition, they become a condition that can potentially cause an event. This is what the second letter of INUS, N (Necessary), means.

On the other hand, this condition is not necessarily assumed to deterministically produce the desired target of increased walking. This is based on prior work suggesting that human behavior is influenced by non-linear, non-predictive, and stochastic influences as mentioned earlier. The third letter of INUS, U (Unnecessary), recognizes this attribute for each INUS factor; each is unnecessary in that it is not the only possible form that could be used to produce the contribution needed for an INUS condition to be met. However, this factor is a sufficient

form for activating the theorized function that the factor is meant to contribute towards the overall INUS Condition. The fourth letter of INUS, S (Sufficient), means this.

Once we get beyond this conceptual level and implement and operationalize each factor, we refer to it as an INUS form, can be combined to operationalized define variations of INUS conditions and, thus, enable the test of our theorized JIT states. While there are subtle differences, for the sake of this thesis will use these terms (i.e., INUS condition and JIT State; and INUS factors and forms) interchangeably, with the INUS factor providing the concept, INUS form, the operationalization of each element of the causal pattern and the JIT State providing the conceptual structure and INUS Condition providing its operationalization.

With this as background, we turn now to how we formulated our dynamic hypotheses. With these terms defined, we can now restate our generic hypothesis into a more testable form. Specifically, we hypothesized that different people may be predictably responsive to different INUS conditions as operationalizations of varied JIT states. Therefore, based on prior research and domain knowledge, we enumerated 32 INUS conditions[1] that we thought would be plausible states that could be identified as observable conditions that could be identified as a reliably predictable condition (or set of conditions; in this work and logic, there is no requirement that only one INUS condition need work for one person) for each individual. We designed an experiment that, via experimental manipulation of different decision policies that define variations in INUS factors that would be experienced and manifest across different theorized critical contexts (weekend vs. weekday and morning vs. afternoon), it enabled us to

[1] **4 timings** x **4 JIT state combinations** x **Whether we intervened (2, T/F) = 32**
- **Timings**: Weekday morning, weekday afternoon, weekend morning, and weekend afternoon
- **JIT state combinations**:
1) All JIT states (Need, Opportunity, Receptivity) are true
2) Need and Opportunity are true
3) Need and Receptivity are true
4) No JIT states are considered

experimentally test the causal impact of these INUS conditions for each individual. Since our primary interest is "Does sending a notification to walk increase steps in this context relative to when we do not send notifications?", we conducted pairwise comparisons of the INUS conditions with and without the intervention (32/2=16 pairs) and estimate the effect of the intervention on behavior within the context of these experimentally manipulated variations of INUS conditions using Bayesian statistics. As we experimentally varied both sending and not sending notifications and our decision policy, these estimates, using counterfactual causal logic, are akin to the causal predictions used in classic RCTs, but, now, with the causal inferences occurring for each individual. To the best of our knowledge, this is the first type such an innovative approach for studying a phenomena with this level of causal complexity has been conducted. Using Bayesian statistical notions of credibility intervals instead of frequentist confidence intervals, we sought to identify INUS conditions whereby there was an 80% or greater likelihood that, over the 9 months of the study, when a given INUS condition was met, a person walked a minimum of 100 additional steps compared to the same INUS condition, but without a notification being sent. With this, we produce an estimate of *the degree* to which a particular INUS condition that includes notifications, can reliably result in increased steps relative to the same INUS condition but without a notification offered based on experimental variation of both the decision policies and notifications.

Just-In-Time Adaptive Intervention (JITAI)

Need of JITAI

JITAI refers to a set of temporal interventions that aim to provide the right kind of intervention when an individual needs it [81]. The definition in the previous sentence is somewhat unclear about what it means to be a "temporal" intervention, but we use the term here to mean that time is accounted for when determining if a person needs a particular type of

intervention. JITAI's basic assumption, as stated in the definition, is that the need for the intervention, or the type and content of the intervention, changes over time. These factors are naturally assumed to vary across individuals, hence the basic premise of personalized intervention delivery. In many cases, this is a two-step process: 1) estimating the need for intervention, the kind of intervention needed, and what a more effective intervention might look like, and then 2) providing that intervention in a timely manner [5,81]. The process of estimating the appropriate intervention often involves measuring or estimating the individual's current state, recent behavior, context, or circumstances.

The sensitive flexibility of interventions described above assumes that interventions will be less effective, or even harmful, if they are delivered non-selectively or if they are not the appropriately matched intervention to a particular context. JITAI is particularly valid for certain behaviors, contexts, and interventions where these assumptions are illustrated or demonstrated. Studies that delve into JITAI have rapidly increased in recent years, particularly in the mobile health field [1,82–85].

Enabling Factors of JITAI

Technological advances, including sensors, wireless communications, and mobile devices, have profoundly impacted how interventions for health behavior change are delivered.

Sensors. The development of sensors has broken new ground regarding objective behavior measurement. The ability to measure various signals at high frequencies, including amplitude, magnitude, frequency, and type of behavior, is an essential shift in the behavior intervention field. The triaxial accelerometer sensor is a prime example. First used in research in the 80s, accelerometer sensors [86] had limitations in terms of price, usability, and accuracy, but their promise was recognized [87], and their use gradually increased [88].

In recent years, the availability of commercial sensors beyond research sensors has also played an essential role, despite their more significant limitations in terms of validity than research sensors [89]. Their use is becoming more widespread for reasons of price and availability [90].

Wireless Communication and Mobile Device. The widespread use of smartphones, with their powerful computational power, portability, and always-on wireless data connectivity, has opened a new frontier for mobile health. Smartphones serve as a platform from which information measured by sensors can be transmitted to the Internet in real time while also serving as a two-way communication channel as the basis for digital interventions to be delivered to individuals. For this reason, smartphones with wireless communication are one of the critical elements of JITAI.

Real-time Signal Processing in the Study Server. Powerful servers and algorithms are not necessarily used in every JITAI, but they can play an important role in increasing its fidelity. Efficient and robust server programming is essential to receive and process data streams generated by sensors in real time, determine the appropriate content of interventions, and respond to them individually.

Challenges of JITAI

JITAI requires the integration of digital technologies, including the use of the internet, wearable sensors, or mobile devices to measure and transmit information about individuals, process and analyze data, and deliver interventions. There is a risk of creating a digital divide [91,92] and burden of use in the aged populations [93] or difficulties for under-resourced providers [94]. There are also concerns about clinical interventions being delayed due to inaccurate risk assessments in mobile apps [95], the potential risk that it may take longer to access existing interventions with established effectiveness [91], and user concerns about privacy

and cybersecurity [96]. Most of these concerns are not unique to JITAI but are shared across digital interventions.

Measurement of Walking Behavior

Conventional Measurement

There are many ways to measure gait, or gait movement. In past studies, direct observation [89], physical activity questionnaires [97], transportation surveys [98], pedometers [99,100], and video measurements [101] have been widely used. Physical activity questionnaires, a type of self-report, are recognized as a survey method with a wide range of validation.

However, these methods have difficulty measuring walking behavior in free-living environments and are unsuitable for supporting JITAI or producing high-resolution time series data. Furthermore, they either cannot be objectively measured or require participants to record them by hand, which creates a high measurement burden to increase the measurement frequency.

However, as Chevance et al. documented in [102], the patterns identified within human dynamical behavior are known to be highly dependent on temporal resolution. In our work [103] and in past studies, our team has learned from observation that human behavior, such as gait, change too rapidly to be characterized by daily or more aggregated data. Depending on the research question, it might be appropriate to utilize such aggregate data (e.g., increased weekly activity as this is aligned with national recommendations of PA). However, suppose we are interested in context-dependent phenomena (e.g., the relationship between contextual variables like weather and day of the week on steps/day) or fast time-scale dynamics (e.g., if a person engaged in steps/min during the 3-hour window after when a notification to walk could feasibly have been sent). For each of these, more fine-grained aggregations of the steps/min data are appropriate.

Sensor-based Measurement

Accelerometer-based sensors have emerged to mitigate these limitations. Pedometers utilize mobile mechanical components such as springs to convert the strong impulses from walking into electrical signals [99,100]. In contrast, electronic semiconductor-based accelerometers use static parts, including piezoresistive, piezoelectric, or differential capacitive accelerometer material-based components to measure the change in acceleration or angle of gravity within an inertial system and convert it into electrical signals [104].

The acceleration changes recorded by the above methods are used in conjunction with a built-in timer to record the position, velocity, and acceleration in the three-dimensional space of the body part where the sensor is worn as a high-frequency time series on the internal storage [105]. Accurately calculating step counts based on these accelerations, velocities, and positions is also a significant challenge [106,107]. While heuristic approaches based on domain knowledge are sometimes used, approaches based on machine learning, such as decision trees, are more common [108–110].

According to the Nyquist-Shannon sampling theorem [111], to avoid aliasing, a type of distortion of a signal whereby the influence of two or more signals cannot be disambiguated, measurements should be made once every half a cycle of the minimum period of the phenomenon of interest. The minimum period length of the phenomenon should be determined by prior research or close observation. In the absence of prior research or where it is technically feasible to increase the period, it is desirable to measure at the fastest time-period possible, as it enables later aggregation to the appropriate timescale [102]. For example, although human gait movements, even at their fastest, are typically no more than 3 Hz, most currently available commercial or research-grade accelerometers offer sampling frequencies of 30-100 Hz [112–114].

Commercial Wearable Sensors

Advances in commercial mobile devices and information technology have made it relatively easy for researchers to gather this level of granularity, at least as it pertains to PA and now increasingly other variables such as heart rate. For example, some device manufacturers, including Fitbit, can send minute-by-minute data to researchers on a regular basis [115], and others, such as Apple Watch, can instantly send the exact time of each gait movement with the millisecond resolution, depending on the researcher's settings [116]. The amount of information collected this way is staggering compared to data gathered from traditional nomothetic research approaches, such as survey-based epidemiology. Although the technical challenges of simply transmitting and storing it can be overcome with modern technology, the challenge of using these data flows for behavioral research lies in analysis and interpretation.

Automatic Data Gathering Through the Internet

The process of automatic data gathering through the Internet is initiated by the establishing the study server (rightmost rectangle of Figure 1.3). When the study server is built on the Internet, its Internet address (e.g., in our case, justwalk.ucsd.edu/update) is registered to the Fitbit server system (arrow 1 of Figure 1.3).

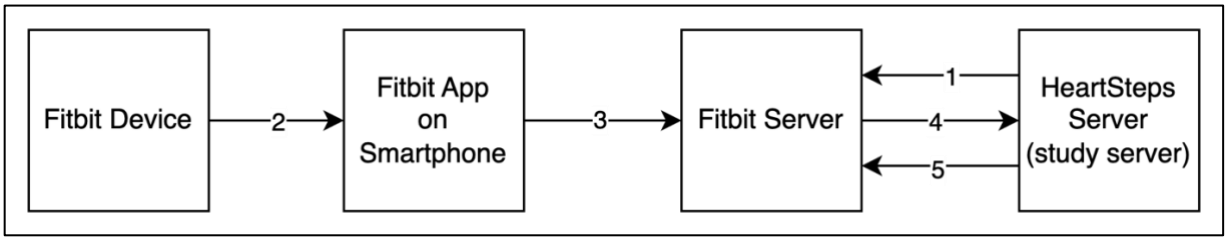


Figure 1.3 Schematics of Data Sharing Between Fitbit Wearable, Server, and Study Server

The data measured through wearable (e.g., Fitbit) are stored in their wearable device's local storage. Then, suppose the device is near the smartphone, and the smartphone is available to synchronize the data with the wearable (e.g., the battery is enough, the phone is not in power-

saving mode, and the Fitbit app is alive in the operating system's background process queues). In that case, the wearable's internal storage data is sent to the smartphone through a wireless connection (arrow 2 of Figure 1.3). In rare cases, some vendors' wearables have a standalone Wi-Fi connection to submit their data directly to the manufacturer's server. However, in our study, Fitbit does not have such capability.

If the data are shared to the smartphones, the data are automatically uploaded to the wearable manufacturer's server (i.e., in our study, Fitbit's server, arrow 3 of Figure 1.3). If the phone has no internet connection (e.g., camping, staying outdoors, on flights, or overseas without a data roaming connection), the data remains in the smartphone indefinitely. If the data connection to the Internet is resumed, all the data are sent to the vendor's server.

When the vendor's server receives the data, it accumulates the data stream for at least 15 minutes. After 15 minutes, Fitbit's server notifies the study server that there is a new update for a particular participant (arrow 4 in Figure 1.3). Then, the study server requests the Fitbit server to return the participant's updates (arrow 5 in Figure 1.3). Although this flow seems complicated, it is an industry-standard information flow for the sensor-measured data over the Internet.

Technical Challenges of the Measurement, Their Impacts on Studies, and Promising Approaches

It is well known that there is potential risk of wearables' data might come to the study server late [117,118]. We anticipated the risk of missing data purely caused by technical reasons, not the participant's intention. Thus, when we need the activity data, we checked the server, and if the phone had not sent the data for over an hour, we sent a regular text message to the participant to request to run the Fitbit app once more to synchronize the data to the server.

Such phenomena occur more frequently in the late evening or early morning when the participant's activity data is needed than at other times. The phenomenon might be related to the

user's phone usage patterns, battery status, charging status, and usage of other apps, but it was difficult to identify a clear pattern. In our lab tests on an iPhone 13 Pro from Apple Inc., we found it difficult to find any settings in the official iOS Settings app that correlated with this behavior.

If the participant did not turn on their phone when they received the text messages, this phenomenon can last multiple days until they turn on the phone. Such data gaps largely impact the real-time JIT state estimation. For example, our operationalization of the JIT states heavily depends on real-time activity data. More importantly, the algorithms are set to rely more on recent days. The estimation may shift inaccurately if the data are missed in recent days. The impact of such data delay will be discussed in the result section (See "Exploratory Aim 2: Post Hoc Analysis of JIT states" on page 162).

The ultimate solution for this is for the devices and smartphones to stay connected as much as possible. Low-power wireless connection technologies, including Bluetooth Low Energy or Bluetooth 5 [119], are introduced. Smartphone operating systems need more powerful batteries and better power management technologies. However, as of 2024, this phenomenon has not been resolved yet.

Statistical Models and Computational Considerations

Mixed Effects Models

Mixed effects models refer to statistical models that combine fixed-effects models, in which the model parameters have fixed values, with random-effects models, in which the model parameters can vary across samples[120].

Repeated measures analysis of variance (ANOVA) is traditionally used when analyzing repeatedly measured data from multiple respondents because the measurements usually do not satisfy the assumption of independence (i.e., measures are correlated), which is not required by

repeated measures ANOVA. However, several limitations have been pointed out, including the inability to model both within- and between-individual differences simultaneously; the widespread use of listwise deletion, which requires eliminating an item from all subjects if a missing value occurs in one respondent; and the fact that it answers whether an effect is significant but does not answer the direction or magnitude of the effect[121].

Mixed-effects models have emerged to overcome these issues[120,122]. They are known to overcome all, or at least partially, the problems mentioned above. Differences within and between individuals can be modeled simultaneously. Information loss is greatly reduced in the case of missing values because only those missing values are eliminated. Also, they not only answer the question of whether there is an effect but also provide quantitative information about the direction and magnitude of the effect. These advantages have led to the widespread use of mixed-effects models[122].

The parameters in the model can be estimated by specifying:

1. Which individuals does each sample belong to?
2. Which variables should vary across individuals but remain constant within individuals (random effects)?
3. Which variables should remain constant across all samples regardless of the individual (fixed effects)?

The number of parameters the model estimates will be *(number of fixed effect variables) + (number of random effects variables) x (number of individuals).* In addition, the hyperparameters that the modeler should determine include the linkage function, which describes the relationship between the variables in the model and the measurements, the statistical

distribution of the measurements (See the next section about count regression), and the type of optimizer used in operating the mixed effects model[120].

Based on this, the mixed effects model accounts for the nomothetic pattern expressed on average across all individuals as fixed effects, and for individual differences that deviate from this common tendency but are maintained within the individual as random effects. The residuals represent the remaining variance that is not explained by these two.

This thesis initially attempts to use the mixed effects model to provide a generalized lens to understand how our intervention is helping people meaningfully, in general, across the study sample as an indicator of a population estimate. The decision points are labeled as to which individuals they belong to (item #1 in the list above). Each individual can have their own baseline 3-hour activity level and linear increasing/decreasing tendency in overall activity level as the individual progresses through the intervention (item #2 in the list above, the random effects). All other factors (e.g., how many more steps are taken in a 3-hour when notifications are given, on weekends, or when more ambitious step goals are provided) were modeled as fixed effects (item #3 in the list above). This model structure was an attempt as a standard and traditional methodology for developing a nomothetic response model to our intervention to answer the question: "did the intervention components 'work', on average, across the sample?".

In addition, intending to optimize the model structure to describe the participants' behavior responses the best in a predictable manner, we performed stepwise regression, which is widely used in introductory statistics practice, in a forward fashion. As a base model, we included the key elements of the intervention:

1. Whether or not the notification was provided,

2. A pseudorandom value between 0-1 used to determine the daily step goal (i.e., goal factor),
3. A dichotomized decision policy (random vs. all other levels) and
4. 2-way interaction terms between them.

In this base model, we examined each variable in the following order: linear univariate terms, quadratic univariate terms, 2-way interaction terms, and 3-way interaction terms. If adding a term decreases Akaike Information Criterion (AIC) values, which are estimators of prediction error, the term is retained.

The log linkage function and ZINB model with constant dispersion and zero-inflation probability were used.

Count Regression

Step counts, or more precisely, the sum of step counts over a time-period, are a measure of the number of times the event of taking a step occurs. The probability distribution of a statistic that counts the number of times an event occurs is called a count distribution, and the regression methods suitable for it are collectively called count regression [123]. Poisson regression is often used as the most basic count model for count regression.

$$Pr(X = k) = \frac{\mu^k e^{(-\mu)}}{k!}$$

Equation 1.1 Probability density function of Poisson Distribution

However, Poisson regression has the constraint that the expected value and variance of the counts must be equal (the mean μ is the only parameter of the Poisson distribution). In practice, this condition is challenging to satisfy in most cases, and our preliminary research showed that the variance of step count data was about 1,000 times the mean. This phenomenon of large variance relative to the mean is called overdispersion [124].

Negative Binomial (NB) regression can be applied for overdispersed count data. The NB distribution is the distribution of the total number of trials until the desired event occurs r times through Bernoulli trials with probability p. It models a similar shape to the Poisson distribution. However, it allows for estimation when the variance is large relative to the mean. $NB(r,p)$ is expressed by the following formula:

$$Pr(X=k)=\binom{k-1}{r-1}p^r(1-p)^{k-r}=\frac{\Gamma(k)}{\Gamma(r)\Gamma(k-r+1)}p^r(1-p)^{k-r}$$

$$p\in[0,1], r>0, r\leq k$$

$E(X)=\frac{r}{p}, Var(X)=\frac{r(1-p)}{p^2}$. On the contrary, we can formulate the parameters with the expectation and variances as follows: $p=\frac{E(X)}{Var(X)+E(X)}, r=\frac{E(X)^2}{Var(X)-E(X)}$.

However, the above description is only a consequentialist interpretation and applied use of the NB distribution. If we use NB (or its zero-inflated variant; see next section), we need to think theoretically about why it works.

Why is Negative Binomial Distribution Appropriate for Walking Behavior?

We analyze the human walking behavior in three phases as follows: 1) each individual has a dynamically changing urge to walk[2]; 2) with each step taken, a small amount of this urge is released[3]; and 3) when all of the urge is released, the individual stops walking. Let us denote by r the amount of urges, the sum of internal desires and needs, that an individual has at any given

[2] It can be viewed as an intentional urge to take a walk, a need for steps in daily life (e.g., walking from the office to the copier, steps needed to get to work or lunch), or a fixed distance that needs to be walked (e.g., from home to the train station). In other words, it is a quantity that needs to be addressed as a walking behavior.

[3] It can be thought of as the average amount resolved per step, or if we assume that a step is a Bernoulli trial, such as a probabilistic coin flip, it can be thought of as the probability p of a Bernoulli trial. In this case, $1/p$ is the expected value of the steps required to resolve one unit of urge. Conceptually, it only makes a little difference to the overall idea.

time, and by p the expected value of the amount resolved with each step. Assuming linear resolution, it would take r/p steps to resolve a total of r. (For convenience, let $k = r/p$)

This conceptualization is precisely consistent with the definition of the NB distribution described above. A sequence of k trials, probability p, and a success event r follows the probability distribution $X \sim NB(r,p), Pr(X = k)$.

Assuming that the amount released per step does not vary significantly over time within a single individual (i.e., p is fixed) and r is a dynamically varying urge, k and r are proportional.

To illustrate this concept in our experiment, the unit duration of our measurement is 3 hours, and r at that point in time can be simply modeled as follows:

$$\beta_0 + \beta_{notification} \times (notification) + \beta_{goal} \times (step\ goal) + \beta_{trend} \times (time\ elapsed)$$

Equation 1.2 Illustrative basic modeling of urge to walk for Negative Binomial model.

This model states that given a notification, urges increase by $\beta_{notification}$, and given an increase in step goal by 1 unit, urges increase by $\beta_{notification}$. Additionally, if there is a long-term trend in base urges (e.g., a gradual increase), the increase per unit time elapsed since the beginning of the trial can be modeled as β_{trend}. This model is arbitrarily assumed, but can be constructed differently depending on the research question, experimental design, and assumptions.

If the model is built as above, it means that the number of steps (i.e., k) has an NB distribution with mean $\frac{r}{p}$ and variance $\frac{r(1-p)}{p^2}$ if the independent variables (i.e., notification, step goal, time elapsed) are constant. Therefore, we can use the average value of the number of steps over a given period ($X \sim NB(r,p)\ with\ E(X) = \frac{r}{p}$) as the dependent variable in the regression.

Zero-Inflated Models

Consider the situation where step count data is modeled as $X \sim NB(r,p), r,p \neq 0$, and the data yields zero steps ($X = 0$). By the definition of NB, $Pr(X = k)$ is defined only for $k \geq r$, because it is impossible to succeed r times without trying r times, so if we observe zeros in the data, we need to find another explanation. However, if we look at real-world data, we see many zeros.

A simple way to model this is to imagine a coin with probability ψ, and assume that if the coin comes up heads, the output of the model has a value of 0, regardless of what the value of $NB(r,p)$ was, and if the coin comes up tails, the model outputs the original value of $NB(r,p)$. For convenience, ψ can be assumed to be a constant, but depending on the assumption, it can also be a variable to be modeled.

In this case, regression methods are named with the prefix zero-inflated, which means that the number of zeros is much higher than expected. In our case, this would be the Zero-Inflated Negative Binomial (ZINB), and similarly, the Zero-Inflated Poisson distribution and Zero-Inflated Geometric distribution are also possible.

$$Pr_{ZINB}(X = k) = \begin{cases} \psi, & k = 0 \\ (1-\psi) Pr_{NB}\,(X = k), & k > 0 \end{cases}$$

With such a model, performing regressions on over-dispersed count data is possible, free from the influence of too many zeros. To describe ψ conceptually, we assume that there is an exogenous barrier to taking a step or an exogenous blockage that prevents a step from being taken but not measured. The total probability of these binomial events occurring is then assumed to be ψ [125,126]. However, we did not hypothesize this term to be significant in our study, so we left it constant and excluded its estimate from the analysis.

Bayesian Regression and Markov Chain Monte Carlo (MCMC)

A detailed mathematical description of Bayesian regression and MCMC is beyond the scope of this thesis. However, since they are central to our research, we briefly introduce the concepts and explain why we adopted them [80,127].

Bayesian regression is a methodology that aims to obtain a probability distribution rather than derive a single, specific value for the parameter to be estimated when performing a regression. It is useful to contrast the concept with frequentist regression, which has a similar goal and similar inputs and outputs. Instead, frequentist regression estimates point estimates.

For example, a conventional frequentist regression for the model $y = \beta_0 + \beta_1 x_1 + \epsilon$ yields one fixed value for each of β_0 and β_1. We also assume that each of these estimates is normally distributed (i.e., has a mean and standard deviation) and calculate a confidence interval and p-value based on our desired significance level. The numerical process for obtaining these values is implementation-dependent but typically utilizes Maximum Likelihood Estimation (MLE).

Bayesian regression aims to obtain a probability distribution for β_0 and β_1. This probability distribution may be normal, but it may not be. Since we do not assume a normal distribution, we can use definite integration to calculate the probability that the estimate falls in any credibility interval we want, which is more informative than a point estimate. This process typically utilizes samples drawn from the stationary state of the Markov Chain, depending on the implementation, to compute multiple times how well the data fits our probability density assumptions and how the assumed probability distribution (called the *prior distribution*) should be updated if we follow what the data tells us. This process results in a new, updated probability distribution (called the *posterior distribution*) obtained through numerous samplings. This process borrows from the principles of Monte Carlo simulation.

As mentioned above, we assumed that step counts follow $ZINB(\psi, \mu, \sigma^2)$ (where μ is the mean and σ^2 is the variance). We also assumed that μ increases significantly given the intervention.

$$\boldsymbol{Steps} = \boldsymbol{ZINB}(\psi, \mu, \sigma^2)$$

ψ: zero-inflation parameter
μ: mean of the steps
σ^2: variance of the steps

$$\mu = (intercept) + \left(effect_{notification}\right) \times (notification) + \left(effect_{goal}\right) \times (goal\ factor) + (effect_{interaction}) \times (notification) \times (goal\ factor) + (trend) \times (elapsed\ time, in\ days)$$

Equation 1.3 Main Regression Model

The resulting probability distribution of each coefficient is called the Posterior Distribution (or *a posteriori* distribution), and the estimated value with the highest probability frequency value (i.e., the most plausible value) is called the Maximum *a posteriori* Point (MAP) (See Figure 1.4).

As the *a priori* distributions of the notification and goal effects, we assigned a non-informative (i.e., no prior knowledge is assumed) distribution, a Cauchy random distribution with the mean of 0 and dispersion parameter of 10,000. Intercept and α can only be positive, so we assumed a Half Cauchy random distribution with the dispersion parameter of 10,000, and ψ can only be between 0 and 1, so we assumed a uniform random distribution in the range [0,1].

Rationale for the Credibility Interval Range

With this methodological foundation, we can estimate the probabilistic distribution of the effect estimate. We decided that 100 steps/3 hours is the minimum value of the MAP intervention effect estimate we would like to see and that an INUS condition is valid only if there is at least an 80% chance of achieving an effect beyond 100 steps/3 hours. These numbers (100 steps/3 hours , 80%) were selected given both the innovative nature of the experiment and a

theorized minimal clinical impact, particularly if multiple states could be identified that could be used repeatedly for an individual to support increases in steps.

Beyond these points and keeping in mind our ultimate goal of informing a future control-system driven intervention, we explicitly sought a credibility interval that is "good enough" to be informative in a future controller-driven intervention. This is based on other work being conducted in the lab (R01CA244777) which involves studying in an RCT, a controller-driven intervention that is using this general COT approach to personalized predictive models for each person at scale. Within that work, we are finding that it can be "good enough" from an intervention perspective, to get estimates of explaining individual variance in steps that are relatively small, in the range of 5-10% of variance of an individual's steps response.

With this, and again, given the innovative nature of our approach and experimental design, we sought to bias towards the detection of plausibly good enough reliability estimates of predictive effects, hence the use of 80% credibility intervals. This was also buttressed by a focus not merely on >0 steps/day estimate (which would be more of a frequentist cut-off) but 80% credibility of at least 100 steps more within the 3-hour increment in comparison to the intervention. Thus, while not the same, this mixed credibility interval with a higher clinical threshold is, arguably, analogous to 5% significance level in frequentist statistics.

This target was based on the recognition that, for some individuals, it might be possible to recognize multiple INUS conditions that could a reliable estimate of effect could be detected. The more INUS conditions that are available to provide "the right support at the right time" the more that 100 steps more can result in meaningful shifts in steps/3 hour increment after a notification, particularly given the general results described earlier that it seems even small increases in movement can have health benefits.

Thus, we used the increment of steps of 100 steps/3 hours (in comparison to when no notification was sent for a given INUS condition) as a base threshold of clinical impact. With that, we do predict the average step increase for each INUS condition and report as such to provide a more robust capacity to interpret the plausible clinical impact of such an intervention. With all of this said, given the innovative nature of the study, we had no clear benchmarking to draw upon to establish these thresholds. Future work could explore other options for this.

Statistical Analysis to Compare INUS Condition Effects

To help understand the logic in more detail, suppose we are given the effect probability distributions corresponding to several INUS conditions, as illustrated in Figure 1.4. The first case is not an effective INUS condition because the MAP estimate is 50 steps/3 hours. The second case has a MAP estimate of 300 steps/3 hours and a 97.72% probability that the effect estimate exceeds 100 steps/3 hours, so it is an effective INUS condition. The third case is a more realistic picture. It has an asymmetric effect probability density function. The interpretation is the same in this case. The MAP estimate is 626 steps/3 hours, and the probability that the effect estimate exceeds 100 steps/3 hours is 100%, so this case is effective. As in the last case, the MAP estimate exceeds 100 steps/3 hours, but the uncertainty is so large that the probability that the effect exceeds 100 steps/3 hours is less than 80%, which is considered invalid. Similarly, the MAP estimate is more important if the probability that the effect estimate exceeds 100 steps/3 hours is greater than 80%. However, if the probability that the effect estimate is greater than 100 steps/3 hours is less than 80%, it does not matter what the MAP estimate, which we treat as an unreliable prediction.

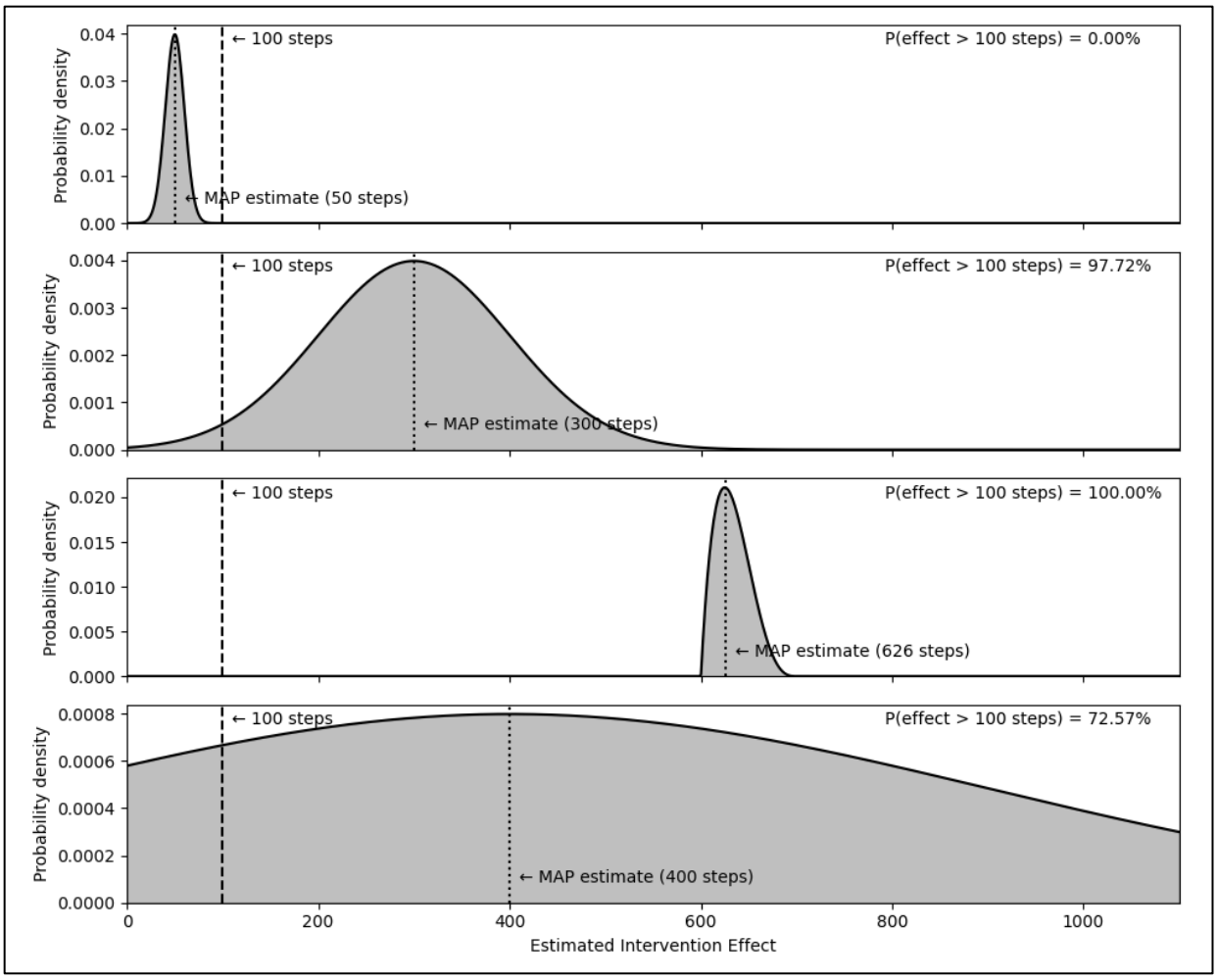


Figure 1.4 An illustrated probability distribution of effect estimated.

Machine Learning Based Regression and Its Benefit

Unlike model-based regression described above, machine learning-based regression only specifies inputs and outputs and does not assume a specific statistical model in advance (e.g., ZINB). Therefore, arbitrary models that describe the relationships in the data, including non-linear models, can be introduced automatically, and model elements that were never expected can be identified[128]. It is helpful for problems challenging to model by hand, such as predicting step counts over time and context at the individual level. In Bayesian regression, to solve this problem, we stratified by the value of each JIT state. ML-based regression, on the other hand, is trained by including each JIT state as an input variable.

Multilayered Perceptron (MLP)

We used the Multilayered Perceptron (MLP) algorithm among the numerous ML algorithms due to its powerful performance (see Appendix 1 for performance comparison). The MLP algorithm is a network model with an input layer that corresponds 1:1 to the input vector, an output layer that corresponds to the output vector (or value), and a few hidden layers in between [128]. Each layer linearly transforms the values of the previous layer with a weight matrix and a bias vector as parameters. The linearly transformed intermediate value vector of each layer typically reflects a non-linear activation function (e.g., hyperbolic tangent [128], sigmoid [129], ReLU [130], or leaky ReLU [130] functions) and is passed to the next layer. Strategies have been developed to make the model more stable by optionally normalizing each layer or introducing regularization, including a drop-out strategy that randomly sets the values of randomly chosen channels to zero [128]. The weights and biases are updated using a backpropagation algorithm [131], which compares the initial output value vector obtained by passing the input data through the initial model with the true output value vector, and adjusts the weights and biases to bring the output closer to the true output value incrementally on a gradient.

Hyperparameter Searching and Optimization

However, the performance of these machine learning models is significantly determined by a large number of hyperparameters (i.e., parameters that determine the structure of the model). See Appendix 1 for the model improvement and its impact. For the MLP described above, these include the number of hidden layers, the size of each hidden layer, the learning speed and strategy during the training process, the type of activation function, the dropout rate and intensity, the initial values of weight and bias, the number of samples per training session, and the performance evaluation metric function.

It is difficult to find the best fit among these hyperparameters manually. An experienced ML scientist can iteratively experiment to find the optimal values. However, since the optimal hyperparameters depend on the input data, it is not practical to do this for many models and input data sets. Therefore, many automated hyperparameter optimization techniques have been developed.

In this study, we used a package called Ray Tune and an implementation of the algorithm called Async Successive Halving (ASHA) [132–135]. These packages and algorithms allow for experiments with different hyperparameter combinations, the sampling of additional hyperparameters near the best-performing combination, and iteration through the experiments to find the optimal combination that performs well enough. ASHA is a widely-used, efficient way to find optimal hyperparameters by training multiple ML models simultaneously and iteratively culling the worst-performing half of the models several times during training.

Parallel and Distributed Computing

Both Bayesian regression and machine learning modeling have enormous computational requirements. In particular, they require exponentially more computational power to achieve more accurate results [128], and parallel and distributed computing is essential to run them effectively.

Semantically and mathematically, changing a program so that the exact computation can be executed on multiple computers is a tedious but challenging engineering task. In this study, the team used a centralized database server to store data, parameters, experimental results, and multiple computers running parallel computing instances, each with its CPU cores. They query the database to see any remaining tasks, perform them if there are, upload the results to the database, and exit. This process is repeated in all instances on all computers until there are no

remaining tasks. Finally, the collected computation results are aggregated on a separate computer for visualization and statistical processing.

Converting Conceptual Terminologies to Operationalized Terms

In order to connect the conceptual discussion in Chapter 1 to the operationalization level starting in Chapter 2, we decided that it was necessary to translate expressions and concepts into everyday terms. For example, terms such as INUS condition and factor are great for establishing theoretical concepts. Still, they can introduce some additional difficulties during interpretation, so we replaced them with everyday terms such as *state* (corresponding 1:1 to INUS condition) and *decision policy* (corresponding 1:1 to INUS condition according to a combinatorial condition based on the psychological conditions of Need, Opportunity, and Receptivity). The *time condition* will now refer to INUS factors related to the time of day, such as weekday, weekend, morning, and afternoon.

The term *effective* will refer to the case that the intervention's effect (i.e., increment of the step count during 3 hours after each decision points) posits higher than the threshold (i.e., MAP value of the effect exceeds 100 steps/3 hours AND the probability of effect exceeding 100 steps/3 hours is over 80%). There would be rare cases that we utilize the frequentist counterpart, where the term *effective* refers to effect estimates over 100 steps/3 hours with p-value less than 0.05.

This translation work aims to make it easier for a broader audience to understand the rest of this thesis without requiring a deep conceptual understanding of these terms. For technical clarity, we will continue to use specific expressions and terms for methodological use.

Aims and Hypothesis[4]

Aim 1 (A1).

To identify which decision policy and time condition where the intervention was effective for each participant, analyzed via both nomothetic and idiographic models.

Hypothesis 1 (H1). Among 16 states (4 decision policies × 4 time conditions), there will not be a single state that is commonly effective for the majority (>50%) of participants.

Rationale. The aim is to answer the central research question that underlies this thesis: "*Do individuals respond differently to interventions based on time conditions and decision policies?*" We will first utilize a traditional method, a mixed effects model, to see if we can find significant common patterns across the entire population (i.e., a nomothetic model). Then, we will use Bayesian regression on a subset of the data stratified by each participant and state (i.e., a combination of time condition and decision policy) to take a closer look at individual response patterns (i.e., an idiographic model).

[4] Since aims and hypothesis section is critically important, the paragraphs are edited as spacious to improve readability.

Aim 2 (A2).

To identify similarities in response patterns across participants within the same time condition.

Hypothesis 2-1 (H2-1). For each of 4 time conditions, participants will not be *evenly distributed* across 16 ($=2^4$) possible combinations with respect to dichotomized effect patterns for 4 decision policies; it should be statistically tested by comparing the participants' actual pattern distribution and uniform distribution.

Hypothesis 2-2 (H2-2). Hypothesis 2-2 (H2-2). For each of 4 time conditions, participants will form significantly fewer numbers of dichotomized effect patterns across 4 decision policies as indication of a different types of participants responding similarly, as tested using Monte Carlo (MC) simulations; it is tested by comparing the numbers of patterns of the actual participants and against patterns produced by MC simulations, designed to produce a random distribution. A significant effect ($p<.05$) is indicative that the actual data from participants includes clusters between participants that is likely not due to pure chance.

Rationale. This aim is to test the hypothesis that the characteristics of individuals found in aim 1 are influenced by differences between people, with the possibility of people clustering together. If we could find the clusters of people, it would be beneficial when we needed to figure out to which cluster a newly enrolled individual belongs, even before we identify the individual's pattern. This will expedite the optimization process, sparing some time to check based on a priori probability density, benefitting from a useful probability-updating framework of Bayesian modeling.

Exploratory Aim 1 (EA1).

To explore the potential value of the multilayer perceptron algorithm in identifying which decision policy and time condition where the intervention was effective for each participant.

Hypothesis 1 for Exploratory Aim 1 (EH1-1). The MLP model of each participant, with its hyperparameters optimized by ASHA, that are trained by first *P*% of decision points (i.e., time condition, decision policy, step goals, day elapsed as input features and steps during 3 hours as outcome) will be able to predict the outcome with significantly small MSE than the null model (i.e., frequentist's linear regression and ZINB models) at the significance level 0.05. The rest of the data (i.e., later portion of $(100 - P)\%$) will be used as test dataset. P = 30, 40, 50, 60, and 70.

Rationale. This aim revisits our core research question, answered in aim 1: "*Do individuals respond differently to interventions based on time conditions and decision policies?*" What conclusions can be drawn using other methodologies, particularly machine learning ones that specialize in detecting irregular, idiographic nonlinearities? How do the results differ from those based on models? If they are consistent, it would increase the importance of that information; if they differ, it is important to see how each can be used, with a more careful understanding of their underlying assumptions, which could provide greater insights on confidence of claims made between each approach and could help guide further theorizing about human behavior (see discussion).

Exploratory Aim 2 (EA2).

To examine the relationship between real time and post hoc states, and the impact on identified individual response patterns on the notifications per the time condition on aim 1.

Organization and transparency note on exploratory aim 2 (EA2): this aim was added and more specified after the trial is finished. For details, see the section on technological limitation on page 29 and the section on how to specifically define the analysis problem on page 125. Also note, the reason why the problem definition of the EA2 is described in Chapter 4 (see page 125), unlike those for the other aims described in this chapter, was that the problem definition of EA2 needs the highly specific details about operationalization of the trials described in Chapter 2 and Chapter 3.

Given the highly exploratory nature of Exploratory Aim 2, we did not create a falsifiable hypothesis. Instead, as we did in the aim 1 hypothesis 1, we will visualize the idiographic patterns across all combinations of post hoc JIT states and timing.

Acknowledgments

Chapter 1, 2, 4, 5, 6, and 7 of this thesis, in part, are currently being prepared for submission for publication of the material. Park, Junghwan; Kim, Meelim; El Mistiri, Mohamed; Kha, Rachael; Banerjee, Sarasij; Gotzian, Lisa; Chevance, Guillaume; Rivera, Daniel E.; Klasnja, Predrag; Hekler, Eric. The dissertation author was the primary researcher and author of this material.

Chapter 2 STUDY DESIGNS AND CONSIDERATIONS

This section outlines the considerations and scientific design intent of the most important parts of protocols before diving into a full-fledged clinical trial protocol (see CLINICAL TRIAL PROTOCOL on page 63). As such, this section only covers some details; the next section will detail the rest.

Prior Studies

JustWalk 1 (2018)

One of our group's initial studies [136] aimed to investigate how a smartphone-based intervention with dynamic step goals and rewards impacts step counts. This study was based on Social Cognitive Theory, similar to the approach adopted in this dissertation research. The results indicated that participants increased their daily step count by 2,650 steps using a linear mixed model. The peak effects were observed during the third cycle (i.e., days 33-48) when employing mixed-effects modeling with a quadratic term fitted using maximum likelihood. While this study initially assumed and analyzed a nomothetic approach, it was one of the early studies to present individualized results on interventional experimentation and was the initial pilot study that informed the proposed dissertation study. There are key differences between the prior study and the one which this thesis is focusing on. Specifically, the prior study included only administration of a single unilateral intervention from researchers to participants (in the dissertation study, we including both daily step goals, modeled after this study, along with provision of notifications within day); it utilized only daily timescale analyses (not within day, which is a key focus of this dissertation); and the prior study provided daily step goals and points as the main interventions (in this, we did not include the points but did add within-day walking suggestions).

HeartSteps 1 (2019)

Our within-day walking suggestions was inspired by another prior study conducted by the group, Klasnja et al. [137]. It was the original study that created the technology platform (HeartSteps) that we adapted to enable this experiment to be conducted. This study along with [136], were highly influential studies for this dissertation, as it established the basic framework for sending notifications to encourage walking at variable frequencies. It was conducted as a Microrandomized Trial (MRT) and found that sending notifications increased the total number of steps participants took in 30 minutes from the decision point by 14%. In addition, the data from this study had the greatest impact on our team's ability to visualize and understand how individuals respond differently to interventions. We also used data from this study to develop the Opportunity JIT state [103] ([103] is reprinted as Appendix 1 on page 188).

Modeling Individual Differences (2018)

Phatak et al. [3] conducted a secondary analysis on [136] of idiographic dynamical systems modeling (the approach to data analysis most commonly used in system identification experiment), to study individual differences using the same data described already [136]. In this intervention study, they identified the factors necessary to predict the level of response to an intervention and noted that they varied considerably between individuals. Among the several differences between this study and the dissertation research, the most important is that we viewed "factors that do not explain behavior" as necessary signals as well. Tagging a person with a single set of factors that best explains their behavior, while useful for clustering purposes, makes it difficult for controllers to operate with a holistic view, and it also makes it difficult to create strategies that are clustered by context. Therefore, the dissertation research generated profiles of effects across contexts and decision policies for each individual, and leveraged these.

Absorbing Nomothetic Pattern using Machine Learning (2019)

The paper [138] is a good reference when considering the analysis method of this study. It compares nomothetic and ideographic approaches by utilizing machine-learning algorithms to understand predictors of stress. In the nomothetic approach, the gaps or similarities between patterns of individuals are modeled by training *the entire data at once* rather than using algorithms that specifically allow for individual differences to be observed first (such as is proposed in this dissertation). This approach has the advantage of being somewhat simpler to write programs for when training machine learning algorithms. However, if we develop a single pattern that absorbs all the data rather than forming patterns for each individual, we will have limited knowledge of what patterns an individual has. One would need to apply the entire model every time, even if the initial data for a particular individual makes it clear which model is appropriate. This is not only computationally expensive but also introduces a bias into the future predictions that will inevitably carry with it the bias of the entire model. (If we knew more about an Individual's model, the bias would be reduced to that Individual's model.) This study is not from our research group, but it was an important inspiration for us to organize this study.

Designing the Intervention

Notification and Decision Points

The focus of this dissertation is the notifications provided through the app, which are sent in the hope that the participant will go for a walk in the next three hours. These notifications can be sent up to four times, three hours apart, starting at the time the participant specified at onboarding (e.g., What time does your day start?).

Notifications were not sent all the time; they were determined by several algorithms that will be described later. Some notifications were sent randomly, regardless of the participant's behavior (Random decision policy), some are sent based on the participant's behavior only (N+O

decision policy), and some are sent based on the participant's behavior and response to the intervention, and the history of the intervention (Full and N+R decision policies).

Messaging

When the notification is sent, one of 50 messages is selected from two categories. These two categories are (1) an invitation for a person to create a micro-implementation intention on when, where, and how to fit in a ≥10-minute walk in the next 3 hours—we label these suggestions as *bout planning* (n=24, 48%)—and (2) an invitation for a person to become aware of interoceptive experiences and signals (e.g., stiff muscles and lethargy) that may inspire them to go for a walk—we refer to these messages as *cultivating an urge* (n=26, 52%). To prevent an individual from receiving repetitive messages, we shuffled the messages in advance and sent them to each individual in a round-robin style. When all 50 messages are exhausted, we started again at number 1. The study server kept track of the last message number a particular individual received. See Supplemental Table 3.4 for a complete list of messages.

Time Condition

As described earlier, each individual has a total of four decision points per day. These were grouped into mornings and afternoons, with the first two decision points (typically 7 am and 10 am) categorized as "morning" and the last two decision points (typically 1 pm and 4 pm) categorized as "afternoon." We also categorized Monday through Friday as "weekdays" and Saturday and Sunday as "weekends" since the intervention takes place daily. Following this categorization, all decision points are categorized into four time conditions: "weekday morning," "weekday afternoon," "weekend morning," and "weekend afternoon."

Daily Step Goals

As an additional intervention element, we assigned daily changing step goals. Step goals for the first 26 days were determined based on data from the baseline period (the first 10 days of

the trial before the intervention began). Step targets were determined using the median number of steps taken during the baseline period as the minimum value, adding a pseudorandom sequence ranging from 0 to 1 (i.e., "goal factor") multiplied by 4000 steps (See page 88 for details). This process was repeated every 26 days for the duration of the intervention, with the N-th 26-day cycle determined by using the median of daily step counts from the (N-1)-th 26-day cycle. If the median value for a particular cycle was less than 2000, we set the minimum value to 2000 and the maximum value to 6000; if the median value was greater than 8000, we set the minimum value to 8000 and the maximum value to 12000.

Based on this design of daily varying step goals, the daily step goals are highly correlated to the current level of physical activity. In the analysis, we used the goal factor as predictor, instead of the raw goal, because the goal factor was the actual means for the experimental manipulations. The interpretation of the goal factor is the likely difficulty of the goal for a person. See Chapter 4 for more details.

Experimentation and Operationalization

Just-in-Time states: Need, Opportunity, and Receptivity

The three most important Just-in-Time states representing dynamic psychological constructs in this study are need, opportunity, and receptivity. Notably, they were all assessed four times a day in real-time, so the concept of "right now" is included in the definition, even if not labeled separately. They were used to jointly form decision policies, as described in detail below.

Need

The concept of need is more precisely "the individual's perceived need for intervention." To estimate need, we calculated the number of steps taken to date relative to a given step goal. The given step goal is included in the denominator because it is a "doable but ambitious"

behavioral goal for the day, a kind of agreement between the individual and the intervention that if one reach this amount, they have exercised enough. Since it is not a self-set goal but a given goal, it falls under goal-setting (1.1) – "agree on ..." in the BCT taxonomy [41].

This step goal was calculated as a "prorated goal" by spreading it linearly over the day (which we viewed as an operational definition of a day as the 12 hours between 7 am-7 pm covered by the four decision points). For example, at 10 am, 3 hours had passed since the start of the day at 7 am, so ¼ of the day had passed, and we evaluated whether we had achieved ¼ of our daily step goal for the day. If the daily step goal for the day is 8000 steps/day, and it is currently 10 am, we have evaluated whether the participant has taken 2000 steps from that morning to now. If the participant has taken more than 2000 steps, we can rate their progress as "good today" at this time point, and their (perceived) need is assumed as false. If we rate their progress as "not good enough," i.e., achievement up to that time point is less than the prorated goal; then the perceived need is assumed as true, meaning that external support (i.e., intervention) may be useful.

Since the prorated goal does not apply to the first decision point of each day, we assessed the need by comparing the total number of steps to the total daily step goal for the previous day.

Opportunity

Opportunity is the answer to the temporally meaningful question, "Is this an opportunity to go for a walk?". Qualitatively, it asks, "Is the likelihood of a walk based on historical records not too low, not too high, but just right, so that if we support it, the likelihood of a walk will be significantly increased within a range?" In the end, it is the construct that plays the most important role in terms of the core pillar of JITAI, "Just-in-Time".

To operationalize this concept, we conducted two phases of experiments. We first developed a model incorporating domain knowledge and a linear model and applied it in a

clinical trial. Then, as a second phase, this model was further developed in the paper in Appendix 1. However, since the output of the second phase was not directly used in the clinical trial, it was not included in the main chapter but placed in the Appendix. The specific operationalization of the first model used in the clinical trial can be found in Appendix 4.

Receptivity

Receptivity is defined by answering two questions: "Did this participant meaningfully respond to the notifications we sent in the last 24 hours?" and "Did we send too many notifications?". The technical definition considers three dichotomous variables as follows:

1. **R1**: Whether or not a notification was sent in the immediately preceding decision point

2. **R2**: Whether the message was sent more than 6 times in the last 3 days, with a threshold of 71 hours to safely exclude cases where the message was sent exactly 72 hours ago.

3. **R3**: Whether any of the notifications sent in the last 24 hours had more than 60 steps taken in a row for more than 5 minutes within 3 hours after the notification was sent.

Utilizing the above three dichotomous variables, Receptivity is positive only in the following two cases:

A. **R1 = False** and **R2 = True** and **R3 = True**

B. **R1 = False** and **R2 = False**

Study Designs to Detect Dynamics of Operationalized Psychological Constructs

The operationalized psychological constructs defined above were assumed to change dynamically on a time axis. One key to the trial was using these changing constructs to make real-time decisions about whether to send or not to send notifications. To do so, it was essential to fetch data from each participant once at short intervals (e.g., every 15 minutes) via Fitbit Inc.'s servers.

Data were automatically uploaded for many participants every 15 minutes as long as the phone was sufficiently charged. The upload process happened without opening the Fitbit app or doing anything else. Data for some participants were intermittently paused uploading for unknown reasons, which had a significant impact on JIT state estimation, which requires utilizing quasi-real-time data (see page 29 for details). To minimize this impact, at each decision point, we checked when the most recent data upload was, and if there was a data gap of more than an hour and the same text message had not been sent within 24 hours, we sent a text message to open the Fitbit app. If the participant opens up the Fitbit app, the data upload was resumed without exceptions.

Study Design to Detect Idiographic Effects

Almost every aspect of this study was designed to detect variation between individuals. In particular, all individuals were to have their data collected with minimal efforts, and activity suggestion notifications were provided individually so that their idiographic response patterns to the varying interventions, over decision policies and time conditions, could be fully captured. Additionally, each individual was given a personalized daily step goal, which further allowed us to see how each individual responded to these intervention components. Specifically, the daily step goal was provided as a random value within a 4,000-step range, with each person's past median step count as the minimum value. Thus, no two people were assigned the same step goal on a single day during the entire study, and each person was assigned a step goal that was purely proportional to their past activity level. The decision policy values were also designed to have a ratio of Full:N+O:N+R:Random of 2:1:1:1, but the order and placement were all personalized. Before the study, 50 pseudorandom sequences were generated, stored on the server, and personalized by "issuing" one to each newly enrolled participant.

System Identification Experimental Design

Decision Policies

In order to create four decision policies of potential intervention strategies, three JIT states (i.e., need, opportunity, and receptivity) were utilized. Each decision policy was assigned to an individual in a different order and daily. All but the random decision policy are algorithms that depend deterministically on the JIT state. Each decision policy was designed with an intention to be used out-of-the-box for future control system design.

Random Decision Policy

This decision policy is a day where we randomly choose to send or not send notifications via coin flip (50:50 chance) without considering any construct. This decision policy was pseudorandomly assigned to 20% of all days.

N+O Decision Policy

Only Need and Opportunity states are considered. If both Need and Opportunity are met (see page 53 for details), a notification is sent; if either Need or Opportunity is not met, no notification is sent. Receptivity is not considered. This decision policy was pseudorandomly assigned on 20% of all days.

N+R Decision Policy

Only Need and Receptivity states are considered. If both Need and Receptivity are met, a notification is sent; if either Need or Receptivity is not, no notification is sent. Opportunity is not considered. This decision policy was pseudorandomly assigned to 20% of all days.

Full Decision policy

Need, Opportunity, and Receptivity states are considered. Notifications were sent when all three states were positive, and no notifications were sent when any of the three states were not met. This decision policy was pseudorandomly assigned to 40% of all days.

Application of Three Decision Policies and Three JIT States

We assumed, as an *a priori* design principle, the Need JIT state was necessary at all circumstances. Given the operationalization of the Need JIT state, if the participant had already had enough level of activity up to that point, we expected the chances of being active is low. Hence, we decided not to bother the participant. Opportunity state was used selectively; in N+R decision policy, we excluded the Opportunity state from consideration. Receptivity state was also used selectively; in N+O decision policy, we excluded the Receptivity state from consideration.

Signal Designs

As mentioned, the input signals used for System Identification must be constructed to satisfy certain mathematical properties (e.g., frequency characteristics; see page 84 for details). Proper consideration of frequency characteristics means, in behavioral science terms, that intervention periods and gaps of varying lengths should be varied so that *maintenance* after the intervention paused or *duration effects* (i.e., association between the intervention effect and the duration of intervention) can be examined. For categorical variables such as decision policy, it also means that the order in which they are alternated should have a sufficiently flat frequency distribution and probability distribution. In order to assign these carefully designed signals, we operationalized them by using pre-generated pseudorandom signals to assign to individuals, as similar to the method we used for the step goals.

Identification of States

Based on the concepts and design described above, in addition to our prior studies and domain knowledge, we identified the following states to examine individual response pattern to the notifications (See Table 2.1).

Table 2.1 The matrix of identified states.

Decision Policies / Time Conditions	Full	N+O	N+R	Random
Weekday Morning	State 1	State 2	State 3	State 4
Weekday Afternoon	State 5	State 6	State 7	State 8
Weekend Morning	State 9	State 10	State 11	State 12
Weekend Afternoon	State 13	State 14	State 15	State 16

We expected some people may react better (i.e., engage in more activity) when they have an opportunity to walk, whereas others may react comparably good enough (or even better) when they did not have an opportunity to walk. In this particular case, removing the Opportunity state is helpful. The assumption supporting this decision was partly influenced by JustWalk 1 study (see page 49); we could find people responding to the stress differently [136]. To test multiple realistic intervention strategies simultaneously, we designed three decision policies (N+O+R, N+O, N+R) that utilizes the subsets of the states.

It is Important to note that we experimented partial decision policies (N+O, N+R), in addition to full decision policy (N+O+R, or Full), to estimate the relative value of considering Receptivity. This is a natural design that we expected Receptivity may act as positive factor to the effect. Thus, instead of experimenting N+O+R^- decision policy (i.e., positive need and opportunity and negative Receptivity), we tested N+O. However, in a scientific perspective, it is still valuable to examine an exhaustive set of parameter configurations (e.g., $\underline{\text{N+O+R}^+}$ vs. $\underline{\text{N+O+R}^-}$ vs. …). Thus, with the exploratory aim 2, we conducted a *post hoc* analysis about this matter. Please refer to the following section and "Post Hoc Analysis of JIT states" section in page 162.

Conceptual Model for Exploratory Aim 2

Since the behavior data can be eventually available from minutes to days *after* passing the decision point, it is possible to reconstruct what the JIT state *was* at the time based on this

post hoc data analysis. From now on, we will refer to JIT states estimated using only the data at the time of the decision point as *real-time* JIT states, and JIT states estimated using the full data after the trial has ended as *post hoc* JIT states.

For example, if step information is missing because the Fitbit and smartphone are not sharing data with the server, Need may be represented as True in real-time (i.e., you have not already walked as much as you need) even though it was False in post hoc (i.e., you have walked as much as you need). Opportunity is determined by summing the frequency of steps in similar situations in the past, so a case that could be estimated as True ("I have an opportunity to walk at this time") if step data were available may be estimated as False ("I don't have an opportunity to walk at this time") in real-time. In the case of Receptivity, it is possible that the person actually took a walk since the preceding notification, so that even if post hoc Receptivity was True, it would be False in real-time because no step data was available.

These possibilities can also work in opposite directions. For example, Opportunity is estimated to be False if the probability of walking at that time is too high (>80%). Thus, a post hoc JIT state can be False, but become True in real-time as data is delayed. See Appendix 6 for an analysis of the degree of misalignment between real-time and post-hoc JIT states due to this delay. In this section, we examine how the response patterns of individuals identified using post hoc JIT states differ from those identified using real-time JIT states.

To examine this, we pursued to examine the response patterns stratified by individual, decision policy, and time condition using Bayesian Regression, similar to the idiographic full model in Aim 1, but with a subtly different settings from Aim 1. In Aim 1, real-time JIT states were notified deterministically, given a Decision policy. For example, a notification was always sent on a day with an N+O decision policy, if both real-time Need and real-time Opportunity

were True. However, in terms of *post hoc* JIT states, a notification could have been sent on a day with an N+O decision policy even if post hoc Need or Opportunity was False, or vice versa.

Thus, the addition of *post hoc* JIT states allows us to test more diverse analytical model (see Table 2.2). However, we are most curious about which strategy is most appropriate for each individual to choose based on the time condition context, i.e., how they respond to each decision policy.

As similar to aim 1, we stratified the data by participant, and time condition. Then, to build progressive models, we iteratively took subset of each stratified dataset, as shown in Table below.

This analysis is valuable because it shows the ideal results of how participants would have responded differently if we had theoretically perfect data (i.e., if the sync delay never existed). This idealized result is challenging to achieve as long as the sync delay exists. However, future technological advances that reduce the sync delay suggest that the results shift closer to this ideal.

Acknowledgments

Chapter 1, 2, 4, 5, 6, and 7 of this thesis, in part, are currently being prepared for submission for publication of the material. Park, Junghwan; Kim, Meelim; El Mistiri, Mohamed; Kha, Rachael; Banerjee, Sarasij; Gotzian, Lisa; Chevance, Guillaume; Rivera, Daniel E.; Klasnja, Predrag; Hekler, Eric. The dissertation author was the primary researcher and author of this material.

Table 2.2 Exhaustive combination of Need, Opportunity, Receptivity sub-setting conditions exploratory aim 2

Case	Need	Opportunity	Receptivity	Similarity with real-time decision policy
1	Controlled	Controlled	Controlled	Random
2	Controlled	Controlled	Positive	
3	Controlled	Controlled	Negative	
4	Controlled	Positive	Controlled	
5	Controlled	Positive	Positive	
6	Controlled	Positive	Negative	
7	Controlled	Negative	Controlled	
8	Controlled	Negative	Positive	
9	Controlled	Negative	Negative	
10	Positive	Controlled	Controlled	
11	Positive	Controlled	Positive	N+R
12	Positive	Controlled	Negative	
13	Positive	Positive	Controlled	N+O
14	Positive	Positive	Positive	Full
15	Positive	Positive	Negative	
16	Positive	Negative	Controlled	
17	Positive	Negative	Positive	
18	Positive	Negative	Negative	
19	Negative	Controlled	Controlled	
20	Negative	Controlled	Positive	
21	Negative	Controlled	Negative	
22	Negative	Positive	Controlled	
23	Negative	Positive	Positive	
24	Negative	Positive	Negative	
25	Negative	Negative	Controlled	
26	Negative	Negative	Positive	
27	Negative	Negative	Negative	

Chapter 3 CLINICAL TRIAL PROTOCOL

NOTE: This is a direct copy of our previously published protocol paper. The core details needed to interpret the data analyses and results were already reviewed in chapters 1 and 2. This protocol paper is included for completeness of the overall work and to provide potential additional desired details for a comprehensive understanding of the overall study.

Title: Advancing Understanding of Just-in-Time States for Supporting Physical Activity (Project JustWalk JITAI): Protocol for a System ID Study of Just-in-Time Adaptive Interventions.

Abstract

Background

Just-in-time adaptive interventions (JITAIs) are designed to provide support when individuals are receptive and can respond beneficially to the prompt. The notion of a just-in-time (JIT) state is critical for JITAIs. To date, JIT states have been formulated either in a largely data-driven way or based on theory alone. There is a need for an approach that enables rigorous theory testing and optimization of the JIT state concept.

Objective

The purpose of this system ID experiment was to investigate JIT states empirically and enable the empirical optimization of a JITAI intended to increase physical activity (steps/d).

Methods

We recruited physically inactive English-speaking adults aged ≥25 years who owned smartphones. Participants wore a Fitbit Versa 3 and used the study app for 270 days.
The *JustWalk JITAI* project uses system ID methods to study JIT states. Specifically, provision of support systematically varied across different theoretically plausible operationalizations of JIT states to enable a more rigorous and systematic study of the concept. We experimentally varied

2 intervention components: notifications delivered up to 4 times per day designed to increase a person's steps within the next 3 hours and suggested daily step goals. Notifications to walk were experimentally provided across varied operationalizations of JIT states accounting for need (i.e., whether daily step goals were previously met or not), opportunity (i.e., whether the next 3 h were a time window during which a person had previously walked), and receptivity (i.e., a person previously walked after receiving notifications). Suggested daily step goals varied systematically within a range related to a person's baseline level of steps per day (e.g., 4000) until they met clinically meaningful targets (e.g., averaging 8000 steps/d as the lower threshold across a cycle). A series of system ID estimation approaches will be used to analyze the data and obtain control-oriented dynamical models to study JIT states. The estimated models from all approaches will be contrasted, with the ultimate goal of guiding rigorous, replicable, empirical formulation and study of JIT states to inform a future JITAI.

Results

As is common in system ID, we conducted a series of simulation studies to formulate the experiment. The results of our simulation studies illustrated the plausibility of this approach for generating informative and unique data for studying JIT states. The study began enrolling participants in June 2022, with a final enrollment of 48 participants. Data collection concluded in April 2023. Upon completion of the analyses, the results of this study are expected to be submitted for publication in the fourth quarter of 2023.

Conclusions

This study will be the first empirical investigation of JIT states that uses system ID methods to inform the optimization of a scalable JITAI for physical activity.

Trial Registration

ClinicalTrials.gov NCT05273437; https://clinicaltrials.gov/ct2/show/NCT05273437

International Registered Report Identifier (IRRID):DERR1-10.2196/52161

Introduction

Background

There is great interest in the promise of just-in-time adaptive interventions (JITAIs) to support behavioral medicine. A JITAI is a behavioral intervention that is designed to (1) provide interventions and support during *just-in-time (JIT) states*, defined as times when a person would have a *need for support*, an *opportunity to act* in accordance with the support, and *be receptive* to support [1], and (2) adapt over time to a person's changing needs with the use of adaptation algorithms that strive toward enabling a person to meet clinically meaningful behavioral targets (e.g., national recommendations for a given behavior) while accounting for the person's current capabilities and constraints. Although there is a lot of interest in this type of intervention, more work is needed to advance the understanding of the foundational concepts implied by JITAIs, particularly the JIT state. The JIT state concept is inherently context dependent, dynamic, and likely to manifest differently for different people over time. Given this complexity, much of the work on JITAIs has focused on either creating interventions that are theory driven in terms of specifying JIT states according to a priori decision rules or through more data-driven approaches such as reinforcement learning. An important gap is the lack of a conceptual understanding of JIT states, which could be achieved by conducting rigorous theory-testing protocols designed to test dynamic hypotheses about JIT states.

The purpose of this paper is to describe a research protocol for a National Institutes of Health–funded Smart and Connected Health study (R01LM013107) explicitly designed to produce rigorous empirical evidence to study JIT states in the context of a physical activity (PA) JITAI. The structure of the paper is as follows. First, background information is provided about JITAIs that is necessary to understand the motivation for our system ID protocol. Next, a

description of the system ID experimental protocol is provided, including the specific goals of the project, experimental design procedures, measurement approach, and analysis plan. Finally, a discussion and the Implications of this work are offered in terms of future research on JITAIs.

Improving Understanding of *Just-in-Time States* Within a Digital Health PA Intervention

There is convincing evidence indicating that PA is valuable for reducing the risk of colon, breast, endometrial, lung, and pancreatic cancers [139,140] and cardiovascular disease [8] and improving glycemic control [10]. With an aging population, step interventions could help prevent chronic diseases, reduce health care costs, and improve functional life years and quality of life [8,10,13–17,19,139–147]. The clinical guidelines for steps suggest 8000 steps per day for adults [24,26], but only one-third of this group meets the guidelines [27–37]. Across PA interventions for adults (e.g., human-delivered and digital), results show increases of 496 steps per day achieved above baseline levels of 5000 steps per day, and even high-impact interventions peak at 1363 steps per day above baseline; both result in activity that is still below the guidelines. Even among interventions that produce an effect, maintenance is rarely measured, and when it is, it is not achieved by many participants [75,148–150]. Our long-term goal is to create a model-predictive controller-driven JITAI to increase walking that, we hypothesize, will be more effective than current PA interventions at supporting individuals in achieving and maintaining national guideline recommendations of at least 8000 steps per day averaged across a week [24,26].

Although there are many possible algorithmic approaches to achieve this, such as reinforcement learning [66] or recommender systems [151], this research effort is focused on the use of a model-predictive controller approach [2]. A model-predictive controller is an adaptation algorithm that uses time-series data from an N-of-1 unit [152,153] to support decisions over time

in dynamic, often complex situations, such as dynamically providing support to a person to increase their PA. As the name implies, a central feature is a computational dynamical model, which is a series of mathematical equations that encode previous domain knowledge and include parameters that are estimated from data derived from each N-of-1 unit (i.e., a person in this context), thus enabling the controller to account for individual differences in predictions. These computational models, much like weather or climate forecasting models, enable rigorous simulations of a person's likely responses to different types of support provided both now and in the future. For example, the model could be used to simulate a person's response to the provision of a notification meant to nudge them to walk within the next 3 hours. The model would generate predictions on the likelihood that a person will walk after receiving the notification at each moment. In addition, the model can be used to simulate the potential synergistic or antagonistic effects that might occur because of different decisions that could be made. For example, using the model, predictions could be made on the potential diminishing effects of providing notifications over time owing to habituation or growing annoyance, particularly if notifications are sent when a person does not need them.

As this description implies, model-predictive controller-driven JITAIs are complex and, thus, are difficult to create using theory alone, which, historically, was the dominant way in which adaptive behavioral interventions were developed [154,155]. Instead, JITAIs require robust experimentation that enables empirical optimization of their elements, particularly the generation of the computational dynamical models that the controller uses to run simulations and, by extension, make dynamic decisions. As described in previous work [3], the empirical estimation and validation of dynamical models occur through system ID.

The system ID study described in this paper had 2 complementary but distinct aims. First, it aimed to gather empirical evidence on the concept of JIT states. By varying whether a notification is provided when the person is thought to be in a state of need, when they have an opportunity to walk, when they are thought to be receptive, or combinations of these 3 states, the experiment collected initial evidence for which aspects of the JIT state are most important for supporting the effectiveness of JIT interventions and whether this changes over time.

Second, the experiment was designed to collect the data needed to optimize a digital health intervention, *JustWalk JITAI*. The goal was to estimate and validate dynamical models that can be used to construct a model-predictive controller that can make decisions on the provision of support in given moments to achieve and sustain clinically meaningful PA targets. Prior work was used as a foundation to achieve these aims, particularly a dynamic model of social cognitive theory (SCT) that encapsulates domain knowledge about behavioral processes that influence PA [40,74,156]. The SCT models were refined using the newly collected data both to help us better understand JIT states and to develop models that can be incorporated into a multitimescale model-predictive controller.

Methods

Overview

Aim

The aim of this study was to conduct a system ID experiment to empirically assess the conceptual elements of a JIT state and estimate and validate dynamical computational models relevant to JIT states. This work was conducted to inform the development of a future model-predictive controller-driven JITAI. We had three broad hypotheses: (1) walking bout planning

prompts that are provided when the system determines that individuals meet all 3 conditions[5] of a JIT state—have a need, have an opportunity to walk, and are receptive to intervention notifications—will be more effective than when such prompts are provided when only some or none of those conditions are met, (2) idiographic computational models (i.e., models developed by and for individual participants) can be produced that are effective at predicting contexts in which suggestions to go for a walk will be effective and how such suggestions and adaptive step goals combine to support a person in achieving both daily step goals and sustained engagement in steps per day, and (3) nomothetic analyses (i.e., insights gleaned from data aggregated across participants) will reveal meaningful clusters for different types of contextual patterns and trajectories of change across participants. These clusters will enable the selection of initial dynamical model parameters and, by extension, the development of a generic semiphysical model that can be used as a starting point for new participants in a future model-predictive controller-driven JITAI. In aggregate, these results will also be used to empirically test the added value of previous domain knowledge, as encapsulated in previous computational models, for improving model prediction and response, with a basic autoregressive model with external input as a reference model that only accounts for previous domain knowledge in the form of variable selection but not the structure of their relationships.

Study Design Overview

Building on prior work, including the mobile health app *HeartSteps* (which was relabeled *JustWalk JITAI* for this study to continue on the control systems side of JITAI development) [137,157], we conducted a system ID experiment designed to study the theoretical concept of JIT states as a tool for fostering behavior change. The system ID experiment focused on two key

[5] In other sections of this thesis, the term "condition" (if it is used with need, opportunity, or receptivity) is replaced with "states" or "JIT states", depending on the contexts. However, since this chapter is a full reprint of a manuscript that is already published, we decided to keep the original terminologies. Apologies for the confusion.

intervention components: (1) notifications delivered up to 4 times per day designed to increase a person's steps within the next 3 hours via either increased awareness of the urge to walk or via bout planning and (2) adaptive daily step goals. Both types of notifications prompting short walks within the next 3 hours were experimentally provided or not across variations of *need* (i.e., whether daily step goals were previously met), *opportunity* (i.e., the next 3 h are a time window when a person has an opportunity to walk based on their previous step data), and *receptivity* (i.e., the person received <6 messages in the last 72 h and walked after notifications were sent). In addition, the suggested daily step goals also varied systematically across time rooted in a person's baseline levels of steps per day (e.g., 4000 steps) and gradually increasing until they met clinically meaningful targets (at least 8000 steps/d on average). Participants wore a Fitbit Versa 3 and used the study app for 270 days.

Technology

Wearable Sensor

The Fitbit Versa 3 is a wrist-worn, watch-style activity tracker that records participants' steps and minutes of moderate or vigorous PA (*active minutes* in the language used by Fitbit) that the tracker detects based on accelerometer and heart rate data. The Fitbit tracker records the step and activity data, automatically synchronizes with the Fitbit server, and pushes to the *JustWalk JITAI* servers using Fitbit's subscription application programming interface (API). It was recommended to participants to set the Fitbit to use one of the market-available watch faces with the following features: (1) always visible information about the current step count, daily step goal, and progress toward meeting the goal and (2) positive reinforcement (in the form of a fireworks display and vibrations) when the daily step goal is met. The list of watch faces that met these requirements was provided by the staff.

Mobile App

The *JustWalk JITAI* app contained (1) *pull* components that participants could access at any time by opening the *JustWalk JITAI* app (Figure 3.1), (2) *push* components that were sent to the participants as app notifications based on system-based rules (these were our key experimental manipulations and are described in the *Interventions and System ID Experimental Design* section and Figure 2), and (3) ecological momentary assessment (EMA) questions (described in the *Measures* section).

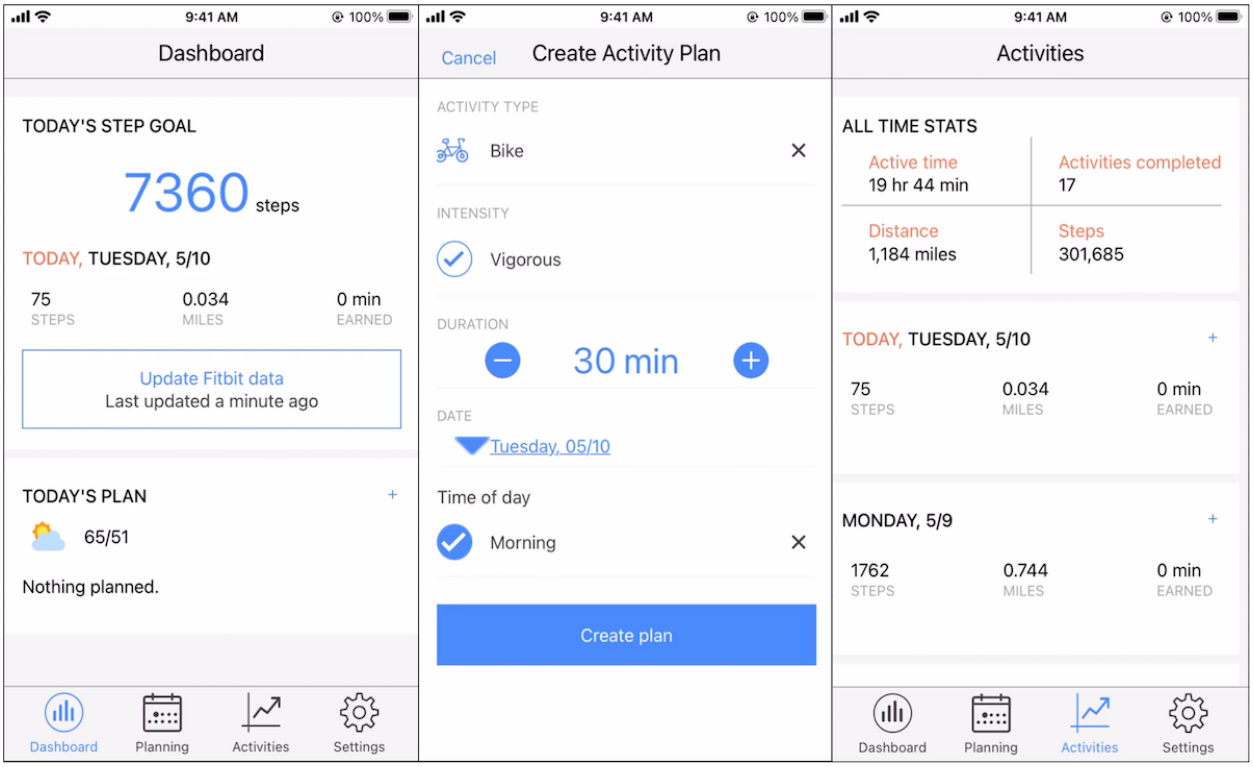


Figure 3.1 JustWalk JITAI app screenshots (left: app dashboard; center: planning tab; right: activity log tab)

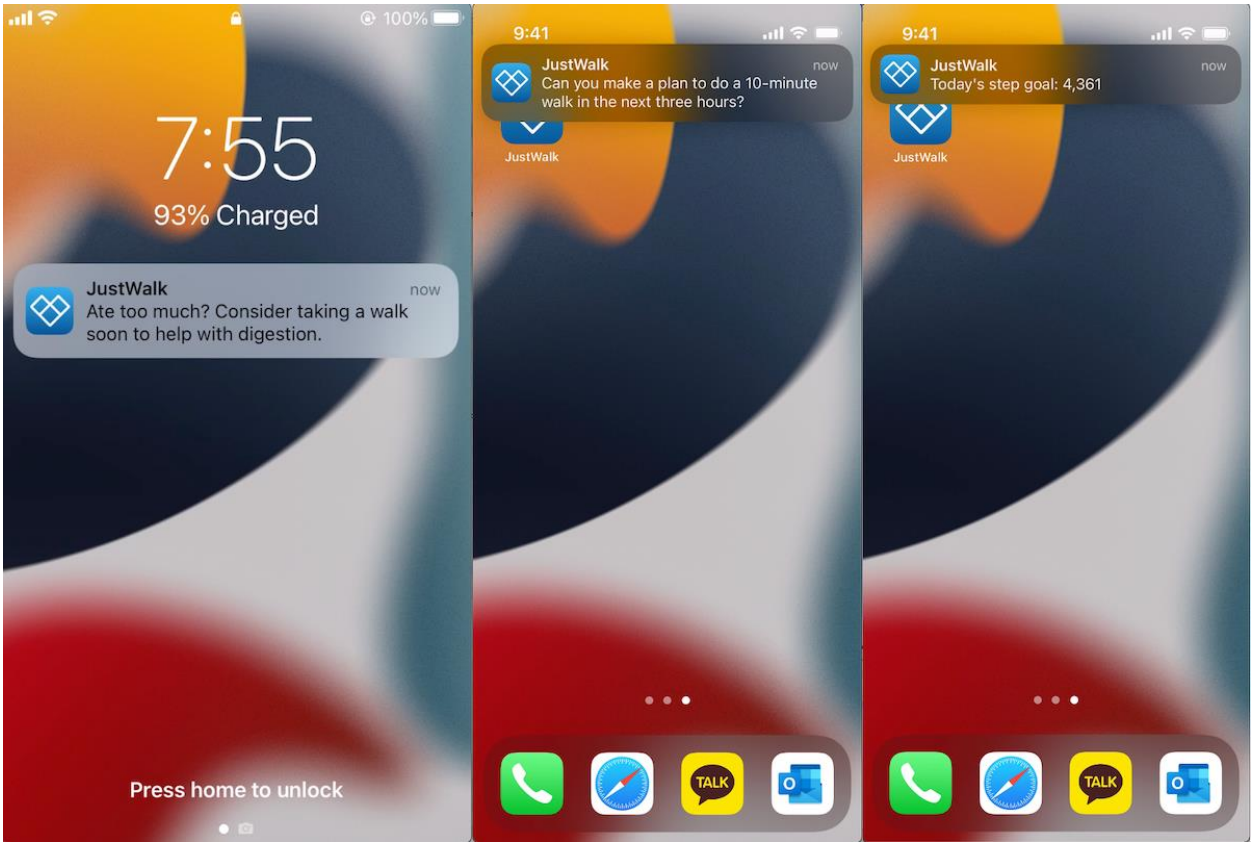


Figure 3.2 JustWalk JITAI app notification screenshot on the locked screen and background status.

The *JustWalk JITAI* app consisted of 3 *pull components,* which were drawn from the *HeartSteps* app [157]. These were accessible through tabs along the bottom of the app screen (Figure 3.1; left):

1. *Dashboard*: the dashboard was the home screen of the app and was displayed whenever a participant entered the app or finished interacting with a *JustWalk JITAI* notification. The dashboard implemented 3 behavior change techniques [158,159]: self-monitoring, feedback on goal progress, and notifications of activity plans. At the center of the dashboard, a participant's progress toward their daily step goal was shown as feedback. If participants created an activity plan for the day, the dashboard also displayed this plan for the participants.

2. *Planning*: through the planning tab (Figure 3.1; center), participants could identify when they would plan to exercise that week. The planning tool was designed to operationalize the behavior change technique implementation intentions [160,161] by enabling participants to identify when they would be active and for how long and identify a specific activity to engage in.
3. *Activity log* (Figure 3.1; right): participants could see an activity summary for the last 2 weeks, including steps walked and distance covered each day, as well as the types of activities that the Fitbit tracker detected automatically or the user logged manually (e.g., running, hiking, walking, and yoga). In addition, the activity log displayed Fitbit-derived active minutes for each day. Finally, the tab displayed the participants' all-time statistics—hours of active time, total distance walked, total counts of activities (detected or manually logged), and total number of steps recorded since the user started using the JustWalk JITAI system. These all-time statistics were intended to provide longer-time-frame feedback on what the participant accomplished over the duration of the study.

Participant Procedures

Recruitment

Participants were recruited nationally using a noncontact approach. Participants were recruited mainly through university mailing lists and word of mouth. The targeted number of participants was 50, with a final fully enrolled sample of 48 as 2 participants never showed up to preintervention meetings.

Eligibility

Inclusion criteria were participants who (1) were aged ≥25 years; (2) were inactive, defined as engaging in <60 minutes per week of self-reported moderate-intensity PA; (3) owned either an iPhone with iOS 11 or above or an Android phone with Android 5.1 or above; (4) stated

a commitment to follow study protocols, including regularly carrying a mobile phone, using the *JustWalk JITAI* app, answering phone-based questionnaires, and wearing the Fitbit Versa 3 activity tracker at least 8 hours a day; and (5) were fluent in English.

The exclusion criteria were participants who (1) were Incapable of providing Informed consent or (2) had a psychiatric disorder that limited their ability to follow the study protocol, including psychosis and dementia.

Screening, Informed Consent, and Onboarding Meeting

All participant interactions occurred remotely. Recruitment materials directed participants to the study website, which included a contact entry form. Upon completion, participants were automatically sent a link to an eligibility screener, which asked about age and PA levels (as reported using the Global Physical Activity Questionnaire [162] and the revised Physical Activity Readiness Questionnaire [163]). The staff reviewed the survey responses to confirm eligibility.

Eligible participants were offered time slots for an informed consent meeting. Ineligible participants were informed of their ineligibility.

During the informed consent meeting, the following activities took place: (1) the study was described in detail via a guided read-through of the consent form, (2) participants were provided with a list of reasons to and reasons not to take part in the study, (3) participants were invited to develop a list of their own pros and cons for taking part in the study, and (4) participants were given time to ask any questions they had. If participants verbally agreed, the study staff asked them to sign the consent form via DocuSign (DocuSign, Inc).

Consented participants were asked for a mailing address to send a Fitbit. Participants were also sent instructions on how to set up the Fitbit and the app via email. Once confirmation

of delivery of the Fitbit was received by the staff, a follow-up email was sent to participants inviting them to pick a time slot for the preintervention meeting.

During the onboarding Internet-based meeting, participants were instructed on the following topics: (1) instructions on how to install and use the Fitbit and study app, (2) information on the 10-day baseline, (3) information on what to expect after the 10-day baseline period, (4) direction to complete a baseline survey, and (5) instructional videos with corresponding notes about how to use and maintain the Fitbit (e.g., strategies to keep it charged and notifications to clean it to reduce skin irritation). All meetings between the participants and the staff took place via Zoom (Zoom Video Communications), and the interviews took place within 2 weeks after the time slots were sent.

Incentives

Participants received the following incentives: (1) Fitbit Versa 3 (received at study enrollment; US $229 in value), (2) US $25 gift cards provided to them up to 3 times (US $75 in total) if they completed at least 80% of the daily EMA items within each 3-month period, and (3) US $25 gift cards if they attended an optional postintervention interview.

Study Timeline

During the 10-day baseline, participants were asked to engage in their normal level of steps or PA and always wear the Fitbit except while charging. No interventions were provided, and no EMA questions were asked during the baseline period. When participants opened the app, 10 circles were shown designating the number of days they had met the minimal wear time requirements (i.e., 8 h/d). If the participant did not wear the Fitbit for at least 8 hours a day, the circles did not fill up. Once all 10 circles were filled, the app automatically transitioned to the intervention phase, displaying a dashboard.

In the intervention phase, all app features were delivered, including the 2 push intervention components (i.e., walking prompts and adaptive suggested step goals) and the daily EMA questions. The participants also gained access to other parts of the *JustWalk JITAI* app such as activity logs and planning support. Participants were asked to interact with the app whenever it sent them notifications and were told that they could open the app at any time if they wanted to access pull components and found them useful. Total interaction time with the app from push interventions and EMA notifications does not exceed 10 minutes each day, but participants may choose to spend more time on the app accessing other features. The interactions participants were prompted to do occurred in response to four types of notifications: (1) daily step goal notifications, (2) walking suggestion notifications, (3) prompts to complete the daily EMA battery, and (4) experience sampling prompts (i.e., if Fitbit detected an activity) to complete EMA items throughout the day and in relation to the notifications to either increase the urge to walk or plan (for details about EMA, see the *Measures* section).

Interventions and System ID Experimental Design

System ID Overview

This study used a system ID approach to manipulate 2 intervention components experimentally: *walking suggestions* and *daily step goals.* To achieve the desired dynamics on the timescale of interest, we used 2 input signals, one for each of the 2 components. Although our study design enables traditional statistical analyses to examine the impact of intervention components on behavioral outcomes, that is not the primary focus of a system ID experiment. The primary goal of a system ID experiment is to estimate and validate dynamical computational models that are validated based on their ability to predict the future responses of each individual's behavior across time. These aims are achieved by having different intervention components—suggestions to walk in the next 3 hours and adaptive goal setting—delivered at

different timescales and orthogonally, that is, statistically independent of each other. Our approach is analogous to a within-person factorial experiment (and, indeed, can be treated as such with all relevant nomothetic statistics used on the developed data set, which the team plans to perform as secondary analyses). The critical difference is that, in system ID studies, the designed input signals achieve statistical independence through orthogonality as verified in the frequency domain. Orthogonality enables separate study and estimation of the dynamics and impact of each intervention component. One can think of frequencies as different repeating rhythms across time, such as the notion of a minute-by-minute, daily, or weekly frequency. The study was designed to ensure that the intervention signals were disambiguated across time (via delivery at different frequencies). This enables a rigorous independent study of dynamical responses to both intervention components within the same experiment and, indeed, within each person, both proximally (e.g., immediate responses following intervention delivery) and distally (e.g., continued or delayed effects up to several days after any notification).

Specifically, both signals are designed to follow the guidelines presented in the study by Rivera et al [164], in which Equation 3.1 is highlighted to define the effective frequency range of the input signal based on a priori knowledge of the dominant system time constant.

Equation 3.1 is the equation used to define the effective frequency range of the input signals of the *JustWalk JITAI* study based on a priori knowledge:

$$\omega_* = \frac{1}{\beta_s \tau_{dom}^H} \leq \omega \leq \frac{\alpha_s}{\tau_{dom}^L} = \omega^* \qquad (3.1)$$

τ_{dom}^L and τ_{dom}^H represent the higher and lower bounds for the estimated dominant time constant of the system, meaning the range in which signals relevant to walking would occur naturalistically. A_s and β_s dictate the input signal's content of high and low frequency, respectively. The orthogonality of our intervention components was confirmed via the use of

cross-correlation analysis (the appropriate approach for testing orthogonality via frequency domains) to the designed input signals.

Sample Size Considerations

The number of participants has little impact on the power estimate for system ID studies as system ID approaches mostly use dynamic models that consider individual-level changes over time [3]. Instead, different methods, such as study length and validation analysis, can be used to establish the equivalent notion of *power* [165]. A multisine signal's *cycle,* or predetermined time interval, serves as the foundation for the power calculation [4]. By dividing the cycles into estimation and validation data sets, this type of design maximizes the signal-to-noise ratio and aids in the evaluation of model fits.

Previous research has demonstrated that 3 independently excited harmonics per cycle can achieve a sufficient excitation to deliver dynamically meaningful data in relation to daily frequencies. This can be performed with 3 sinusoids per cycle, resulting in a cycle that lasts at least 12 days [166]. Estimating and comprehending the multitimescale dynamics of behavior change are the goals of this effort. A total of 9 excited harmonics were revealed through simulations to be necessary to provide persistence of excitation across relevant frequencies [167]. The result is that each cycle lasts 26 days. From this, the final study length was set at 10 cycles (260 days).

Intervention Design

Overview

An overview of the *JustWalk JITAI* intervention elements is provided in Table 3.1.

Table 3.1 Summary of the JustWalk JITAI intervention elements

Wearable sensor

- Fitbit Versa 3

Mobile app

- *HeartSteps* [137] (renamed *JustWalk JITAI*)

Walking notifications

- Walking notifications were pushed up to 4 times a day starting at the participant-set time (e.g., 7 AM or 8 AM), with 3-hour gaps between each possible notification (e.g., 7 AM, 10 AM, 1 PM, and 4 PM as possible decision points for sending notifications).
- Notification texts were randomly chosen from a library of 50 messages that included 24 messages meant to inspire participants to plan a time when they would walk in the next 3 h and 26 messages meant to invite participants to become aware of internal urges that could inspire them to walk (Figure 3.2 and Supplemental Table 3.4).
- The experimentation setting used 4 just-in-time (JIT) definitions: (1) full JIT (need [N], opportunity [O], and receptivity [R]), (2) N+R, (3) N+O, and (4) random, with each element defined as follows:
    - *N*: on track to meet the daily step goal accounting for time of day when assessed (e.g., 50% of steps accrued halfway in a person's day, using their self-selected start time as a reference and assuming 12-h windows)
    - *O*: next 3-h time window was predicted to have an 80% likelihood that someone could take steps using a previously published algorithm [103]
    - *R*: participant received <6 notifications and responded (i.e., walked) to at least 50% of notifications sent to them within the previous 72 h

Adaptive step goals

- Each morning, participants were provided with a suggested daily step goal.
    - The notification also included a single-item ecological momentary assessment whether it was helpful or not.
- Daily step goals were calculated using the following procedures:
    - For the first cycle, median steps/d were used as a personalized reference to guide adaptive step goal suggestions. For cycle 1, a participant's personalized reference (i.e., median steps/d) was calculated from their 10-d baseline period excluding nonwear days.
    - For all subsequent cycles, their personalized reference (i.e., median steps/d) was calculated from the previous 26-d cycle (e.g., cycle 2's median steps/d were calculated using all step/d data from cycle 1 excluding nonwear days).
    - Participants were provided with a step goal that ranged between their personalized reference (median steps/d) up to their personalized reference+4000 steps.

Table 3.1 Summary of the JustWalk JITAI intervention elements, continued

○ A multisine signal design that ranged from 0 (personalized reference) to 1 (personalized reference+4000 steps) was used. For example, if a person's median steps/d during the baseline period were 5000 steps/d, they would receive step goal suggestions between 5000 and 9000 steps/d. ○ Maximum step goals were set at 12,000 steps/d, and minimum step goals were 2000 steps/d.

Walking Notifications

Overview

The first component of the *JustWalk JITAI* was notifications meant to inspire short (e.g., ≥10 min) walking bouts within 3 hours after receiving the notification. This component had two variations that targeted different behavioral processes: (1) an invitation for a person to create a microimplementation intention on when, where, and how to fit in a ≥10-minute walk in the next 3 hours—we label these suggestions as *bout planning*—and (2) an invitation for a person to become aware of interoceptive experiences and signals (e.g., stiff muscles and lethargy) that may inspire them to go for a walk—we refer to these messages as *cultivating an urge.* Both types of walking notifications were drawn from a library of 50 messages (n=24, 48% on bout planning and n=26, 52% on cultivating an urge). Walking notifications were provided in the form of push notifications from the *JustWalk JITAI* app. The notification could be sent 4 times a day (i.e., decision points) starting at the user-defined *start of day* time, which was gathered during the onboarding processes. Starting from the participant's self-described *start of the day,* the walking notification decision points occurred every 3 hours. For example, if a participant's day started at 8 AM, their decision points would be 8 AM, 11 AM, 2 PM, and 5 PM. For each participant, on each day of the study at each of the 4 decision times, the *JustWalk JITAI* system decided whether to send a walking notification based on the system ID procedure described in the following section.

Operationalization of JIT States

JIT states were experimentally varied via the use of different rules to define a *just-in-time (JIT)* state. By nudging participants when they were in *JIT* states, the hypothesis is that the effect of the walking notifications should increase while maintaining a low level of burden on participants, thus minimizing notification fatigue [43,44].

A JIT state has been previously conceptualized [1] as a state in which a person is receptive to support (e.g., if a notification is sent, a person would appreciate receiving said notification) and has the opportunity to engage in the desired behavior (or vulnerability to a negative behavior). Building on this theoretical formulation, a third theoretical parameter was added: the need for intervention support. For example, if someone is already meeting their daily step goals, they likely do not need additional intervention prompts to walk. For the purposes of this study, JIT states were operationalized as follows:

1. *Need (N)*: a person is defined as in a state of need if they did not meet the previous day's step goal (for the first decision point) or if they are not making steady progress toward that day's goal (for all other decision points). Sufficient progress was defined as the goal prorated to the current time of day as in Equation 3.2:

$$(daily\ goal) \times \frac{(time\ elapsed\ since\ the\ first\ decision\ point)}{12\ hours} \qquad (3.2)$$

2. *Opportunity (O)*: a person is deemed to be in a state of opportunity when they can feasibly walk. To operationalize this, a predictive algorithm described in the study by Park et al [103][6] used a threshold of 80% probability that, within the next 3 hours, a person may walk. We used the high threshold of 80% so that even a slight nudge to walk could effectively achieve short-term behavior change (note: whether notification is

[6] This study is included in this thesis as Appendix 1.

needed at such a high moment of opportunity is a question that we will be able to study retrospectively).

3. *Receptivity I*: a person is deemed to be receptive when they have received ≤6 messages in the last 72 hours *and* have responded (i.e., walked in the following 3 h) after ≥50% of the walking notifications sent in that period.

The operationalizations of these 3 facets of a JIT state allow us both to define a full JIT state—that is, when need, opportunity, and receptivity are *all* present—and to empirically test how different operationalizations of JIT states (e.g., states when only some of these components are present) influence walking notification effectiveness. Specifically, daily decision rules were tested that embodied four different decision policies of being in a JIT state:

1. *Full JIT state:* need, opportunity, and receptivity are present.
2. Partial JIT state (2 forms): *N+O* (only need and opportunity are present) and *N+R* (only need and receptivity are present).
3. *Not in a JIT state (random):* notifications are randomized each time at 50% probability.

How and when each of these rules for defining JIT states was varied experimentally is described in the following section.

Previous Observations and Theoretical Considerations That Guided Our Study Design to Test JIT States

This specific study design was created based on data from the original *HeartSteps* trial [137] followed by engaging with previous domain knowledge, including behavioral theory and our previously developed SCT dynamical model [168], to guide the final study design such that this study could provide robust data for supporting computational model testing.

Concerning previous observations, in the *HeartSteps* trial related to the notifications designed to inspire bouts of walking, 3 empirical observations guided our understanding of JIT states. First, as reported previously [137], notifications had a diminishing proximal impact on the total number of steps taken within the 30-minute window after the notification. These results indicated a theorized diminishing value-to-burden ratio of the prompts, namely, a dynamic concept that balances, for each instance, a person's perceived value that they receive from an intervention compared with the perception of the level of burden of the intervention. This dynamic hypothesis, which conformed to the data, was that the value-to-burden ratio would diminish over time, with initial notifications being perceived as more valuable than burdensome; however, by the end of the study, this would shift toward a low or negative value in relation to the burden.

Second, it was observed that, if <2 notifications were sent on a given day, even later in the 6-week trial, then the bout notifications resulted in significantly improved steps taken within the 30-minute window after the notification. We interpreted this dynamic observation as representing a hypothesized *auto-recovery* that could take place on a person's value-to-burden ratio. One could think of this as analogous to a neuron. Once a neuron fires, if new signals come in, the neuron will not fire again until it has sufficient time to recover, but this recovery process is automatic. It was hypothesized that a similar dynamic takes place regarding notifications. Namely, if notifications are sent at a rate that is faster than a person's autorecovery rate, habituation will set in and the notifications will be ignored (again, similar to a neuron not firing). If, in contrast, *sufficient time* has passed for autorecovery to take place (e.g., such as a neuron re-establishing itself as ready for the next signal), then a notification sent would be more likely to be attended and reacted to by a participant. Previous data guided us to a population-

based starting point of, on average, 2 notifications in a day, providing sufficient time for autorecovery. With that said, it was postulated that this autorecovery may vary among individuals. This study design enables us to study these individual differences in temporal responses.

Finally, it was observed that there was a trend in the daily timescale or frequency. Specifically, it was observed that there was an overall trend of increased steps per day over the 6-week intervention period. This third observation was translated into a hypothesized *internalization* process of the knowledge, skills, and practices that the intervention was meant to cultivate. This third dynamic hypothesis is the most critical target for designing an effective JIT intervention. Specifically, the goal is to create a JITAI that would enable a person to develop internalized knowledge, skills, and practices that could be maintained after the cessation of the intervention while accounting for the likely diminishing value-to-burden ratio and the need for *recovery* between notifications. This complex, interactive dynamic hypothesis, which postulates 3 different underlying dynamics that interact together, is what is primarily being studied in this experiment. Most critically, it was hypothesized that *internalization*, observed in the form of increases in steps per day in a time series, would take place more often when notifications or interventions were provided using JIT states compared with times when notifications were offered without taking account of JIT states.

System ID Experiment Design via Simulation Studies

With these empirical observations as a foundation, previous behavioral literature was reviewed to (1) look for previous domain knowledge that could be used to guide the understanding of these dynamics and (2) support us in better operationalizing the dynamic expectations we hypothesize to be observed, particularly if interventions could be provided

consistently taking account of JIT states. A computational model was developed guided by these empirical observations and building on principles drawn from operant learning and cognitive science, which is described elsewhere [169]. A set of simulations was run to model anticipated responses to receiving PA notifications during positive and negative JIT states. A key focus of the simulation work was to determine whether the models could produce the dynamics observed in *HeartSteps*, described previously, and to guide the anticipated length of time needed to observe a possible overall step per day increase across days when notifications are repeatedly delivered during positive JIT states. In this context, a positive outcome was operationalized as a person taking at least 1000 steps (as a proxy for 10 min) within the 3-hour window after receiving a walking notification. Overall, it was hypothesized that a greater number of positive outcomes when using the full JIT state operationalization (N+O+R) would be observed, with increased overall steps per day occurring across days during those times (accumulative internalization). In contrast, it was hypothesized that a relatively steady steps per day response would occur during times when notifications were sent at random (which was a replication of the original *HeartSteps* study and intentionally did not consider JIT states; thus, it was hypothesized that the random signals would replicate the observations from the original study). A set of additional simulations was run based on the SCT model [168], with the results of the simulation reported elsewhere [167] to further refine our study design.

On the basis of the simulation results from both models, a system ID study was devised that experimentally varied the use of different definitions of JIT states but did so in a way that would enable the study of possible accumulation or degradation of the dynamic effects across days. Specifically, 4 days was set as the minimal length of days needed to observe the effects of successive full JIT rules. It was anticipated that stabilization to degradation of effects would

start to occur within 1 day of sending non-JIT notifications based on our simulation studies. With that said, given the highly novel study design and limited robust empirical data to guide this subtle study of dynamics, longer periods were used, particularly for the full JIT state (N+O+R). In other words, the experiment compared decision rules that range from not trying to intervene in a JIT state to trying to intervene in a full JIT state over a sustained period that, based on simulation studies, would be sufficiently long to detect accumulative effects if they occurred.

This resulted in a categorical 4-level design. To construct this categorical 4-level input signal, a pseudorandom binary sequence (PRBS) was used (for full justification and details, see the study by El Mistiri et al [167]). This base signal compares JIT with some form of partial JIT or random (i.e., non-JIT) states. To incorporate the exploratory examination of differences between JIT operationalizations, a random multilevel sequence was superimposed over one of the PRBS binary levels to compare the 2 incomplete JIT decision rules (N+O and N+R) with the randomly sent walking notifications [167]. The input signal design parameters for the PRBS were chosen as $\tau_{dom}^{L} = 3\ days$, $\tau_{dom}^{L} = 3.5\ days$, $\alpha_s = 2$, $\beta_s = 2$, which was done to cover the frequency range of interest based on the guidelines provided in Equation 3.1. This resulted in a 60-day cycle with n_r=4 shift registers and switching time T_{sw} = 4 days (Figure 3.3). This 60-day cycle enabled the team to (1) study the hypothesized dynamic, positive accumulative effect on steps within 3 hours of notification times, and steps per day when walking notifications were sent during theoretically defined JIT states; (2) compare these dynamics with the hypothesized dynamic degradation across days when walking notifications were delivered during partial or negative JIT states (i.e., at random); and (3) as an exploratory aim, study if the dynamics vary across different JIT state operationalizations. In total, 4 cycles of a 60-day PRBS signal were generated to support both estimation and validation of the dynamical models that operationalized

the hypothesized dynamics, which results in a 240-day period followed by a final period of full JIT state level to match the full study period (260 days), which was constrained by adaptive step goal cycles (26 days × 10 cycles).

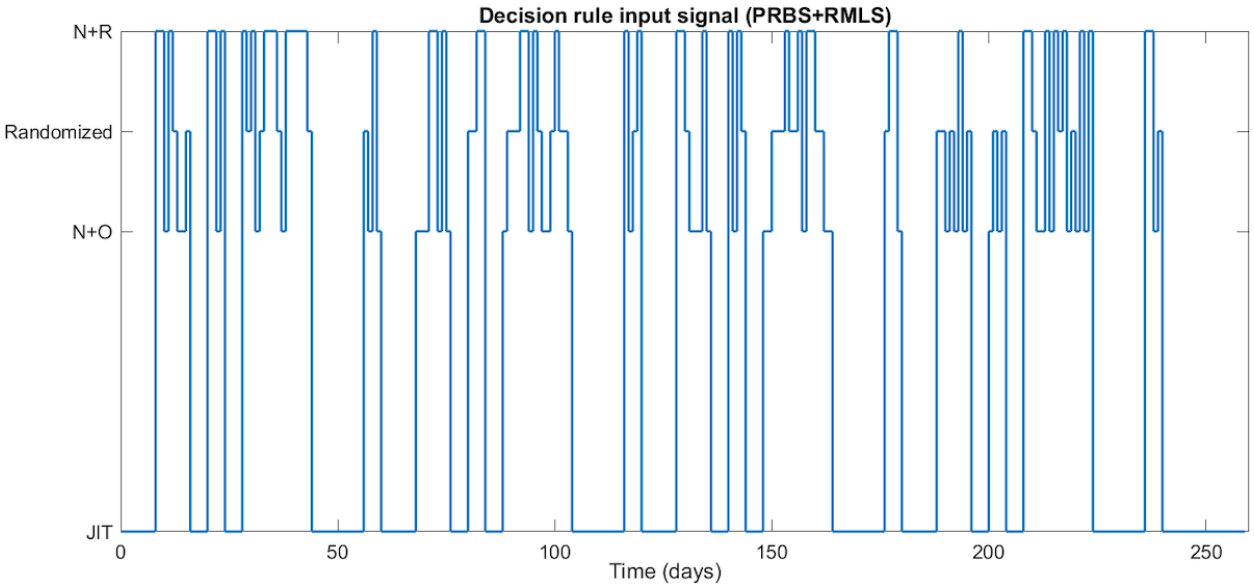


Figure 3.3 The designed decision rules signal for the walking notification component of the intervention in the time domain in the JustWalk JITAI study.[7]

Figure 3.4 provides a visualization of the spectral power density as it relates to the walking notifications. This visualization provides insights into the degree to which the theorized dynamics will be appropriately excited, enabling the detection of effects if they occur across various frequencies and, thus, the study of the proposed dynamic hypotheses. The results suggest sufficient persistent excitation by the number of harmonics included in the effective frequency range between 0.14 and 0.67 rad per day. This frequency range, determined by the time constant guidelines in Equation 3.1, ensures that the appropriate slow dynamics (i.e., low frequencies) and fast dynamics (i.e., high frequencies) of the system are captured.

[7] Each level represents one of the decision rules. The 4 signal levels were obtained by superimposing a 3-level random multilevel sequence (RMLS) signal on the base pseudorandom binary sequence (PRBS) signal. JIT: just-in-time; N+O: need and opportunity; N+R: need and receptivity.

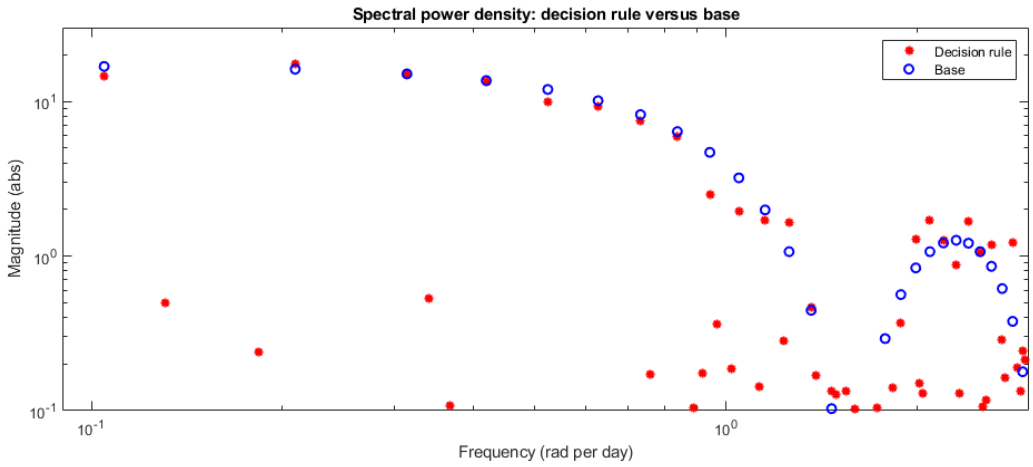


Figure 3.4 Spectral power density of the designed decision rule input signal of the JustWalk JITAI study.[8]

Adaptive Step Goals Intervention Component

Overview

The adaptive daily step goal component follows a similar structure to that of a previous system ID experiment whereby participants were given a specific suggested step goal to target each day [5]. Similar to this previous design, a key consideration is to design a *cycle*, which is a deterministic, repeatable pattern that defines the provision of intervention options to an individual. Intervention options can be provisioned to mimic randomness via pseudorandom signal designs that achieve the valuable properties of randomness for causal inference while still being deterministic and, thus, repeatable. This provides valuable properties for a system ID experiment as it enables a more robust comparison between cycles (for more details, see prior work [2]). In this study, the same basic logic of prior work was followed, specifically, using a pseudorandom signal design that varied step goals between an *achievable* target up to a plausibly

[8] It is shown to determine whether sufficient excitation across key frequencies is established within the trial.

ambitious target (how goals are assigned each day is described in greater depth in prior work [167]).

The definition of an *achievable* step goal was personalized to each participant, which was labeled as a personalized reference, defined as a person's median steps per day calculated from the previous 26-day cycle period [167]; note that, for the first cycle only, the personal reference was the 10-day-baseline period. Each morning, participants received a notification informing them of their targeted step goal for the day. The updated goal was also available to them on the *JustWalk JITAI* dashboard and was automatically synced by the *JustWalk JITAI* server to the participant's Fitbit account so that the feedback on the Fitbit app and the participant's Fitbit tracker always showed the correct goal progress each day. To further facilitate goal pursuit, participants were instructed to install a watch face providing the step goal number and a goal progress bar to enable always visible goal progress feedback. Fitbit's native visual and haptic feedback was used when the participant completed the daily step goals (i.e., fireworks animation and vibrations).

Experimental Manipulation: Input Signal Design

To define a cycle for this component, a multisine signal was used. The input signal design parameters ($\tau_{dom}^{H} = 1\ day$, $\tau_{dom}^{H} = 2\ days$, α_s=2, β_s=2), as described in Equation 3.1, were chosen based on the results from previous work and the simulation studies we conducted in preparation for staging this system ID experiment [167]. The design parameters result in a cycle length of 26 days, as shown in Figure 3.5. For each participant, a personalized realization of the multisine signal generates the daily goals throughout the 260-day intervention across 10 cycles by determining, for each day, the factor by which the 4000 steps per day range is multiplied and then added to the participant's personalized reference (i.e., median steps/d), as described in Table 3.1.

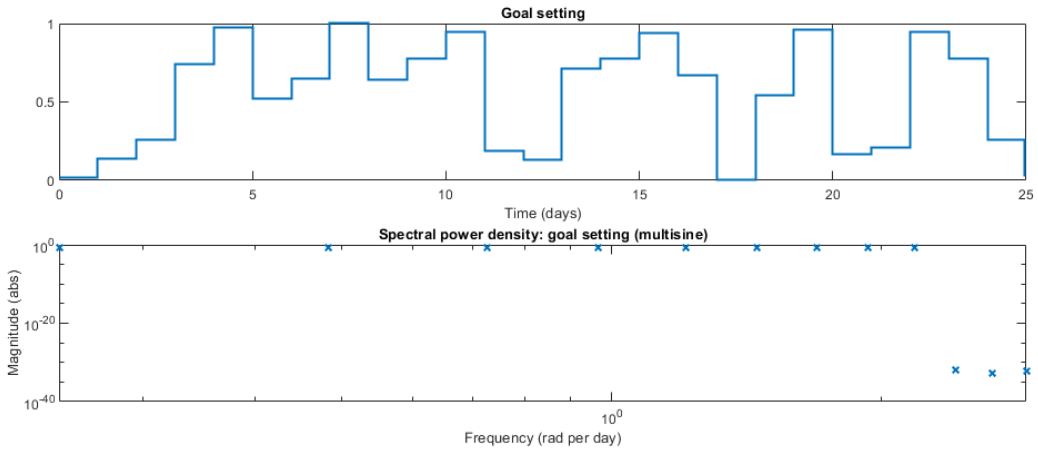


Figure 3.5 One cycle of the designed multisine input signal for the goal setting component of the JustWalk JITAI intervention in both the time (top) and frequency (bottom) domains.[9]

The effective frequency range of the signal is related to the design parameters through Equation 3.1, which yields the persistence of excitation between $\omega_* = 0.25$ rad per day and $\omega^* = 2$ rad per day approximately for the designed multisine signal, as it is highlighted in the power spectrum of the signal shown in Figure 3.5. This showcases that the designed input signal for goal setting creates variability in the relevant dynamical ranges of interest.

Measures

Baseline Survey

The baseline survey includes (1) demographic information, including age, height, weight, ethnicity, gender, race, marital status, household size, employment status, and level of education; (2) personal characteristics related to the study (i.e., how the participants spend their time and information about their routine and their neighborhood); (3) self-perception (i.e., personality [170] and perceived stress [171]); and (4) life habits [172] and PA [173] (i.e., how they feel

[9] The multiplier factor varies between 0 and 1 over time. The spectral power density plot highlights the number of excited harmonics at the frequencies of interest.

about PA, how they engage in exercise, how much they intend to exercise, and how they notice the effects of exercise).

Continuous Measurement of Activity and Heart Rate

Steps per minute and minute-level heart rate were measured using the Fitbit Versa 3, a wrist-worn, consumer-level activity tracker that uses triaxial accelerometry to measure movement. Further details are provided in the next subsection. Moderate or vigorous PA was measured in four ways: (1) automatically triggered objective measurement (the Fitbit Versa 3 automatically detects vigorous movement [174] if the activity is sufficiently vigorous and long), (2) manually initiated objective measurement (the Fitbit Versa 3 or the Fitbit app on the smartphone has an *activity* tab to log activity manually), (3) manual logging in the Fitbit app, and (4) manual logging on the study app using the *activity* tab (which was also populated with any reporting on the Fitbit app). The Fitbit assesses steps, PA intensity levels, energy expenditure, start or end time point and type of activity, distance traveled, and number of floors. Prior work shows that Fitbits likely underestimate heart rates and, by extension, total activity, but they do so reliably, thus establishing a meaningful within-person comparison [89], which is the focus of this study.

EMA Items

Psychological constructs and process variables were asked daily for inclusion in our targeted, dynamic models via EMA conducted at 7 PM local time. The EMA items included concepts of (1) self-efficacy for walking, (2) self-efficacy for problem-solving, (3) positive context for walking, (4) negative context for walking, (5) supportive routine, (6) drive to walk, (7) relationships supportive of walking, (8) interoceptive awareness of cues that could inspire walking (e.g., stiffness and fatigue), (9) negative reinforcement of walking, (10) behavioral repertoire, and (11) typical supportiveness for walking. Detailed items are included in

Supplemental Table 3.1 through 3. Items 1 to 6 were asked daily, items 7 to 10 were asked every 4 days to minimize the burden of responding, and item 11 was asked every day for the first week of each month (i.e., 7 times/mo).

In total, 2 other types of EMA question items were sent. Triggered by completed PA, the participants were asked if they felt healthy, fatigued, energized, and discomfort. We also asked if the participants thought that they could meet the daily goal.

Weather

Daily weather data (current weather and weather forecasts) were gathered from the public weather database API [175]. No actual location GPS data were gathered; instead, a self-reported home zip code was used to gather weather data.

Postintervention Interviews

After the 260-day intervention, participants were asked to fill out a brief postintervention survey and were also given the option to participate in a postintervention interview. During the postintervention interviews, participants were asked about their overall reactions, including both positive and negative aspects of using the app and any suggestions to improve the intervention. The interviews were audio recorded and transcribed.

The participants were also given a choice to either stop the use of the *JustWalk JITAI* intervention or continue to use it, in which case the data past the study end date would not be used in the analyses.

Treatment Fidelity Monitoring Procedures

Mobile App Use Logs

Mobile app use was recorded with time stamps for every page viewed in the *JustWalk JITAI* app, including opening notifications, opening the app, viewing pages within the app, and opening surveys. The one piece of information that could not be logged owing to operating

system limitations was whether notifications (e.g., walking notifications) were seen without being opened, such as when they automatically expanded on the iOS lock screen.

Monitoring the JustWalk JITAI Systems

The *JustWalk JITAI* server was automatically monitored every 10 minutes throughout the study period using a separate program to check for 5 performance and stability targets: the web server, the database server, the security firewall, the software framework for the server, and the Fitbit API. If the server stopped working or took too long to respond (>3000 ms), the program sent SMS text messages and emails to the study staff. The monitoring program was separately overseen by another program to ensure that monitoring was conducted properly. If any data operations failed (e.g., if the Fitbit server was not responding), the study staff were immediately notified via email. If there was an error in the Fitbit data synchronization, when the data connection resumed, all the missing data were refetched to fill up any missing period.

Data Collection Monitoring

Data collection was monitored by the study staff on a weekly basis with automated visualizations to ensure that there were no technical errors that may compromise the study.

Study Adherence

Study adherence was monitored automatically using the *JustWalk JITAI* server. During the preintervention meetings, participants were asked to wear the Fitbit for a minimum of 8 hours a day, but it was suggested that they wear the Fitbit all day, even at night. Fitbit devices typically synchronize with the Fitbit server via the Fitbit phone app every 15 minutes. This synchronization stops if the Fitbit device runs out of battery or is not worn for several days. The *JustWalk JITAI* server regularly checks whether a participant's device has stopped syncing with the Fitbit server, and if so, it sends an adherence SMS text message to the participant.

As an operationalized protocol, at 15 minutes before the first decision point for walking notifications (e.g., 6:45 AM local time for most participants), if a participant had no Fitbit app synchronization records for 60 minutes (e.g., since 5:45 AM local time for most participants), the server sent an automated adherence SMS text message including an approximate length of the period for which the updates were missed (e.g., a few hours, a day, or a while) to invite participants to recharge and synchronize their Fitbit on the Fitbit app.

This approach helped avoid making the mistake of responding too immediately to the problem of data drops caused by accidental battery discharges. As it can be assumed that people do not carry their Fitbit charger around during the day, sending an immediate *charge it now* message when data updates stop during the day is unlikely to be an effective remedy. In addition, given that it only takes approximately 30 minutes to charge a Fitbit from fully depleted to lasting more than a day, it was assumed that sending these notifications before the start of the day would give participants a chance to charge their Fitbit.

Modeling and Data Analysis[10]

A series of system ID estimation approaches will be used to analyze the data and obtain control-oriented dynamical models to study JIT states. General data analysis will start with examining the cross-correlation of the data to verify the hypothesized structure of the system as operationalized via the computational models described previously. Nonparametric estimation methods such as correlational analysis and spectral analysis [176] will be used to obtain preliminary information about the responses of each individual (i.e., time constants, gains, and orders). The knowledge gained from the nonparametric estimation methods will be used to obtain ideographic models through prediction error modeling approaches such as autoregressive

[10] This section is directly quoted from the published manuscript, which targets an overarching set of analyses. Hence, it does not precisely reflect this thesis' analytical methods. Please refer to Chapter 4 for the analytical methods for this thesis.

with external input and output error [176] estimation and more elaborate gray box methods using the SCT model structure. In addition, the model-on-demand [177,178] estimation will be used to estimate more flexible models that address nonlinearities in the system. The estimated models from all approaches will be contrasted with one another, and the advantages or disadvantages of each will be assessed to inform future efforts.

Ethical Considerations

The study was approved by the University of California San Diego, institutional review board (protocol 800132) and was preregistered on ClinicalTrials.gov (NCT05273437).

Results

Simulations

The input signal design for the 2 intervention components in this study involved an iterative procedure that relied on a priori knowledge and simulation results for different types of anticipated participants to guide the efforts. The simulations were based on a dynamic SCT model derived from a fluid analogy, which provided the means to guide specific conditions for the JIT states considered in the decision rules to ensure that both the number of notifications sent per day and the overall number of notifications sent throughout the intervention were not burdensome for the participants. Furthermore, the simulation framework with diverse scenarios provided insights into the *ambitious yet achievable* range of adaptive goals provided in each goal setting cycle. A detailed account of the model used in the simulations, technical details of the designed input signals, and simulation results that guided the design are provided in the study by El Mistiri et al [167]. In this section, the results for a hypothetical adherent participant are presented to illustrate the dynamic nature by which the daily goals adapt to the participant's performance, as well as the effectiveness of the JIT decision rules in limiting the provision of support only to times that are hypothesized to be beneficial.

Figure 3.6 [167] shows the implementation of the designed daily goal signal in a simulation setting. In this case, the goals in each cycle are adjusted to the performance of the participant in the previous cycle, as described in Table 3.1. Note that, in this case, a hypothetical adherent participant is capable of achieving the daily goals given to them in each cycle; consequently, the median of the participant's performance increases, which leads to an increase in the goals provided in subsequent cycles. As a result, the daily goals gradually increase over the span of the intervention, from a low of 2000 steps per day in the first 2 cycles of the intervention (the first 52 days) to a high of 12,000 steps per day in the last 5 cycles. This simulation result illustrates that the input signal design for this component is working as intended by adapting the daily goals to each participant in a personalized manner while nudging the participant toward higher levels of PA through a combination of *ambitious* and *achievable* goals.

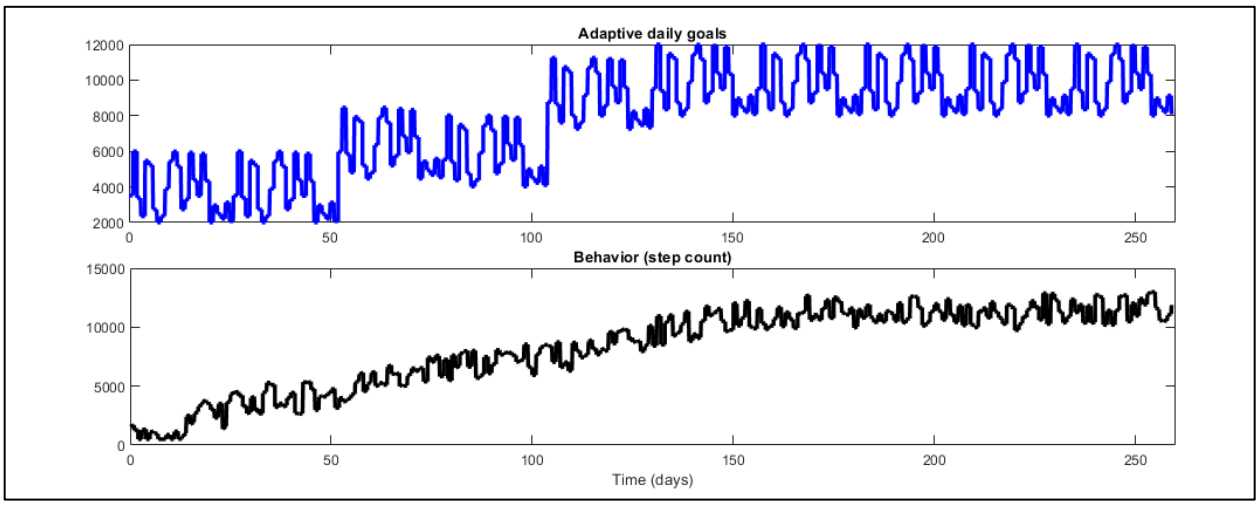


Figure 3.6 Simulation results illustrating the implementation of adaptive daily goals (top) in reaction to the performance of a hypothetical adherent participant in terms of daily step count (bottom) in the JustWalk JITAI study (adopted from the study by El Mistiri et al [167]).

Figure 3.7 [167] shows the walking notification component of the intervention in the simulated scenario for the hypothetical adherent participant. As shown in the figure, the decision rules work as intended in terms of dictating the nature of the notifications sent to the participant.

At the beginning of the intervention, when the participant does not achieve the daily goals (hence, the need condition of the decision rules is met), the number of notifications sent to the participant is high across all levels[11] of the decision rules. Later in the intervention, as the participant adopts healthier behaviors and meets the daily goals, the number of walking notifications sent on a daily basis is significantly lower. Furthermore, note that, on days when the receptivity condition is considered, the number of notifications sent follows the notification budget mentioned in Table 3.1.

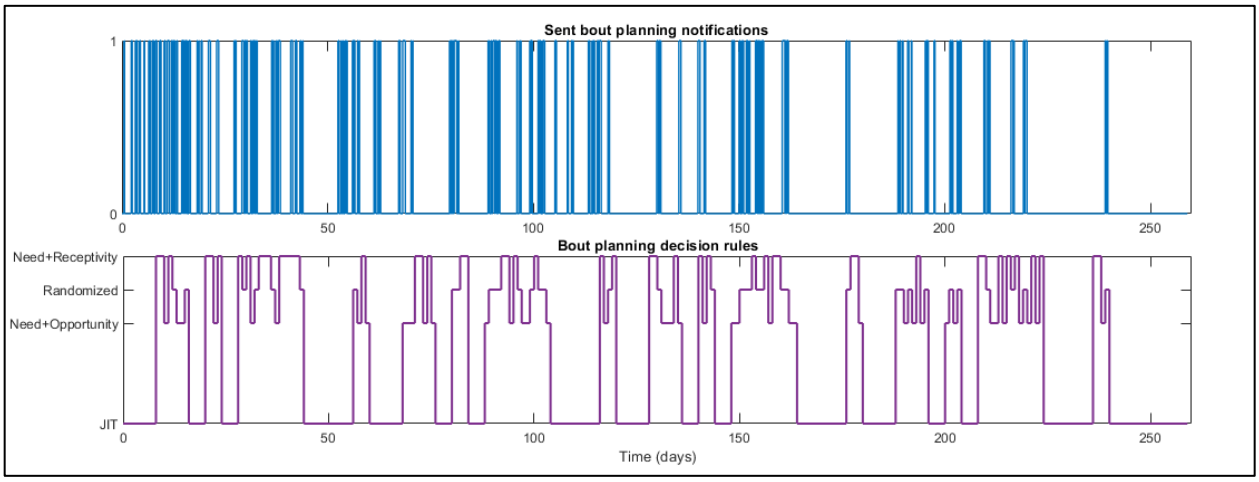


Figure 3.7 Simulation results for a hypothetical adherent participant illustrating the expected walking notifications (top) sent based on the designed decision rules signal (bottom) of the JustWalk JITAI study[12].

Finally, as the need condition is not met by the participant toward the end of the intervention, walking notifications are only sent on days with fully randomized notifications. This design allows for comparing the impact on the participant of fully randomized notifications with that of notifications that are guided by JIT state conditions that should make them more

[11] In other sections of this thesis, the term "level" is replaced by "decision policy" for better interpretation. However, since this chapter is a full reprint of a manuscript that is already published, we decided to keep the original terminologies. Apologies for the confusion.

[12] For the sent notifications signal at the top, a 0 value implies that no notifications were sent at that decision point, whereas a value of 1 implies that a notification was sent (adapted from the study by El Mistiri et al [167]). JIT: just-in-time.

beneficial. From these simulation results, the rate at which notifications are sent (i.e., notifications or decision point) on full JIT state days is the lowest at 0.084, followed by days of need and opportunity (N+O) conditions at 0.148 and need and receptivity (N+R) conditions at 0.176. The highest rate of notifications is observed on days with fully randomized walking notifications at 0.488.

Recruitment

Enrollment began in March 2022 and ended in July 2022. In total, 761 potential participants submitted a letter of interest, and 48 (6.3%) were enrolled in the study. Figure 3.8 shows the CONSORT (Consolidated Standards of Reporting Trials) diagram [179]. The intervention was completed in April 2023. The data were gathered without major incidents. The source code for the server and the app is publicly available on the project's GitHub repository [180].

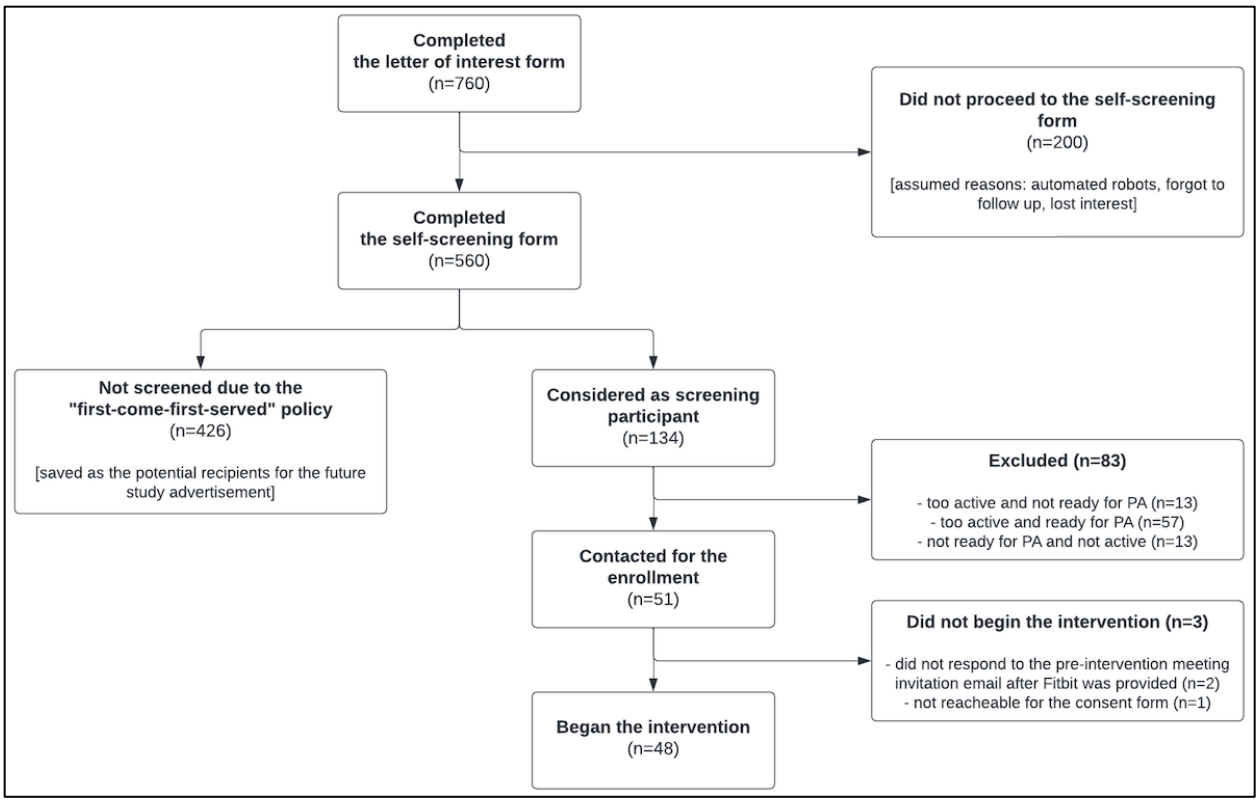


Figure 3.8 CONSORT (Consolidated Standards of Reporting Trials) recruitment diagram for the JustWalk JITAI study. PA: physical activity.

Discussion

This is a study protocol to investigate 3 JIT states (i.e., need, opportunity, and receptivity) empirically and enable the empirical optimization of a JITAI intended to increase PA (steps/d) in adult populations with an inactive lifestyle.

There is well-documented evidence suggesting that digital health interventions to date have not lived up to their intended potential [181–183]. From issues of poor adherence; results that only produce limited effects; and questionable scalability, particularly among those with less access to digital devices, the potential of digital health interventions has not yet translated to a real-world impact [181,182,184]. A pathway for improving this is to focus on producing evidence directly targeting and seeking to improve the fundamental shortcomings of digital health interventions [182,183].

In this study, our primary focus was to rigorously examine the notion of JIT states. To date, JIT states used to formulate intervention decision rules have either been assumed to be correctly defined—typically based on guidance from behavioral theory—but not empirically tested or they have been derived using data-driven approaches such as reinforcement learning whereby previous domain knowledge and understanding is underemphasized and, instead, there is hope that useful Insights about Intervention time condition will emerge from data collected in intervention studies [84]. Although we believe that both of these paths have merit, this study protocol offers a middle ground whereby previous domain knowledge is encapsulated into computational models, which, through simulation studies, can then be used to guide the careful generation of evidence that can test dynamic hypotheses about the nature of JIT states.

This is important as a JIT state, conceptually, is an inherently nonlinear causal phenomenon [76]. There is no one causal factor that makes a given moment a JIT state, but instead, it is a mixture of different factors such as time of day, a person's current motivational levels, their relationships, recent experiences with the types of support given, and the degree to which support is well matched to a person's current needs. Such factors combine at a given moment to influence the decision to engage—or not—in the target behavior. This study protocol recognizes the inherent nonlinear causal nature of the phenomenon under study and provides a rigorous approach to gathering the data needed to make progress in the context of such complexity. By varying whether a notification is provided when the person is thought to be in a state of need, when they are thought to have an opportunity to walk based on their personal historical step data, and when they are thought to be receptive, or combinations of these 3, the experiment will collect initial evidence for which aspects of the JIT state are most important for supporting the effectiveness of JITAIs and whether this changes over time. Using this

information, particularly when linked with slow dynamic processes of change (i.e., daily goal setting), the experiment produces data needed to empirically optimize a digital health intervention, *JustWalk JITAI*. Specifically, this work will result in individualized, empirically validated dynamical models that can be used to predict each individual's response to the intervention options offered to them. These individualized or idiographic dynamical models can be applied to optimal personalized behavioral interventions through sophisticated control algorithms such as model-predictive control [3,185]. These model-predictive control–driven JITAIs could have the potential to work more effectively than previous digital health interventions.

The key limitations of this study stem from the high novelty of the overall experiment and its approach. To the best of the authors' knowledge, no system ID experiment of this sophistication for studying human behavior has ever been conducted. On the basis of this, there was very little robust previous evidence and examples that we could draw upon to guide study design decisions. Although we did compensate for this by drawing on some relevant data (largely from our own work, as already described) and via a number of simulation studies, overall, there are potential risks and limitations to our approach. For example, given the novelty of this experiment, it is unknown how well the assumptions we used to guide the experiment will hold up. With this, it is unclear exactly how informative those data will be for studying JIT states. Second, given the novelty of this experiment, it was unclear what an appropriate sample size should be. This point is critical for determining the degree to which any patterns or insights gleaned about JIT states from this sample will be transportable to other populations or settings.

With that said, the primary focus of any system ID experiment is the study and articulation of computational models that are predictive and foster robust control decisions for

each *system*. In this context, a *system* is a person. This is critical to note because, as flagged previously, the notion of *statistical power* as is used in the classic frequentist inferential statistics used most commonly by health scientists does not have any direct translation or use within system ID experiments. Indeed, the key focus of system ID experiments is to work within each system to gain a deep understanding of its dynamics. This focus makes sense particularly for a concept such as JIT states, which, definitionally, will likely manifest idiographically. The critical question is not whether some general pattern of JIT states can be inferred but, instead, whether the same algorithmic development processes can be conducted ideographically and in a replicable and scalable fashion to enable the insights that the algorithm can produce to guide intervention decision-making. This is the primary focus of our work. Thus, the limitation is less one of *statistical power* and more akin to what arises with regard to the right training data sets for machine learning algorithms. It is unclear at this time what variations across people, places, and time could occur in real-world contexts that would render our approach nonfunctional. With a sample of only 48 participants, a key limitation is that we very likely did not have diversity across variations in people and places where this type of algorithm could be used to test the robustness of our approach. With that said, given the great novelty of our overall approach, we contend that this is an appropriate trade-off. Most critically, it is likely that, even in the sample of 48 participants, we will discover some individuals from whom we can create computational models that are informative and others from whom we cannot. That will be the type of initial data we could use to then develop more rigorous hypotheses about the transportability of our methods, which can then guide future experimentation.

Overall, this work could feasibly be a key step in filling the gap between the hope of digital tools and current realities in terms of limited long-term impact and engagement based on

the evidence. Although this is all still quite hypothetical, this trial is a critical step in testing the potential benefits of this overall approach for intervention optimization.

Acknowledgments

Chapter 3, in full, is a reprint of the material as it appears in Park, Junghwan, Meelim Kim, Mohamed El Mistiri, Rachael Kha, Sarasij Banerjee, Lisa Gotzian, Guillaume Chevance, Daniel E. Rivera, Predrag Klasnja, and Eric Hekler. 2023. "Advancing Understanding of Just-in-Time States for Supporting Physical Activity (Project JustWalk JITAI): Protocol for a System ID Study of Just-in-Time Adaptive Interventions." *JMIR Research Protocols* 12 (September): e52161. The dissertation author was the primary investigator and author of this paper. The study described in this chapter was funded by the National Library of Medicine (R01LM013107).

Chapter 1, 2, 4, 5, 6, and 7 of this thesis, in part, are currently being prepared for submission for publication of the material. Park, Junghwan; Kim, Meelim; El Mistiri, Mohamed; Kha, Rachael; Banerjee, Sarasij; Gotzian, Lisa; Chevance, Guillaume; Rivera, Daniel E.; Klasnja, Predrag; Hekler, Eric. The dissertation author was the primary researcher and author of this material.

Data Availability

The study materials and analysis code will be available through a public repository. The data set generated in this study will be available from the corresponding author upon reasonable request and with completion of the data user agreement.

Authors' Contributions

MEM, RK, LG, GC, DER, PK, and EH performed theoretical foundation work and simulations. JP, MK, MEM, RK, DER, PK, and EH developed the intervention. JP and PK developed the technical systems, including the server and mobile app. Patrick Neggie, Rahul

D'Costa, and Nick Reid supported the development and operation of the technical systems. JP, MK, RK, and SB conducted recruitment and interviews and worked on the trial logistics (e.g., sending out Fitbits). Michael Higgins and Shadia J Assi supported the operation of the clinical trial. JP and MK monitored the participants, communicated with them if necessary, and executed the adherence protocol. MEM, RK, and DER developed system ID programs. DER, PK, and EH provided guidance at each stage of the study. All authors contributed to the writing of the manuscript.

Conflicts of Interest

JP is an employee of the Ministry of Health and Welfare, Korean National Government. LG is an employee of Lufthansa Industry Solutions.

Supplemental Table 3.1 Ecological Momentary Assessment Items: Daily EMA (asked at 7 pm local)

Index	Question Text	Frequency	Answering Options
1	Being active is **a top priority** tomorrow.	Daily	Likert (Not at all – Completely)
2	**Circumstances will help me** to walk tomorrow (e.g., nice weather, getting in nature, free time).	Daily	Likert (Not at all – Completely)
3	My **schedule makes it easy** to be active tomorrow.	Daily	Likert (Not at all – Completely)
4	I **expect obstacles** (e.g., no time, unsafe, poor weather) to being active tomorrow.	Daily	Likert (Not at all – Completely)
5	I know how to **solve any problems** to being active tomorrow.	Daily	Likert (Not at all – Completely)
6	I am confident I can **overcome obstacles** to being active tomorrow.	Daily	Likert (Not at all – Completely)
7	**No matter what**, I'm going to be active tomorrow.	Daily	Likert (Not at all – Completely)
8-1	In general, my **friends help me** to be active.	Daily, but one of the three question items was asked per day	Likert (Not at all – Completely)
8-2	I regularly feel **urges to** be active.		Likert (Not at all – Completely)
8-3	I am active because it **helps me feel better** (e.g., reduce stress, stiffness, or fatigue).		Likert (Not at all – Completely)
9	My **typical Sunday includes being active.**	Daily for the first week of each month	Likert (Not at all – Completely)

(Bold texts were shown as bold on the app)

Supplemental Table 3.2 Ecological Momentary Assessment Items: Activity Triggered EMA (asked within 15 minutes when a physical activity is detected)

Index	Question Text	Frequency	Answering Options
1	Are you **feeling healthy** now? Agile, fit, limber, strong...	Activity triggered	Likert (Not at all – Completely)
2	Are you **feeling fatigued** now? Tired, exhausted...	Activity triggered	Likert (Not at all – Completely)
3	Are you **feeling energized** now? Awake, lively, vigor...	Activity triggered	Likert (Not at all – Completely)
4	Are you **feeling discomfort** now? Tired, aches, sweat...	Activity triggered	Likert (Not at all – Completely)

(Bold texts were shown as bold on the app)

Supplemental Table 3.3 Ecological Momentary Assessment Items: Daily Step Goal EMA (asked individually set morning time (i.e., start of a day))

Index	Question Text	Frequency	Answering Options
1	Today's step goal: x,xxx I think I can meet today's goal	Local time (start of day)	No / Maybe / Yes

Supplemental Table 3.4 Bout planning notification messages and their classifications into two categories of messages

Messages	**Category**[1]
How are you feeling? Think a 30-minute walk in the next three hours could make you feel better?	1
Feeling lonely? Could you walk and call a friend or ask a friend to walk with you in the next three hours?	1
Ate too much? Consider taking a walk soon to help with digestion.	2
Do you love podcasts? Any chance you could plan a walk and listen to a short podcast (even part of a long one) in the next 3 hours?	1
Feeling productive? If not, going for a short walk soon could feed your brain and help your focus and energy.	2
Feeling low? Even a 15-minute walk can often help. Can you take a walk in the next 3 hours?	1
Ate recently? Going for a walk is great for averting the food coma.	2
Feeling stressed? Do you think a quick walk soon might help?	2
Are you at a place with stairs? Would taking a few minutes to go up and down the stairs help you feel more energized?	2
Feeling anxious? Do you think a walk might help to clear your thoughts?	2
Need to get something? Is it nearby? If yes, consider walking there to get your activity and get something done at the same time.	2
Feeling sleepy? Taking a quick walk soon will help wake you up and feel better.	2
Having trouble sleeping at night? Do you think taking a walk in the next few hours may help you sleep better tonight?	2
Do you have a plan on when to be active today? If not, consider finding time in the next 3 hours.	1
Feeling bored? Would taking a walk soon help with reconnecting to your priorities?	2
Want to catch up with a friend? Text them to see if they could take a 10-minute walking break with you soon (phone or in-person).	2
Is there any point in the next three hours when you could take a break to take a 10-minute walk?	1
Feeling stiff? Would going for a walk soon help?	2
Is there a beautiful place nearby where you could go for a quick walk in the next few hours?	1

Supplemental Table 3.4 Bout planning notification messages and their classifications into two categories of messages, continued

Messages	**Category**[1]
Feeling confused or conflicted about something? Would taking a quiet walk soon help clear your mind and give you some clarity?	2
Have you gotten enough activity today? If not, could you go for a short walk in the next two to three hours?	1
Want to live a long and healthy life? Being regularly active really helps. Could you take a walk in the next few hours?	1
How is your schedule looking today? Is there a free window in the next three hours when could walk?	1
Muscles feeling tight? If so, consider taking a walk in the next couple of hours to loosen up your body.	2
Craving a snack? Are you really hungry or just need to change things up? If the latter, consider taking a walk soon rather than eating.	2
Need to get something done but can't focus? A walk could help you clear your head so you can get things done more efficiently.	2
Are you feeling stuck about something? Consider taking a walk in the next few hours to clear your head and maybe get unstuck.	2
Can you make a plan to do a 10-minute walk in the next three hours?	1
Did you know regular activity is a key protective factor for dementia? Can you go for a walk in the next three hours?	1
Feeling frustrated? Would taking a short walk help you get some perspective?	2
Remember good being active can feel? Do you need that right now? If so, consider going for a walk soon.	2
Ate too much? Consider taking a walk to help you digest and not get sleepy.	2
Did you know regular activity helps you sleep better? Could you take a walk in the next few hours to help your sleep tonight?	1
Upset about something? Do you think taking a walk would help you calm down?	2
Can you see any gaps in your calendar in the next three hours when you may be able to take a 10+-minute walk?	1
Feeling like you can't get things done today? Consider clearing your head with a walk.	2
Trouble deciding on something? Consider letting your mind wander while walking. The answer just might come to you.	2
Feeling overwhelmed? Taking a walk may help you gain perspective and feel better.	2
Regular activity is an important part of keeping a healthy weight. Could you squeeze in a walk in the next few hours?	1
When the cravings come, consider short-circuiting them by getting away from food and going for a walk instead.	2
Feeling cold? Could a walk help warm you up?	2
Look at your calendar. See any times for a 10+ minute walk soon?	1
Look at your calendar. See any times for a 10+ minute walk soon?	1

Supplemental Table 3.4 Bout planning notification messages and their classifications into two categories of messages, continued

Messages	**Category**[1]
Can you think of anything beautiful near you, like nice scenery or architecture? Do you think you could walk over there soon?	1
Any upcoming meetings? Can any of them be walking meetings (including walking phone meetings)?	1
Are you going to be eating with a colleague, friend, or a family member in the next few hours? Could you take a post-meal walk together?	1
Do you have meetings in the next few hours? Could you end one 10 minutes early so you could go for a walk?	1
Will you be finishing a task in the next few hours? If so, could you go for a walk to reward yourself for a job well done (even if it's just answering email)?	1
Have you spent enough time with your family? How about taking a walk together (including a walking call) in the next few hours to catch up?	1
Thinking of watching TV soon? If so, could you replace a bit of that time with walking instead? We promise you you'll feel better.	1

[1]Categories:
1) messages designed to inspire participants to plan a time when they would walk in the next 3 hours
2) messages designed to invite participants to become aware of internal urges that could inspire them to walk

Chapter 4 ANALYSIS PROTOCOLS

Data preprocessing and cleaning

After the end of the clinical trial in April 2023, the entire database on the server where all measurements were stored backed up locally. The HeartSteps systems, built on the Django Web Framework [186], was organized in a way that divides and stores information in multiple relational database tables. We prepared the base data by extracting only the data we needed from the database, which consists of 195 tables in total.

We excluded the participants who 1) never turned on the app or 2) turned on the app but did not use it (i.e., non-adherent). The system logged the participants' activity with the local time along with time zone information. If the participants travelled across the border of the time zone, there was a chance that the person's that day could be longer than, or shorter than 24 hours. If the participant reported the same local time stamp more than once (i.e., travelled to west), we added the step count on the same time stamp. If the participant travelled to east, the time gap was handled as missing data.

Fidelity Check

We decided to check the distribution of the data, whether the key variables were distributed as we expected, and whether the experiment was conducted in a way consistent with our experimental design. We looked at the following:

1. The number of decision points per individual and contexts,
2. The number of notifications received per individual and contexts,
3. The number of steps taken by each participant per individual and contexts,
4. The total amount of time each participant wore their Fitbit per individual, and
5. The number of hours each participant wore their Fitbit per individual.

For detailed results, please see Appendix 5 on page 224.

Aim 1: Individual Response Patterns to Intervention

Aim 1. To identify which decision policy and time condition where the intervention was effective for each participant, analyzed via both nomothetic and idiographic models.

To delve into aim 1, we used two sets of statistical models.

Nomothetic models.

To extract a common pattern that covers the participants widely, we used frequentist mixed effects model with ZINB distribution (see page 36). The decision point was the unit of analysis. Within this analysis, we compared three meaningful nomothetic models:

1. **Nomothetic Null Model**: based on the null hypothesis that the response pattern to the intervention does not depend on any other covariates other than intervention components, the regression model only consisted of the intervention components and their interactions.

$$\Pr(steps = K) = ZINB(K; \psi, \mu, \sigma^2)$$

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{1,2} X_1 X_2 + \beta_{1,3} X_1 X_3 + b_1$$

steps: step count during 3 hours after each decision point
β_0: intercept (mean steps of the cases without intervention components)
X_1: Notification (provided = 1, not provided = 0)
X_2: Goal factor (daily step goal = median value of previous 26-day cycle + 4000 × goal factor, see page 52)
X_3: Dichotomized decision policy (Random = 0, Full, N+O, or N+R=1). In this nomothetic model, we hypothesized that the participants may walk more when they were provided the customized decision policy of any type.
b_1: Random effects for the participants' baseline step count. Each participant may represent their own base level of PA.

Special notes on terms:
- $\beta_{1,2} X_1 X_2$: the interaction effect between the notifications and goal factor. We hypothesized their synergistic effect. (i.e., $\beta_{1,2} > 0$)
- $\beta_{1,3} X_1 X_3$: the interaction effect between the notifications and decision policy. We hypothesized their synergistic effect.

2. **Nomothetic Full Model**: based on the theory-based hypothesis that the response pattern to the intervention may depend on all measured covariates, in addition to the effect

caused by intervention, the regression model consisted of the intervention components, all measured covariates, and their interactions. Also, based on the prior study [136], we included the quadratic term of days elapsed since the beginning of the intervention.

$$\Pr(steps = K) = ZINB(K; \psi, \mu, \sigma^2)$$

ψ, σ are assumed as constant for all participants.

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{4,4} X_4^2 + \beta_5 X_5 + \beta_6 X_6 + \beta_{1,2} X_1 X_2 + \beta_{1,3} X_1 X_3 + \beta_{1,4} X_1 X_4 + \beta_{1,5} X_1 X_5 + \beta_{1,6} X_1 X_6 + \beta_{2,3} X_2 X_3 + \beta_{2,4} X_2 X_4 + \beta_{2,5} X_2 X_5 + \beta_{2,6} X_2 X_6 + \beta_{5,6} X_5 X_6 + b_1 + b_2 X_4$$

steps: step count during 3 hours after each decision point
β_0: intercept (mean steps of the cases without intervention components)
X_1: Notification (provided = 1, not provided = 0)
X_2: Goal factor (daily step goal = median value of previous 26-day cycle + 4000 × goal factor, see page 52)
X_3: Dichotomized decision policy (Random = 0, Full, N+O, or N+R=1). In this nomothetic model, we hypothesized that the participants may walk more when they were provided the customized decision policy of any type.
X_4: days elapsed since the beginning of the intervention.
X_5: whether this decision point was in the afternoon (afternoon=1, morning=0)
X_6: whether this decision point was on the weekend (weekend=1, weekday=0)

b_1: Random effects for the participants' baseline step count. Each participant may represent their own base level of PA.

Special notes on terms:

- $\beta_4 X_4 + \beta_{4,4} X_4^2$: based on the prior study [136], we hypothesized the nomothetic change of steps in quadratic terms of days elapsed
- $\beta_{1,2} X_1 X_2$: the interaction effect between the notifications and goal factor. We hypothesized their synergistic effect. (i.e., $\beta_{1,2} > 0$)
- $\beta_{1,3} X_1 X_3$: the interaction effect between the notifications and decision policy. We hypothesized their synergistic effect.
- $\beta_{1,4} X_1 X_4$: the interaction effect between the notification and days elapsed since the beginning of the intervention. We hypothesized their synergistic effect.
- $\beta_{1,5} X_1 X_5$: the interaction effect between the notification and afternoon. We hypothesized the variation in the effect of notification between morning and afternoon, but without the direction. (i.e., $\beta_{1,5} \neq 0$)

- $\beta_{1,6}X_1X_6$: the interaction effect between the notification and weekend. We hypothesized the variation in the effect of notification between weekday and afternoon, but without the direction. (i.e., $\beta_{1,6} \neq 0$)
- $\beta_{2,3}X_2X_3$: the interaction effect between the goal factor and dichotomized decision policy. We hypothesized the variation in the effect of goal factor between the cases of JIT states considered and not, but without the direction. (i.e., $\beta_{2,3} \neq 0$)
- $\beta_{2,4}X_2X_4$: the interaction effect between the goal factor and days elapsed since the beginning of the intervention. We hypothesized their antagonistic effect. (i.e., $\beta_{2,4} < 0$)
- $\beta_{2,5}X_2X_5$: the interaction effect between the goal factor and the afternoon. We hypothesized the variation in the effect of goal factor between morning and afternoon, but without the direction. (i.e., $\beta_{2,5} \neq 0$)
- $\beta_{2,6}X_2X_6$: the interaction effect between the goal factor and the weekend. We hypothesized the variation in the effect of goal factor between weekday and weekend, but without the direction. (i.e., $\beta_{2,6} \neq 0$)
- $\beta_{5,6}X_5X_6$: the interaction effect between the weekend and afternoon. We hypothesized the variation in the step level differences between morning and afternoon on weekends, but without the direction. (i.e., $\beta_{5,6} \neq 0$)
- b_2X_4: Random effects for the participants' linear change of the step count over the study period. Each participant may represent their own slope of PA change.

3. **Nomothetic Stepwise Regression Model**: to explore the useful models that explains participants response pattern well enough, we used stepwise regression with forward selection [187]. Considering widely known criticism on stepwise regression including [188], we used this method as an auxiliary information to explore models between the hypothesis-based null and full model[189]. We used Akaike Information Criterion (AIC) [190]. From the null model, when a term that was not included in the model but included in the full model is newly introduced, if both of the metrics decreased, we decided to keep the term in the current model. We iterated this process until 1) no terms to further include left, or 2) no addition introduces the AIC decreases. Based on the guidelines by [187–190], we compared the selected model with theory-based models (full and null), along with a few final candidates that are not chosen in model selection.

Idiographic Models.

As a main triangulation trial to build the idiographic models to examine how each individual responded to the intervention under diverse JIT states, we *stratified* each user's data (A. idiographic null model). Then, we also conducted multiple, hierarchical analysis with further stratification by 1) time condition (B. idiographic time-based model), 2) decision policy (C. idiographic decision-policy-based model), and 3) time condition + decision policy (D. idiographic full model).

1. **Idiographic Null Model**: we firstly hypothesized that some individuals have their own response pattern regardless of decision policy (X_3), day elapsed since the beginning of the intervention (X_4) or time condition (X_5). Thus, we stratified the dataset by participant ID only, then controlled for X_3, X_4 and X_5. We focused on β_1 (notification's effect), β_2 (goal factor's effect), $\beta_{1,2}$ (interaction effect between notification and goal factor). We conducted a Bayesian Regression for all available participants, then visualized the results. (N x 1 table, N: number of participants) This model can be interpreted as follows: participant X had a general significant effect over 100 steps/3 hour increment of walking with over 80% of probability, after controlling for time condition, decision policy, and days elapsed since day 1.

$$\Pr(steps = K) = ZINB(K; \psi, \mu, \sigma^2)$$

Idiographic null model:

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_{1,2} X_1 X_2$$

Stratification: by participant ID only

steps: Step count during 3 hours after each decision point
β_0: Intercept (mean steps of the cases without intervention components)
X_1: Notification (provided = 1, not provided = 0)
X_2: Goal factor (daily step goal = median value of previous 26-day cycle + 4000 × goal factor, see page 52)

X_3: Categorical decision policy variable transformed into 3 dummy variables (reference: Random). In this nomothetic model, we hypothesized that the participants may walk more when they were provided the customized decision policy of any type.
X_4: Days elapsed since the beginning of the intervention. Based on the prior study [136], we hypothesized the individualized change of steps in quadratic terms of days elapsed. However, for the simplicity of the model, we only used the linear term.
X_5: Categorical time condition variable transformed into 3 dummy variables (reference: weekday morning)

Special notes on terms:

- $\beta_{1,2}X_1X_2$: The interaction effect between the notifications and goal factor. We hypothesized their synergistic effect. (i.e., $\beta_{1,2} > 0$)
- $\beta_3X_3, \beta_4X_4, \beta_5X_5$: Since we are most interested in β_1 (notification's effect), β_2 (goal factor's effect), $\beta_{1,2}$ (interaction effect between notification and goal factor), and we did not stratify by X_3 or X_5, we controlled for X_3, X_4 and X_5.

2. **Idiographic Time-based Model**: secondly, we hypothesized that some individuals have their own time-specific response pattern regardless of decision policy (X_3) and day elapsed since the beginning of the intervention (X_4). We did not hypothesize the direction of the effect variation. Thus, we stratified the dataset by participant ID and time condition, then controlled for X_3 and X_4. We focused on β_1 (notification's effect), β_2 (goal factor's effect), $\beta_{1,2}$ (interaction effect between notification and goal factor). We conducted a Bayesian Regression for all available participants, then visualized the results. (N x 4 table, N: number of participants) This model can be interpreted as follows: participant X had a significant effect over 100 steps/3 hour increment of walking with over 80% of probability for a specific time condition Y, after controlling for decision policy and days elapsed since day 1.

$$\Pr(steps = K) = ZINB(K; \psi, \mu, \sigma^2)$$

Idiographic null model:

$$\mu = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_{1,2}X_1X_2$$

Stratification: by participant ID and time condition X_5.

Steps: Step count during 3 hours after each decision point
β_0: Intercept (mean steps of the cases without intervention components)
X_1: Notification (provided = 1, not provided = 0)
X_2: Goal factor (daily step goal = median value of previous 26-day cycle + 4000 × goal factor, see page 52)
X_3: Categorical decision policy variable transformed into 3 dummy variables (reference: Random). In this nomothetic model, we hypothesized that the participants may walk more when they were provided the customized decision policy of any type.
X_4: Days elapsed since the beginning of the intervention. Based on the prior study [136], we hypothesized the individualized change of steps in quadratic terms of days elapsed. However, for the simplicity of the model, we only used the linear term.

Special notes on terms:

- $\beta_{1,2}X_1X_2$: The interaction effect between the notifications and goal factor. We hypothesized their synergistic effect. (i.e., $\beta_{1,2} > 0$)
- β_3X_3, β_4X_4: Since we are most interested in β_1 (notification's effect), β_2 (goal factor's effect), $\beta_{1,2}$ (interaction effect between notification and goal factor), and we did not stratify by X_3, we controlled for X_3 and X_4.

3. **Idiographic Decision-Policy-Based Model**: we hypothesized that some individuals have their own decision-policy-specific response pattern regardless of day elapsed since the beginning of the intervention (X_4) and time conditions (X_5). We did not hypothesize the direction of the effect variation. Thus, we stratified the dataset by participant ID and decision policy, then controlled for X_4 and X_5. We focused on β_1 (notification's effect), β_2 (goal factor's effect), $\beta_{1,2}$ (interaction effect between notification and goal factor). We conducted a Bayesian Regression for all available participants, then visualized the results. (N x 4 table, N: number of participants) This model can be interpreted as follows: participant X had a significant effect over 100 steps/3 hour increment of walking with over 80% of probability for a specific decision policy Z, after controlling for time condition and days elapsed since day 1.

$$\Pr(steps = K) = ZINB(K; \psi, \mu, \sigma^2)$$

Idiographic null model:

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_5 X_5 + \beta_{1,2} X_1 X_2$$

Stratification: by participant ID and time condition X_5.

Steps: Step count during 3 hours after each decision point
β_0: Intercept (mean steps of the cases without intervention components)
X_1: Notification (provided = 1, not provided = 0)
X_2: Goal factor (daily step goal = median value of previous 26-day cycle + 4000 × goal factor, see page 52)
X_4: Days elapsed since the beginning of the intervention. Based on the prior study [136], we hypothesized the individualized change of steps in quadratic terms of days elapsed. However, for the simplicity of the model, we only used the linear term.
X_5: Categorical time condition variable transformed into 3 dummy variables (reference: weekday morning)

Special notes on terms:

- $\beta_{1,2} X_1 X_2$: The interaction effect between the notifications and goal factor. We hypothesized their synergistic effect. (i.e., $\beta_{1,2} > 0$)
- $\beta_4 X_4, \beta_5 X_5$: Since we are most interested in β_1 (notification's effect), β_2 (goal factor's effect), $\beta_{1,2}$ (interaction effect between notification and goal factor), and we did not stratify by X_5, we controlled for X_4 and X_5.

4. **Idiographic Full Model**: we hypothesized that some individuals have their own time and decision-policy-specific response pattern regardless of day elapsed since the beginning of the intervention (X_4). We did not hypothesize the direction of the effect variation. Thus, we stratified the dataset by participant ID, time condition, and decision policy, then controlled for X_4. We focused on β_1 (notification's effect), β_2 (goal factor's effect), $\beta_{1,2}$ (interaction effect between notification and goal factor). We conducted a Bayesian Regression for all available participants, then visualized the results. (N x 16 table, N: number of participants) This model can be interpreted as follows: participant X had a significant effect over 100 steps/3 hour increment of walking with over 80% of probability for a specific time condition Y and decision policy Z, after controlling for days elapsed since day 1.

$$\Pr(steps = K) = ZINB(K; \psi, \mu, \sigma^2)$$

Idiographic null model:

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_{1,2} X_1 X_2$$

Stratification: by participant ID, decision policy X_3, and time condition X_5.

Steps: Step count during 3 hours after each decision point
β_0: Intercept (mean steps of the cases without intervention components)
X_1: Notification (provided = 1, not provided = 0)
X_2: Goal factor (daily step goal = median value of previous 26-day cycle + 4000 × goal factor, see page 52)
X_4: Days elapsed since the beginning of the intervention. Based on the prior study [136], we hypothesized the individualized change of steps in quadratic terms of days elapsed. However, for the simplicity of the model, we only used the linear term.

Special notes on terms:

- $\beta_{1,2} X_1 X_2$: The interaction effect between the notifications and goal factor. We hypothesized their synergistic effect. (i.e., $\beta_{1,2} > 0$)
- $\beta_4 X_4$: Since we are most interested in β_1 (notification's effect), β_2 (goal factor's effect), $\beta_{1,2}$ (interaction effect between notification and goal factor), and we stratified by X_3 and X_5, we controlled for only X_4.

The decision point was the unit of analysis. For the Bayesian Regression, we used MCMC (see page 37 for technical details about MCMC). Four sampling chains were used when performing Bayesian modeling. The number of warm-up samples per sampling chain was 1000, the number of estimation samples was initially set as 500 (i.e., the total number of estimation samples was $(1000 + 500) \times 4 = 6000$), and the target acceptance rate was set to 0.8 as an initial value. The number of estimation samples and target acceptance rate were gradually increased until numerical stability was achieved. The number of estimation samples was increased by multiplied by 2, as advised by [191], depending on the ratio of convergence diagnostics $\hat{R}$ > 1.1. The target acceptance rate was increased at each step according to the following formula:

$$t_{i+1} = 1 - \frac{1 - t_i}{r}$$

t_i: i-th target acceptance rate ($i \geq 0$)
r: division factor, we used 2.

The conditions for terminating the iteration were as follows:

1. The convergence diagnostics $\hat{R}$ [191] of all variables did not exceed[13] 1.1, AND more than 95% of the samples converged[14],
2. The target acceptance rate for MCMC sampling exceeded 99.99%[15] or
3. The sampling operation did not terminate within 1 hour (most sampling operations terminated in 4-6 minutes).

For the condition 3, we allowed a couple more hours as a buffer.

We used 100 steps increase during 3 hours as the effect threshold value (see page 38 for details). We assumed that there is an effect only if the estimated effect is more than 80% probable (i.e., the credibility interval is over 100 steps/3 hours) and the Maximum A Posteriori Point (MAP) of the effect is more than 100 steps/3 hours. Otherwise, there is no effect. Among the models with effects, we clipped to 1000 steps/3 hours as a maximum if the MAP was greater than 1000 steps/3 hours.

[13] It means that the individual dimensions (i.e., variables) converge for the majority of the samples.
[14] It means that all dimensions of the sample converge when considered together.
[15] It means that too high target acceptance rate makes the sampling process extremely inefficient. The usual range is 80-95%.

Aim 2: Examining the Distribution of Individual Response Patterns

Aim 2. To identify similarities in response patterns across participants within the same time condition.

This section aims to examine the distribution of the individual response patterns across individuals x decision policies that we performed in Aim 1 for each time condition. First, for each time condition (e.g., weekday morning), we dichotomized the responses to the four decision policies across the 44 individuals. The reasoning of dichotomization was that, since our goal was to develop the control systems, it is less critical what MAP value each individual had for each context than whether they had a meaningful effect in that context. Thus, all effects were coded as either 0 (no effect) or 1 (did have an effect).

First, we grouped the response patterns distributions into 16 groups (=2^4, from no effect in any decision policy to effect in all decision policy). Then, as a hypothesis testing (see page 46), we compared the distributions with uniform distribution with Chi-square test, to test the null hypothesis that the distributions are not significantly different from the uniform distribution. (Hypothesis 2-1)

If the distribution is significantly different from the uniform distribution, we tested the distribution is significantly different from the binomial distribution of the same probability of being effective. For each time condition, a Monte Carlo (MC) simulation was conducted to test it. Average probability of being effective was calculated by dividing the total number of effective states by the total number of states (e.g., 4×44). This average probability was used as binomial distribution parameter, with the assumption of the being effective for each state was independent, to simulate the hypothetical effect response patterns across 44 participants and 4 decision policies. Then, the hypothetical effect response patterns were grouped as we did with

the observed data, and the number of groups were counted. This experiment was conducted 1,000 times to calculate p-value of the observed group count. The portion of the simulations with the numbers of groups equal to or smaller than the observed was used to report the p-value.

Exploratory Aim 1: Discovery of Individual Response Pattern Using Machine Learning

> **Exploratory Aim 1.** To explore the potential value of the multilayer perceptron algorithm in identifying which decision policy and time condition where the intervention was effective for each participant.

Overview

Exploratory Aim 1 is an attempt to replicate the research questions and answers from Aim 1 with machine learning (specifically, multilayer perceptron, MLP). If similar patterns can be detected with slightly different methodologies (i.e., Bayesian modeling and MLP), it would provide additional evidence that we have detected meaningful patterns. It would also be an indirect anecdotal observation that supports the claim that there may be relational validity between the two methodologies.

Methodologically, this exploratory analysis is a natural extension of the opportunity condition operationalization study included in Appendix 1. The study in Appendix 1 averaged across individuals (i.e., nomothetic), took as input hourly activity over the past five weeks, and predicted predictive activity over the next three hours with 82% accuracy.

What made this exploratory analysis different from the studies in Appendix 1 was that it considered intervention factors and context, that it aimed for an idiographic model by building an individual-level model, and that the target variable was the exact number of steps taken over a three-hour period, rather than whether or not they were active.

We included as input variables whether the intervention component was provided or their dose, the number of days elapsed since the beginning of the intervention, and the context of the time condition. We also built idiographic models using the stratified data for each individual. Finally, since step counts follow a ZINB distribution (i.e., a count distribution with a large fraction of zeros), special consideration was needed for this.

Table 4.1 Modeling specification for exploratory aim 1

Input Variables - Whether or not the notification was provided (dichotomous, not sent: -1, sent: 1) - decision policy (4 levels, dummy variable, each category is positive: +1, otherwise: -1) - time condition (4 levels, dummy variable, each category is positive: +1, otherwise: -1) - days elapsed since the beginning of intervention (day0: -1, day 259: +1) **Output Variables** - steps during 3 hours after each decision point (divided by 10,000 for faster training) **Stratification**: Per Participant **Optimizer**: Adam algorithm [192] with the fixed learning rate (the rate will be determined by hyperparameter search) **Validation Set**: Among the training data, the last 20% of samples with respect to the days elapsed were used to 1) detect overfitting, 2) evaluate the model during the training, 3) determine the training should continue or halted, and 4) evaluate the score of the hyperparameters to decide the next set of hyperparameters

Evaluation

For evaluation, we conducted a modified K-Fold Cross Validation for time series data [193]. For time series data, the traditional K-Fold Cross Validation [193,194] may lead to *look-ahead bias* [193]. Look-ahead bias refers to the performance inflation that can occur when information that was not yet available at the time of the prediction(e.g., future information), is brought into the past to make a prediction[193]. To avoid this, it is appropriate to evaluate performance based on realistic assumptions, where the time series data is broken into steps from the front, and the data before a certain point in time is used to predict the data after that point. In

this study, we evaluated a total of 260 days of data by training on the first 80, 100, 120, ..., 180 days and testing the performance on the remaining 180, 160, 140, ..., 80 days.

While this is a good way to evaluate time-series model performance without look-ahead bias, we can also expect it to help give us a rough answer to how much data we need to achieve our desired performance. For example, if we see that for less than 140 days, the performance is significantly short of what we expect, but after 140 days, if the performance is in line with our expectations, we can know that our model needs at least 140 days of data.

Consideration of Zero-inflated Step Distribution

Zero-inflated count distributions have excessive zeros, making ordinary loss functions (e.g., mean-squared errors) behave inadequately. Using MSE in the traditional way, if the zeros are properly fitted, they do not train well for other non-zero values because 1) the non-zeros are minority, so they are not well sampled, and 2) it is not easy for the optimizer to skip barriers because even a small deviation from zero increases the loss as the square of the value.

Recognizing this problem, as a methodological preliminary to this study, we conducted a number of methodological attempts to develop MLPs for zero-inflated step distributions, and we adopted the method that showed the best predictive performance.

Two-stage model

The model consisted of two main parts: Zero Part and Value Part. The Zero parts was only trained about whether the output will be zero or not based on the input. The output channel of the training data (i.e., steps) was mapped to 0 and 1 (i.e., whether the individual walked at least one step or not). The model used a binary cross-entropy loss function and was trained solely on 0 and 1.

The next step, the Value Part, was trained on what the value will be, assuming the output is non-zero. We implemented this by eliminating samples in the training data that have an output

channel of zero (i.e., excluding decision points when the participant did not walk at all), and only get trained on samples that have a positive value. After eliminating the 0-valued samples, we used Mean Squared Error as the loss function for the value part. To account for the activation function of the hidden layer (e.g., hyperbolic tangent function with the range of (-1, 1) or similar restricted range), we used the number of steps divided by 10000 for fast training. Later, after the training is over, when we run the predictions, we multiply the results by 10000.

After the training was over, the paired models were merged into a single model. First, the zero part was applied to the input to identify samples that will be zero, and then the value part was applied again to the non-zero input samples to determine their value. In our implementation, we ran the predictions by both models on the entire input, applied a 0.5 threshold to the zero part model, and then computed the predicted value by pairwise multiplication.

This method mimics the construction of ZINB (binomial distribution + negative binomial distribution). The binomial distribution part is simulated using a binary cross-entropy loss function, and the negative binomial distribution part is simulated using a mean-squared error loss function. And the order of their application is preserved, applying the zero part first (i.e., giving it numerical superiority through multiplication), followed by the value part.

Modeling in this way, even with the best computing resources we have, about 2.5-3 hours were required to train on one participant's data (including modified K-fold cross-validation and hyperparameter search). Therefore, in the results chapter, we presented only a few example outcome data from a few participants.

Hyperparameter Searching

We utilized Ray Tune and ASHA [132,133] to derive the optimal hyperparameters for each of the zero and value parts. The variables included in the search space were shown below.

Table 4.2 Hyperparameters to search for each model components.

1. the number of hidden layers 2. the size of each hidden layer 3. activation functions of the hidden layers 4. whether to include a dropout layer and dropout rate 5. whether to include a normalization layer 6. weight and bias initializing function of the hidden layer 7. batch size for training 8. learning rate

Comparison with the Null Model

The predictive models were to be compared to the null models, frequentist's linear regression and ZINB models. Then the Diebold-Mariano tests [195,196] followed to test whether the machine learning model provides significantly better prediction accuracy measured by MSE than the null model.

Simulation

Assuming that the models are of adequate quality, the developed models can be used to simulate how many steps each individual will take when exposed to the hypothetical situations. We briefly showed how many steps this person (an example participant) would take as the notification and step goal factors change, under a certain time condition and decision policy.

Exploratory Aim 2: *post hoc* Analysis of JIT states

Exploratory Aim 2. To examine the relationship between real time and post hoc states, and the impact on identified individual response patterns on the notifications per the time condition on aim 1.

This section aims to explore the impact of step data that was not delivered at the decision point due to technical limitations (see page 29).

To estimate the impact of delayed sync of step data, we re-estimated the JIT states with the post hoc step data (i.e., after long time passed from the decision point, given we attained all the dataset from the participant), at the end of the trial. Then we built the predictive models for the participants' responses to the intervention per time condition.

The unit of analysis was the decision point; the independent variables are whether notification was provided, goal factor, and their interaction. Also, the model controlled for or stratified by *post hoc* JIT states (27 exhaustive cases ($= 3 \times 3 \times 3$), e.g., N+, N-, O+, O-, …, N+/O+, N+/O-, …. See Table 2.2 for the full list) and controlled for days elapsed since the beginning of the intervention. The dependent variable is the number of steps taken in the 3 hours after each decision point. The model was a Zero-Inflated Negative Binomial (see page 36), and Bayesian Regression using MCMC was performed.

We hypothesized that some individuals have their own response pattern that are associated with time condition and *post hoc* JIT state after controlling for days elapsed since the beginning of the intervention (X_4). We did not hypothesize the direction of the effect variation.

This model can be interpreted as follows: participant X had a significant effect over 100 steps/3 hour increment of walking with over 80% of probability for a specific time condition Y and decision policy Z, after controlling for days elapsed since day 1.

$$\Pr(steps = K) = ZINB(K; \psi, \mu, \sigma^2)$$

Idiographic null model:

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_{1,2} X_1 X_2 + \sum \beta_{6,i} X_{6,i}$$

Stratification: by participant ID, post hoc JIT states X_6, and time condition X_5.

Steps: Step count during 3 hours after each decision point
β_0: Intercept (mean steps of the cases without intervention components)
X_1: Notification (provided = 1, not provided = 0)
X_2: Goal factor (daily step goal = median value of previous 26-day cycle + 4000 × goal factor, see page 52)
X_4: Days elapsed since the beginning of the intervention. Based on the prior study [136], we hypothesized the individualized change of steps in quadratic terms of days elapsed. However, for the simplicity of the model, we only used the linear term.
$X_{6,i}$: Three JIT state variables (N, O, R). Depending on the configuration denoted in Table 2.2, they can be used as stratification conditions, or controlling variables. See page 62.Table 2.2

Special notes on terms:

- $\beta_{1,2} X_1 X_2$: The interaction effect between the notifications and goal factor. We hypothesized their synergistic effect. (i.e., $\beta_{1,2} > 0$)
- $\beta_4 X_4$: Since we are most interested in β_1 (notification's effect), β_2 (goal factor's effect), $\beta_{1,2}$ (interaction effect between notification and goal factor), and we stratified by X_3 and X_5, we controlled for only X_4.

The following details were the same as aim 1. Four sampling chains were used when performing Bayesian modeling. The number of warm-up samples per sampling chain was 1000, the number of estimation samples was initially set as 500 (i.e., the total number of estimation samples was $(1000 + 500) \times 4 = 6000$), and the target acceptance rate was set to 0.8 as an initial value. The number of estimation samples and target acceptance rate were gradually increased until numerical stability was achieved. The number of estimation samples was increased by multiplied by 2, as advised by [191], depending on the ratio of convergence diagnostics $\hat{R}$ > 1.1. The target acceptance rate was increased at each step according to the following formula:

$$t_{i+1} = 1 - \frac{1 - t_i}{r}$$

t_i: i-th target acceptance rate ($i \geq 0$)
r: division factor, we used 2.

The conditions for terminating the iteration were as follows:

1. The convergence diagnostics $\hat{R}$ [191] of all variables did not exceed[16] 1.1, AND more than 95% of the samples converged[17],
2. The target acceptance rate for MCMC sampling exceeded 99.99%[18] or
3. The sampling operation did not terminate within 1 hour (most sampling operations terminated in 4-6 minutes).

For the condition 3, we allowed a couple more hours as a buffer.

We used 100 steps increase during 3 hours as the effect threshold value (see page 38 for details). We assumed that there is an effect only if the estimated effect is more than 80% probability (i.e., the credibility interval is over 100 steps/3 hours) and the Maximum A Posteriori Point (MAP) of the effect is more than 100 steps/3 hours. Otherwise, there is no effect. Among the models with effects, we clipped to 1,000 steps/3 hours as a maximum if the MAP was greater than 1,000 steps/3 hours, for the visualization purposes.

Acknowledgments

Chapter 1, 2, 4, 5, 6, and 7 of this thesis, in part, are currently being prepared for submission for publication of the material. Park, Junghwan; Kim, Meelim; El Mistiri, Mohamed; Kha, Rachael; Banerjee, Sarasij; Gotzian, Lisa; Chevance, Guillaume; Rivera, Daniel

[16] It means that the individual dimensions (i.e., variables) converge for the majority of the samples.
[17] It means that all dimensions of the sample converge when considered together.
[18] It means that too high target acceptance rate makes the sampling process extremely inefficient. The usual range is 80-95%.

E.; Klasnja, Predrag; Hekler, Eric. The dissertation author was the primary researcher and author of this material.

Chapter 5 RESULTS

Summary

As expected, results of our nomothetic statistical analyses suggested that our intervention strategies as implemented, on average and across the population, were not effective at producing significant increases in steps/3 hours. These results, which could be thought of as akin to a multiphase optimization trial screening experiment to examine the usefulness of intervention components, suggests that, as implemented, our intervention components, as they were delivered, should not be used as an optimized intervention package. This includes our general theorized approach for defining JIT states via experimentally varying a decision policy that either considered need, opportunity, and receptivity when sending notifications or delivered notifications at random as the notification × decision policy interaction within the nomothetic statistical analyses was non-significant. In line with our a priori hypothesis, using idiographic Bayesian statistics, we found that it was feasible to identify individualized states whereby individuals would reliably increase steps/3 hours post support (compared to no support given in the same state relevant for each of our three intervention variations described earlier. Specifically, we found that we could identify at least one JIT state for 91% (40/44) of participants with sufficient data (83% using an intent to treat approach, 40/48). The pooled effect size of the interventions impact was an increase of 372 steps/3 hours relative to the appropriate comparator for each intervention strategy described above within the same state, which is a general effect size of .62 (normalized mean difference; the ratio between effect and baseline SD). Given that this estimate is for a non-normative targeted timescale of steps/3 hours, we calculated and inferred likely steps/day effects that would be observed when the "right" intervention support is provided for a person at the "right time". The inferred daily effect was an increase of 1,486 steps/day (effect size=0.62, normalized mean difference; ratio between effect

and baseline SD). Results from our secondary analyses generally revealed limited capacity to identify meaningful clusters of types of people responding in similar ways. Further, results from the machine learning analyses generally suggested that a machine learning approach produced limited, yet promising informative insights for guiding further intervention optimization. Finally, exploratory simulation analyses suggested that, if we allowed our decision policies to vary need, opportunity, and receptivity independently, it is likely that we would have identified successful JIT states for 43 out of 44 (98%) of our participants. Full details justifying these key "take home messages" provided in the remainder of this chapter.

Recruitment and Enrollment

Enrollment began in March 2022 and ended in July 2022. In total, 761 potential participants submitted a letter of interest, and 48 (6.3%) were enrolled in the study (with the majority not recruited, via a first-come-first served basis). Figure 5.1 shows the CONSORT (Consolidated Standards of Reporting Trials) diagram [179]. The intervention was completed in April 2023. The data were gathered without major incidents including no reported adverse events from study participants.

Among the 48 participants, 4 participants did not provide sufficient data to enable analyses. This lack of data occurred immediately for 3 of the 4 non-responding participants and 23 days into the study for the 4th participant. Specifically, one participant did finish the pre-intervention meeting, but did not turn on the app. Since the Fitbit account is connected during the onboarding process after turning on the app, the Fitbit account was never connected and thus, no data were available due to technical reasons. Another participant did turn on the app, connected the Fitbit account to study server, but did not wear the Fitbit, during the study period. We did reach out to the participants, but they did not respond. Thus, we had a sample the final data are available from 44 participants.

Based on this, we offer two denominators for determining the percentage of participants that we could identify predictable pattern, a conservative intent-to-treat approach (N=48) and providing enough data (i.e., more than 23 days of Fitbit data; N=44).

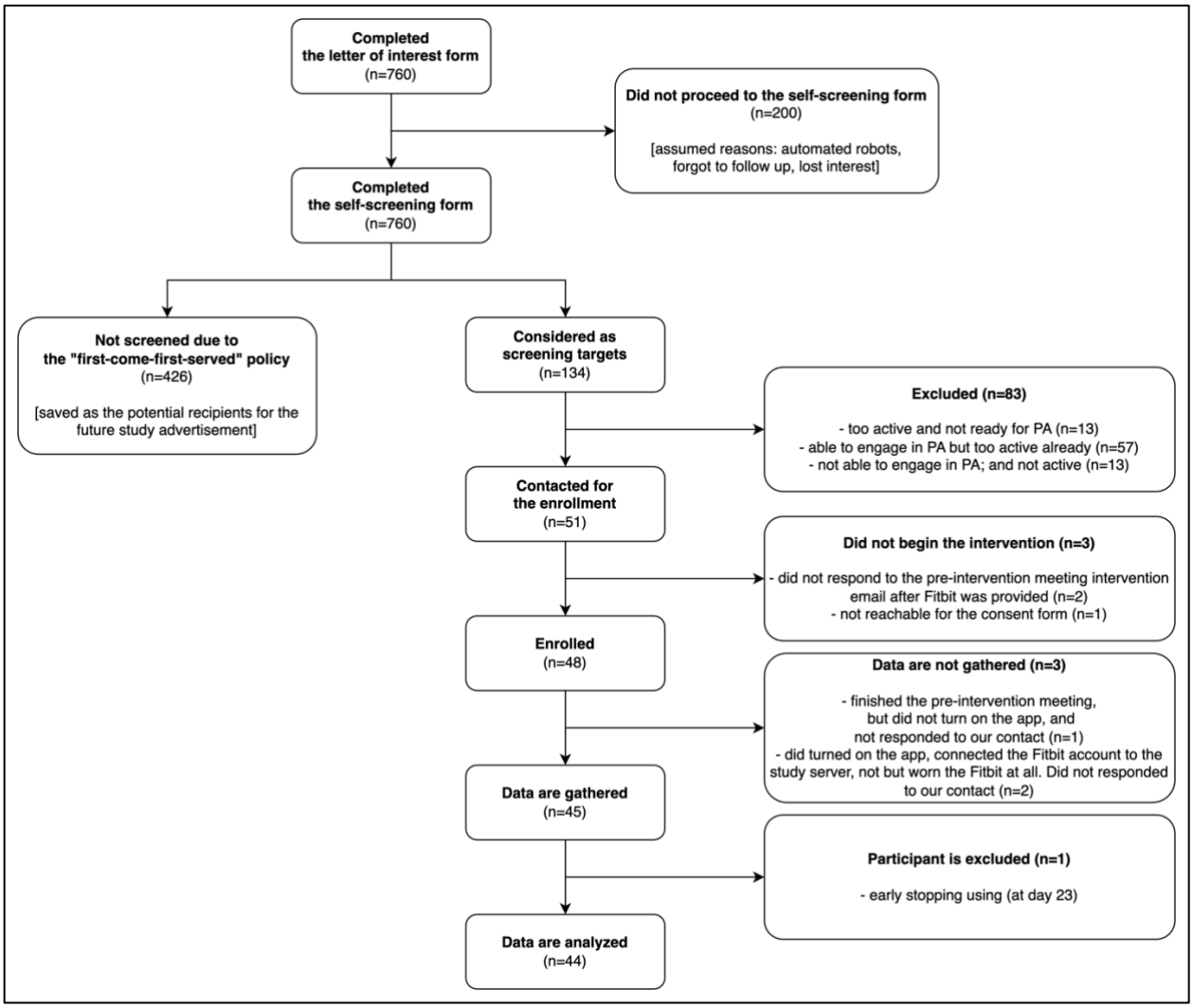


Figure 5.1 CONSORT (Consolidated Standards of Reporting Trials) recruitment diagram for the JustWalk JITAI study.[19]

[19] PA: physical activity. This figure is an updated version of Figure 3.8.

Participant Characteristics and Baseline

Descriptive information about the participants is presented in Table 5.1. 67% of the participants responded that they were female. Over 40% of the participants were between 25 and 35. Average weight and height were 86.0 kg and 172.1 cm, respectively. 16% of participants identified as Hispanic/Latino. Individuals who identified as Asian or White each was 30%+ of the study sample. Over 70% of the participants were in a relationship. Over 93% of the participants lived with someone else in the household, and 47% lived with children up to 12. Only one participant responded as "unemployed," while all others responded as some employment. Three participants responded as "some college or 2-year college." Potentially due to the recruitment source (e.g., university employee mailing lists), the demographics did not reflect the national population census.

Table 5.1 Characteristics of the participants

Sample characteristic	N (%) (n=44)	Mean (SD[1])	Missing[2]	N (%) (n=48)	Mean (SD[1])	Missing[2]
Gender			1		-	3
Female	29 (67%)			29 (64%)		
Male	14 (33%)			16 (36%)		
Age (years)		38.4	0		38.4	2
25-35	19 (43%)	(9.5)		20 (43%)	(9.3)	
35-45	14 (32%)			15 (33%)		
45-55	6 (14%)			6 (13%)		
55-65	5 (11%)			5 (11%)		
Weight (kg[1])		85.6	1		86.2	3
40-50	1 (2%)	(22.0)		1 (2%)	(21.7)	
50-60	4 (9%)			4 (9%)		
60-70	8 (18%)			9 (20%)		
70-80	6 (14%)			6 (13%)		
80-90	6 (14%)			6 (13%)		
90-100	7 (16%)			7 (15%)		
100-110	7 (16%)			8 (17%)		
110-120	1 (2%)			1 (2%)		
≥120	3 (7%)			3 (7%)		
Height (cm[1])		172.0	0		172.4	2
150-160	4 (9%)	(16.0)		4 (9%)	(15.8)	
160-170	16 (36%)			16 (35%)		
170-180	16 (36%)			18 (39%)		
180-190	6 (14%)			6 (13%)		
190-200	1 (2%)			1 (2%)		
200-210	0 (0%)			0 (0%)		
≥210	1 (2%)			1 (2%)		
Hispanic or Latinx		-	1		-	3
Yes	7 (16%)			8 (18%)		
No	36 (84%)			37 (82%)		

Table 5.1 Characteristics of the participants, continued

Sample characteristic	N (%) (n=44)	Mean (SD[1])	Missing[2]	N (%) (n=48)	Mean (SD[1])	Missing[2]
Employment[3]		-	0		-	2
Full-time (including self-employed)	38 (86%)			40 (87%)		
Part-time (including self-employed)	2 (5%)			2 (4%)		
In school or vocational training	3 (7%)			3 (7%)		
In school or vocational training, unemployed, laid-off or looking for work	1 (2%)			1 (2%)		
Highest level of education		-	0		-	2
Some college or 2-year degree	2 (5%)			3 (7%)		
College graduate	21 (48%)			21 (46%)		
Some graduate school	1 (2%)			1 (2%)		
Graduate Degree	20 (45%)			21 (46%)		

[1] Abbreviation: SD, Standard Deviation; kg, kilograms; cm, centimeters.
[2] All items were presented in the survey as optional. Unanswered items are marked as "missing."
[3] These items allowed multiple selections. The responses are concatenated with commas.

Fidelity Checks

Data checks on the distribution of the data were conducted to determine whether the key variables were distributed as we expected, and whether the experiment was conducted in a way consistent with our experimental design. In summary, our fidelity checks confirmed that our experiment was conducted with sufficient fidelity to enable interpretable results. For detailed results, please see Appendix 5 on page 224.

Aim 1: Individual Response Patterns to Intervention

Organization of This Section

This section describes the results of analyses for Aim 1. The "Overview" subsection describes all the results in summary. Then, a series of incrementally developing models from "Nomothetic Models", "Idiographic Null Model", "Idiographic Time-based Model", "Idiographic Decision-Policy-Based Model", and lastly, "Idiographic Full Model" are discussed in each subsection.

This organization is used to identify the potential additional benefit of identifying predictable JIT states for each person as we include more elements of our theorized 16 paired

INUS condition pairs and to provide the "take home message" first, following by a detailed justification of this core message via the results of each sub-set analysis.

Overview

The core findings for the study are summarized in Table 5.2 first on page 135. Since the nomothetic models were not intended to capture idiosyncratic responses, they were not included in Table 5.2.

Each column denotes model types, with a progressive model building approach whereby additional variables are included incrementally. Each row denotes different sources of intervention effects (i.e., notifications, goal factor, the interaction between them, and the interaction between notification and our decision policy), and their summation.

Table 5.2 Summary of the effects of intervention components in idiographic models

Metric	Idiographic Null Model	Idiographic Time-based Model	Idiographic Decision-Policy-based Model	Idiographic Full Model
Total number of participants with a predictive JIT state (n, %)	8 (18)	29 (66)	32 (73)	40 (91)
Total number of participants with predictive JIT state for notifications (n, %)	1 (2)	12 (27)	16 (36)	25 (57)
Total number of participants with a predictive JIT state related to goals (n, %)	2 (5)	12 (27)	16 (36)	23 (52)
Total number of participants with a predictive JIT state related to interaction between notification and goal factor (n, %)	5 (11)	17 (39)	17 (39)	30 (68)
Average effect of all types of intervention components[a] (steps/3 hours)	212	422	457	598
Average effect of notification only (steps/3 hours)	262	435	509	667
Average effect of daily step goal only (steps/3 hours)	180	403	395	621
Average effect of both notification and daily step goals[b] (steps/3 hours)	214	426	473	522

a. Average of all cases that are effective (notification only, daily step goal only, and both components are provided)
b. Average of the cases that are effective only by both components, not individual components alone

Nomothetic Models

Table 5.3 shows the regression results of the nomothetic null model, nomothetic full model, and nomothetic stepwise regression model. Comparing the null model, full model and three stepwise models, we find:

1. The effects of notification and goal factors alone were not apparent in average (i.e., nomothetic) models. In the majority of models, the association between the notification or decision policy with the step count was not significant. The association between goal factors and step count was significant, but in a negative way (i.e., if we provide more ambitious goals, the participants tend to walk less, on average, which fits with prior work [197]).

2. In all models, the interaction between notification and goal factor was significant ($p<0.001$), and the effect was positive (186-190). This means that, on average, participants walked 186 steps during 3 hours after the decision points when they were given high goals (i.e., goal factor = 1) and a notification.

By introducing stepwise model building, we could not find any new information other than slight variations of effect estimates, but it was useful to check the results across a meaningful number of alternative regression models with the low AIC value.

Table 5.3 Summary of the regression results of nomothetic models

	Null model	Full model	Stepwise model 1	Stepwise model 2	Stepwise model 3
Model attributes					
N	40,690	40,668	40,674	40,675	40,676
Random Effect's Variance					
Variance of intercepts	0.3	0.3	0.3	0.3	0.3
Variance of estimates of effect of day elapsed		0.0	0.0	0.0	0.0
Modeling results					
AIC	635,329.0	634,914.2	634,904.5	634,906.4	634,909.0
BIC	635,397.9	635,103.7	635,042.4	635,035.6	635,029.6
Dispersion parameter	1.7	1.7	1.7	1.7	1.7
Estimates for effects (+: step gain, -: step loss)					
Intercept	1,063***	1,055***	1,055***	1,061***	1050***
Notification	-83*	-46	-55	-76	-74
Goal factor	-48*	-54	-50*	-50*	-50*
Dichotomized decision policy	0	16	12	11	12
Days elapsed since day 1		-1**	-1**	-1**	-1**
Quadratic term of days elapsed since day 1		0**	0**	0**	0**
Afternoon		**78****	**88*****	**89*****	**110*****
Weekend		**76***	**86****	**65*****	**102*****
Interactions between notification and					
goal factor	**189*****	**186*****	**189*****	**190*****	**188*****
decision policy	12	-3	-4	-4	-6
days elapsed		0			
Weekend		-70*	-70*		
afternoon		37			
Interactions between weekend and afternoon		73*	73*	72*	
Interactions between goal factor and					
weekend		19			
afternoon		0			
days elapsed		0			
decision policy		9			

[1] Significance codes. ***: p-value < 0.001, **: p-value < 0.01, *: p-value < 0.05

Idiographic Null Model

Figure 5.2 shows the visual representation of the results of the Bayesian regression from the idiographic null model that tests individual responses for each person with comparison made between notifications sent or not, high vs. low goal factors, and the interaction between notifications and high goals compared to low goals and no notification, after controlling for time conditions and decision policies.

Each cell represents one participant. The cells are colored to indicate whether the intervention was effective for the participant in general. White indicates no or weak effect (i.e., Maximum A Posteriori Point (MAP) Effect estimates of 100 steps/3 hours or less, or no more than an 80% chance of effect exceeding 100 steps/3 hours), black indicates MAP Effect estimates exceeding 1000 steps/3 hours with an 80% chance of effect exceeding 100 steps/3 hours. In between, gray indicates MAP Effect estimates between 100 and 1000 steps/3 hours (gray cells are only shown if at least an 80% chance of effect exceeding 100 steps/3 hours). In other words, a completely black cell means that for that cell (i.e., specific time of day, specific intervention strategy), participants would probably take about 1000 more steps during 3 hours when they received the in-app notification, or a high step goal (i.e., goal factor = 1), compared to when they did not receive the notification, or received a low step goal (i.e., goal factor = 0). Participants were placed in order of recruitment date.

The numbers in the cell denote the effect estimates of each intervention component. "Noti" means the effect of the notifications, "Goal" means the effect of daily step goals (i.e., goal factor), and "N+G" means the joint effect of the notifications and daily step goals.

Participant ID	Effect estimates	Participant ID	Effect estimates	Participant ID	Effect estimates
1		16		31	Noti: 262
2		17	N+G: 354	32	
3		18		33	
4		19		34	
5		20		35	
6		21		36	
7		22		37	N+G: 149
8		23		38	N+G: 173
9	Goal: 170	24	N+G: 229	39	
10		25		40	
11		26		41	
12	N+G: 165	27		42	
13		28		43	
14		29		44	
15	Goal: 190	30			

1. Noti: notification effect's MAP value
2. Goal: step goal factor effect's MAP value
3. N+G: a predicted effect if both notification and high goal (i.e., goal factor of 1) were given

Figure 5.2 Summary of effect of interventions estimated by idiographic null models.

Among 44 participants, overall, 8 participants (18%) had a predictive intervention component identified. These overall predictive intervention components were identified in relation to providing a notification (vs. no notification) alone for one participant (2%), two participants (5%) were supported by being provided a high goal factor (vs a low goal factor) alone, and 5 participants (11%) were supported when a notification and high goal factor was present (compared to no notification and a low goal).

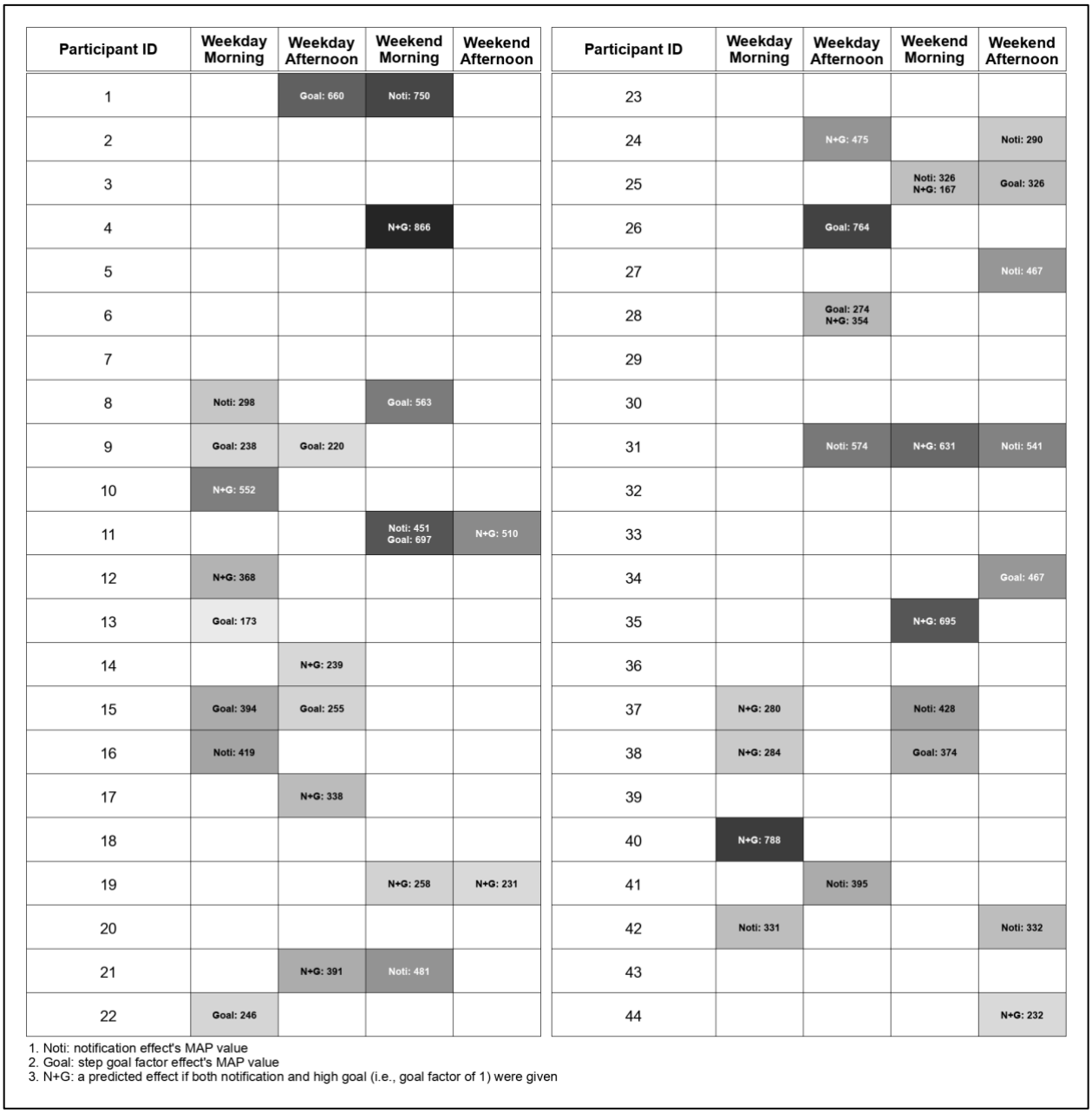
The overall average MAP across all intervention components was 212 steps/3 hours, meaning that, when the "right" support is offered, people would walk, on average over the 9 month period, 212 steps/3 hours more compared to when an intervention was not provided. The average MAP effect of each intervention component was 262, 180, and 214 steps/3 hours for notification only, daily step goal only, and both, respectively.

Idiographic Time-based Model

Figure 5.3 shows the visual representation of the results of the Bayesian regression from the idiographic time-based model that tests our hypothesized 4 time conditions for each person with comparison made between notifications sent or not, high vs. low goal factors, and the interaction between notifications and high goals compared to low goals and no notification, after controlling for decision policies. The regression was conducted for each participant and stratified by 4 time conditions (weekday morning, weekday afternoon, weekend morning, weekend afternoon).

Each cell represents one time condition, and 4 cells in a row represents a participant. The cells are colored to indicate whether the intervention was effective for the participant for that time condition. White indicates no or weak effect (i.e., Maximum A Posteriori Point (MAP) Effect estimates of 100 steps/3 hours or less, or no more than an 80% chance of effect exceeding 100 steps/3 hours), black indicates MAP Effect estimates exceeding 1000 steps/3 hours with an 80% chance of effect exceeding 100 steps/3 hours. In between, gray indicates MAP Effect estimates between 100 and 1000 steps/3 hours (gray cells are only shown if at least an 80% chance of effect exceeding 100 steps/3 hours). In other words, a completely black cell means that for that cell (i.e., specific time of day, specific intervention strategy), participants would

probably take about 1000 more steps during 3 hours when they received the in-app notification. Participants were placed in order of recruitment date.



Participant ID	Weekday Morning	Weekday Afternoon	Weekend Morning	Weekend Afternoon
1		Goal: 660	Noti: 750	
2				
3				
4			N+G: 866	
5				
6				
7				
8	Noti: 298		Goal: 563	
9	Goal: 238	Goal: 220		
10	N+G: 552			
11			Noti: 451 Goal: 697	N+G: 510
12	N+G: 368			
13	Goal: 173			
14		N+G: 239		
15	Goal: 394	Goal: 255		
16	Noti: 419			
17		N+G: 338		
18				
19			N+G: 258	N+G: 231
20				
21		N+G: 391	Noti: 481	
22	Goal: 246			

Participant ID	Weekday Morning	Weekday Afternoon	Weekend Morning	Weekend Afternoon
23				
24		N+G: 475		Noti: 290
25			Noti: 326 N+G: 167	Goal: 326
26		Goal: 764		
27				Noti: 467
28		Goal: 274 N+G: 354		
29				
30				
31		Noti: 574	N+G: 631	Noti: 541
32				
33				
34				Goal: 467
35			N+G: 695	
36				
37	N+G: 280		Noti: 428	
38	N+G: 284		Goal: 374	
39				
40	N+G: 788			
41		Noti: 395		
42	Noti: 331			Noti: 332
43				
44				N+G: 232

1. Noti: notification effect's MAP value
2. Goal: step goal factor effect's MAP value
3. N+G: a predicted effect if both notification and high goal (i.e., goal factor of 1) were given

Figure 5.3 Summary of effect of interventions estimated by idiographic time-sensitive models.

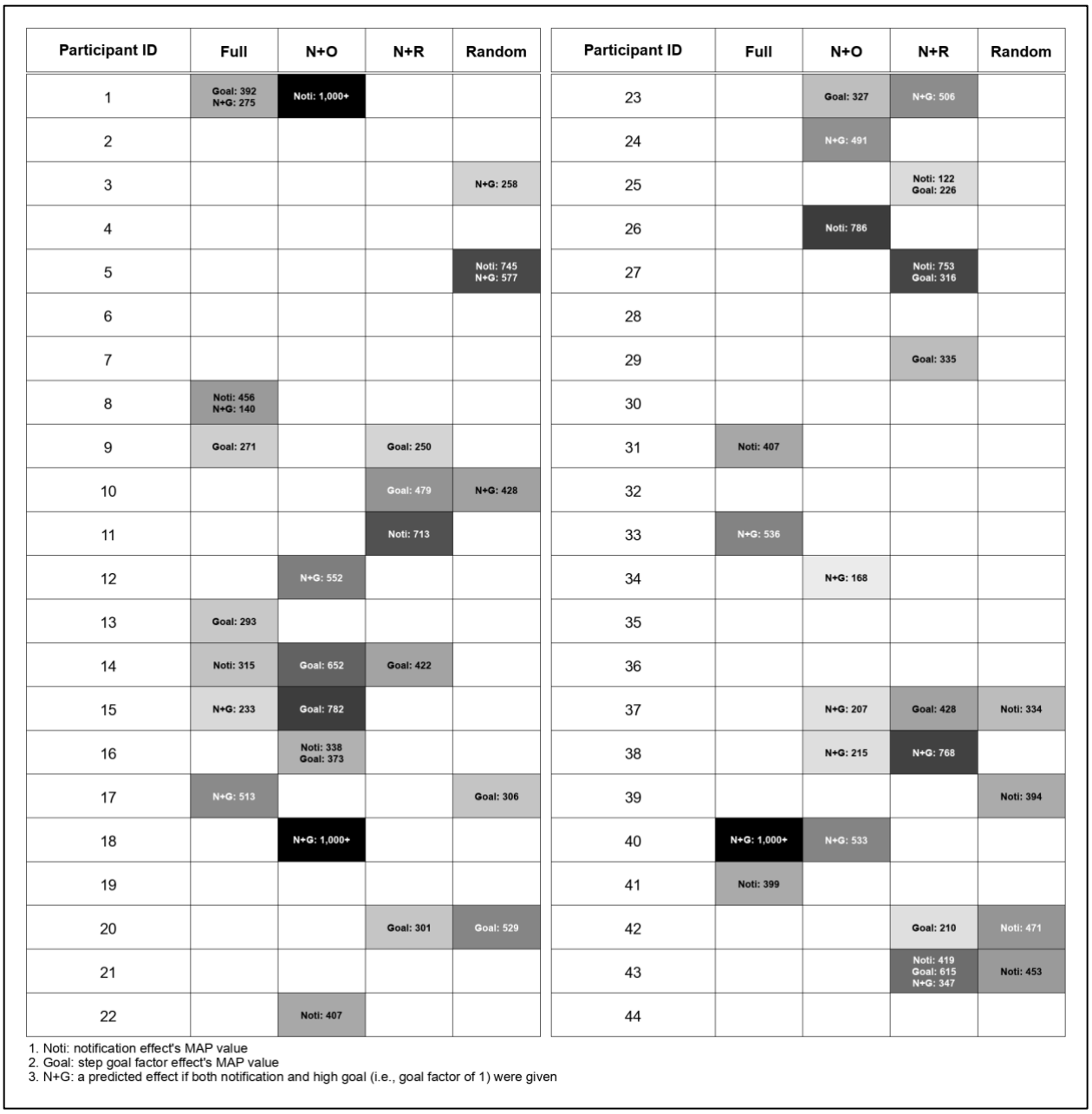
The numbers in the cell denote the effect estimates of each intervention component. "Noti" means the effect of the notifications, "Goal" means the effect of daily step goals (i.e., goal factor), and "N+G" means the joint effect of the notifications and daily step goals.

Among 44 participants, overall, 29 (66%) participants had a predictive time condition identified. These overall predictive JIT states were identified in relation to providing a notification (vs. no notification) alone 12 (27%), 12 (27%) of them were supported by being provided a high goal factor (vs a low goal factor) alone, and 17 (39%) of them were supported when a notification and high goal factor was present (compared to no notification and a low goal).

The overall average MAP across all intervention components was 422 steps/3 hours in a targeted JIT state, meaning that, in each JIT state when the "right" support is offered, people would walk, on average over the 9 month period, 422 steps/3 hours more compared to when an intervention was not provided. The average MAP effect of each intervention component was 435, 403, and 426 steps/3 hours for notification only, daily step goal only, and both, respectively.

Idiographic Decision-Policy-based Model

Figure 5.4 shows the visual representation of the results of the Bayesian regression from the idiographic decision-policy-based model that tests our hypothesized 4 decision policies for each person with comparison made between notifications sent or not, high vs. low goal factors, and the interaction between notifications and high goals compared to low goals and no notification, after controlling for time conditions. The regression was conducted for each participant and stratified by 4 decision policies (Full, N+O, N+R, and Random).



Participant ID	Full	N+O	N+R	Random	Participant ID	Full	N+O	N+R	Random
1	Goal: 392 N+G: 275	Noti: 1,000+			23		Goal: 327	N+G: 506	
2					24		N+G: 491		
3				N+G: 258	25			Noti: 122 Goal: 226	
4					26		Noti: 786		
5				Noti: 745 N+G: 577	27			Noti: 753 Goal: 316	
6					28				
7					29			Goal: 335	
8	Noti: 456 N+G: 140				30				
9	Goal: 271		Goal: 250		31	Noti: 407			
10			Goal: 479	N+G: 428	32				
11			Noti: 713		33	N+G: 536			
12		N+G: 552			34		N+G: 168		
13	Goal: 293				35				
14	Noti: 315	Goal: 652	Goal: 422		36				
15	N+G: 233	Goal: 782			37		N+G: 207	Goal: 428	Noti: 334
16		Noti: 338 Goal: 373			38		N+G: 215	N+G: 768	
17	N+G: 513			Goal: 306	39				Noti: 394
18		N+G: 1,000+			40	N+G: 1,000+	N+G: 533		
19					41	Noti: 399			
20			Goal: 301	Goal: 529	42			Goal: 210	Noti: 471
21					43			Noti: 419 Goal: 615 N+G: 347	Noti: 453
22		Noti: 407			44				

1. Noti: notification effect's MAP value
2. Goal: step goal factor effect's MAP value
3. N+G: a predicted effect if both notification and high goal (i.e., goal factor of 1) were given

Figure 5.4 Summary of effect of interventions estimated by idiographic decision-policy-sensitive models.

Each cell represents one decision policy, and 4 cells in a row represents a participant. The cells are colored to indicate whether the intervention was effective for the participant for that decision policy. White indicates no or weak effect (i.e., Maximum A Posteriori Point (MAP) Effect estimates of 100 steps/3 hours or less, or no more than an 80% chance of effect exceeding 100 steps/3 hours), black indicates MAP Effect estimates exceeding 1000 steps/3 hours with an

80% chance of effect exceeding 100 steps/3 hours. In between, gray indicates MAP Effect estimates between 100 and 1000 steps/3 hours (gray cells are only shown if at least an 80% chance of effect exceeding 100 steps/3 hours). In other words, a completely black cell means that for that cell (i.e., specific time of day, specific intervention strategy), participants would probably take about 1000 more steps during 3 hours when they received the in-app notification. Participants were placed in order of recruitment date.

The numbers in the cell denote the effect estimates of each intervention component. "Noti" means the effect of the notifications, "Goal" means the effect of daily step goals (i.e., goal factor), and "N+G" means the joint effect of the notifications and daily step goals.

Among 44 participants, overall, 32 (73%) participants had a predictive JIT state identified. These overall predictive JIT states were identified in relation to providing a notification (vs. no notification) alone 16 (36%), 16 (36%) of them were supported by being provided a high goal factor (vs a low goal factor) alone, and 17 (39%) of them were supported when a notification and high goal factor was present (compared to no notification and a low goal).

The overall average MAP across all intervention components was 457 steps/3 hours in a targeted JIT state, meaning that, in each JIT state when the "right" support is offered, people would walk on average over the 9 month period, 457 steps/3 hours more compared to when an intervention was not provided. The average MAP effect of each intervention component was 509, 395, and 473 steps/3 hours for notification only, daily step goal only, and both, respectively.

Idiographic Full Model

Figure 5.5 shows the visual representation of the results of the Bayesian regression from the idiographic full model that tests our hypothesized pairwise 16 INUS conditions for each

person with comparison made between notifications sent or not, high vs. low goal factors, and the interaction between notifications and high goals compared to low goals and no notification. The regression was conducted for each participant and stratified by 4 time conditions (weekday morning, weekday afternoon, weekend morning, and weekend afternoon) and 4 decision policies (Full, N+O, N+R, and Random).

Each cell represents one decision policy, and 16 cells in a row represents a participant, aligned with our 16 INUS Conditions as operationalizations of our theorized JIT states. The cells are colored to indicate whether the intervention was effective for the participant for that JIT state. White indicates no or weak effect (i.e., Maximum A Posteriori Point (MAP) Effect estimates of 100 steps/3 hours or less, or no more than an 80% chance of effect exceeding 100 steps/3 hours), black indicates MAP Effect estimates exceeding 1000 steps/3 hours with an 80% chance of effect exceeding 100 steps/3 hours. In between, gray indicates MAP Effect estimates between 100 and 1000 steps/3 hours (gray cells are only shown if at least an 80% chance of effect exceeding 100 steps/3 hours). In other words, a completely black cell means that for that cell (i.e., specific time of day, specific intervention strategy), participants would probably take about 1000 more steps during 3 hours when they received the in-app notification. Participants were placed in order of recruitment date.

The numbers in the cell denote the effect estimates of each intervention component. "Noti" means the effect of the notifications, "Goal" means the effect of daily step goals (i.e., goal factor), and "N+G" means the joint effect of the notifications and daily step goals.

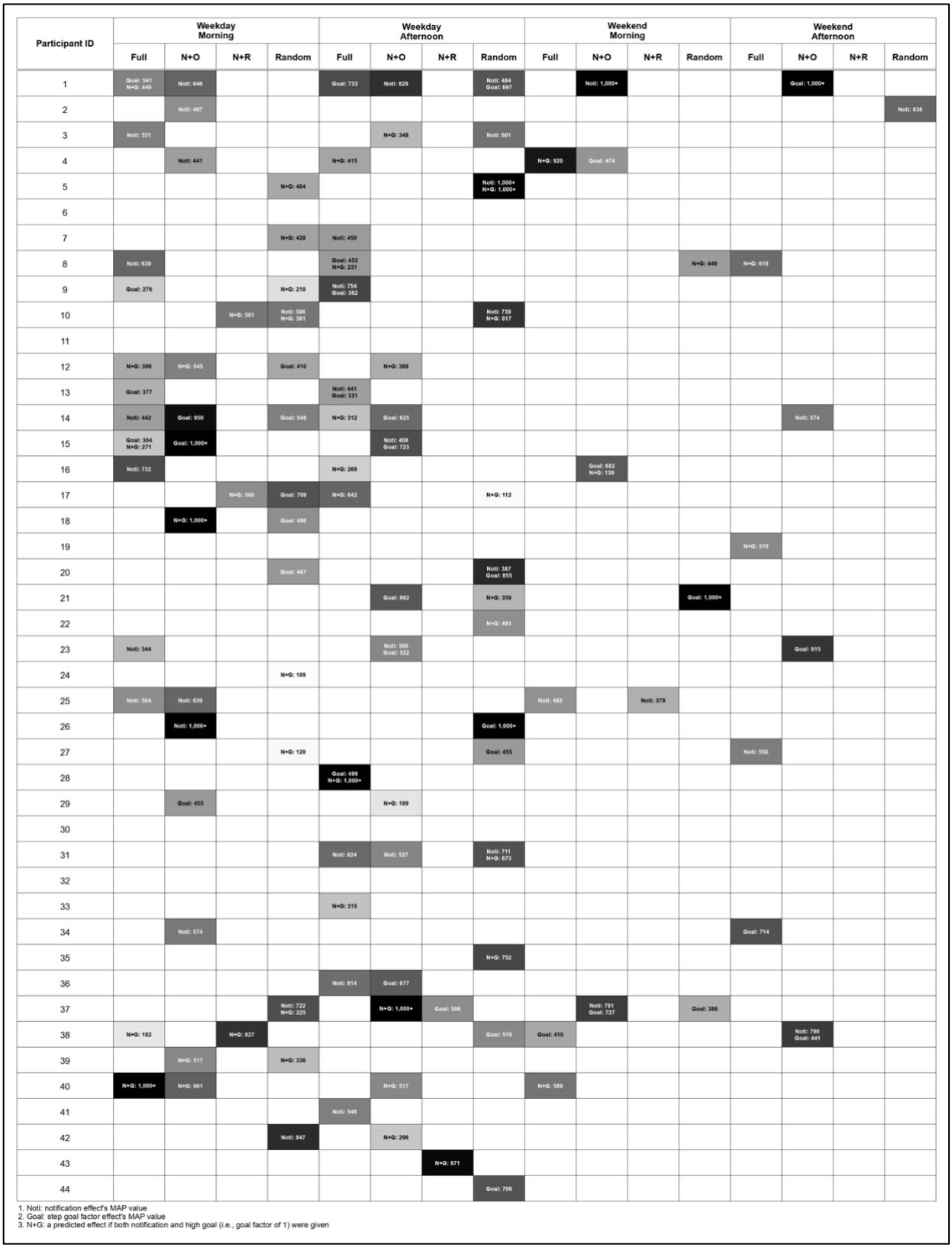


Figure 5.5 Summary of effect of interventions estimated by idiographic full models.

Among 44 participants, overall, 40 (91%) participants had a predictive JIT state identified. These overall predictive JIT states were identified in relation to providing a notification (vs. no notification) alone 25 (57%), 23 (52%) of them were supported by being provided a high goal factor (vs a low goal factor) alone, and 30 (68%) of them were supported when a notification and high goal factor was present (compared to no notification and a low goal).

The overall average MAP across all intervention components was 598 steps/3 hours in a targeted JIT state, meaning that, in each JIT state when the "right" support is offered, people would walk, on average over the 9 month period, 598 steps/3 hours more compared to when an intervention was not provided. The average MAP effect of each intervention component was 667, 621, and 522 steps/3 hours for notification only, daily step goal only, and both, respectively.

Individual Effect Sizes

The average daily effect of hypothetical, optimized intervention based on the idiographic full model was that +1,486.14 steps per day, with a more pronounced impact observed on weekdays (+1,827.16 steps/day) compared to weekends (+633.60 steps/day), resulting in an overall average effect size of 0.62. To determine the hypothetical effect sizes of the intervention, we assumed the intervention would be applied only under timing and decision policy conditions when the favorable responses was predicted. We first analyzed the data using idiographic full models to identify the strongest decision policy for each user at each decision point. This was followed by aggregating the effects for each day to calculate the daily sum, as shown in columns A and B of Table 5.4. Subsequently, we computed the daily average over a week from these daily sums, expressed as a weighted sum in column C. Additionally, we extracted the average daily step count and the standard deviation from the baseline period, shown in columns D and E,

respectively. Each individual's effect size was then calculated by dividing the weighted sum of effects by the baseline standard deviation, as outlined in column F.

Table 5.4 Individual effect sizes of hypothetical optimized intervention

Participant	Daily Effects on Weekday (A)	Daily Effect on Weekend (B)	Average Daily Effect (C, weighted sum of A and B, 5:2)	Daily Step Count During Baseline (D and E) (Mean (SD))	Individual Average Effect Size (F, C/E)
1	2,949.1	6,796.8	4,048.4	6,096.3 (3,612.0)	1.12
2	993.5	1,276.7	1,074.4	7,024.6 (3,706.1)	0.29
3	2,304.1	-	1,645.8	3,877.3 (2,556.7)	0.64
4	1,713.1	1,839.2	1,749.1	4,810.0 (871.0)	2.01
5	4,187.8	-	2,991.3	9,564.4 (4,431.4)	0.68
6	-	-	-	6,809.1 (3,360.2)	-
7	1,756.5	-	1,254.6	7,356.6 (4,837.3)	0.26
8	2,182.5	2,132.8	2,168.3	2,942.0 (1,579.9)	1.37
9	2,060.1	-	1,471.5	5,214.5 (1,930.2)	0.76
10	2,805.3	-	2,003.8	12,304.1 (3,741.1)	0.54
11	-	-	-	11,260.9 (3,999.6)	-
12	1,866.2	-	1,333.0	2,774.9 (4,121.7)	0.32
13	1,635.8	-	1,168.5	6,660.3 (2,230.0)	0.52
14	3,149.0	1,147.5	2,577.2	8,296.2 (3,866.5)	0.67
15	3,784.6	-	2,703.3	6,014.6 (1,787.1)	1.51
16	2,001.7	1,363.6	1,819.4	6,706.3 (2,219.2)	0.82
17	2,702.4	-	1,930.3	5,100.3 (3,416.4)	0.56
18	3,120.6	-	2,229.0	2,740.4 (2,487.9)	0.90
19	-	1,032.7	295.0	6,791.9 (2,796.2)	0.11
20	2,642.9	-	1,887.8	2,821.2 (1,638.7)	1.15
21	1,383.1	2,013.6	1,563.3	11,646.9 (2,479.6)	0.63
22	986.9	-	704.9	5,806.5 (3,283.3)	0.21
23	1,732.3	1,630.7	1,703.3	4,573.3 (889.3)	1.92
24	218.9	-	156.4	2,866.7 (1,486.3)	0.11
25	1,277.0	963.7	1,187.5	3,748.9 (1,874.6)	0.63
26	5,177.0	-	3,697.9	5,997.0 (4,122.1)	0.90
27	1,148.3	1,112.9	1,138.1	9,822.9 (3,838.2)	0.30
28	2,058.3	-	1,470.2	4,325.0 (3,639.4)	0.40
29	1,286.6	-	919.0	6,299.3 (2,559.3)	0.36
30	-	-	-	2,643.2 (1,331.1)	-
31	1,422.3	-	1,015.9	12,487.7 (6,670.8)	0.15
32	-	-	-	3,314.9 (1,790.0)	-
33	629.7	-	449.8	7,097.6 (3,999.6)	0.11
34	1,148.0	1,428.9	1,228.2	4,782.8 (1,472.9)	0.83
35	1,503.8	-	1,074.1	3,884.1 (1,963.8)	0.55
36	1,353.9	-	967.1	9,214.2 (3,056.2)	0.32
37	3,474.9	1,562.9	2,928.6	5,836.2 (1,659.7)	1.76
38	2,692.0	2,399.2	2,608.3	6,625.2 (2,101.0)	1.24
39	1,033.5	-	738.2	9,440.2 (4,234.9)	0.17
40	3,277.6	1,177.2	2,677.5	3,674.9 (4,464.8)	0.60
41	1,096.9	-	783.5	11,992.1 (3,342.1)	0.23
42	2,284.5	-	1,631.8	5,219.5 (1,774.9)	0.92
43	1,942.4	-	1,387.4	7,164.5 (6,085.0)	0.23
44	1,411.6	-	1,008.3	4,727.6 (2,889.6)	0.35
Average	1,827.16	633.60	1,486.14	6,326.3 (2,959.0)	0.62

Hypothesis Testing

Hypothesis 1 (H1). Among 16 states (4 decision policies × 4 time conditions), there will not be a single state that is commonly effective for the majority (>50%) of participants.

Testing results (H1): According to idiographic full models fit to every individual (noted in Figure 5.5), no single JIT state was commonly effective for participants. It was found that no just-in-time (JIT) state proved universally effective across participants. The maximum number of participants for whom any single state was effective tied at 14 (32%), specifically during weekday afternoons under both Full and Random decision policies, and weekday mornings under the Random decision policy.

Aim 2: Examining the Distribution of Individual Response Patterns

This section presents results of the examination of the distribution of individual response patterns to the intervention across time conditions and decision policies, under the assumption that there would be similarity between individuals. The assumption of similarities is that some types of response patterns will be more common than others, rather than JIT states being uniformly distributed as possibilities across all participants.

Distribution of the Response Patterns

This section shows the distribution of the dichotomized response patterns. For each time condition, since there are only 4 decision policies, the number of possible combinations of decision policies are 16 ($=2^4$). Table 5.5 shows the frequency table of the individual response patterns to *notifications* across the combination of decision policies including non-existent cases.

Table 5.5 Frequency table of the individual response patterns to notifications across the combinations of decision policies

Effective pattern across decision policies	**Weekday Morning**	**Weekday Afternoon**	**Weekend Morning**	**Weekend Afternoon**
None	30	31	41	40
Full only	5	5	0	1
N+O only	5	2	2	2
N+R only	0	0	0	0
Random only	3	4	0	1
Full / N+O	1	0	0	0
Full / N+R	0	0	1	0
Full / Random	0	0	0	0
N+O / N+R	0	0	0	0
N+O / Random	0	1	0	0
N+R / Random	0	0	0	0
Full / N+O / N+R	0	0	0	0
Full / N+O / Random	0	1	0	0
Full / N+R / Random	0	0	0	0
N+O / N+R / Random	0	0	0	0
All decision policies	0	0	0	0

Table 5.6 shows the frequency table of the individual response patterns to *any intervention components* across the combination of decision policies including non-existent cases.

Table 5.6 Frequency table of the individual response patterns to any intervention components across the combinations of decision policies

Effective pattern across decision policies	**Weekday Morning**	**Weekday Afternoon**	**Weekend Morning**	**Weekend Afternoon**
None	15	11	35	35
Full only	5	9	2	4
N+O only	5	6	2	4
N+R only	0	1	0	0
Random only	7	9	2	1
Full / N+O	4	2	1	0
Full / N+R	1	0	1	0
Full / Random	1	1	0	0
N+O / N+R	0	1	0	0
N+O / Random	2	2	1	0
N+R / Random	2	0	0	0
Full / N+O / N+R	0	0	0	0
Full / N+O / Random	2	2	0	0
Full / N+R / Random	0	0	0	0
N+O / N+R / Random	0	0	0	0
All decision policies	0	0	0	0

Statistical Testing of Distribution of Patterns

Applying our *a priori* null hypothesis of no prominent response patterns within each time condition, we first needed to test whether all 44 people were evenly distributed across all possible patterns of 16 (4 dichotomous decision policies, $2^4 = 16$), per time condition.

Table 5.7 shows the test statistics of Chi-square test for the comparison between the observed response patterns of notifications and uniform distribution, per time condition. Table 5.8 shows the same test statistics but for the any intervention components. For both of the cases, we could find the distribution is not uniform, meaning there is some level of a pattern of increasingly more likely JIT states compared to others.

Table 5.7 Test statistics of Chi-square test for the comparison between the observed response patterns of notifications and uniform distribution, per time condition

Time Condition	X^2 Statistics	p-value
Weekday Morning	305.09	<0.001
Weekday Afternoon	322.55	<0.001
Weekend Morning	569.09	<0.001
Weekend Afternoon	540.00	<0.001

Table 5.8 Test statistics of Chi-square test for the comparison between the observed response patterns results of any intervention components and uniform distribution, per time condition.

Time Condition	X^2 Statistics	p-value
Weekday Morning	84.73	<0.001
Weekday Afternoon	77.45	<0.001
Weekend Morning	406.91	<0.001
Weekend Afternoon	413.45	<0.001

Monte Carlo Simulation on Number of Patterns

While these results rule out the possibility of a uniform distribution (null hypothesis), they do not provide a clear signal on if there are predictable clusters that could be used for more informed decision-making via the identification of different types of people with meaningful clusters of JIT states.

The goal was to determine if the number of unique patterns observed was significantly smaller than expected, which would suggest the presence of meaningful clusters in the data rather than random variance. To explore this, we performed Monte Carlo (MC) simulations under four different time conditions. We looked at two types of interventions: notification only, and those involving additional components. For each scenario, we calculated the average likelihood of observing effects across participants and decision policies. For instance, during weekday mornings, if 40 effective responses are observed among 44 participants across 4 decision policies, the average effect probability is 22.7% (= $40/(4 \times 44)$, the average chance for an effect to occur). In these MC simulations, we assumed the same 44 hypothetical participants,

each associated with four decision policies. These decision policies were represented by dichotomous random variables—either showing an effect or not—based on a binomial distribution with a 22.7% probability of showing an effect. We then counted the number of unique response patterns that emerged.

Figure 5.6 and Figure 5.7 shows the results of MC simulation for response group count of the effects to the notifications and any intervention components, respectively. All subplots denotes each time condition. Black dashed lines mean observed number of response patterns, and gray bars denote the simulated density via MC simulations. P-values are noted on the top right corners.

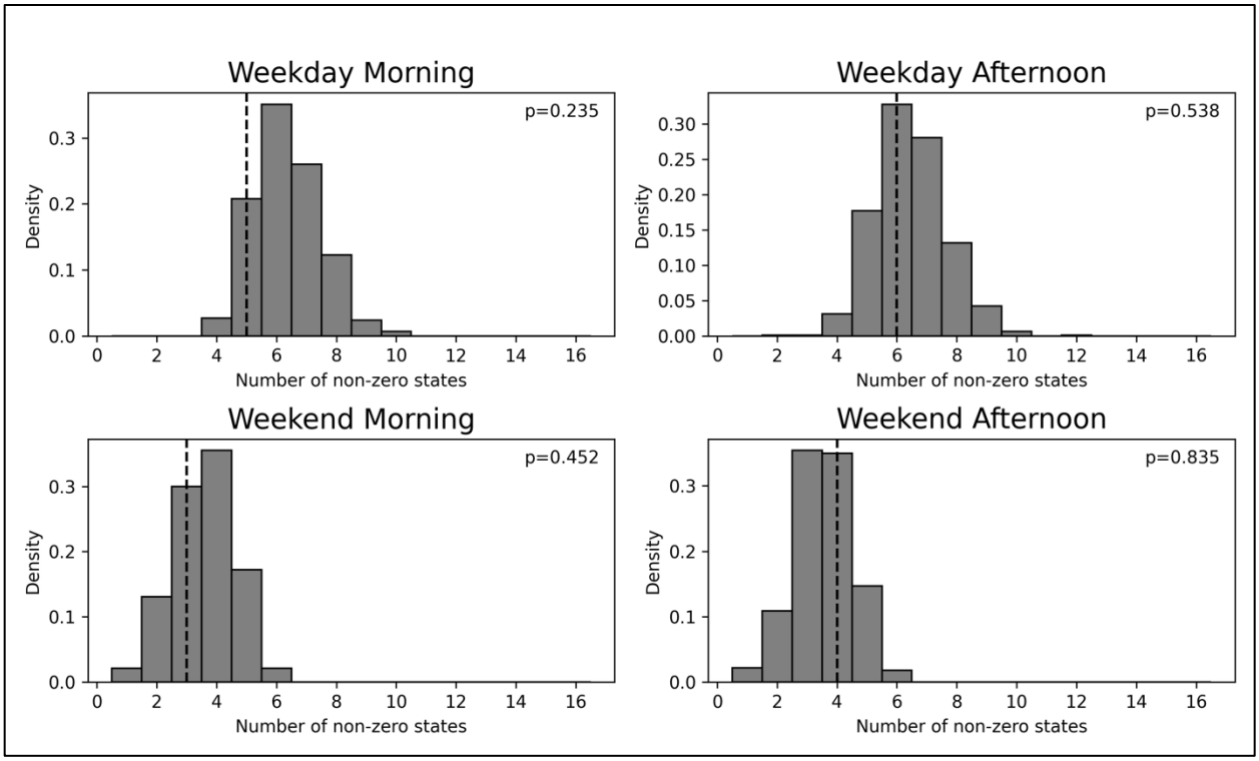


Figure 5.6 Results of MC simulations for response group count of the effects to the notifications.

According to the MC simulations, at the significance level of 0.05, the number of groups of the response patterns per time condition was not significantly outside of the expected range of

random distribution with the same probability of effectiveness (i.e., no p-values were less than 0.05). Thus, there does not appear to be any possible clustering of effects across participants.

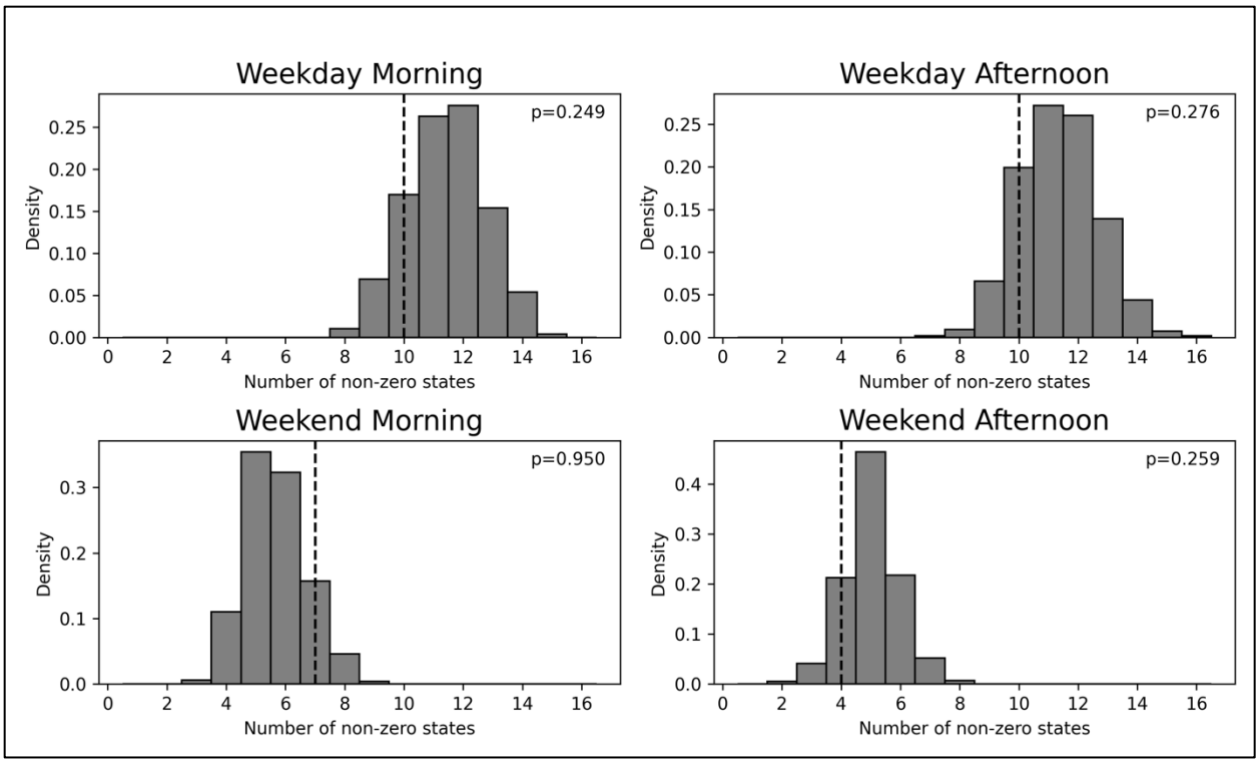


Figure 5.7 Results of MC simulations for response group count of the effects to the any intervention components.

Hypothesis Testing

Aim 2. To identify similarities in response patterns across participants within the same time condition.

Hypothesis 2-1 (H2-1). For each of 4 time conditions, participants will not be *evenly distributed* across 16 ($=2^4$) possible combinations with respect to dichotomized effect patterns for 4 decision policies; it should be statistically tested by comparing the participants' actual pattern distribution and uniform distribution.

Testing Results (H2-1). According to the statistical test results shown in

Table 5.7 and Table 5.8, the null hypothesis that the response patterns are uniformly distributed to 16 possible categories was rejected. We could find some combinations of decision policies that are common than others.

Hypothesis 2-2 (H2-2). For each of 4 time conditions, participants will form *significantly fewer numbers* of dichotomized effect patterns across 4 decision policies as indication of a different types of participants responding similarly, as tested using Monte Carlo (MC) simulations; it is tested by comparing the numbers of patterns of the actual participants and against patterns produced by MC simulations, designed to produce a random distribution. A significant effect ($p<.05$) is indicative that the actual data from participants includes clusters between participants that is likely not due to pure chance.

Testing Results (H2-2). According to the statistical test results shown in Figure 5.6 and Figure 5.7, the alternative hypothesis that the observed response patterns will form significantly fewer numbers of dichotomized effect patterns across 4 decision policies is rejected. It did not form a fewer number of groups than the pure chances of the same probability of positive effective cases.

Exploratory Aim 1: Discovery of Individual Response Pattern Using Machine Learning

Exploratory aim 1 is to explore whether an ML algorithm can be built for each individual to predict their walking behavior (i.e., step counts and dichotomized walk) during the three hours after the decision points, based on the time condition, decision policy, whether or not they received a notification, the daily step goal factor, and the days elapsed since the beginning of the intervention. To do so, we built machine learning models using each participant's initial data (truncated in 20-day increments from day 80 (~30%) to day 180 (~70%)) and examined how well they could predict the remaining periods.

Model Performance

Table 5.9 shows the summary predictive performance for two example participants. For both participants, and for all training dataset lengths, the fit was poor. Figure 5.8 shows two anecdotal examples of the multilayer perceptron predictions with the poor fit. Case A of Figure 5.8 shows the wide dispersion across the range of true values. For many examples, data points with high effect (e.g., > 2000) are predicted as non-effective (e.g., <500), and vice versa. Case B shows condensed band, which denotes predicted values vary less than the actual value.

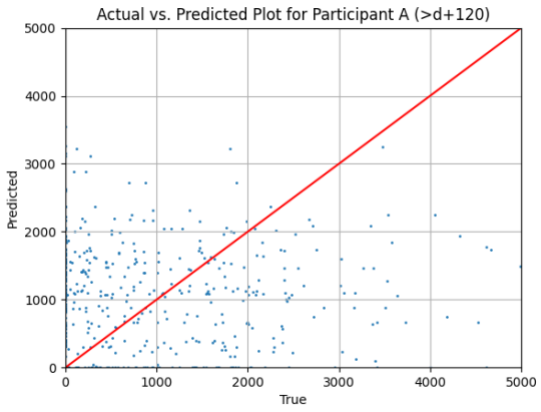
Table 5.9 Summary of the predictive performance of machine learning models for two example participants[20]

Length of Training Dataset (days)	Participant A		Participant B	
	RMSE[21]	MAE[22]	RMSE	MAE
80	1,456	1,084	1,537	978
100	1,695	1,188	1,819	944
120	1,445	1,079	1,776	1,165
140	1,654	1,372	1,442	914
160	1,401	1,027	1,552	833
180	1,393	1,108	1,544	1,013

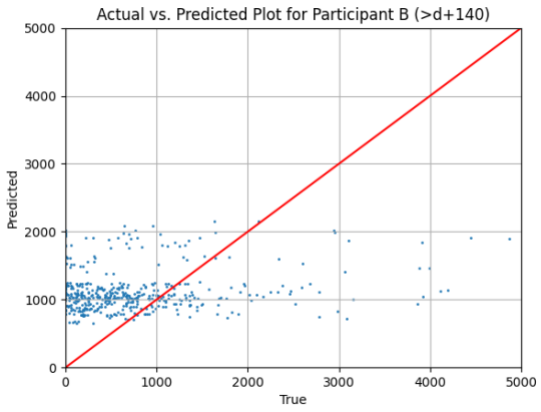
[20] Underlined cases were shown in detail in Figure 5.8.

[21] **Root Mean Square Error.** Square root of sum of squared errors divided by number of samples. $\frac{\sqrt{\Sigma_i \left(y^i_{actual} - y^i_{predicted}\right)^2}}{n}$

[22] **Mean Absolute Error.** Sum of absolute values of errors divided by number of samples. $\frac{\Sigma_i \left| y^i_{actual} - y^i_{predicted} \right|}{n}$



Case A: Participant 1, length of training: 120 days



Case B: Participant 2, length of training: 100 days

Figure 5.8 Two anecdotal examples for the poorly fitted machine learning models.

The potential reasons of poor fit may include that there are other factors that are not measured (or collapsed) information to explain the large variances with respect to the number of steps during 3 hours after decision points. Dichotomous variables takes up 9 out of 11 dimensions of input features, whereas goal factor and days elapsed since the beginning of intervention were the only continuous factors, which may not be enough to explain all the variations of step count.

As an indirect observation about this, the model shown as Case A of participant 1 in Figure 5.8 is further examined. Among 32 combinations of 3 categorical variables (4 decision policies, 4 time conditions, 2 notification provisions, $4 \times 4 \times 2 = 32$), 3 sample groups were arbitrarily chosen to show the distribution within the groups. Figure 5.9 shows the phenomena: large variations of actual values (i.e., measured step counts during 3 hours after decision points) are observed within the group, and small variations in predicted values are observed. Each group forms wide horizontal band, which hints that low reliability of the estimated values.

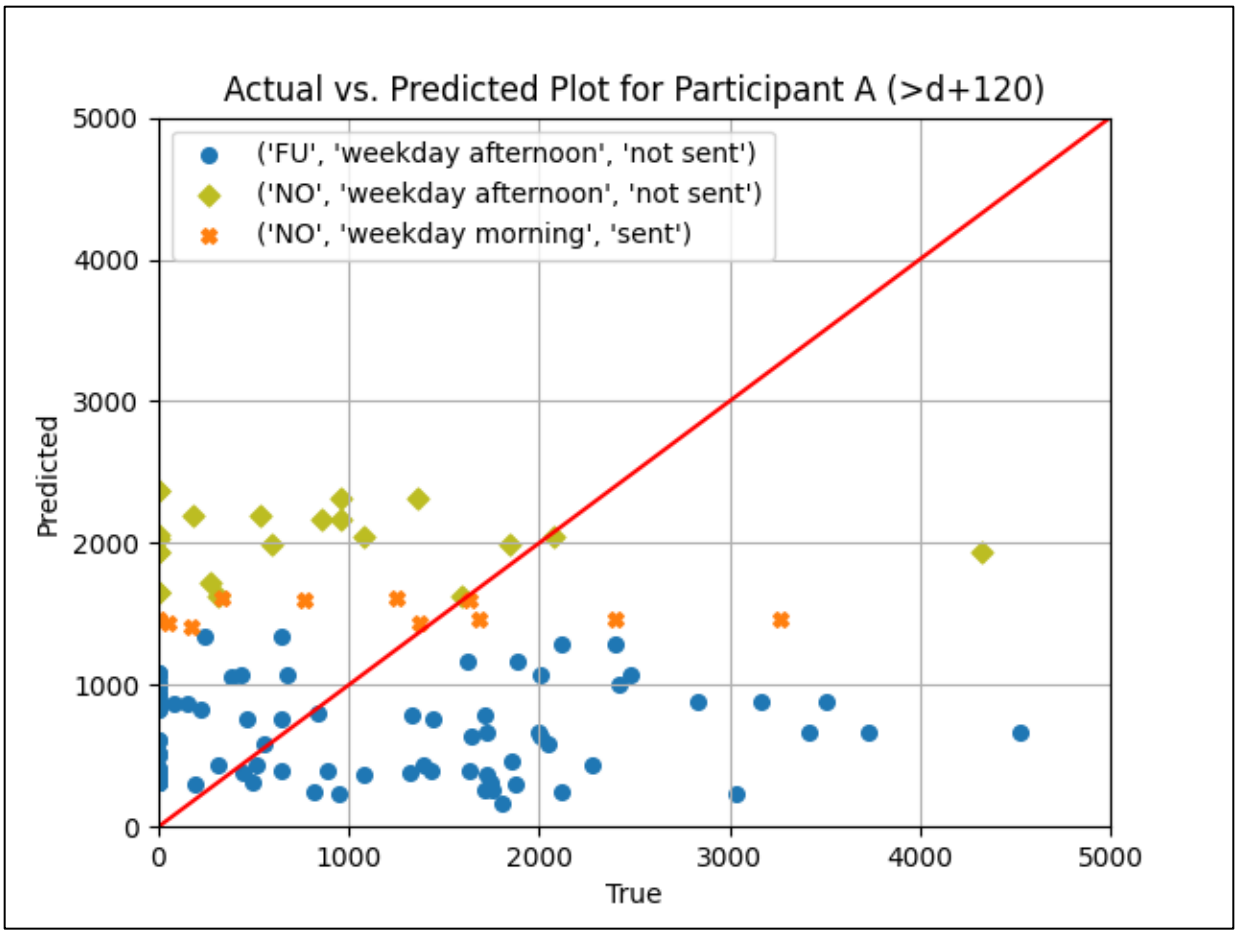


Figure 5.9 Example combinations of categorical variable values with large variations within the group

Comparison with the Null Models

Regardless of the initial fit quality, we conducted a series of statistical tests, including the Diebold-Mariano tests [195,196], to evaluate whether the machine learning models provided added informational value in predicting participants' responses compared to traditional, extremely simplistic models. These tests demonstrated that, despite their suboptimal validity, machine learning models significantly outperformed the baseline null models—frequentist linear regression and Zero-Inflated Negative Binomial (ZINB) regression—in *all cases* presented in Table 5.9, achieving significance at the 0.05 level.

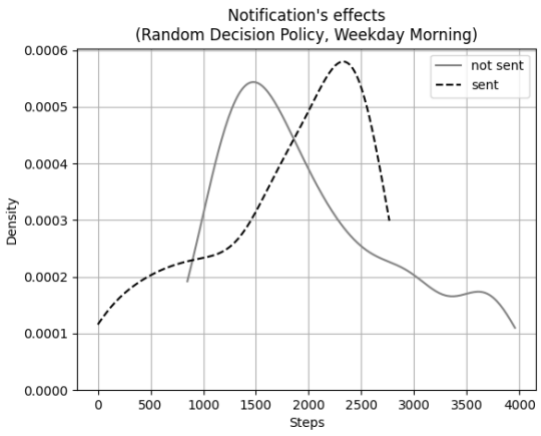
This analysis was essential to determine if the machine learning approach, even with its noted limitations in model fit and methodological challenges such as data stochasticity

(discussed further in the discussion section), still offered substantial improvements in predicting responses to interventions. Therefore, this machine learning model serves as an intermediate method in our research to develop idiographic models that effectively capture the response patterns to interventions.

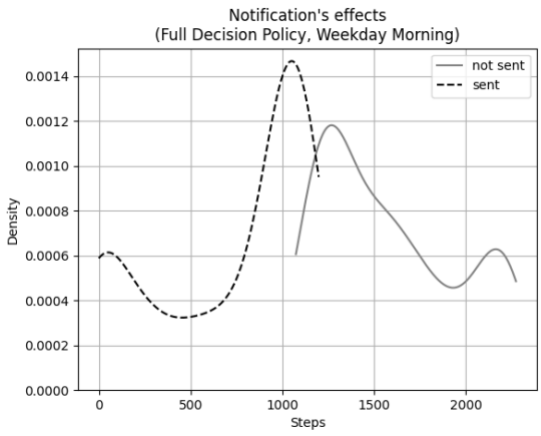
Simulation

The simulations from exploratory aim 1 illustrate a proposed strategic approach to behavioral interventions using machine learning models. By simulating hypothetical scenarios for Case A of Participant 1 (as shown in Figure 5.10; this is the same case of Figure 5.8 left), we examined how different intervention strategies—such as the use of notifications under random and full decision policies on weekday mornings—affect their expected behavior. Under the Random decision policy with notifications, Participant 1 is likely to increase walking activity (Figure 5.10, left), while under the Full decision policy, notifications might decrease their activity (Figure 5.10, right).

This case study underscores the potential of machine learning models to not only predict individual responses to different interventions but also to inform which interventions might be most effective or should be avoided. Overall, these results advocate for the use of idiographic ML models to optimize interventions, tailoring strategies to individual needs and contexts to enhance the efficacy of behavioral interventions.



Random Decision Policy on Weekday Morning



Full Decision Policy on Weekday Morning

Figure 5.10 Simulated response of the case A of participant 1 (the one shown in Figure 5.8 left) in random and full decision policy on weekday morning per notification provision.

Exploratory Aim 2: Post Hoc Analysis of JIT states

According to our internal test, for up to roughly 15% of cases, the smartphone's operating system automatically terminated the Fitbit app after certain time of inactivity (i.e., the app stays in the background). The phenomena were observed in both Google's Android phones and Apple's iPhones. These events happened consistently for a small number of participants and never happened to others. It was unclear why the phone consistently terminates the app. We were unable to figure out why. This section is to partly analyze the implication of this sync delay to the intervention optimization and operation.

The JIT states utilized in the previous aims' analysis were estimated using only the information available at the time of the decision point. Thus, they represent what was the actual experimental manipulation, including what was definable, in real-time regarding our decision policies to infer JIT states taking into account our INUS factors of need, opportunity, and receptivity. As described in our fidelity checks (see page 224), while the study did have sufficient fidelity to support interpretation of these results, the fidelity checks also highlighted mismatches that occurred based on inherent limits to current digital technologies (e.g., lags in data transfers between the Fitbit and the smartphone app; see page 29 for the problem definition and page 228 for the misalignment analysis). Overall, what this means is that there were times that were defined, in real-time on the app, as being one of our targeted INUS Condition definitions of JIT states. During those times though, based on replicating our INUS condition analyses using complete data (i.e., run post hoc), there were instances when the real-time algorithm did not make the "right" decision. This impacted 33% of time overall (12%, 46%, 47% for need, opportunity, and receptivity, respectively) across the entire study, which, as stated earlier, was determined as being of high enough quality to warrant running our *a priori* tests. Further, all the above results represent what would be possible in a real-world deployment with

on real-time determinations of JIT states using currently available technology. Thus, even if fidelity were more problematic than we found, there would still be real value in testing the effects.

With all this acknowledged, in the next set of analyses, we sought to replicate our primary Idiographic Bayesian tests, but this time, using INUS factors defined using the complete dataset, instead of the inaccurate (from a theoretical standpoint) definitions that were available to the system in real-time. These analyses provide an indication of the plausibility of our capacity to identify JIT states, in ideal conditions, when no technological issues to doing this type of work would exist. Thus, while our primary analyses in aim 1 are indicative of what is possible using technology of today, these analyses represent what is theoretically possible if these technical challenges to data syncing could be overcome.

Figure 5.11 through Figure 5.13 below show the time condition-specific effect estimates using post hoc JIT states for each individual. It can be interpreted similarly to Figure 5.5. Each cell represents one of 27 cases of complete combinations of Need, Opportunity, and Receptivity including not considering, positively considering, and negatively considering. Each row represents one time condition (weekday morning, weekday afternoon, weekend morning, weekend afternoon), four rows are noting one participant.

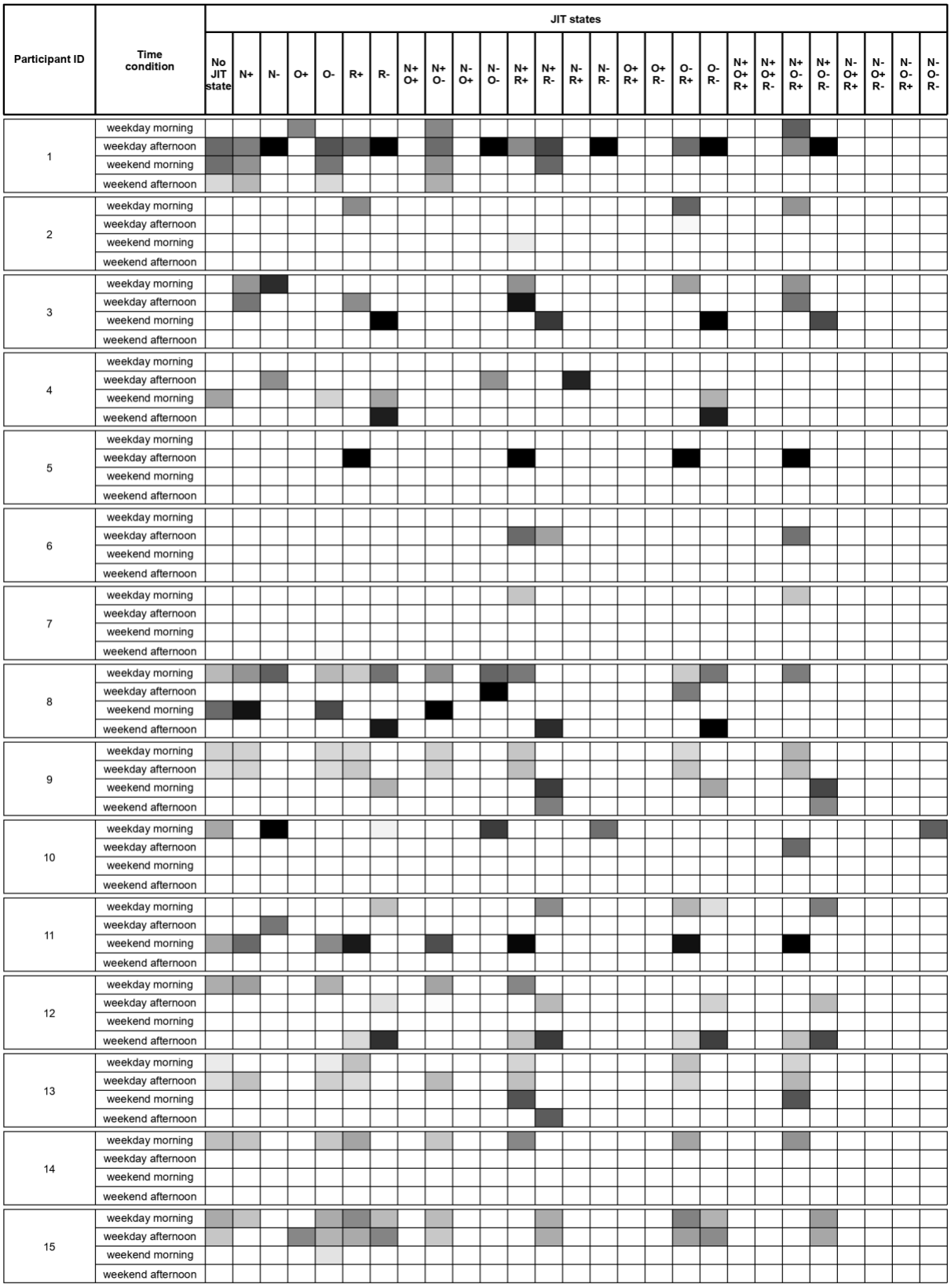


Figure 5.11 Summary of post hoc response patterns to intervention (participant 1 through 15)

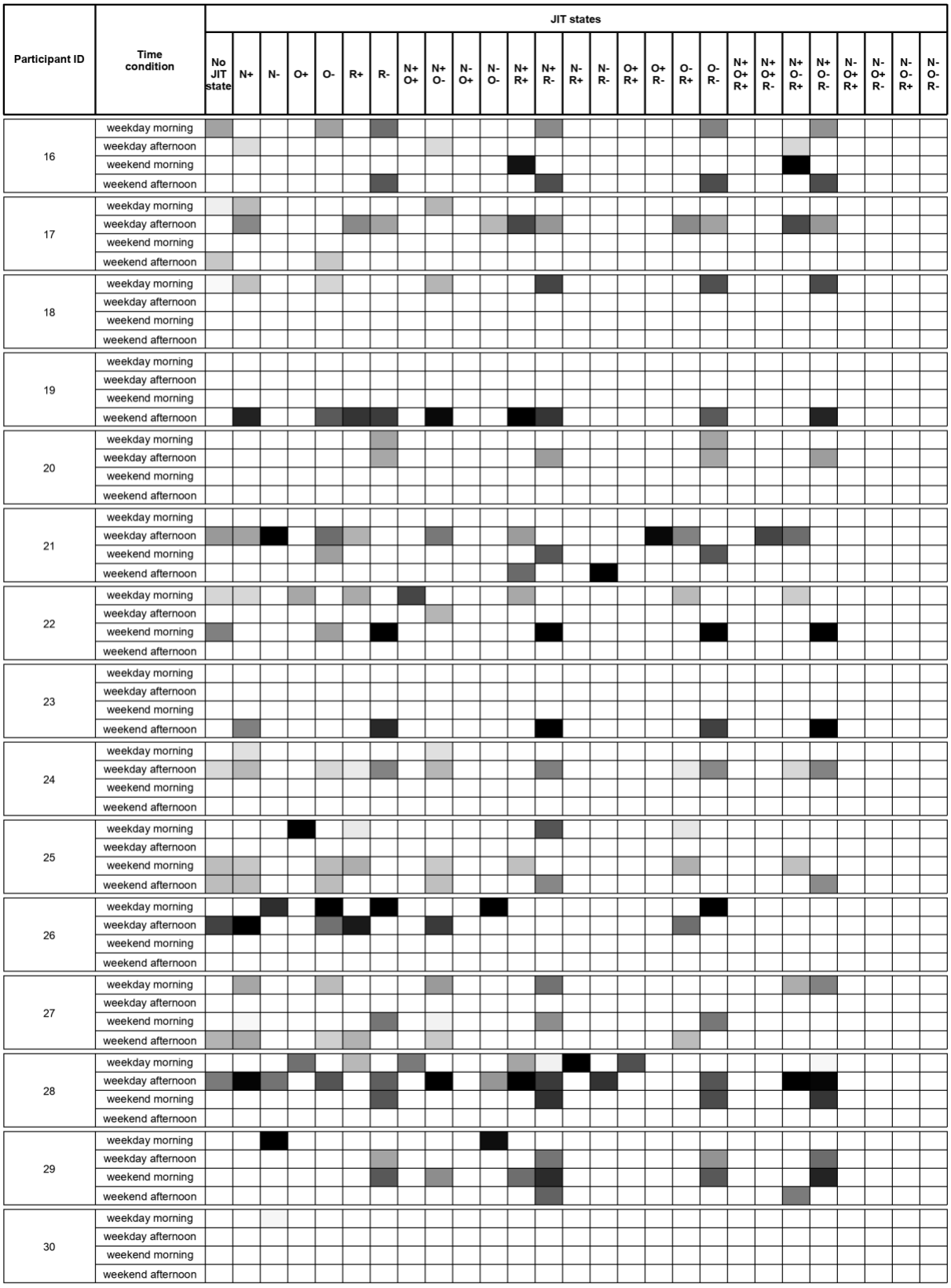


Figure 5.12 Summary of post hoc response patterns to intervention (participant 16 through 30)

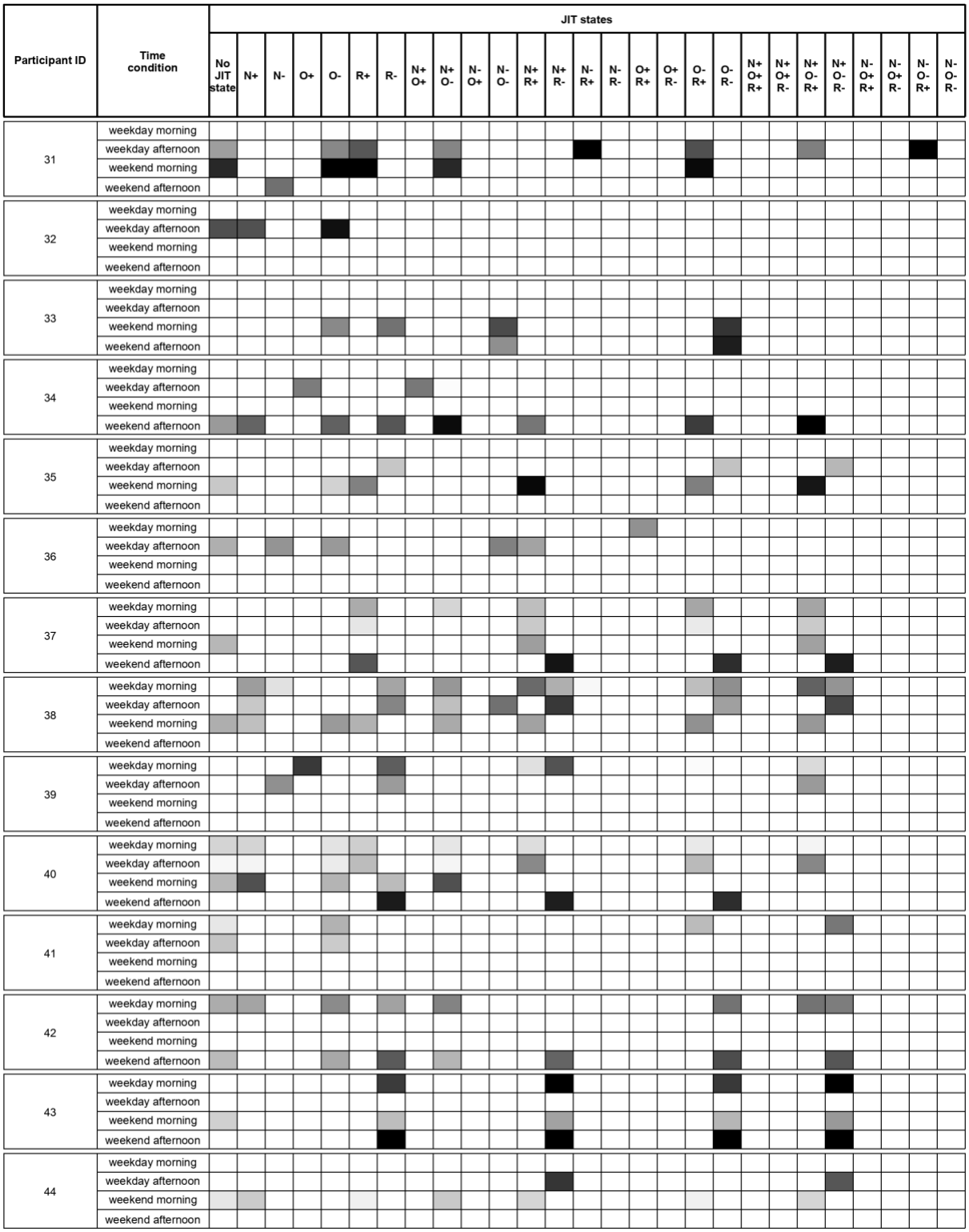


Figure 5.13 Summary of post hoc response patterns to intervention (participant 31 through 44)

Each cell is colored to indicate whether the intervention was effective for that case. If one of the notification, goal factor, or the interaction of them is effective, we use the effect size of the maximum value for the color. White indicates no or weak effect (i.e., Maximum A Posteriori Point (MAP) Effect estimates of 100 steps/3 hours or less, or no more than an 80% chance of effect exceeding 100 steps/3 hours), black indicates an 80% chance of effect exceeding 100 steps/3 hours, and MAP Effect estimates exceeding 1000 steps/3 hours. In between, gray indicates MAP Effect estimates between 100 and 1000 steps/3 hours (gray cells are only visible if at least an 80% chance of effect exceeding 100 steps/3 hours).

In summary, we were able to detect at least one meaningful intervention strategies for 43 (98%) of the individuals. It is worth to note that this is not directly comparable to the results of the full model of aim 1 (shown in Figure 5.5) because, this is a full model in the sense we considered decision policies and time conditions, but we expanded the number of decision points from 4 to 27. Figure 5.14 shows the directly comparable results, and it covers 91% (40/44) of participants, which is the same portion of participants that we could attain from aim 1.

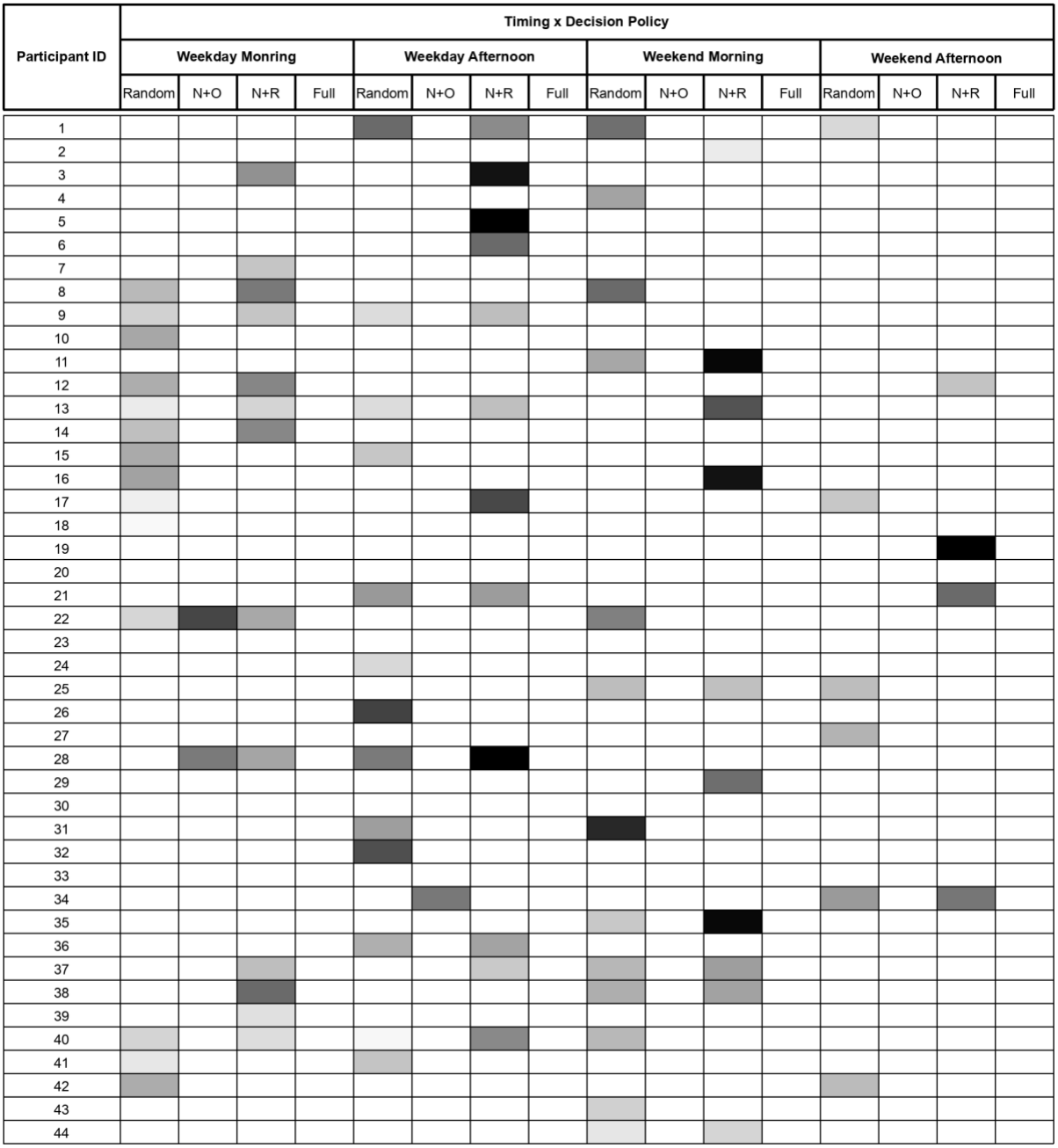


Figure 5.14 Summary of post hoc response patterns to intervention for the selected decision policies

Recap

Given the complexities of these results, I provide the summary of results one final time prior to moving to the discussion to clearly call out the key "take home messages." As expected, results of our nomothetic statistical analyses suggested that our intervention strategies as

implemented, on average and across the population, were not effective at producing significant increases in steps/3 hours. These results, which could be thought of as akin to a multiphase optimization trial screening experiment to examine the usefulness of intervention components, suggests that, as implemented, our intervention components, as they were delivered, should not be used as an optimized intervention package. This includes our general theorized approach for defining JIT states via experimentally varying a decision policy that either considered need, opportunity, and receptivity when sending notifications or delivered notifications at random as the notification × decision policy interaction within the nomothetic statistical analyses was non-significant. In line with our a priori hypothesis, using idiographic Bayesian statistics, we found that it was feasible to identify individualized states whereby individuals would reliably increase steps/3 hours post support (compared to no support given in the same state relevant for each of our three intervention variations described earlier. Specifically, we found that we could identify at least one JIT state for 91% (40/44) of participants with sufficient data (83% using an intent to treat approach, 40/48). The pooled effect size of the interventions impact was an increase of 372 steps/3 hours relative to the appropriate comparator for each intervention strategy described above within the same state, which is a general effect size of .62 (normalized mean difference; the ratio between effect and baseline SD). Given that this estimate is for a non-normative targeted timescale of steps/3 hours, we calculated and inferred likely steps/day effects that would be observed when the "right" intervention support is provided for a person at the "right time". The inferred daily effect was an increase of 1,486 steps/day (effect size=0.62, normalized mean difference; ratio between effect and baseline SD). Results from our secondary analyses generally revealed limited capacity to identify meaningful clusters of types of people responding in similar ways. Further, results from the machine learning analyses generally suggested that a machine

learning approach produced limited, yet promising informative insights for guiding further intervention optimization. Finally, exploratory simulation analyses suggested that, if we allowed our decision policies to vary need, opportunity, and receptivity independently, it is likely that we would have identified successful JIT states for 43 out of 44 (98%) of our participants.

Acknowledgement

Figure 5.1 of Chapter 5, in part, is a reprint of the material as it appears in Park, Junghwan, Meelim Kim, Mohamed El Mistiri, Rachael Kha, Sarasij Banerjee, Lisa Gotzian, Guillaume Chevance, Daniel E. Rivera, Predrag Klasnja, and Eric Hekler. 2023. "Advancing Understanding of Just-in-Time States for Supporting Physical Activity (Project JustWalk JITAI): Protocol for a System ID Study of Just-in-Time Adaptive Interventions." *JMIR Research Protocols* 12 (September): e52161. The dissertation author was the primary investigator and author of this paper.

Chapter 1, 2, 4, 5, 6, and 7 of this thesis, in part, are currently being prepared for submission for publication of the material. Park, Junghwan; Kim, Meelim; El Mistiri, Mohamed; Kha, Rachael; Banerjee, Sarasij; Gotzian, Lisa; Chevance, Guillaume; Rivera, Daniel E.; Klasnja, Predrag; Hekler, Eric. The dissertation author was the primary researcher and author of this material.

Chapter 6 DISCUSSION

The overall goal of this study was to determine if JIT states, meaning a moment when an intervention delivered at the right time for that person results in significant increases in steps/3 hours relative to the same state, but no intervention provided, can be reliably detected for each person. If this is possible for a majority of participants who took part in the trial, it would provide strong empirical evidence to justify the next step in this systematic line of research, the development of a multi-timescale control-system-driven JITAI that could utilize detection of these JIT states in providing highly personalized and adaptive support to individuals to help them increase their physical activity.

Results from this study provide strong empirical evidence to suggest that it is possible to detect JIT states for the vast majority of individuals (i.e., 91% using our a priori specified approach to defining need, opportunity, and receptivity in an aliased way and up to 98% of participants who used the intervention and assuming no technology limitations and that we considered need, opportunity, and receptivity independently) in a way that could be used by a future control-system-driven JITAI. In addition, our results suggest, that our general hypothesis that people are different, context matters, and things change but in a way that is predictable using idiographic methods is true. Specifically, our results suggest that if we only used nomothetic statistical approaches (the standard approaches commonly used in behavioral science research to date), we would have concluded that our intervention components of notifications, goals, and notifications interacting with goals and our decision policy, had, in general, limited impact on supporting individuals in increasing steps/3 hour periods.

With that said, the results of our idiographic models provided clear evidence of the capacity to identify JIT states, when examined on a case-by-case basis. Finally, our exploratory

results suggest that there did not appear to be meaningful clustering of people in terms of common patterns of JIT states. In addition, our machine learning exploratory analyses suggested that these effects were not detectable using a more model-free approach to analyses, which required the system to engage in more of the "learning" needed to detect these individual patterns. This last result, while not hypothesized per se a priori, does fit with expectations as our primary analyses of the idiographic Bayesian models incorporated prior domain knowledge that likely enabled these a priori JIT state hypotheses to be detected. Overall, these results provide strong empirical evidence to justify the team's approach to JIT identification and provides strong empirical justification for the development of a future control-system-driven JITAI.

Turning now to each set of analyses, we start with the nomothetic analyses. These results were most akin to a MOST factorial screening experiment, but with the use of a within-person factorial experiment (i.e., MRT) used instead of a between-person factorial experiment. These results are valuable to determine if the intervention components, as implemented, would result in improved steps/3 hours at the targeted decision points. The results suggest that the interventions, as implemented, had limited impact individually or in combination on steps/day and even seemed to result in undesirable effects on average, such as goal factors resulting in fewer steps within a 3 hour period, though this result should be interpreted with caution given that the goal factor was a daily intervention construct. Most critical for this line of thinking, if we were using a classic "no dead weight" criterion for judging if an intervention component should be used or not, these results would suggest that our theory-driven approach to defining JIT states within a decision policy (compared to notifications sent purely at random) was ineffective. If these were the only results available, which would likely have been the case if we had conducted this trial as a pure MRT, the conclusion we would have drawn is that our approach was ineffective at

developing a JITAI. Further, these results would suggest that the intervention, as deployed, would not be appropriate to be deployed to support individuals increase their physical activity.

Of course, within this study, we explicitly did not create an experiment that used these analyses as our primary analyses. These are presented to provide a comparator to what would normatively be the focus and lessons learned that would be drawn from traditional nomothetic statistics. As described in detail in chapter 2, we anticipated that differences in context, timing, and individual differences would reduce the utility of any results produced using population-based statistics. This was based on our assumption, which was confirmed, of the likely non-ergodic nature of our results and aligned with the use of our INUS condition causal logic. Further, the team, a priori, did not conduct this study to test the impact of each intervention component, even though, technically, that can be estimated, as we have done here. Instead, the team explicitly conducted a system identification optimization trial, with the goal of testing if it is possible to detect JIT states for each individual, with the a priori expectation that if these JIT states can be detected for each person, the system ID experiment could be embedded in a future COT study and corresponding multi-timescale control-system-driven intervention. With this, overall, the relatively limited effects observed via nomothetic statistics were expected, a priori.

Turning now to our primary idiographic Bayesian regression modeling analyses, we found that it was possible to identify JIT states whereby individuals would reliably increase steps/ 3 hours when the intervention was offered vs not, specifically when: 1) a notification was sent compared to not in an otherwise similar JIT state, 2) a high daily step goal was provided compared to a low daily step goal in an otherwise similar JIT state; or 3) a notification was sent during a day with a high step goal compared to no notification sent on a low goal day in an otherwise similar JIT state. Using the most conservative intent-to-treat approach that included

participants who consented to take part but used the Fitbit for 0 to 23 days (N=48), we could identify a reliable JIT state for 83% of our participants. This number increased to 91% when looking at participants who engaged with the intervention. Finally, using a re-calculation of our decision policies using complete data (instead of what was available in real-time), we could identify a predictable JIT state for 98% of our participants. These results provide strong evidence for our general hypothesis that people are different, context matters, and things change, in a predictable way that could be useful for developing more robust behavioral interventions. They also provide strong evidence to justify the next step in this systematic line of work, the development of a multi-timescale control optimization trial that can operationalize the system ID experiment used in this study to identify these patterns and then, have the control systems use these predictions to provide support only during these JIT states for each person. The team has already demonstrated their capacity to deploy their proposed COT approach at scale, as they are currently testing a more simplified single time-scale control-system intervention to increase physical activity in a randomized controlled trial with a targeted N of 386. As of time of reading, the team has recruited more than 300 participants with the COT approach fully deployed and running in an automated fashion. With the results from this dissertation and the demonstrated feasibility of the COT approach to be deployed at scale, it is highly likely that, with an approach like this, the dream of a robust JITAI whereby support is only provided when it is needed can likely be created, though that does require conducting the follow up COT trial that incorporates insights from this study.

Related Work

The potential for Just-In-Time Adaptive Interventions (JITAIs) to effectively enhance physical activity (PA) has been demonstrated in various methodological frameworks [83], particularly through the use of Micro-Randomized Trials (MRTs). For instance, a recent study

by Figueroa et al. [198] on the efficacy of the daily motivational text message intervention to promote PA employing MRT increased 729 step counts daily among 93 participants, indicating that timely and contextually relevant interventions could lead to meaningful improvements in physical activity levels.

Another recent JITAI optimization study was conducted by Klasnja et at. [199], with a specialized intervention for a post-bariatric surgery with daily varying step goals and motivational text messages utilizing MRTs. It resulted in an 1,866 steps increase from baseline. This study experimented with various goal-setting strategies (e.g., 60^{th} percentile) and rest days (i.e., days without daily step goals).

Despite these advancements in emerging studies, they highlight several limitations that the current JITAIs are exposing. Notably, many JITAI studies, including those by Figueroa et al. [198], have not sufficiently applied when they decide the intervention the behaviors of participants or the contexts in which they are in. While they provide valuable insights into the immediate effectiveness of intervention components, we expect there is room for even stronger effects by applying the information about people's Just-In-Time states. This gap underscores the need for more dynamic and contextually aware experimental designs that can adapt to individuals' varying psychological states and environmental conditions.

Furthermore, the integration of Machine Learning (ML) techniques in supporting JITAIs, while promising, is still relatively uncommon [66,200–202]. The complexity and opacity of ML models often lead to a substantial burden, and the computational load required for these systems mirrors the implementation challenges we have encountered in this study [203,204]. These factors collectively contribute to the limited adoption of ML in JITAIs, pointing to the need for advancements in technology and methodology to enhance their feasibility and effectiveness.

Value of This Study

The value of this study is easy to see when we bring it back to the context of our ultimate goal of implementing a JITAI. Traditionally, knowing how and when to optimize notifications for each participant is challenging, even in commercial applications [43,88,95,203]. This is a key reason for the experiment of notifications and other types of intervention support being provided to people when they do not need it, do not have the opportunity to act favorably to the support, and/or are not receptive to the support. Over time, this scenario produces the all-to-common experience of notification fatigue whereby people simply tune out completely to any support being provided [44]. Our results highlight not only a theoretical reason for this (i.e., people are different, context matters, and things change) but also point to a methodological approach that can address this challenge. With this, we now turn to a more methodological comparison of the results, with the goal of describing the differing assumptions used in each method and, from that, clarifying when, where, for what goals, and for what types of phenomena are each of these methods appropriate.

Previous studies have refined contextualized personalization strategies by targeting 1) more generally receptive participants on an individual level, 2) more meaningful exogenous contexts where the participant would respond more favorably in general, or 3) using both of them [3,64,66,102]. However, this study is meaningful because we pursued even higher effects by appending another axis, an individual's psychological state.

The following paragraphs are the notes on the limitations of this study regarding the methodologies, or the potential reasons why we attained unexpected results in each aim.

Nomothetic Modeling: Mixed Effects Modeling

Conventional linear mixed effects modeling assumes 1) normality of random effects, 2) independence of random effects and residuals, 3) homoscedasticity of residuals, 4) normality of

residuals, and 5) independence of observations and random effects. However, count data modeling needs different formulation and interpretation on 2, 3, 4, and 5 [123,124,205] (See the footnote for the details). They were tested, and the data were generally within the acceptable range of assumptions.[23]

Based on this, the nomothetic modeling using ZINB on our dataset was considered legitimate. However, given the little or negative main effects of notification and goal factor alone, there are two explanations: 1) there were effects for each individual, but the model could not detect the meaningful main effect because of the effect dilution and cancelling out across individuals, or 2) there was actually no main effect. Considering we could find the individual main effects, it is more likely that the former is the case.

Non-full Idiographic Models: Still-existing Dilution of Effects

A similar argument can be applied to the idiographic models. We could only identify 8 participants who benefited from our intervention approach, if we take the average model for each individual after controlling for the time condition and decision policy. But we could identify more meaningful patterns if we include more information, which was in line with our a priori hypotheses about INUS condition causal logic to guide operationalization of our JIT states. Therefore, one way to interpret the limited results of the main effects ideographically, is that, causally speaking, the effect manifests when INUS conditions are met. This is in contrast to the more traditional linear causal assumption commonly used in studying intervention effects that assumes all of the causal contributions come from intervention alone. These results highlight the

[23] The ZINB distribution has definitive relations between the mean and variance (See page 53). $E(X) = \frac{r}{p}$, $Var(X) = \frac{r(1-p)}{p^2}$. Thus, if the mean increases, the variance also increases, and the ratio ($\frac{1-p}{p}$) is assumed as constant because the p is assumed as constant (see page 54). Thus, if the variances and means are correlated, and its ratio is normally distributed, it is a comparable trait to normal distribution of variances in linear models [205]. Also, residuals, usually defined as the departure of predicted value from the actual value, naturally correlates with the expected value. Based on these adjusted definitions of the assumptions, the data were tested.

value of incorporating INUS condition causal thinking into studying the actions and activities of human behavior. Further, the results suggest that the use of a priori well-designed stratification may be useful to identify the effective yet appropriately conditioned decision policies for each participant, as we did in our full model ones that incorporated all of our 16 paired INUS conditions.

Multilayer Perceptron Models

Multilayer perceptron (MLP) models showed suboptimal performance in modeling our data as we can observe in Figure 5.9. We assumed that this could be caused at least one of the following:

1) the MLP model provided the deterministic estimation only based on the input; or

2) the last layer of MLP did not properly take the count distribution into account.

Multilayer perceptron models are operated by deterministic functions and numerical calculations. Although there are numerous variations or additional features to improve predictive performance or stability[128], in most cases, multilayer perceptron models are used as deterministic regressors. Thus, given the input values and the model, the predicted results are single values calculated by the network architectures, weights, biases, and other hyperparameters including activation functions, dropout layers, or normalization layers.

With this, we had relatively small amount of input variance. Multiplying possible values for each categorical variables (2 notification provision x 4 time conditions x 4 decision policies) to get 32 groups of data, and utilizing only two continuous variables. While this targeting of possible decision options is highly valuable, particularly within a control systems context (a point we will return to), it is problematic from a machine learning perspective, as it highlights an imbalance between input variance and output variance. To achieve the maximum performance in such problem, as a deterministic model, the training process becomes equivalent to the

optimization problem to pick up a single point for a group of samples. As we could indirectly observe in Figure 5.9, the average value line for the actual (vertical) and the predicted (horizontal) meets around x = y line for each group, and the data points locates around it forming a band, not a square (i.e., significantly greater variance in actual than the predicted). Since MLP does not allow uncertainty in prediction, this limitation may be inevitable.

On the other hand, the use of Bayesian regression, used in the aim 1, produces a probabilistic distribution in the context of high stochasticity within the data. Hence, if modeled properly, the stochasticity may be parameterized. Further, the Bayesian regression approach allows for prior domain knowledge to be used to establish, a priori, the key inputs and their variations (i.e., the intervention options a future controller could make), to really focus on producing estimations that could inform when, where, and for whom to make differing decisions. Thus, these Bayesian models provide parsimonious and human-interpretable results that are, simultaneously, can be incorporated and used in a future control system to support real-time on the fly decision-making via the controller. Returning to the machine learning approaches, there is a set of variations of the machine learning models that allow uncertainty in their prediction [206–209]. It uses the same methodologies as Bayesian regression, but with the layered architecture. It was considered to address the issue, but this approach produces a new concern, which is greatly increasing the computation load [206–209], thus rendering this approach impractical to use in a future control system intervention. With this, these results flag the value of utilizing prior domain knowledge from behavioral science to carefully design plausible variations of JIT states, operationalized as experimentally varied and observable INUS conditions, and the use of Bayesian regression to model this as a practical approach that is not easily achieved using machine learning techniques.

Another potential cause of non-conclusive findings from the machine learning models is the activation function of the last layer. In our implementation, the last layer used ReLU function.

$$ReLU(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

This function collapses all negative values to zero, but does not impact on the positive values. It only guarantees non-negativeness. Since our step counts are scaled down by a factor of 10,000 to fit generally in [0, 1], ReLU can cover the magnitude of the true output's range, fast enough[24]. However, the linearity of the last layer may not reflect the dispersed nature of count distribution. This may be improved by introducing a negative binomial activation function, along with its gradient and corresponding network architecture [210]. We decided to defer this improvement to the future study.

Time Condition and Just-In-Time States, in a Broader Sense

In this trial, we designed specific definitions of time condition and JIT state a priori, but it is important to take a step back and think about what they mean conceptually.

Time Condition represents a completely *exogenous* environment to which any individual may be commonly exposed. We have broken it down into weekdays, weekends, mornings, and afternoons, but it could be broken down further. Moreover, expanding on exogenous variables could include weather, geographic location, and even social factors outside of the individual's control, like COVID-19. These are the environmental and contextual conditions that any individual is subject to, which can change the individual and moderate their response to an intervention, i.e., the effect of the intervention. Using INUS Condition causal logic, these

[24] If the output's range is significantly larger than 1, the MLP can model it, but it needs more iterations to search proper weight matrices.

exogenous variables need to be incorporated as INUS factors contributing to the possible manifestation (or not) of a hypothesized INUS condition that will result in the targeted behavior (i.e., steps/3hours in our study).

Decision Policies represent *endogenous* states. It is a useful to understand JIT states as multidimensional space that can be operationalized into endogenous inputs in the form of decision policies. In the multi-dimensional space of psychological constructs, an individual posits in the coordinate of $(X_{1,t}, X_{2,t}, \dots, X_{K,t})$ if $X_{i,t}$ denotes a construct *i* of time t. As the time passes, the individual may change their position. We have assumed three dimensions that we could operationalize into a decision policy: Need, Opportunity, and Receptivity, but we are not necessarily limited to these. We assume these internal states may moderate how each individual responds to the intervention, or each component of intervention (e.g., notification, goal factor, or the interaction of them). We have shown above that this assumption was true, at least in our trial. We can further refine our intervention by defining additional state definitions (i.e., increase the number of dimensions).

Objective Measurement of JIT States

This study was designed with both exogenous (e.g., time condition) and endogenous (i.e., decision policies that were experimentally varied) being measured objectively in relation to provision of support, notifications and daily suggested step goals. There are other ways to measure the exogenous and endogenous variables. Our research group is also conducting a 2-arm RCT that includes as the intervention condition, a COT in which endogenous states are measured via daily ecological momentary assessment (EMA) and passed to the intervention controller for modeling and then intervention. (ClinicalTrials.gov ID: NCT05598996)

In this study, while we did include EMA in the overall study, in this dissertation, we did not incorporate those insights. The reason for this was a focus on produced practical results that

have a high likelihood of being incorporated into future digital health platforms. To do this, the benefits of focusing on objectively endogenous and exogenous information is clear. The idea is that individuals interacting with an app or wearable device require less information to be entered and spend less energy to experience the intervention. However, whether this is an advantage or disadvantage may depend on the context around the intervention and the degree to which key information that only a person can provide via EMA is needed to operationalize INUS conditions. For example, if carefully designed, entering this information into the EMA can be seen as self-reflection tool, which may reinforce behavior change. Further, EMA could be used to support individuals in having more control over these interventions by being able to do things such as request that an intervention component be turned off for some time. The role of incorporating participants perceptions and responses, with the ultimately goal of both minimizing burden upon them and maximizing their control and agency over the intervention is a critical area of future research.

Another important question to ask is, "What is the appropriate time unit to look at JIT states?" In our prior work, we arbitrarily assigned one day as the appropriate time unit for analysis. However, our results showed that a one-day JIT decision policy diluted effective intervention strategies, and a half-day time condition and weekday/weekends was able to identify effective interventions. Therefore, future studies may consider assigning even shorter time intervals than half day but grouping similar types of time units is also possible.

Limitations and Strengths

The limitation of this study mostly originates from a slight departure of this dissertation thesis analysis from "classical" approach to analyzing data within system identification experiments as used in control systems engineering. Specifically, the classical approach for analyzing data within system identification experiments is dynamical system modeling, which

seeks to study the dynamic patterns (e.g., lags, accumulating effects, feedback loops) that can be observed within each "system" (in this case, each person). These classical analyses, which are being conducted by other members of the study team, guided the overall design of the system identification experiment. With this, the design was developed with a strong focus on the possibility of detecting accumulating effects if a person continually only receives intervention support in the "right time" for them. With this, the system ID study varied decision policies across time, but, within a pre-specified day's decision policy, such as full JIT (need, opportunity, and receptivity taken into account) vs. randomly sending notifications, the notifications were sent in alignment with those. What this means is that, during full JIT state days, notifications were sent deterministically if it was detected. Interestingly enough, inherent technological limitations, in terms of spotty data transfer, limited data processing, and the like, integrated stochasticity and randomness even in these otherwise deterministic just-in-time signals. It occurred at such a rate that it actually resulted in a key benefit for the primary analyses reported in this dissertation. Namely it produced a large number of observations that occurred at random and distributed across the nine months of the trial when each of the deterministic decision policies should have sent a notification but did not due to technical problems. With this, we had sufficient temporal variation to enable the analyses conducted here, including all of our pairwise tests.

With this those, this is a key limitation for concluding the results. In hindsight and to support the analyses we conducted here most rigorously, incorporation of an "exploration" phase whereby, even in days that were defined as times to deterministically sent notifications, we should have incorporated a small portion, such as 20% to still be sent at random. This would have produced a more balanced data sample within each decision policy. With that said, tested for impact of imbalance, sufficient sample sizes within each JIT state pair, and examined

possible issues of the technology variations occurring in more non-random ways, with all results of these fidelity tests providing compelling justification for our capacity to run the primary analyses in this study. In future work, the system identification experiment could benefit from more thoughtful incorporation of the reinforcement learning algorithm notion of "explore/exploit" whereby "exploration" is always still included in via random signal, even when there could be benefit in "exploiting" prior knowledge (in this case, theorized ways of defining JIT states) in intervention delivery.

Similarly, the trial focused on the intervention period to the decision policies that are likely to bring meaningful changes in behavior. However, if we exhaustively explored all possible decision policies, even though they were expected to be ineffective, we may be able to double check the sanity of our assumptions, or discover a peculiar pattern. Again, this was our choice, but it is partly limiting the results of this analysis.

Since we recruited the participants through university mailing list, the demographics of the participants were not coherent to national census in education, occupation, age, sex, race, ethnicity, and geographical location. Even though we did not focus on group-based aggregated model or estimates, the concentration of participants in a certain population group may limit the impact the result of this study.

Another limitation of this study is that some of the new methods were used without rigorous validation studies. For example, the preprocessing for detecting intentional walks was designed based on past theoretical [211] and empirical [10,23,24,68,212–215] research but was not subjected to a separate validation study. The zero-inflated negative binomial distribution is theoretically recognized as an important regression method for matching steps [123–125], but specific validation studies for the use of ZINB to model steps measured by commercial

wearables in an intervention context were hard to find. However, as these methods are based on theoretically sound research, they were adopted in this study despite the need of validation as a separate study.

A strength of this study is that, as noted above, it is specifically designed to observe dynamic behaviors and responses while maintaining an idiographic perspective from conceptualization and design through to analysis. Further, a strength of our approach is the use of INUS causal logic to guide our causal theorizing and our innovative approach for utilizing both counterfactual causal logic (e.g., use of random and pseudorandom signal design) coupled with this INUS Causal logic to provide a robust way of studying this causally complex domain.

In particular, this study explores the causal relationships among INUS factors, forms, and conditions in the context of physical activity in a way that can be targeted to be used in the practical systems including commercial ones. Building on this foundation, we can utilize high-frequency (minute-level) time-series wearable data to explore causal factors. This approach provides a basis for better understanding of the relationship between physical activity and interventions and, offers an opportunity to synthesize transportable patterns. Furthermore, by using this study as a case study, we have developed and demonstrated methods to explore idiographic, dynamic INUS condition causality in intensive longitudinal data. Our system identification methodology, in particular, provides theoretical and empirical support for this process.

Acknowledgments

Chapter 1, 2, 4, 5, 6, and 7 of this thesis, in part, are currently being prepared for submission for publication of the material. Park, Junghwan; Kim, Meelim; El Mistiri, Mohamed; Kha, Rachael; Banerjee, Sarasij; Gotzian, Lisa; Chevance, Guillaume; Rivera, Daniel

E.; Klasnja, Predrag; Hekler, Eric. The dissertation author was the primary researcher and author of this material.

Chapter 7 CONCLUSION

The goal of this study was to individually determine if an intervention delivered at the right time for that person results in significant increases in steps relative to the same condition without the intervention, for each person. Results of our primary analyses confirmed our hypotheses. First, as expected, the results from the nomothetic analyses suggest that the intervention components, as implemented within the trial, would not be appropriate for implementing at scale as, on average across the study sample, they were not effective. These results also suggest that our theoretically driven approach to define JIT states, on average and across the population was not effective, thus, again, suggesting the decision policies, as implemented in this trial, should not be used on a population level to support people in increasing steps/day. These null population-level results were in line with a priori expectations. Second, and our more true primary analyses, this system identification optimization trial provides the hypothesized evidence needed to demonstrate the capacity of our approach for identifying individualized states whereby each person could benefit from receiving support at the "right time" and in the "right place" for them for most of our target sample. The trial provides evidence to suggest that significant increases in steps/day would likely be observed if intervention options are provided at the right time and place for each person. Further, these results demonstrate that our system identification approach provides a viable, deployable, and scalable approach for identifying those JIT states for individuals. Together, these results provide strong justification for the next step in this systematic line of research whereby we would integrate this system identification optimization trial into a control optimization trial (COT) that enables these insights to be used in real-time and at scale in a future JITAI to support increases in physical activity.

Chapter 8 APPENDICES

Appendix 1. Opportunity Condition Operationalization Study

Title

Development and Validation of Multivariable Prediction Algorithms to Estimate Future Walking Behavior in Adults: Retrospective Cohort Study

Abstract

Background

Physical inactivity is associated with numerous health risks, including cancer, cardiovascular disease, type 2 diabetes, increased health care expenditure, and preventable, premature deaths. The majority of Americans fall short of clinical guideline goals (i.e., 8000-10,000 steps per day). Behavior prediction algorithms could enable efficacious interventions to promote physical activity by facilitating delivery of nudges at appropriate times.

Objectives

The aim of this paper is to develop and validate algorithms that predict walking (i.e., >5 min) within the next 3 hours, predicted from the participants' previous 5 weeks' steps per minute data.

Methods

We conducted a retrospective, closed cohort, secondary analysis of a 6-week microrandomized trial of the *HeartSteps* mobile health physical-activity intervention conducted in 2015. The prediction performance of 6 algorithms was evaluated, as follows: logistic regression, radial-basis function support vector machine, eXtreme Gradient Boosting (XGBoost), multilayered perceptron (MLP), decision tree, and random forest. For the MLP, 90 random layer architectures were tested for optimization. Prior 5-week hourly walking data, including missingness, were used for predictors. Whether the participant walked during the next 3 hours was used as the outcome. K-fold cross-validation (K=10) was used for the internal validation.

The primary outcome measures are classification accuracy, the Mathew correlation coefficient, sensitivity, and specificity.

Results

The total sample size included 6 weeks of data among 44 participants. Of the 44 participants, 31 (71%) were female, 26 (59%) were White, 36 (82%) had a college degree or more, and 15 (34%) were married. The mean age was 35.9 (SD 14.7) years. Participants (n=3, 7%) who did not have enough data (number of days <10) were excluded, resulting in 41 (93%) participants. MLP with optimized layer architecture showed the best performance in accuracy (82.0%, SD 1.1), whereas XGBoost (76.3%, SD 1.5), random forest (69.5%, SD 1.0), support vector machine (69.3%, SD 1.0), and decision tree (63.6%, SD 1.5) algorithms showed lower performance than logistic regression (77.2%, SD 1.2). MLP also showed superior overall performance to all other tried algorithms in Mathew correlation coefficient (0.643, SD 0.021), sensitivity (86.1%, SD 3.0), and specificity (77.8%, SD 3.3).

Conclusions

Walking behavior prediction models were developed and validated. MLP showed the highest overall performance of all attempted algorithms. A random search for optimal layer structure is a promising approach for prediction engine development. Future studies can test the real-world application of this algorithm in a "smart" intervention for promoting physical activity.

Keywords

mobile health; mHealth; physical activity; walk; prediction; classification; multilayered perceptron; microrandomized trial; MRT; just-in-time adaptive intervention; JITAI

Introduction

Physical inactivity is associated with numerous chronic diseases, including cancer, cardiovascular disease, type 2 diabetes [17,146,216], increased health care expenditure [38], and

preventable, premature deaths [38]. Insufficient physical activity (PA) cost $53.8 billion worldwide in 2013. Clinical guidelines indicate 8000-10,000 steps per day [18]; nevertheless, the majority of Americans fall short of this goal [217].

In order to increase the level of PA, more than 300 commercial mobile apps have been developed [218]. The recent development of information technologies enabled mobile apps to deliver behavior change support when the users need this the most or when the utility (e.g., how much the amount of PA was increased by the in-app notification) is predicted to be high. This new, promising type of intervention is called a just-in-time adaptive intervention (JITAI) [83].

JITAIs are not widely used (e.g., 2.2% in 2018 [218]) by commercially available apps. However, it has been shown that JITAIs have the capacity to improve adherence and efficacy [1,219,220]. In addition, health behavior theories that commonly work as theoretical foundations for JITAIs [219], including social cognitive theory [221] and goal setting theory [222], emphasize the importance of timely feedback and anticipatory intervention [221,223–225]. Adaptation to individual, time-varying needs is theorized to be an effective strategy [223] for implementing time-accurate feedback and anticipatory intervention [225]. Since the opportunity window to intervene depends on the individual's environment, a fully automatic, predictive algorithm that can be run repeatedly is one of the key components of JITAI apps [223]. Thus, developing accurate algorithms to empower JITAIs to promote PA is a central task in overall JITAI development.

Prior JITAI studies used pure randomizations [226], condition-triggered Boolean logic [227,228], a combination of manually designed logics [229], or models that reveal the mathematical relationships between input factors and the behavior (e.g., system identification [165]) so that researchers could understand which factors are predictive of the behavior. In this

study, the models were evaluated mainly focusing on predictive accuracy rather than explainability [230]. Time series data of walking behavior (i.e., steps per minute) measured by a wearable sensor was used to predict future walking behavior. Multiple algorithms were compared using various metrics, including accuracy, Mathew correlation coefficient (MCC), sensitivity, and specificity. If these algorithms can be produced, it would be a critical step toward JITAIs that are cost-efficient and fully autonomous (i.e., without human couch interventions), and thus, it could be a valuable part of overall approaches for improving population health. To ensure the model's cost-efficiency and real-time usage feasibility, the training computation time was measured in the standardized computing environment.

Methods

Source of Data

This study used the deidentified Jawbone walking data (i.e., steps per minute) from the *HeartSteps* study [137], conducted in the United States from August 2015 to January 2016.

Ethical Considerations

The original study [137] was approved by the University of Michigan Social and Behavioral Sciences Institutional Review Board (HUM00092845) for data collection. As the data in this study were deidentified prior to being provided, the study was deemed as non-human subject research by the University of California San Diego Institutional Review Board. This study adhered to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) statement on reporting development and validation of the multivariate predictive models [231] (Supplemental Table 8.1).

Study Design and Data Processing Protocol

Exclusion and Data Transformation

Minute-by-minute walking data (i.e., number of steps per minute) were preprocessed in the following three steps: (1) excluded the participants who have the data of less than 10 days,

(2) excluded the data if the participant was inactive (i.e., 0 step per minute) or partially active (i.e., less than 60 steps per minute) during the minute, and (3) excluded short walks lasted less than 5 minutes. Then, walk data were used to decide whether the participant was active or not during the hour. If there was one or more walks (i.e., more than 5 consecutive walking minutes) during the hour, it was marked as an "active hour." Then, the data were transformed to fit the machine learning algorithms (i.e., from the time-series DataFrame objects of *Pandas* library to numerical array objects containing vector objects of *NumPy* library).

Training of Machine Learning Algorithms

The hourly walk data of the 5 prior weeks were used to predict the outcome (i.e., whether the participant will walk or not during the next 3 hours). The following 6 sets of algorithms were used: logistic regression, radial basis function support vector machine [232], XGBoost [233], multilayered perceptron [234], decision tree, and random forest [235] (Figure 8.1). We used the implementation of the open-source projects named "scikit-learn" [236], Keras [237], XGBoost [233,238], and "Sci-Keras" [239] for each algorithm.

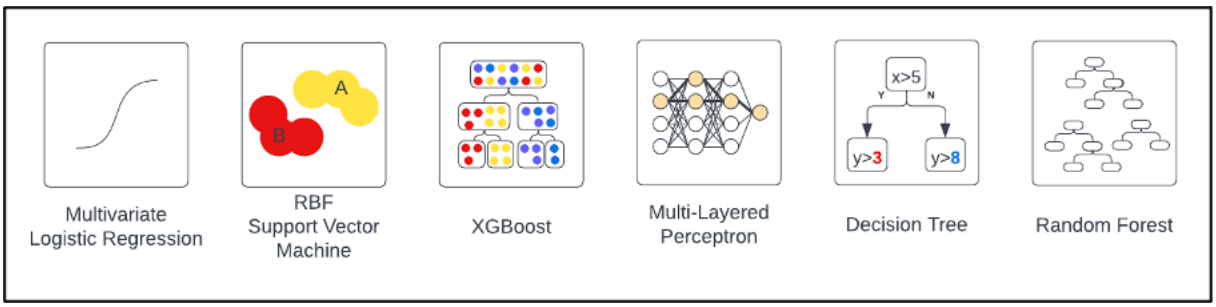


Figure 8.1 Brief algorithm descriptions of classification models

Target Imbalance

Due to sleeping hours and sedentary hours, nonactive hours usually outnumbered active hours. In machine learning algorithms, the phenomena are called "target imbalance" [240,241]. They usually critically reduce the performance of the prediction algorithm. Thus, in this study, we randomly sampled the nonactive hours to attain the same number as that of active hours.

K-fold Validation

After balancing the targets, the data were shuffled to perform K-fold validation [194] (Figure 8.2). We used K=10 in this study. We divide the shuffled data into 10 parts. Then, 1 part was separated out for validation purposes, and 1 part was separated out for performance evaluation. The 8 out of 10 parts were used for machine learning algorithm training, and 1 part is used to reduce the risk of overfitting of the training data [194]. The process is iterated for 10 times, traversing each part for validation. The method allows us to internally validate the performance of the prediction engine. K (=10) sets of results were compared across the algorithms.

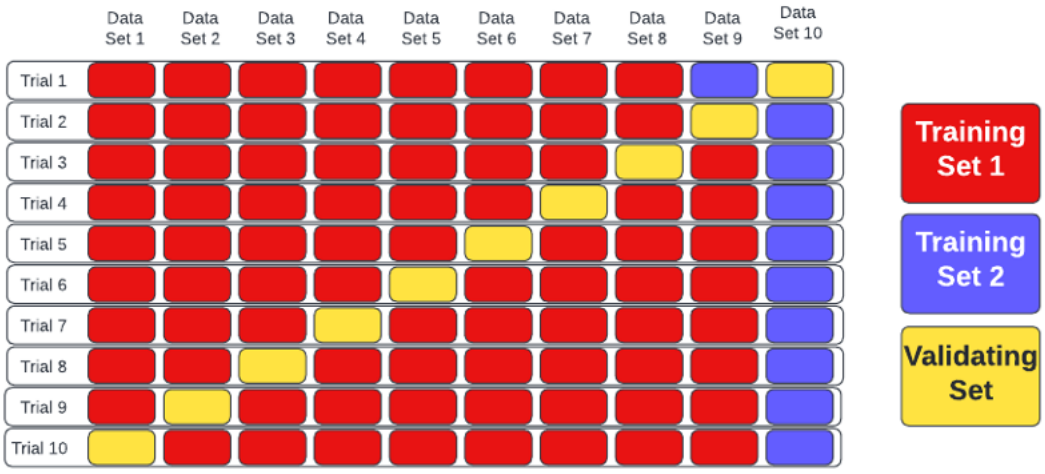


Figure 8.2 Brief description of K-fold validation method (e.g., K=10)

Outcomes

Hourly data were generated during the preprocessing step. For the outcome variable, the activity data for 3 hours were merged. If the participant walked during the 3 hours, the outcome was assigned as "walked."

Predictor Variables

In addition to 5 weeks' hourly walking data, the variables noting the current date and time were used as predictors (Table 8.1). Each variable was encoded by the "One-hot-encoding"

method [242]. It was a commonly used method to represent categorical (including ordinal or finite scale) variables in machine learning. The method converts the categorical variables (i.e., N possible options) into an N-dimensional vector. Integers such as a current hour or current month were also converted into vectors. Each element of the vector can be ones or zeros. Each position in the vector denotes a particular value of options, and if a certain position was 1, the original value was mapped correspondingly. In a single vector, only one "1" was allowed. Since the encoding method enables the machine learning algorithm to train fast, it was commonly used. The discussion on the impact of the method on prediction performance was inconclusive [242].

Table 8.1 Variables used in classification algorithms.

Predictor variables

- Current hour (24 dichotomous variables, one-hot-encoded)
- Today's day of the week (7 dichotomous variables, one-hot-encoded)
- Current month (12 dichotomous variables, one-hot-encoded)
- Current day of the month (31 dichotomous variables, one-hot-encoded)
- Five Weeks' hourly walking (Yes/No/Missing, 3 dichotomous variables, one-hot-encoded)

Outcome Variable

- Whether the individual will walk during the next 3 hours (Yes/No, 1 dichotomous variable)

Random Search for Multilayered Perceptron Model Structure

Unlike other algorithms in this study, the multilayered perceptron (MLP) algorithm uses layer architectures as one of the critical performance factors. Optimization techniques such as evolutionary programming [243] or random search or grid search [244] may be used. A random search was used to maximize the implementation burden while not losing too much performance (Table 8.2).

Table 8.2 Pseudocode for searching optimal model structure.

```
K = 10, MAX_LAYER = 10, MIN_N = 10, MAX_N = 1000

db = initialize_db()
For k = 1 to K:                          # experiment K times
For n = 1 to MAX_LAYER:                  # increase number of layers
   model = initialize_model()            # initialize the model
   For i = 1 to n:                       # for each layer
      n_neuron = random(MIN_N, MAX_N)    # decide number of neurons
      model.add_layer(n_neuron)          # add a layer
   model.train(train_data)               # train the model
   metric = model.test(test_data)        # measure the performance
   db.insert(model, metric)              # save the performance metric
```

Validation of the Models

The internal validation was performed by the K-fold validation methods. We used K=10. Individual test results were used to calculate the performance metrics such as accuracy, specificity, sensitivity, or MCCs. Data separation for the K-fold validation was conducted beforehand, which allows us to compare the metrics across the algorithms.

Mathew Correlation Coefficient

Mathew Correlation Coefficient [245] was defined as follows in Equation 8.1:

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \qquad \text{(Equation 8.2)}$$

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

MCC was sometimes used as an optimization metric. In this study, we measured MCCs as a performance metric, not the optimization metric. Since we have balanced the output (see the Target Imbalance section), accuracy was used as the optimization metric.

Computation Time

To conduct fair comparisons for the computation time, each model was trained in an isolated, standardized computing environment so that the system clock could measure the time elapsed. The system was reset every time a single execution was completed to minimize the fallout of the previous execution to the upcoming execution. Elapsed times were averaged and analyzed per algorithm.

Results

Study Population and Baseline Characteristics

A total of 41 (93%) out of 44 participants were included in the analysis [137]. The population's average age was 35.9 years. Of the 44 study participants, 31 (71%) were female, 26 (59%) were White, and 13 (30%) were Asian, with 36 (82%) having college degree or more. Moreover, 27% (n=12) of the participants had used a fitness app or activity tracker (Table 8.3Table 8.3 Baseline characteristics of participants at study entry.).

Table 8.3 Baseline characteristics of participants at study entry.

Variable	Value
Gender, n (%)	
Female	31 (71)
Male	13 (30)
Race, n (%)	
White	26 (59)
Asian	13 (30)
Black or African American	2 (5)
Other	3 (7)
Education, n (%)	
Some college	8 (18)
College degree	13 (30)
Some graduate school or graduate degree	23 (52)
Married or in a domestic partnership, n (%)	15 (34)
Have children, n (%)	16 (36)
Used fitness app before HeartSteps, n (%)	12 (27)
Used activity tracker before HeartSteps, n (%)	10 (22)
Phone used for study app, n (%)	
Used personal phone	21 (48)
Used study-provided phone	23 (52)
Age (years), mean (SD)	35.9 (14.7)

Data Summary for Predictor and Outcome Variables

On average, participants had available walking data for 43.3 (SD 9.1) days and 145.7 (SD 44.6) minutes per day. The average number of walking minutes per participant per day was reduced to 53.3 (SD 26.1) minutes after filtering with the threshold of 60 steps per minute (Methods section). Participants had 2.6 (SD 1.7) walks (i.e., 5 or more consecutive walking

minutes) every day (Methods section). Average length of each walk was 10.3 (SD 8.0) minutes. In hourly view, the participants had 0.6 (SD 0.1) "walking hours" (i.e., the hours in which the participant walked) per day (Figure 8.3). Missing data were also used as a predictor state (Methods section). There were 18.1 (SD 13.4) missed days on average per participant, equivalent to 36.9% (SD 26.3%) of total days per participant. In the matter of outcome variable, as training and validating data set, 8129 "walking hours" and 37,711 "non-walking hours" (e.g., nighttime or sedentary hours) were prepared (Methods section). Across the data, 17.7% of the time included participant activity. Thus, inactive time is 4.64 times more common than active time. The target imbalance was handled by undersampling (Methods section).

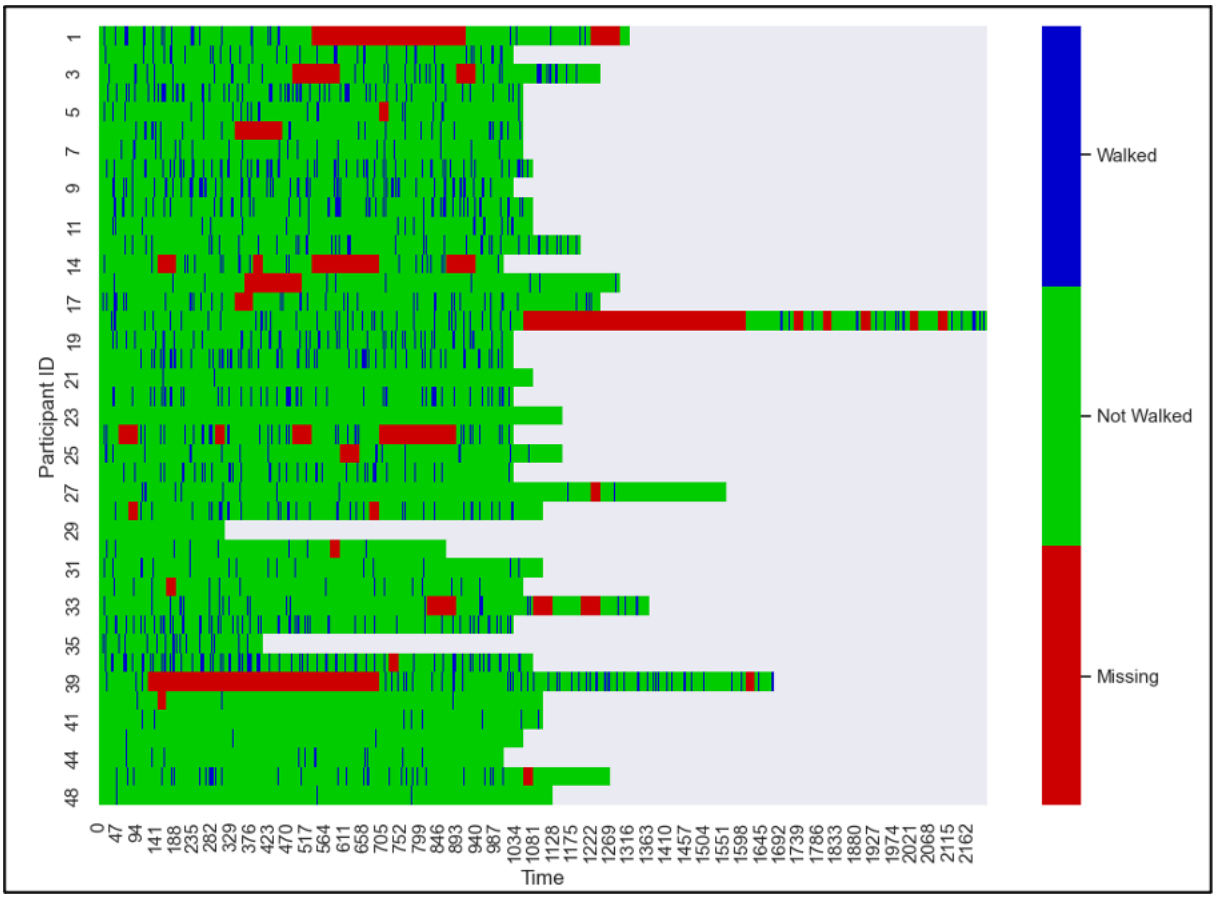


Figure 8.3 Overall distribution of walking data (one narrow cell=one hour)

Development of Prediction Algorithms

The calculation time vastly varied (Table 8.4). The radial basis function support vector machine algorithm and multilayered perceptron algorithm took the longest period to run. Tree-based algorithms such as decision tree and random forests were shorter than others. Random search to discover the optimal layer structure was tried. The optimization process improved the accuracy of the MLP algorithms from 49.8% to 82.1%. The process also improved all other metrics (Figure 8.4).

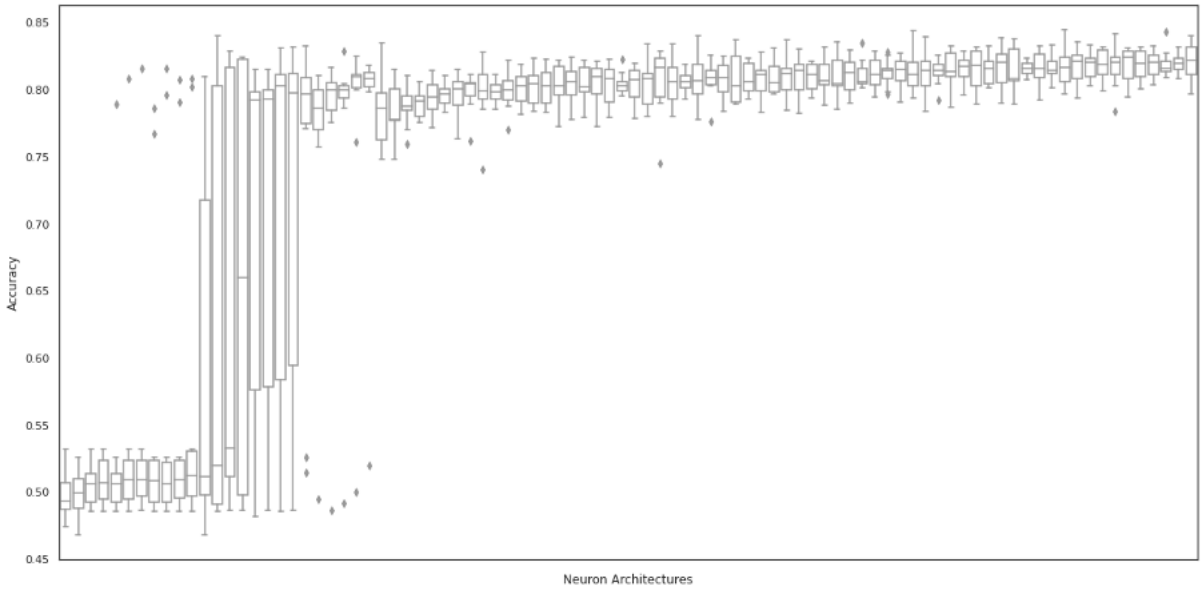


Figure 8.4 Performance of tried neuron architectures (90 trials)

Table 8.4 Performance metrics of tried algorithms.

Algorithms	Accuracy, mean (SD)	MCC[a], mean (SD)	Sensitivity, mean (SD)	Specificity, mean (SD)
Logistic regression	0.772 (0.012)	0.545 (0.024)	0.795 (0.015)	0.749 (0.023)
RBF[b] SVM[c]	0.693 (0.010)	0.389 (0.020)	0.746 (0.022)	0.641 (0.017)
XGBoost	0.763 (0.015)	0.530 (0.030)	0.816 (0.010)	0.711 (0.030)
Multilayered perceptron	0.820 (0.011)	0.643 (0.021)	0.861 (0.030)	0.778 (0.033)
Decision tree	0.636 (0.015)	0.281 (0.026)	0.509 (0.075)	0.762 (0.049)
Random forest	0.695 (0.010)	0.396 (0.023)	0.776 (0.019)	0.614 (0.018)

[a]MCC: Mathew correlation coefficient.
[b]RBF: radial basis function.
[c]SVM: support vector machine.

Validation and Model Performance

The reference algorithm (logistic regression) showed 77.2% (SD 1.2%p) accuracy. XGBoost showed 76.3% (SD 1.5%p), radial basis function support vector machine showed 69.3% (SD 1.0%p), decision tree showed 63.6% (SD 1.5%p), and random forest showed 69.5% (SD 1.0%p), respectively. MLP performance largely varied from 49.8% (SD 1.7%p) to 82.1% (SD 1.3%p). Only 3 MLP architectures with the highest accuracies were included (Table 8.4, Table 8.5; Figure 8.5). Sensitivities, specificities, and MCC showed similar patterns to the accuracies. The decision tree algorithm generally showed the lowest performance overall, except on the dimension of specificity. MLP showed the highest performance across metrics (82.0% accuracy, 86.1% sensitivity, and 77.8% specificity).

Table 8.5 Average confusion matrix of each model of K-fold validation for the validation data set.

	True positive, mean (SD)	True negative, mean (SD)	False positive, mean (SD)	False negative, mean (SD)
Logistic regression	646.3 (27.3)	609.0 (30.6)	203.5 (18.8)	166.2 (11.7)
RBF[a] SVM[b]	606.3 (25.4)	520.3 (18.3)	292.2 (19.4)	206.2 (19.5)
XGBoost	663.0 (18.3)	577.6 (33.3)	234.9 (24.7)	149.5 (12.3)
MLP[c]	699.9 (35.2)	632.6 (34.7)	180.0 (27.5)	112.6 (24.2)
Decision tree	413.8 (65.4)	619.7 (52.5)	192.8 (39.1)	398.7 (56.5)
Random forest	630.3 (13.6)	499.0 (18.2)	313.5 (20.9)	182.2 (20.7)

[a]RBF: radial basis function.
[b]SVM: support vector machine.
[c]MLP: multilayered perceptron.

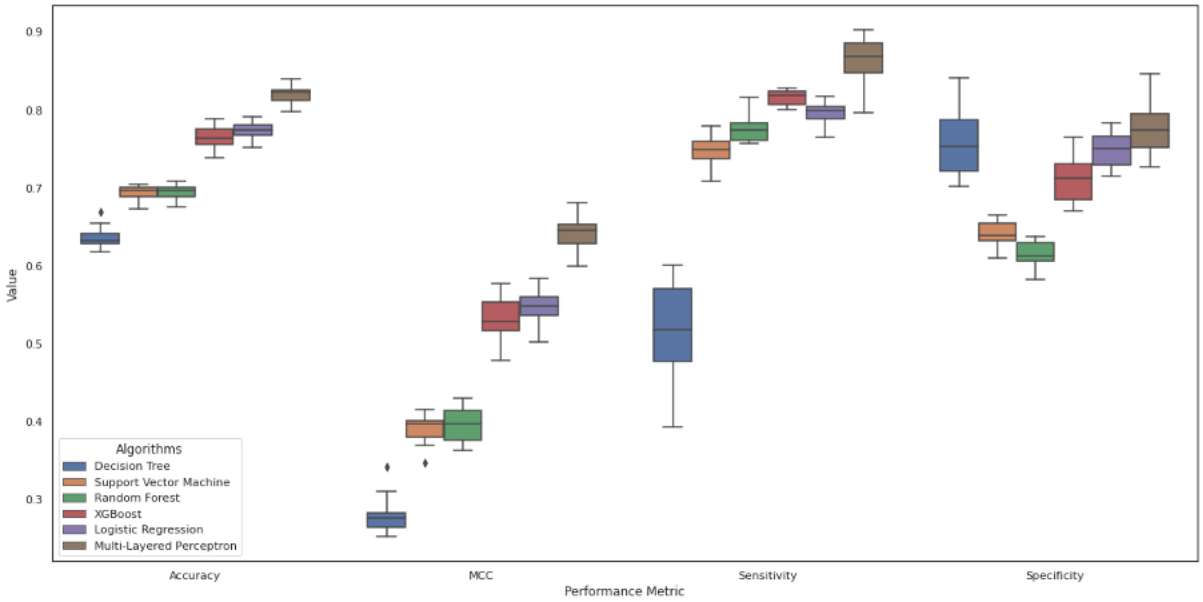


Figure 8.5 Performance metrics of the tried models[1]

[1] Top three architectures were chosen among MLP engines.

Computation Time

In all the tested performance indicators, the optimized MLP showed the best performance and showed the second-longest training time of 225 seconds on average (Table 8.6). If we add up the total training time of all 90 optimization experiments, it took 56 hours. It was feasible to consistently evaluate training speed, accuracy, MCC, sensitivity, and specificity within the standardized performance evaluation framework. Through 90 random experiments, multiple MLP algorithms with optimized performance were obtained. The development, validation, and evaluation protocols can be used for similar prediction or classification problems.

In the matter of computation cost-efficiency (i.e., predictive performance vs computation time), each algorithm showed characteristic results. The logistic regression had reasonable prediction performance and relatively low average computation time cost, whereas MLP showed generally higher prediction performance but had the second highest average computation cost (Figure 8.6).

Table 8.6 Computation time to reach optimally trained status (seconds[a]).

Algorithms	Minimum	Maximum	Mean (SD)	CI
Logistic regression	20.73	24.89	22.37 (1.50)	19.43-25.31
RBF[b] SVM[c]	413.09	683.62	496.57 (94.58)	311.19-681.96
XGBoost	63.92	73.75	67.79 (4.33)	59.30-76.27
Multilayered Perceptron	172.14	300.36	225.35 (38.83)	149.24-301.46
Decision tree	3.30	13.20	5.89 (2.68)	0.65-11.14
Random forest	4.32	13.42	6.63 (2.53)	1.68-11.57

[a]Computation was done in Google Colaboratory Pro+ (High-RAM mode with GPU hardware accelerator); 8 cores of Intel Xeon CPU 2.00 GHz, 53.4GB Memory, Tesla P100-PCIE-16GB.
[b]RBF: radial basis function.
[c]SVM: support vector machine.
Python 3.7.3, Sci-Kit Learn 1.0.2, Numpy 1.21.6, and Pandas 1.3.5, Tensorflow 2.8.0, xgboost 0.90, keras 2.8.0 were used.

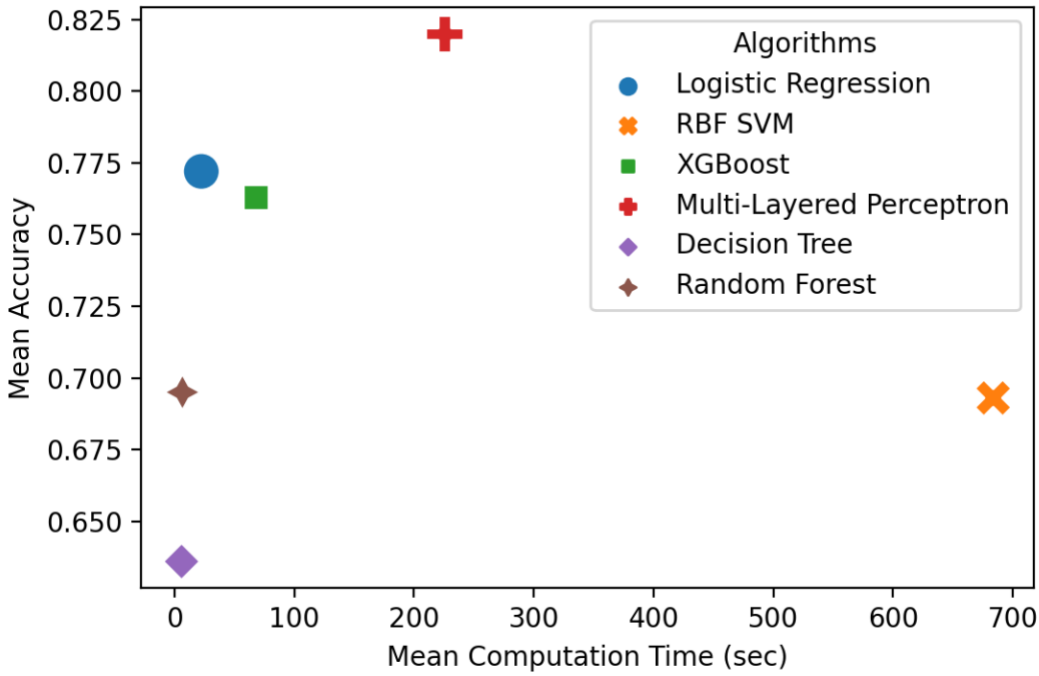


Figure 8.6 The comparisons between algorithms in the matter of mean computation time and mean prediction accuracy.

It was feasible to consistently evaluate training speed, accuracy, MCC, sensitivity, and specificity within the standardized performance evaluation framework. Through 90 random experiments, multiple MLP algorithms with optimized performance were obtained. The

development, validation, and evaluation protocols can be used for similar prediction or classification problems (Figure 8.7).

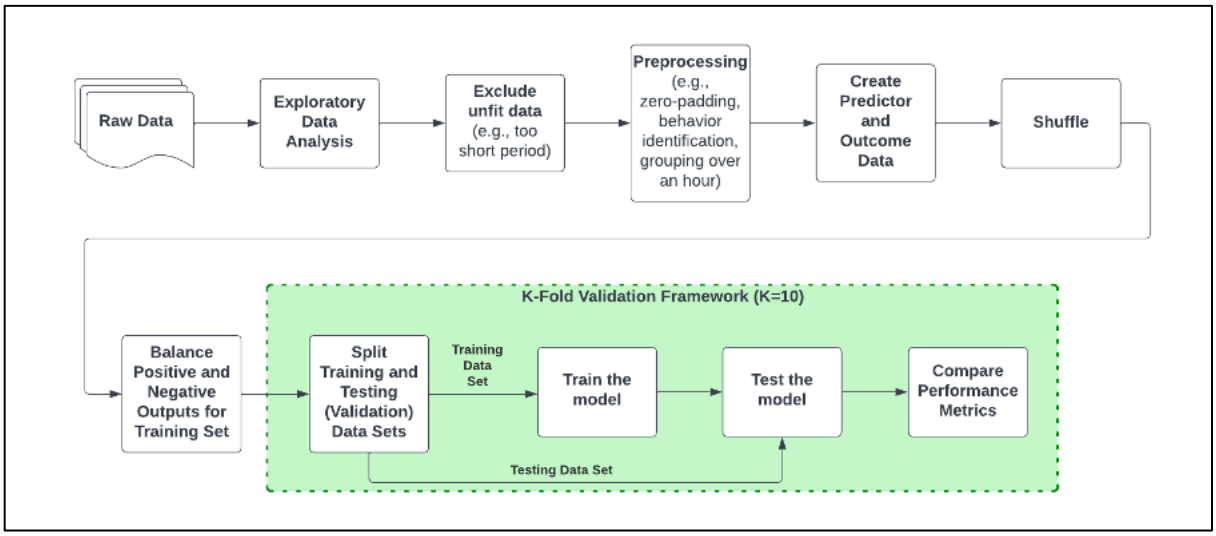


Figure 8.7 The data processing protocol

Discussion

Key Implications

The high-level focus of our work is to develop approaches for using data from individuals themselves to create more individualized and adaptive support via digital technologies. In this paper, our goal was to test if predictive models could be generated that would be useful in terms of sensitive and specific probability estimates of the likelihood that someone will walk within an upcoming 3-hour window and that it could be done in a computationally efficient fashion. The latter part is important as computational efficiency is needed to enable the predictive models to be incorporated into future just-in-time adaptive interventions (JITAIs) that could use these predictive models to guide future decision-making. To support robust, automated decision-making within a JITAI to increase walking, our goal was to test if it would be feasible to produce predictive models that are informative for individuals in terms of identifying moments when a person has some chance of walking as opposed to either times when a person will clearly walk and thus does not need support, or times when there was near-zero probability that, in a given 3-

hour window, a person will walk. If a predictive model could be produced that would provide this information, it would enable a JITAI that could incorporate these individualized predictions as a signal that could be used for making decisions on whether a given moment would be a *just-in-time* moment to provide a suggestion to go for a walk, with the predictive model used to predict the likelihood that, within the next 3 hours, the person would have the *opportunity* to walk while also having a need for a suggestion (i.e., a person would not need a suggestion to walk if they are very likely to walk anyway). Our results, overall, suggest it is possible to generate said models in a scalable fashion, which could then be incorporated into a future JITAI that incorporates these individualized predictive models. Central to this work, the models produced here are definitionally idiographic in nature and thus appropriate for each individual. Thus, the results from the model should not be generalized to other samples. Instead, the key transportable knowledge from this work is the overall approach used for selecting models to guide individualized decision-making in future JITAIs (Figure 8.7).

Principal Findings

We developed 6 models (one of which was a group of models, and we chose the best 3 model architectures) for predicting future walking behavior within the subsequent 3-hour period using the previous 5 weeks' hourly walking data. MLP algorithm showed the best performance across all 4 metrics within this sample. A random search for MLP architecture produced an optimal model with the best performance. Using predictive engines to decide how to configure JITAIs could enable the mobile physical activity app to deliver more timely, appropriate intervention components such as in-app notifications. To the best of our knowledge, interventions that use predictive models to adjust to participant's behavior are still uncommon. Thus, our study makes a significant contribution by introducing the use of predictive algorithms for optimizing JITAIs.

Methodological Considerations and Comparison with Prior Work

In this study, we designed a protocol to develop and validate a predictive model for walking behavior. While developing the model, we had a few common issues that should be handled as follows.

Small Data Sets and the Potential Risk for Low External Validity

Despite the effort to validate the model with the K-fold cross-validation, since we are using a small number of short time-series data, high levels of external validity are not assumed. However, since the model we developed in this study did not assume any prior knowledge or variability (i.e., nonparametric), additional training data are theorized to harness better performance. The model also did not use the pretrained coefficients; we used randomized coefficients. This leaves room for better performance and higher computation efficiency when we use the pretrained model from this study to extend the training. Publicly available lifestyle data, including the All-of-Us project [246] and the ones available on the public data platforms [247], will be a good way to extend the data set.

Target Imbalance

Target imbalance is defined as a significantly unequal distribution between the classes [240]. In numerous clinical [248,249] and behavioral [240] data modeling studies, target imbalance is a common issue. Although a few oversampling methodologies to tackle unbalanced output data have been developed [250], this study used an undersampling approach due to potential concerns of exaggerated accuracy [241]. The separate analysis with oversampling of the same data and methodologies showed 5%-10% increases in the accuracy. It is suspected that the underlying individual behavior patterns in the training samples are partly included in the test and validation samples.

Performance Metrics

Accuracy is the most commonly used performance metric to evaluate classification algorithms. However, the *accuracy* metric is also known to have the inability to distinguish between type 1 and type 2 errors [251]. The metrics of sensitivity and specificity are also commonly used to overcome the limitation of accuracy. The information represented by both metrics is partial (i.e., both are addressing either type of error). MCC [252] is used more commonly in recent publications due to its statistical robustness against target imbalance, which is a common issue of clinical and behavioral data. Considering the imbalance of the classification problem of interest, we included MCC as a performance metric.

Limitations of This Study

The original study was designed for the purpose of pilot-testing and demonstrating the potential of microrandomized trials. Thus, these analyses are all secondary in nature. Further, the initial study was a small study, with only a minimum amount of data (n=41) used. Additionally, since the participants were recruited in a homogeneous environment and demographic groups, the external validity of the algorithms may be limited. With that said, the overall approach for formulating predictive models and their selection could feasibly be used in the future and, thus, it is more of our protocol and approach that is likely to be generalizable and generally useful for JITAIs compared to any specific insights from the models we ran. We contend that, for any targeted JITAI, a precondition for this type of approach is the appropriate data available, and that, for any JITAI, it is more valuable to build algorithms that match localized needs and contexts than seek to take insights from some previous samples that are different from a target population and assume they will readily translate. This, of course, can be done with careful tests of transportability using strategies such as directed acyclic graphs to guide the production of estimands [253] that would create formalized hypotheses of

transportability. However, this is a much higher bar for transportability that, while valuable, can often be prohibitive for fostering progress in JITAIs. Within our proposed approach, the strategy involves gleaning *good enough* data to enable a localized prediction algorithm appropriate for the targeted population to be produced, with subsequent deployment factoring in strategies and approaches for updating and improving the algorithms as new insights emerge.

Implication and Future Work

The results of our study show that prediction algorithms can be used to predict future walking behavior in a fashion that can be incorporated into a future walking JITAI. In this study, we modeled without contextual information other than the date, time, or day of the week. However, if the machine learning algorithm is trained using the other contextual information such as intervention data (e.g., whether the in-app notification message is sent or not, which type of message is sent, and which sentiment is used to draw attention), the prediction engine would be capable of simulating how the intervention components might change the behavior in the multiple hypothetical scenarios. This capability would enable us to use the prediction algorithms uniquely, that is, comparing two or more possible scenarios to decide the optimal intervention mode of a JITAI. We could decide whether to send a message, which message should be sent, or what sentiment we could use to draw attention to our intervention. A pragmatic study that assesses the efficacy of such an approach is necessary.

The search methods for the optimal architectures of MLP could be improved. Evolutionary programming [254] and weight-agnostic neural network [243] are promising approaches. Such improvement could find the MLP architectures' better performance in shorter computation time.

Conclusion

The protocol for developing and validating a prediction engine for health behavior was developed. As a case study, walking behavior classification models were developed and validated. MLP showed the highest overall performance of all tried algorithms, yet it needed relatively higher computation time. A random search for optimal layer structure was a promising approach for prediction engine development.

Acknowledgments

Appendix 1, in full, is a reprint of the material as it appears in Park, Junghwan, Gregory J. Norman, Predrag Klasnja, Daniel E. Rivera, and Eric Hekler. 2023. "Development and Validation of Multivariable Prediction Algorithms to Estimate Future Walking Behavior in Adults: Retrospective Cohort Study." *JMIR mHealth and uHealth* 11 (January): e44296. The National Library of Medicine (R01LM013107) funded JP's stipend.

Authors' Contribution

JP conceptualized the research question, analyzed the data, and wrote the manuscript. PK provided the data. DR provided the program code library to assist the analysis. EH provided guidance at each stage of study. All authors contributed to the writing of the manuscript.

Conflicts of Interest

JP is an employee of Korean National Government, the Ministry of Health and Welfare. GJN is an employee of Dexcom, Inc.

Supplemental Table 8.1 TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	188
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	188
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	189
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	189
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	191
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	191
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	191
	5b	D;V	Describe eligibility criteria for participants.	191
	5c	D;V	Give details of treatments received, if relevant.	N/A
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	193
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	191
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	193
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	191
Sample size	8	D;V	Explain how the study size was arrived at.	191
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	191
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	6
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	192
	10c	V	For validation, describe how the predictions were calculated.	193
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	195
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	N/A
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	N/A
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	195

Supplemental Table 8.1. TRIPOD Checklist: Prediction Model Development and Validation, continued

Section/Topic	Item		Checklist Item	Page
Results				
	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	196
Participants	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	196
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	196; N/A due to K-fold CV
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	196
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	N/A
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	199
	15b	D	Explain how to the use the prediction model.	198
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	199
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	N/A
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	205
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	N/A
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	202
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	202
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	N/A
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	207

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.

Appendix 2. Recruitment Materials

Study Website

Sitting too much and looking for help to get more active?

Want a free Fitbit Versa ($229 value) & $75 for participating in research?

If yes, click below to check your eligibility or read on for more information.

I'M INTERESTED!

Background: It is common for smartphone apps to send too many notifications when trying to help someone be active. This can result in a person ignoring the notifications and not benefiting from the app's support.

Purpose: The purpose of this study is to gather data to support the development of a future smartphone app meant to help a person increase their physical activity that only provides support when and where it is needed for each person.

Who is this study for? This study is for you if you:

- currently engage in l**ess than 60 minutes of exercise** per week;
- **Want to get more active** by increasing the steps you take throughout your day;
- are **at least 21 years old**;
- are **healthy enough to participate**;
- **live in the United States**;

- **own a smartphone** (iPhone or Android)

What will happen during the study? The entire study will be conducted remotely. You will first undergo an interview with one of our staff to determine if you are eligible for the study. If you are eligible, you will be asked to complete a baseline survey and to also take part in an onboarding session to help you get set up with your free Fitbit Versa 3 and the study app (available for either Android or iPhone). We will then ask you to wear the Fitbit for 10 days without any support provided. This will allow the app to learn about your normal activity patterns, so that it can provide personalized support. After 10 days, the app will provide personalized suggested daily step goals and also send notifications with suggestions on building up your skills in fitting in brief, 10-15 minute walking bouts throughout your day. Each evening, the app will ask you to fill out a brief survey, which was specifically designed to help you think about what will happen the next day. Prior research shows filling out a survey like this can be valuable itself, for helping you increase your physical activity. After the intervention period, we will ask you to participate in one final interview.

How much time will this take? The study will last for approximately **9 months (270 days total)** and, for each day, you will only be asked to engage with the app for **5-10 minutes each day**. In addition, there will be three interviews: Enrollment, Pre-Study, and Post-Study. Each interview will last 30-60 minutes.

What benefits will I get from participating? You will be **given a Fitbit Versa 3 to keep ($229 value)** as long as you participate in at least the first 3 months of the study. In addition, you will be provided $25 if you complete at least 80% of the phone surveys sent to you over a 3-month period. There will be a total of 3 3-month periods, hence you have the opportunity to **receive $75 for participation**. If you miss one period, you are still eligible for the $25 from the next time window. Last, while it cannot be guaranteed, it is possible that participating in the study will help you to develop a regular physical activity routine.

Who is conducting the study? Dr. Eric Hekler, who is a Professor and Interim Associate Dean for Community Partnerships at the Herbert Wertheim School of Public Health and Human Longevity Sciences at UC San Diego and the Director of the Center for Wireless and Population Health Systems is leading this project along with his associates, Drs. Daniel Rivera from Arizona State University and Predrag "Pedja" Klasnja from the University of Michigan.

What if I change my mind later?

- **Research is voluntary** – whether or not you join is your decision. You can discuss your decision with others (such as family, or friends).
- You can say **yes but change your mind later**.
- If you say no, we will not hold your decision against you.
- Your decision will not affect any other benefits you may be entitled to.
- Please ask questions or mention concerns before, during, or after the research.

I'm in! What's next?

Click HERE and fill out the form to check if you are eligible.

Our staff will contact you soon.

I'm interested but have questions. Who do I contact?

Please send an email to justwalk@ucsd.edu. Study staff will respond within 1-2 business days.

UCSD IRB#: 800132

Appendix 3. Consent Form

UCSD IRB#: 800132

University of California, San Diego
Consent to Act as a Research Subject

"Just Walk"
-A research study to create and test a smartphone app intervention that provides the right support at the right time

Introduction

Dr. Eric Hekler and associates are conducting this research and asking for your consent to participate. This section provides a summary of important information about the study. The rest of the form provides additional details.

- Research is voluntary - whether or not you join is your decision. You can discuss your decision with others (such as family, or friends).
- You can say yes but change your mind later.
- If you say no, we will not hold your decision against you.
- Your decision will not affect any other benefits you may be entitled to.
- Please ask questions or mention concerns before, during, or after the research.

It is common for smartphone apps that try to help someone change their behavior to send a lot of notifications. The problem is these notifications can become overwhelming, particularly when they are not sent when they would be most helpful. This study is the first step to develop a smart mobile app to nudge people to walk more with alerts only sent when they will be beneficial. The study compares a variety of different ways to figuring out when to send messages, such as reducing the total number sent during the day, or only sending them if a person is not meeting their goals. The data collected through this study will help us to understand strategies that can be used to make sure behavioral support is only provided when it would actually be useful. If successful, the strategy will be shared with the public, thus providing guidance to digital health companies on how to provide the right support at the right time, minimizing unhelpful alerts on your phone from a fitness app.

You will first undergo an interview with one of our staff to determine if you are eligible for the study. If you are eligible, you will receive a wearable device ("Fitbit") and install our mobile app on your phone. The study will be about 9 months long (270 days total). During that time, you will be guided to actively use the app and answer a few questions on the app, which will require 5-10 minutes of your attention each day. In addition, there will be three interviews: Enrollment, Pre-Study, and Post-Study. Each interview will last 1-2 hours. For your participation, you will be given a Fitbit Versa 3 to keep after study completion ($229 value). In addition, you will be provided $25 for each 3-month period of participation if you complete the interview and at least 80% of phone surveys (takes less than 1.5 hours in total to complete the surveys over 3 months)

during the 3-month window. If you miss one period, you are still eligible for the $25 from the next time window.

This app only nudges you to walk occasionally. The level of risk is low and is similar to the same risks you experience from walking every day. As with any exercise, it is possible you might trip or fall while walking or you might experience mild lightheadedness or some other soreness or discomfort. Although it is highly unlikely for you to experience any severe health risks during this study, please share your concerns during the enrollment interview.

Additional, detailed information about this research is provided below. Please feel free to ask questions before signing this consent.

Why have you been asked to participate, how you were selected, and what is the approximate number of participants in the study?
You have been asked to participate in this study because you own a smartphone, you reported engaging in less than 60 minutes per week of exercise, you are interested in starting to walk more, you are at least 25 years old, and you are healthy enough to participate in the study. There will be approximately 60 participants.

What will happen to you in this study?
In addition to the information at the beginning of this form, here are some additional details about what will happen if you agree to be in this study,

You will take part in the following activities:

- A 15-minute consent and enrollment virtual meeting (this meeting)
- A Pre-intervention virtual interview and survey (30-60 minutes). During this interview we will discuss baseline survey, how to use the study app, the details of the intervention and the way to get tech support (if needed).
- The first 10 days are called “the baseline phase”. During this time, you will be asked to go about your normal everyday activities while wearing the Fitbit, day and night. You will also be prompted each day to complete a short (less than 5 minute) survey through the app (10 days, ≤5 minutes/day interaction with app = 50 minutes in total across 10 days).
- After the first 10 days, “the study phase” will start and will last for 260 days, or just over 8 months. During this time, you will wear the Fitbit, day and night, and the app will provide suggestions and support to help you increase your steps per day. Every day, you will be prompted to interact with the app for between 5-10 minutes (≤10 minutes/day interaction with app = 2,600 minutes (approx. 43.3 hours) across 260 days).

- A 20 minute post-intervention virtual interview whereby we will ask you about your experiences with using the app. At the end of this interview, we will ask if we can continue to monitor your natural use of Fitbit for up to 1-year post intervention.
    - If you opt in, there will be no expectations upon you to use the app, use the intervention, or even to continue to use the Fitbit. We would monitor what you naturally do with a Fitbit.
    - If you opt out, we will disconnect monitoring your Fitbit.

All meetings' audio will be temporarily recorded, then transcribed without personal information. After transcription, all audio files will be deleted.

You will be asked to wear the Fitbit, day and night, for 270 days (just over 8 months).

While taking part in the study, you will receive:

- A Fitbit activity tracker;
- a suggested step goal each day that will strive to challenge you but still be doable for you, based on the number steps per day measured earlier in the study;
- notifications meant to help you find ways to plan short walks in the next 3 hours, sent sometime between 0 to 4 times per day
- a daily survey which consists of approximately half a dozen multi-choice questions and takes less than one minute to complete,
- a quick reflection survey after your activity which consists of approximately half a dozen multi-choice questions and takes less than one minute to complete, and
- a weekly reflection survey, which consists of about one dozen multi-choice questions and takes approximately two to three minutes to complete, at your preferred time.

How much time will each study procedure take and how long will the study last?
This study consists of two phases: the baseline and the study phase.

1. **The baseline phase** will last 10 days. You will not receive any notifications focused on increasing your steps during this period. The app will gather data about your usual level of physical activity and ask you questions. The data gathered in this phase will be used to adjust initial support to your personal needs.
2. **The study phase** will last 260 days (approximately 9 months). As explained in the last section, you will use the app to measure your physical activity and answer a few daily questions. These questions are about how you are feeling today in general, and specifically about physical activity. These questions are asked to help us improve our capacities and providing support only during times that would actually be beneficial for someone.

What risks are associated with this study?

Participation in this study may involve some added risks or discomforts. In addition to the risks described at the beginning of this form, there are a few additional risks.

1. During the enrollment interview, subjects may feel anxiety and/or embarrassment when answering personal questions on medical and lifestyle history and completing the questionnaires, including possible feelings of inadequacy if they cannot meet their criteria.
2. Walking, even if it is a voluntary activity, may produce light-headedness, fatigue, nausea, chest discomfort, or delayed-onset muscle soreness, especially under severe weather.
3. Subjects may feel
    a. feelings of inadequacy or embarrassment if unable to succeed at the steps goal,
    b. concern for privacy related to divulging personal information and security of providing personal information through the Fitbit app,
    c. injury during physical activity,
    d. the rare occurrence of a cardiovascular event during physical activity, and
    e. physical discomfort related to wearing the Fitbit Versa.
4. Participating in the study might involve a level of effort or burden (likely in the realm of 2-3 minutes per day), especially since the mobile applications will be designed to help participants think about their health and health behaviors more frequently.
5. Some participants might worry about the security of their data on study servers and on their mobile phones. During the study, even with our best effort, there is a risk to lose confidentiality of your data.

The risks involved in this study are no more than minimal and are reasonable in relation to the potential knowledge that may result from this study.

Because this is a research study, there may be some unknown risks that are currently unforeseeable. You will be informed of any significant new findings.

Are there risks to the reproductive system or a developing fetus?

A moderate level of walking is safe for pregnancy unless there are specific medical reasons. This study does not push you to vigorous exercise which could be potentially harmful to a developing fetus under certain circumstances. However, if there is a concern about being unable to meet your will to exercise due to ongoing reproductive procedures or pregnancy, please share this information during the enrollment interview.

Participation in this study does not create risks for the reproductive system.

What benefits can be reasonably expected?
In addition to the benefits listed at the beginning of this form, the investigators may also learn more about the ways in which people react to app notifications, or not, depending on the circumstances. This knowledge will help researchers to create a better app that reduces unnecessary notifications and, hopefully maximizes benefits.

Further, it also gives us the knowledge to help people to prevent chronic diseases such as diabetes or heart problems. Since this study introduces a new engineering concept called control systems engineering into the mobile health sector, it will develop health science in general. The tools for this project will be published as open-source to be used freely by other researchers.

What happens if you change your mind about participating?
If you decide that you no longer wish to continue in this study, you will be requested to notify the study team and respond to a "reasons to terminate" survey. It will only take less than 5 minutes. You may retain the provided Fitbit device.

You will be told if any important new information is found during the course of this study that may affect your wanting to continue.

Can you be withdrawn from the study without your consent?
You may be withdrawn from the study for the following reasons:

1. In the view of the PI, participating in the study poses a risk to your health or welfare.
2. Stop responding to study staffs' contact regarding things like not recharging your smartwatch, not using the app, and not responding to survey questions.

You may also be withdrawn from the study if you do not follow the instructions given to you by the study personnel. You may retain the provided Fitbit device even if you have withdrawn.

Will you be compensated for participating in this study?
For your participation, you will be given a Fitbit Versa 3 to keep after study completion ($229 value). In addition, you will be provided $25 for each 3-month period of participation if you complete the interview and at least 80% of phone surveys during the 3-month window, potentially $75. If you miss one period, you are still eligible for the $25 from the next time window.

Are there any costs associated with participating in this study?
There will be no cost to you for participating in this study. The download size of the apps for this study is approximately 200MB and it constantly connects to the internet. It will download less than a few megabytes even with extensive use of the app. However, depending on your mobile contract, this amount of data transfer can cost you. This cost will not be reimbursed. If available,

it is recommended that you download the apps when you can access cost-free internet such as wi-fi.

Along with the study app, you will also be requested to install the Fitbit app which includes paid premium services. You may purchase this service, however it is not required for this study. If you do choose to purchase this service, you will NOT be reimbursed for it.

What about your confidentiality?
The following data will be collected:

- Interview audio (temporary, deleted after transcription) and transcript
- Survey responses
  - Gathered from the meetings
  - Daily inputted online survey
- Activity data
  - Number of steps
  - Heartrates
  - Activity intensity: light, moderate, or vigorous
  - Estimated Calories used by physical activities
  - Distances of walking, running, or participant inputted activities
  - Number of floors you climb
  - Sleep hours
- Study App usage data
  - When and what page of our app is opened by participant

Research records will be kept confidential to the extent allowed by law.
All measurements will be de-identified (i.e., processed or partially deleted to make it hard to know whose data it is) and treated confidentially. Written consent forms will be kept in separate locked files cabinets separate from participant data on the study's secure database so that individuals are not easily connected to the study results. All data from the Fitbit Versa will be continuously, passively, and securely streamed to the Fitbit website. It will then be retrieved using software developed by Fitbit Inc (https://www.fitbit.com) and stored securely on their servers. All data that is subsequently downloaded from Fitbit Inc. servers for analytic purposes will remain de-identified and will be stored on secure, password-protected UCSD servers.

Research records may be reviewed by the UCSD Institutional Review Board and Dr. Hekler's associates at the University of Michigan (Dr. Predrag Klasnja) and Arizona State University (Dr. Daniel E. Rivera).

After this study is over, information and data from this study will be de-identified (i.e., delete all personal identifying information) and maintained indefinitely. It could be used in future research studies or distributed to another investigator without your additional consent. If this were to occur, it will be remained unidentifiable status. For example, step totals from this study could be shared with another investigator but there wouldn't be any way to tie those step totals back to a specific person. The investigator will not be given your personal information including name, or your contact.

A description of this clinical trial will be available on http://www.ClinicalTrials.gov, as required by U.S. Law. This Web site will not include information that can identify you. At most, the Web site will include a summary of the results. You can search this Web site at any time.

Will you receive any results from participating in this study?

Study staff will use the results of this study to present our findings via presentation(s) at scientific meetings or in scientific publications. When results are made public, your identity will not be shared. Additionally, after the data is collected and we have had time to analyze the data, we will share your personal results with you, if you'd like. We will be in touch after study completion to offer you the opportunity to learn more about your personal results. This will include the option for you to retain your personal data, if you so choose, as well as any personalized results developed about and for you. You may opt-out from this communication. Once the study is completed and we have had the chance to offer all participants personalized results, we will produce a final de-identified dataset that will be retained. We will then permanently delete all identifying information and capacities to link the study data with individuals, including yourself.

Who can you call if you have questions?

This study has been explained to you and your questions have been answered by the signing witnesses. If you have other questions or research-related problems, you may reach Eric Hekler at (858) 429-9370. Or, send an email to us (justwalk@ucsd.edu).

You may call the Human Research Protections Program Office at 858-246-HRPP (858-246-4777) to inquire about your rights as a research subject or to report research-related problems.

UCSD IRB#: 800132

Your Signature and Consent
You have received a copy of this consent document to keep.

You agree to participate.

Full name of the subject (print)	Signature	Date
_____	_____	_____
Full name of the person conducting the informed consent discussion (print)	Signature	Date

The research team may contact you following the study to ask additional questions and/or regarding potential participation in other research. Please mark whether you agree to be recontacted following the conclusion of the study.

Yes, you agree to be recontacted.

No, you do not agree to be recontacted

_____	_____
Subject's signature	Date

Appendix 4. Opportunity Condition Operationalization

The specific operationalization of the opportunity conditions used in the clinical trial consists of three steps.

1. Selection of the calculation period and data acquisition

Depending on the length of time since the start of the intervention for each individual, the selection of the period to be calculated or the criteria to be utilized for calculating the Opportunity condition varies. In the early days of the intervention (days 0-7), we utilized the past three days of data. We split the past days into weekends and weekdays so that only weekday data was used to calculate the opportunity condition for weekdays, and only weekend data was used to calculate the opportunity condition for weekends.

For example, if a participant started the intervention on Thursday, January 1, and the opportunity condition for Tuesday, January 6 needs to be calculated, January 6 is a weekday, so data from the three closest past weekdays would be used: 1/5 (Mon), 1/2 (Fri), 1/1 (Thu). If an opportunity condition needs to be calculated for Sunday, January 4, it uses data from the three closest past weekends: 1/3 (Sat), 12/28 (Sun), and 12/27 (Sat).

For each day, the calculation uses data from 1440 steps per minute, assigning a value of 1 if the number of steps in each minute is greater than 60 and 0 if it is less than 60. This process is the most essential step in determining whether a participant took steps intentionally during each minute.

Then, during the mid-intervention period (days 8-21), we utilized the past 8 days of data, divided into weekends and weekdays. In the late intervention period (days 22 and beyond), we utilized the past 5 days of data, but only from the same day of the week.

2. Eliminating tiny gaits and treating short breaks between exercises as exercise

We utilized a moving average. We used 7 minutes as the window size for the moving average, averaging the 1440 minutes of each day in a 7-minute rolling window, i.e., averaging the minutes from 0 to 6 (i.e., whether the person walked more than 60 steps/minute, 1 if they did, 0 if they did not), and marking the minutes as 1 if the average was above 0.55, i.e., if they walked more than 4 out of 7 minutes. So, within the 7-minute window, if they walked for more than 4 minutes, even if it was intermittent, they were marked as having walked.

To understand the effect of these procedures, consider an ideally simple case. Suppose that on a given day, a participant took a continuous walk of 60 or more steps/minute for a total of s minutes, from the k-th minute to the $(k + s - 1)$-th minute, with a total of s minutes.

Window Coordinate	**Window Start**	**Window End**	**# of Active Minutes (>60)**	**Activity**
k-1	k-4	k+2	3	0
k	k-3	k+3	4	1
k+1	k-2	k+4	5	1
...				
k+s-2	k+s-5	k+s+1	5	1
k+s-1	k+s-4	k+s+2	4	1
k+s	k+s-3	k+s+3	3	0

This process will not change the data as long as the active minutes are continuous. However, if there are 1-3 inactive minutes in the middle of a continuous walk, the inactive minutes will be populated with 1. On the other hand, if a walk only lasts 1-3 minutes, the active minutes will be ignored if this procedure is performed.

3. Only selecting the moderate possibilities

For the minute-by-minute activity we have, we extract only the 3 hours we are currently interested in. (For example, if the decision point index is 0, only utilize data from hours 7-10, depending on their choice). Suppose any of these 3 hours had at least 1 minute of ACTIVITY (only consecutive walks of 4 or more minutes survived step 2, with consecutive walks of 3 minutes or less being eliminated). In that case, we consider that **3 hours were ACTIVE**.

We divide the total daily activity (note that we only target the same decision points, i.e., the same time of day) by the number of days covered, i.e., we average over three days if the intervention is 7 days or less, eight days if it is 21 days or less, and five days if it is more than 21 days. The Opportunity condition will be positive if the average value is greater than or equal to 0.55 but less than or equal to 0.8. The specific but illustrative requirements for an active day to satisfy these conditions are as follows.

Period	Sample Size	Too low opportunity (O=0)	"Just right" opportunity (O=1)	Too high opportunity (O=0)
d=0-7	3 days	0-1 active days	2 active days	3 active days
d=8-21	8 days	0-4 active days	5-6 active days	7-8 active days
d=21-	5 days	0-2 active days	3-4 active days	5 active days

Appendix 5. Fidelity Check Results

Overall Distribution

The overall distribution of the experimentation and measurement data points are shown in Table 8.7. The intervention period was given as 260 days for all participants and 4 decision points per day were planned, the total number of planned decision points per participants were 1040 times. However, due to various reasons including withdrawal (n=2), extremely low adherence (n=2), study server system's spontaneous malfunction, and premature termination of intervention due to staff error (n=14, 4.95 days on average), daytime travels that crosses the time zone borders, there were participants who had less than 1,040 decision points. The average decision points were approximately 924.8 and the median value of numbers of the decision points was 993.5 (95.5% of 1040).

On average, the participants received approximately 249.6 walking suggestion notifications during the 260-day intervention period. As our intervention algorithms are designed to heavily depend on their behavior patterns, the variation was high (SD=61.3).

The average number of steps per day of all participants was 5,852.7 steps/day. The wearing time were estimated via the heart rates measured by Fitbit.

The percentage of the time of a day wearing the Fitbit during the intervention across all the participants was 76.0%. Median value was quite high (82.1% of 24 hours), which potentially means that most of participant wore the Fitbit even during nighttime, as recommended during the pre-intervention meeting.

Among 260 days of each participant, on average, the participants wore 210.7 days over 8 hours a day. The standard deviation of this number was high (57.7), which means that the variance between individuals was high (i.e., most participants wore the Fitbit well, whereas some did not.) The median value was 229.5 days, and 5 percentile was 66.8 days.

The decision policies were assigned with the identical portions across participants (i.e., Full: ~50%, N+O: ~17%, N+R: ~16%, Random: ~16%). However, due to the reasons that limited the total decision points as described above (e.g., withdrawal, human and system errors), the numbers of the decision points per decision policy were slightly off from the original planned numbers. However, within a person, the portions were well maintained, because of the pseudo-randomized signal. (See "System ID Overview" section on page 76)

Table 8.7 Overall distribution of the data.

Characteristics	**Mean (SD) across individuals**	**Distribution (5%, 50%, 95%)**
Number of decision points	924.8 (167.9)	(613.3, 993.5, 1037.0)
[Percentage of planned decision points]	88.9%	(59.0%, 95.5%, 99.7%)
Number of walking suggestion notification	249.6 (61.3)	(139.4, 262.0, 319.6)
Number of steps per day of all participants during the baseline	6,326.3 (2,792.1)	(2,833.1, 5,928.0, 10,705.6)
Number of steps per day of all participants during the intervention	5,852.2 (2,290.1)	(1,308.5, 5,457.8, 11,277.2)
Percentage of time of a day wearing the Fitbit during the intervention (%, %p)	76.0 (20.7)	(33.1, 82.1, 95.6)
Number of days out of 260 when the participants wore Fitbit at least 8 hours	210.7 (57.7)	(66.8, 229.5, 259.0)
Number of decision points for each JIT decision policy assignment per user		
Full (N+O+R) JIT	458.8 (96.0)	(267.2, 495.5, 537.9)
N+O JIT	163.8 (32.2)	(119.8, 168.0, 200.0)
N+R JIT	151.1 (30.1)	(114.2, 158.0, 190.8)
Random	151.1 (32.4)	(105.2, 154.0, 191.4)
[Percentage of planned decision points]		
Full (N+O+R) JIT	49.6%	(28.9, 53.6, 58.2)
N+O JIT	17.7%	(12.9, 18.2, 21.6)
N+R JIT	16.3%	(12.3, 17.1, 20.6)
Random	16.3%	(11.4, 16.7, 20.7)

Step Distribution Across Decision Policies

The participants, in overall, did not show significant difference between assigned JIT decision policies (Table 8.8 and Figure 8.8).

Table 8.8 Distribution of average steps during 3 hours after decision points per user stratified by assigned JIT decision policies.

Decision policies	Average Number of decision points	Average Number of Notifications[1] Mean (SD)	Average Number of Steps Mean (SD) (steps/3hr)	Distribution of Average Number of Steps (5%, 50%, 95%) (steps/3hr)
Full (N+O+R) JIT	458.8 (96.0)	75.3 (29.7)	1167.3 (535.0)	(75.1, 919.5, 3099.1)
N+O JIT	163.8 (32.2)	70.0 (29.1)	1138.0 (494.5)	(56.5, 901.1, 3062.3)
N+R JIT	151.1 (30.1)	28.0 (16.3)	1133.9 (467.3)	(61.0, 890.5, 3061.5)
Random	151.1 (32.4)	76.3 (16.7)	1146.5 (506.2)	(58.2, 881.3, 3120.4)

[1] Notifications are expected to trigger walking behavior, average numbers of notifications were also shown. No significant difference of the notification density (i.e., number of notifications per day) was found in pairwise t-test between decision policies.

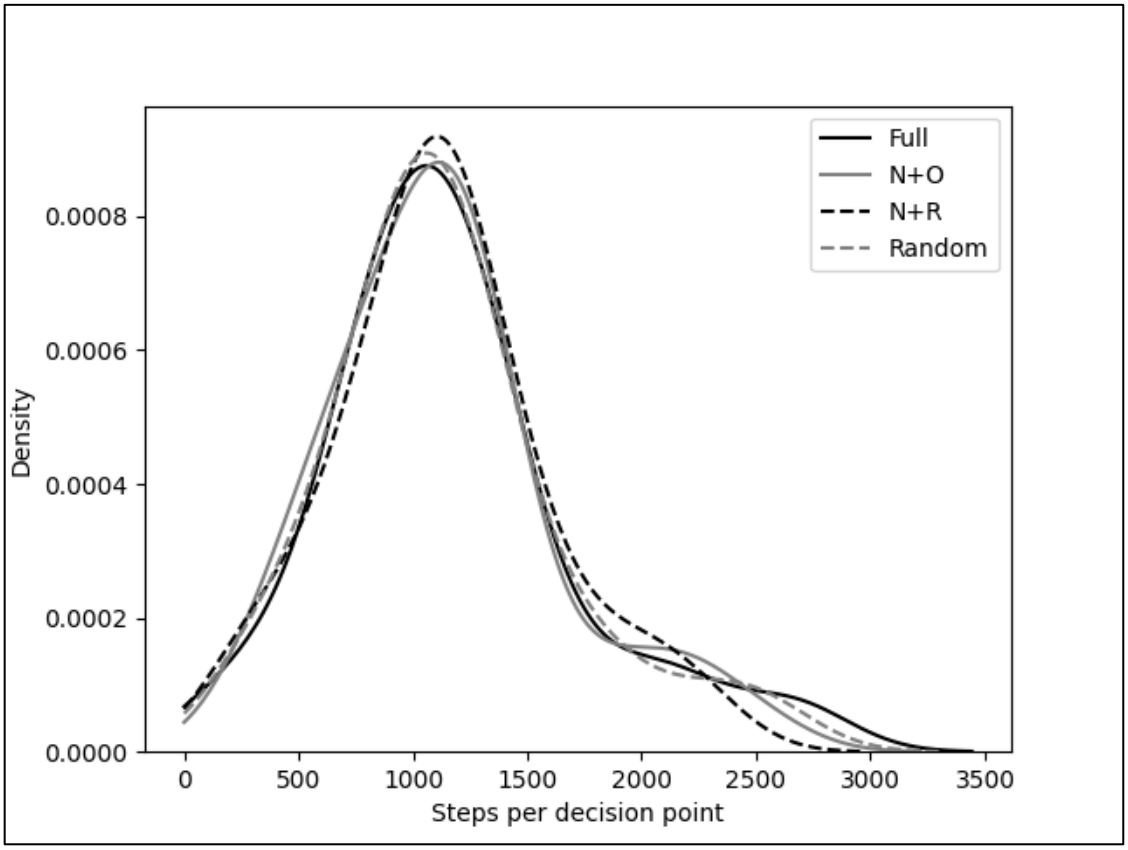


Figure 8.8 Visual representation of the distribution of average steps during 3 hours after decision points per user per assigned JIT decision policies.

Distribution Across Time Conditions

The decision points were grouped into four categories: weekday morning, weekday afternoon, weekend morning, weekend afternoon. Monday through Friday are categorized as

weekdays. First two decision points of each day are categorized as morning. To estimate the level of activity after each decision points, step counts during 180 minutes after the decision points were used.

Across the time conditions, the probability of receiving notifications were highly similar. (See Table 8.9) On the other hand, the general activity level of weekend afternoon was significantly higher than all other 3 time condition.

Table 8.9 Distribution of average notification and short-term step count per user stratified by time condition.

Time Condition	**Average Probability of Receiving Notifications** % (SD, %p)	**Average Number of Steps in 180 minutes** Mean (SD) (steps)	**Distribution of Average Steps in 180 minutes** (5%, 50%, 95%) (steps)
Weekday Morning	25.7 (7.0)	1078.8 (504.9)	(67.4, 859.1, 2897.6)
Weekday Afternoon	28.0 (8.4)	1153.4 (500.1)	(79.0, 947.9, 2905.2)
Weekend Morning	27.3 (8.7)	1179.7 (709.0)	(85.5, 915.6, 3163.8)
Weekend Afternoon	27.3 (8.1)	1321.8 (649.7)	(131.4, 1071.1, 3364.0)

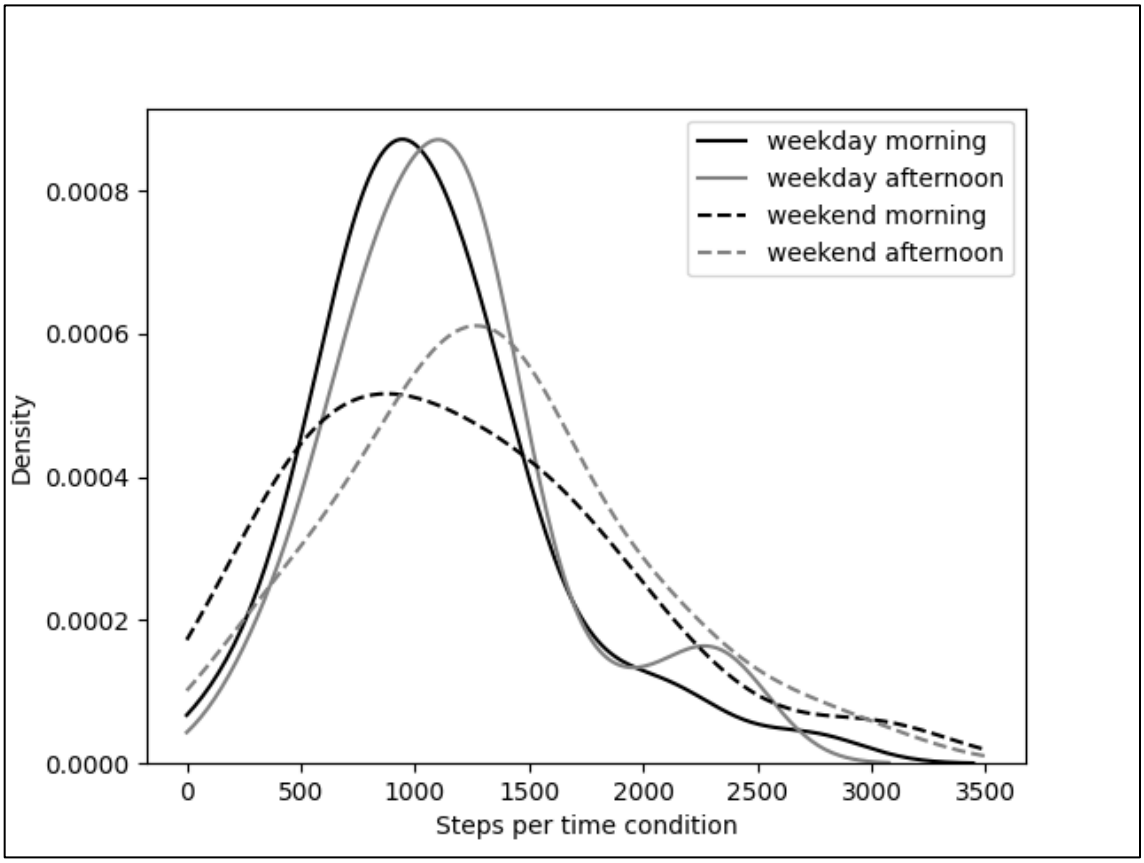


Figure 8.9 Distribution of average steps per user per time condition

Appendix 6. Analysis on Misalignment Between Real Time JIT States and Post Hoc JIT States

Table 8.10 Summary of misalignment between real-time and *post hoc* Need states

***post hoc* Need** **Real-time Need**	**Positive**	**Negative**	**Total per real-time estimation**
Positive	26,142 (85%, 83%)[25]	4,596 (15%, 34%)	30,738 (100%, 69%)
Negative	80 (1%, 0%)	6,740 (99%, 50%)	6,820 (100%, 15%)
Not evaluated then[26]	5,178 (71%, 16%)	2,097 (29%, 16%)	7,275 (100%, 16%)
Total per *post hoc* estimation	31,400 (70%, 100%)	13,433 (30%, 100%)	44,833 (100%, 100%)

- Portion of the aligned samples of all samples: 73%
- Portion of the aligned samples of real-time estimates: 88%
- Recall (Portion of the true positive among real-time positive): 85%
- Precision (Portion of real-time positive among *post-hoc* positive): 83%
- False discovery rate (Portion of false positive among *post-hoc* positive): 0%
- False negative rate (Portion of false negative among real-time positive): 15%

[25] Row-wise portion, then column-wise portion.

[26] During the intervention, the JIT states that are not necessary at the moment were not evaluated, depending on the decision policies of the day (e.g., Random).

Table 8.11 Summary of misalignment between real-time and post hoc Opportunity states

***post hoc* Opportunity** **Real-time Need**	**Positive**	**Negative**	**Total per real-time estimation**
Positive	1,977 (13%, 45%)[27]	12,944 (87%, 32%)	14,921 (100%, 33%)
Negative	972 (6%, 22%)	14,116 (94%, 35%)	15,089 (100%, 34%)
Not evaluated then[28]	1,424 (10%, 33%)	13,399 (90%, 33%)	14,823 (100%, 33%)
Total per *post hoc* estimation	4,374 (10%, 100%)	40,459 (90%, 100%)	44,833 (100%, 100%)

- Portion of the aligned samples of all samples: 36%
- Portion of the aligned samples of real-time estimates: 54%
- Recall (Portion of the true positive among real-time positive): 13%
- Precision (Portion of real-time positive among *post-hoc* positive): 45%
- False discovery rate (Portion of false positive among *post-hoc* positive): 22%
- False negative rate (Portion of false negative among real-time positive): 87%

[27] Row-wise portion, then column-wise portion.

[28] During the intervention, the JIT states that are not necessary at the moment were not evaluated, depending on the decision policies of the day (e.g., Random).

Table 8.12 Summary of misalignment between real-time and post hoc Receptivity states

***post hoc* Opportunity** **Real-time Need**	**Positive**	**Negative**	**Total per real-time estimation**
Positive	12,027 (82%, 40%)[29]	2,648 (18%, 18%)	14,675 (100%, 33%)
Negative	11,079 (75%, 37%)	3,736 (25%, 26%)	14,815 (100%, 33%)
Not evaluated then[30]	7,213 (47%, 24%)	8,130 (53%, 56%)	15,343 (100%, 34%)
Total per *post hoc* estimation	30,319 (68%, 100%)	14,514 (32%, 100%)	44,833 (100%, 100%)

- Portion of the aligned samples of all samples: 35%
- Portion of the aligned samples of real-time estimates: 53%
- Recall (Portion of the true positive among real-time positive): 82%
- Precision (Portion of real-time positive among *post-hoc* positive): 40%
- False discovery rate (Portion of false positive among *post-hoc* positive): 37%
- False negative rate (Portion of false negative among real-time positive): 18%

[29] Row-wise portion, then column-wise portion.

[30] During the intervention, the JIT states that are not necessary at the moment were not evaluated, depending on the decision policies of the day (e.g., Random).

REFERENCES

1. Nahum-Shani I, Hekler EB, Spruijt-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. Health Psychol American Psychological Association; 2015;34(S):1209.

2. Hekler EB, Rivera DE, Martin CA, Phatak SS, Freigoun MT, Korinek E, Klasnja P, Adams MA, Buman MP. Tutorial for Using Control Systems Engineering to Optimize Adaptive Mobile Health Interventions. J Med Internet Res 2018 Jun 28;20(6):e214.

3. Phatak SS, Freigoun MT, Martín CA, Rivera DE, Korinek EV, Adams MA, Buman MP, Klasnja P, Hekler EB. Modeling individual differences: A case study of the application of system identification for personalizing a physical activity intervention. J Biomed Inform 2018 Mar;79:82–97. PMID:29409750

4. Martín CA, Rivera DE, Hekler EB. Design of Informative Identification Experiments for Behavioral Interventions. IFAC-PapersOnLine 2015 Jan 1;48(28):1325–1330.

5. Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A, Murphy SA. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. Health Psychol 2015 Dec;34S:1220–1228. PMID:26651463

6. Schmid C, Yang J. Bayesian Models for N-of-1 Trials. Harv Data Sci Rev 2022 Sep 8;2022(SI3). PMID:38283071

7. McTiernan A, Friedenreich CM, Katzmarzyk PT, Powell KE, Macko R, Buchner D, Pescatello LS, Bloodgood B, Tennant B, Vaux-Bjerke A, George SM, Troiano RP, Piercy KL, 2018 PHYSICAL ACTIVITY GUIDELINES ADVISORY COMMITTEE*. Physical Activity in Cancer Prevention and Survival: A Systematic Review. Med Sci Sports Exerc 2019 Jun;51(6):1252–1261. PMID:31095082

8. Murtagh EM, Nichols L, Mohammed MA, Holder R, Nevill AM, Murphy MH. The effect of walking on risk factors for cardiovascular disease: an updated systematic review and meta-analysis of randomised control trials. Prev Med 2015;72:34–43. PMID:25579505

9. Kokkinos P. Physical activity and cardiovascular disease prevention: current recommendations. Angiology journals.sagepub.com; 2008 May 28;59(2 Suppl):26S–9S. PMID:18508850

10. Karstoft K, Winding K, Knudsen SH, Nielsen JS, Thomsen C, Pedersen BK, Solomon TP. The effects of free-living interval-walking training on glycemic control, body composition, and physical fitness in type 2 diabetic patients: a randomized, controlled trial. Diabetes Care 2013;36(2):228–236. PMID:23002086

11. Shah SZA, Karam JA, Zeb A, Ullah R, Shah A, Haq IU, Ali I, Darain H, Chen H. Movement is Improvement: The Therapeutic Effects of Exercise and General Physical Activity on Glycemic Control in Patients with Type 2 Diabetes Mellitus: A Systematic

Review and Meta-Analysis of Randomized Controlled Trials. Diabetes Ther 2021 Mar;12(3):707–732. PMID:33547579

12. Physical Activity Guidelines Advisory Committee report, 2008. To the Secretary of Health and Human Services. Part A: executive summary. Nutr Rev Oxford University Press (OUP); 2009 Feb;67(2):114–120. PMID:19178654

13. Tudor-Locke C, Craig CL, Thyfault JP, Spence JC. A step-defined sedentary lifestyle index: <5000 steps/day. Appl Physiol Nutr Metab 2013;38(2):100–114. PMID:23438219

14. Bauer UE, Briss PA, Goodman RA, Bowman BA. Prevention of chronic disease in the 21st century: elimination of the leading preventable causes of premature death and disability in the USA. Lancet Elsevier BV; 2014 Jul 5;384(9937):45–52. PMID:24996589

15. Owen N, Bauman A, Brown W. Too much sitting: a novel and important predictor of chronic disease risk? British Journal of Sports Medicine 2009;43(2):81–82.

16. King AC, Sallis JF. Why and how to improve physical activity promotion: lessons from behavioral science and related fields. Preventive Medicine 2009;49(4):286–288.

17. Haskell WL, Blair SN, Hill JO. Physical activity: Health outcomes and importance for public health policy. Preventive Medicine. Academic Press; 2009. p. 280–282. PMID:19463850

18. Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, George SM, Olson RD. The Physical Activity Guidelines for Americans. JAMA 2018 Nov 20;320(19):2020.

19. Services, US Department of Health Human. Physical activity guidelines advisory committee report, 2008. Washington, DC; 2008.

20. Saint-Maurice PF, Troiano RP, Bassett DR Jr, Graubard BI, Carlson SA, Shiroma EJ, Fulton JE, Matthews CE. Association of Daily Step Count and Step Intensity With Mortality Among US Adults. JAMA 2020 Mar 24;323(12):1151–1160. PMID:32207799

21. Kelly P, Kahlmeier S, Götschi T, Orsini N, Richards J, Roberts N, Scarborough P, Foster C. Systematic review and meta-analysis of reduction in all-cause mortality from walking and cycling and shape of dose response relationship. Int J Behav Nutr Phys Act 2014 Oct 24;11:132. PMID:25344355

22. Ku P-W, Hamer M, Liao Y, Hsueh M-C, Chen L-J. Device-measured light-intensity physical activity and mortality: A meta-analysis. Scand J Med Sci Sports 2020 Jan;30(1):13–24. PMID:31545531

23. Tudor-Locke C, Craig CL, Brown WJ, Clemes SA, De Cocker K, Giles-Corti B, Hatano Y, Inoue S, Matsudo SM, Mutrie N, Oppert J-M, Rowe DA, Schmidt MD, Schofield GM, Spence JC, Teixeira PJ, Tully MA, Blair SN. How many steps/day are enough? For adults. Int J Behav Nutr Phys Act Springer Nature; 2011 Jul 28;8(1):79. PMID:21798015

24. Tudor-Locke C, Hart TL, Washington TL. Expected values for pedometer-determined physical activity in older populations. International Journal of Behavioral Nutrition and Physical Activity 2009;6.

25. U.S. Department of Health and Human Services. 2008 Physical Activity Guidelines for Americans. 2008.

26. Tudor-Locke C, Craig CL, Aoyagi Y, Bell RC, Croteau KA, De Bourdeaudhuij I, Ewald B, Gardner AW, Hatano Y, Lutes LD, Matsudo SM, Ramirez-Marrero FA, Rogers LQ, Rowe DA, Schmidt MD, Tully MA, Blair SN. How many steps/day are enough? For older adults and special populations. International Journal of Behavioral Nutrition and Physical Activity 2011;8(1):1–19.

27. Belanger M, Gray-Donald K, O'Loughlin J, Paradis G, Hanley J. Influence of weather conditions and season on physical activity in adolescents. Annals of Epidemiology 2009;19(3):180–186.

28. Brandon CA, Gill DP, Speechley M, Gilliland J, Jones GR. Physical activity levels of older community-dwelling adults are influenced by summer weather variables. Applied Physiology, Nutrition, & Metabolism 2009;34(2):182–190.

29. Centers for Disease, Control, Prevention. Monthly estimates of leisure-time physical inactivity--United States, 1994. MMWR - Morbidity & Mortality Weekly Report 1997;46(18):393–397.

30. Hirvensalo M, Lintunen T. Life-course perspective for physical activity and sports participation. European Review of Aging and Physical Activity 2011;8(1):13–22.

31. Kern ML, Reynolds CA, Friedman HS. Predictors of Physical Activity Patterns Across Adulthood: A Growth Curve Analysis. Personality and Social Psychology Bulletin 2010;36(8):1058–1072.

32. Levin S, Jacobs DR, Ainsworth BE, Richardson MT, Leon AS. Intra-individual variation and estimates of usual physical activity. Annals of Epidemiology 1999;9(8):481–488.

33. Matthews CE, Ainsworth BE, Thompson RW, Bassett DR. Sources of variance in daily physical activity levels as measured by an accelerometer. Medicine and Science in Sports and Exercise 2002;34(8):1376–1381.

34. Pivarnik JM, Reeves MJ, Rafferty AP. Seasonal variation in adult leisure-time physical activity. Medicine and Science in Sports and Exercise 2003;35(6):1004–1008.

35. Rauner A, Jekauc D, Mess F, Schmidt S, Woll A. Tracking physical activity in different settings from late childhood to early adulthood in Germany: the MoMo longitudinal study. Bmc Public Health 2015;15:391.

36. Small M, Bailey-Davis L, Morgan N, Maggs J. Changes in eating and physical activity behaviors across seven semesters of college living on or off campus matters. Health Education & Behavior 2012;1090198112467801.

37. Tucker P, Gilliland J. The effect of season and weather on physical activity: a systematic review. Public Health 2007;121(12):909–922.

38. Ding D, Lawson KD, Kolbe-Alexander TL, Finkelstein EA, Katzmarzyk PT, Van Mechelen W, Pratt M, Committee LPAS 2. E. The economic burden of physical inactivity: a global analysis of major non-communicable diseases. Lancet Elsevier; 2016;388(10051):1311–1324.

39. Glanz K, Rimer BK, Viswanath K. Health Behavior: Theory, Research, and Practice. John Wiley & Sons; 2015. ISBN:9781118628980

40. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice-Hall; 1986. ISBN:9780138156145

41. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, Eccles MP, Cane J, Wood CE. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: Building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81–95. PMID:23512568

42. Lawson PJ, Flocke SA. Teachable moments for health behavior change: a concept analysis. Patient Educ Couns Elsevier; 2009 Jul;76(1):25–30. PMID:19110395

43. Czerwinski M, Gilad-Bachrach R, Iqbal S, Mark G. Challenges for designing notifications for affective computing systems. Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct New York, NY, USA: Association for Computing Machinery; 2016. p. 1554–1559.

44. Cvach M. Monitor alarm fatigue: an integrative review. Biomed Instrum Technol array.aami.org; 2012 Jul-Aug;46(4):268–277. PMID:22839984

45. Bouchard C, Rankinen T. Individual differences in response to regular physical activity. Med Sci Sports Exerc paulogentil.com; 2001 Jun;33(6 Suppl):S446-51; discussion S452-3. PMID:11427769

46. Bryan AD, Nilsson R, Tompkins SA, Magnan RE, Marcus BH, Hutchison KE. The Big Picture of Individual Differences in Physical Activity Behavior Change: A Transdisciplinary Approach. Psychol Sport Exerc Elsevier; 2011 Jan;12(1):20–26. PMID:21278837

47. Hecksteden A, Kraushaar J, Scharhag-Rosenberger F, Theisen D, Senn S, Meyer T. Individual response to exercise training - a statistical perspective. J Appl Physiol journals.physiology.org; 2015 Jun 15;118(12):1450–1459. PMID:25663672

48. Hecksteden A, Pitsch W, Rosenberger F, Meyer T. Repeated testing for the assessment of individual response to exercise training. J Appl Physiol journals.physiology.org; 2018 Jun 1;124(6):1567–1579. PMID:29357481

49. Mann TN, Lamberts RP, Lambert MI. High responders and low responders: factors associated with individual variation in response to standardized training. Sports Med Springer; 2014 Aug;44(8):1113–1124. PMID:24807838

50. Schulhauser KT, Bonafiglia JT, McKie GL, McCarthy SF, Islam H, Townsend LK, Grisebach D, Todd C, Gurd BJ, Hazell T. Individual patterns of response to traditional and modified sprint interval training. J Sports Sci Taylor & Francis; 2021 May;39(10):1077–1087. PMID:33283662

51. Hrubeniuk TJ, Bonafiglia JT, Bouchard DR, Gurd BJ, Sénéchal M. Directions for Exercise Treatment Response Heterogeneity and Individual Response Research. Int J Sports Med thieme-connect.com; 2022 Jan;43(1):11–22. PMID:34399428

52. Box AG, Feito Y, Brown C, Petruzzello SJ. Individual differences influence exercise behavior: how personality, motivation, and behavioral regulation vary among exercise mode preferences. Heliyon cell.com; 2019 Apr;5(4):e01459. PMID:31065599

53. Sudeck G, Jeckel S, Schubert T. Individual Differences in the Competence for Physical-Activity-Related Affect Regulation Moderate the Activity–Affect Association in Real-Life Situations. Journal of Sport and Exercise Psychology Human Kinetics; 2018 Aug 1;40(4):196–205.

54. Bonafiglia JT, Rotundo MP, Whittall JP, Scribbans TD, Graham RB, Gurd BJ. Inter-Individual Variability in the Adaptive Responses to Endurance and Sprint Interval Training: A Randomized Crossover Study. PLoS One journals.plos.org; 2016 Dec 9;11(12):e0167790. PMID:27936084

55. Onnela J-P. Opportunities and challenges in the collection and analysis of digital phenotyping data. Neuropsychopharmacology Springer Science and Business Media LLC; 2021 Jan;46(1):45–54. PMID:32679583

56. Dlima SD, Shevade S, Menezes SR, Ganju A. Digital Phenotyping in Health Using Machine Learning Approaches: Scoping Review. JMIR Bioinformatics and Biotechnology JMIR Bioinformatics and Biotechnology; 2022 Jul 18;3(1):e39618.

57. Robertson MC, Green CE, Liao Y, Durand CP, Basen-Engquist KM. Self-efficacy and Physical Activity in Overweight and Obese Adults Participating in a Worksite Weight Loss Intervention: Multistate Modeling of Wearable Device Data. Cancer Epidemiol Biomarkers Prev 2020 Apr;29(4):769–776. PMID:31871110

58. Li LC, Feehan LM, Xie H, Lu N, Shaw C, Gromala D, Aviña-Zubieta JA, Koehn C, Hoens AM, English K, Tam J, Therrien S, Townsend AF, Noonan G, Backman CL. Efficacy of a Physical Activity Counseling Program With Use of a Wearable Tracker in People With

Inflammatory Arthritis: A Randomized Controlled Trial. Arthritis Care Res 2020 Dec;72(12):1755–1765. PMID:32248626

59. Peacock OJ, Western MJ, Batterham AM, Chowdhury EA, Stathi A, Standage M, Tapp A, Bennett P, Thompson D. Effect of novel technology-enabled multidimensional physical activity feedback in primary care patients at risk of chronic disease - the MIPACT study: a randomised controlled trial. Int J Behav Nutr Phys Act 2020 Aug 8;17(1):99. PMID:32771018

60. Bryan A, Hutchison KE, Seals DR, Allen DL. A transdisciplinary model integrating genetic, physiological, and psychological correlates of voluntary exercise. Health Psychol 2007 Jan;26(1):30–39. PMID:17209695

61. Vellers HL, Kleeberger SR, Lightfoot JT. Inter-individual variation in adaptations to endurance and resistance exercise training: genetic approaches towards understanding a complex phenotype. Mamm Genome Springer; 2018 Feb;29(1–2):48–62. PMID:29356897

62. Gehrman PR, Ghorai A, Goodman M, McCluskey R, Barilla H, Almasy L, Roenneberg T, Bucan M. Twin-based heritability of actimetry traits. Genes Brain Behav 2019 Jun;18(5):e12569. PMID:30916437

63. Wilson KE, Dishman RK. Personality and physical activity: A systematic review and meta-analysis. Pers Individ Dif Elsevier; 2015 Jan 1;72:230–242.

64. Qian T, Walton AE, Collins LM, Klasnja P, Lanza ST, Nahum-Shani I, Rabbi M, Russell MA, Walton MA, Yoo H, Murphy SA. The microrandomized trial for developing digital interventions: Experimental design and data analysis considerations. Psychol Methods 2022 Oct;27(5):874–894. PMID:35025583

65. Rabbi M, Philyaw-Kotov M, Li J, Li K, Rothman B, Giragosian L, Reyes M, Gadway H, Cunningham R, Bonar E, Nahum-Shani I, Walton M, Murphy S, Klasnja P. Translating Behavioral Theory into Technological Interventions: Case Study of an mHealth App to Increase Self-reporting of Substance-Use Related Data. arXiv [csHC]. 2020. doi: 10.1145/nnnnnnn.nnnnnnn

66. Liao P, Greenewald K, Klasnja P, Murphy S. Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020;4(1). PMID:34527853

67. Hekler EB, Michie S, Pavel M, Rivera DE, Collins LM, Jimison HB, Garnett C, Parral S, Spruijt-Metz D. Advancing Models and Theories for Digital Behavior Change Interventions. Am J Prev Med 2016 Nov;51(5):825–832. PMID:27745682

68. Tudor-Locke C, Rowe DA. Using cadence to study free-living ambulatory behaviour. Sports Med 2012 May 1;42(5):381–398. PMID:22462794

69. Fisher AJ, Medaglia JD, Jeronimus BF. Lack of group-to-individual generalizability is a threat to human subjects research. Proc Natl Acad Sci U S A 2018 Jul 3;115(27):E6106–E6115. PMID:29915059

70. Riley WT, Rivera DE. Methodologies for optimizing behavioral interventions: introduction to special section. Transl Behav Med 2014 Sep;4(3):234–237. PMID:25264463

71. Collins LM, Trail JB, Kugler KC, Baker TB, Piper ME, Mermelstein RJ. Evaluating individual intervention components: making decisions based on the results of a factorial screening experiment. Transl Behav Med 2014 Sep;4(3):238–251. PMID:25264464

72. Almirall D, Nahum-Shani I, Sherwood NE, Murphy SA. Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. Transl Behav Med 2014 Sep;4(3):260–274. PMID:25264466

73. Lei H, Nahum-Shani I, Lynch K, Oslin D, Murphy SA. A "SMART" design for building individualized treatment sequences. Annu Rev Clin Psychol 2012;8:21–48. PMID:22224838

74. Riley WT, Martin CA, Rivera DE, Hekler EB, Adams MA, Buman MP, Pavel M, King AC. Development of a dynamic computational model of social cognitive theory. Transl Behav Med 2016 Dec;6(4):483–495. PMID:27848208

75. Hekler EB, Buman MP, Poothakandiyil N, Rivera DE, Dzierzewski JM, Aiken Morgan A, McCrae CS, Roberts BL, Marsiske M, Giacobbi PR. Exploring Behavioral Markers of Long-Term Physical Activity Maintenance: A Case Study of System Identification Modeling Within a Behavioral Intervention. Health Educ Behav 2013 Oct 1;40(1_suppl):51S-62S.

76. Mackie JL. Causes and Conditions. Am Philos Q [North American Philosophical Publications, University of Illinois Press]; 1965;2(4):245–264.

77. Twardy CR, Korb KB. A Criterion of Probabilistic Causation. Philos Sci Cambridge University Press; 2004 Jul;71(3):241–262.

78. Richard N. Burnor. Rethinking Objective Homogeneity: Statistical versus Ontic Approaches. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition Springer; 1993;71(3):307–325.

79. Pearl J. Causality. Cambridge University Press; 2009. ISBN:9780521895606

80. Kay M, Nelson GL, Hekler EB. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems New York, NY, USA: Association for Computing Machinery; 2016. p. 4521–4532.

81. Spruijt-Metz D, Nilsen W. Dynamic Models of Behavior for Just-in-Time Adaptive Interventions. IEEE Pervasive Comput IEEE; July-Sep 2014;13(3):13–17.

82. Wang L, Miller LC. Just-in-the-Moment Adaptive Interventions (JITAI): A Meta-Analytical Review. Health Commun 2020 Nov;35(12):1531–1544. PMID:31488002

83. Hardeman W, Houghton J, Lane K, Jones A, Naughton F. A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. International Journal of Behavioral Nutrition and Physical Activity; 2019;

84. Perski O, Hébert ET, Naughton F, Hekler EB, Brown J, Businelle MS. Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. Addiction 2022 May;117(5):1220–1241. PMID:34514668

85. Teepe GW, Da Fonseca A, Kleim B, Jacobson NC, Salamanca Sanabria A, Tudor Car L, Fleisch E, Kowatsch T. Just-in-Time Adaptive Mechanisms of Popular Mobile Apps for Individuals With Depression: Systematic App Search and Literature Review. J Med Internet Res 2021 Sep 28;23(9):e29412. PMID:34309569

86. Montoye HJ, Washburn R, Servais S, Ertl A, Webster JG, Nagle FJ. Estimation of energy expenditure by a portable accelerometer. Med Sci Sports Exerc 1983;15(5):403–407. PMID:6645869

87. Chen KY, Bassett DR Jr. The technology of accelerometry-based activity monitors: current and future. Med Sci Sports Exerc 2005 Nov;37(11 Suppl):S490-500. PMID:16294112

88. Troiano RP, McClain JJ, Brychta RJ, Chen KY. Evolution of accelerometer methods for physical activity research. Br J Sports Med bjsm.bmj.com; 2014 Jul;48(13):1019–1023. PMID:24782483

89. Chevance G, Golaszewski NM, Tipton E, Hekler EB, Buman M, Welk GJ, Patrick K, Godino JG. Accuracy and Precision of Energy Expenditure, Heart Rate, and Steps Measured by Combined-Sensing Fitbits Against Reference Measures: Systematic Review and Meta-analysis. JMIR Mhealth Uhealth 2022 Apr 13;10(4):e35626. PMID:35416777

90. Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, Buote R, Van Heerden D, Luan H, Cullen K, Slade L, Taylor NGA. Reliability and Validity of Commercially Available Wearable Devices for Measuring Steps, Energy Expenditure, and Heart Rate: Systematic Review. JMIR mHealth and uHealth 2020 Sep;8(9):e18694. PMID:32897239

91. Bryan CJ, Wastler H, Allan N, Khazem LR, Rudd MD. Just-in-Time Adaptive Interventions (JITAIs) for Suicide Prevention: Tempering Expectations. Psychiatry. 2022. p. 341–346. PMID:36344469

92. Saeed SA, Masters RM. Disparities in Health Care and the Digital Divide. Curr Psychiatry Rep 2021 Jul 23;23(9):61. PMID:34297202

93. Müller AM, Blandford A, Yardley L. The conceptualization of a Just-In-Time Adaptive Intervention (JITAI) for the reduction of sedentary behavior in older adults. Mhealth 2017 Sep 12;3:37. PMID:29184889

94. Larrison CR, Xiang X, Gustafson M, Lardiere MR, Jordan N. Implementation of Electronic Health Records Among Community Mental Health Agencies. J Behav Health Serv Res 2018 Jan;45(1):133–142. PMID:28439789

95. Foster KR, Torous J. The Opportunity and Obstacles for Smartwatches and Wearable Sensors. IEEE Pulse 2019;10(1):22–25. PMID:30872210

96. Torous J, Wisniewski H, Liu G, Keshavan M. Mental Health Mobile Phone App Usage, Concerns, and Benefits Among Psychiatric Outpatients: Comparative Survey Study. JMIR Ment Health 2018 Nov 16;5(4):e11715. PMID:30446484

97. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, Pratt M, Ekelund ULF, Yngve A, Sallis JF. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exercise LWW; 2003;35(8):1381–1395.

98. Pucher J, Renne JL. Socioeconomics of Urban Travel: Evidence from the 2001 NHTS. Transp Q unknown; 2001 Jan 1;57(3). Available from: http://dx.doi.org/

99. Bassett DR Jr, Ainsworth BE, Leggett SR, Mathien CA, Main JA, Hunter DC, Duncan GE. Accuracy of five electronic pedometers for measuring distance walked. Med Sci Sports Exerc 1996 Aug;28(8):1071–1077. PMID:8871919

100. Schneider PL, Crouter SE, Lukajic O, Bassett DR Jr. Accuracy and reliability of 10 pedometers for measuring steps over a 400-m walk. Med Sci Sports Exerc 2003 Oct;35(10):1779–1784. PMID:14523320

101. Kadaba MP, Ramakrishnan HK, Wootten ME. Measurement of lower extremity kinematics during level walking. J Orthop Res 1990 May;8(3):383–392. PMID:2324857

102. Chevance G, Perski O, Hekler EB. Innovative methods for observing and changing complex health behaviors: four propositions. Transl Behav Med 2021 Mar 16;11(2):676–685. PMID:32421196

103. Park J, Norman GJ, Klasnja P, Rivera DE, Hekler E. Development and Validation of Multivariable Prediction Algorithms to Estimate Future Walking Behavior in Adults: Retrospective Cohort Study. JMIR Mhealth Uhealth 2023 Jan 27;11:e44296. PMID:36705954

104. Godfrey A, Conway R, Meagher D, OLaighin G. Direct measurement of human movement by accelerometry. Med Eng Phys 2008 Dec;30(10):1364–1386. PMID:18996729

105. Bassett DR Jr, Mahar MT, Rowe DA, Morrow JR Jr. Walking and measurement. Med Sci Sports Exerc 2008 Jul;40(7 Suppl):S529-36. PMID:18562970

106. Kwapisz JR, Weiss GM, Moore SA. Activity recognition using cell phone accelerometers. SIGKDD Explor Newsl New York, NY, USA: Association for Computing Machinery; 2011 Mar 31;12(2):74–82.

107. Karantonis DM, Narayanan MR, Mathie M, Lovell NH, Celler BG. Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. IEEE Trans Inf Technol Biomed ieeexplore.ieee.org; 2006 Jan;10(1):156–167. PMID:16445260

108. Farrahi V, Niemelä M, Kangas M, Korpelainen R, Jämsä T. Calibration and validation of accelerometer-based activity monitors: A systematic review of machine-learning approaches. Gait Posture Elsevier; 2019 Feb;68:285–299. PMID:30579037

109. Migueles JH, Cadenas-Sanchez C, Ekelund U, Delisle Nyström C, Mora-Gonzalez J, Löf M, Labayen I, Ruiz JR, Ortega FB. Accelerometer Data Collection and Processing Criteria to Assess Physical Activity and Other Outcomes: A Systematic Review and Practical Considerations. Sports Med Springer; 2017 Sep;47(9):1821–1845. PMID:28303543

110. de Almeida Mendes M, da Silva ICM, Ramires VV, Reichert FF, Martins RC, Tomasi E. Calibration of raw accelerometer data to measure physical activity: A systematic review. Gait Posture Elsevier; 2018 Mar;61:98–110. PMID:29324298

111. Shannon CE. Communication in the Presence of Noise. Proceedings of the IRE IEEE; 1949 Jan;37(1):10–21.

112. Redenius N, Kim Y, Byun W. Concurrent validity of the Fitbit for assessing sedentary behavior and moderate-to-vigorous physical activity. BMC Med Res Methodol 2019 Feb;19(1):29. PMID:30732582

113. Accelerometer Sensor Guide. Available from: https://dev.fitbit.com/build/guides/sensors/accelerometer/ [accessed Nov 8, 2023]

114. Brønd JC, Arvidsson D. Sampling frequency affects the processing of Actigraph raw acceleration data to activity counts. J Appl Physiol 2016 Feb 1;120(3):362–369. PMID:26635347

115. How do Fitbit devices sync their data? Available from: https://archive.is/1jkt2/image [accessed Apr 4, 2024]

116. Apple Inc. HKUpdateFrequency.immediate. Apple Developer Documentation. Available from: https://archive.is/cnacM/image [accessed Apr 4, 2024]

117. Adams MA, Hurley JC, Todd M, Bhuiyan N, Jarrett CL, Tucker WJ, Hollingshead KE, Angadi SS. Adaptive goal setting and financial incentives: a 2 × 2 factorial randomized controlled trial to increase adults' physical activity. BMC Public Health 2017 Mar;17(1):286. PMID:28356097

118. McCurdy AJ, Normand MP. The effects of a group-deposit prize draw on the step counts of sedentary and low active adults. Behav Interv Wiley; 2022 Jul;37(3):700–712.

119. Bluetooth® Core Specification Version 5.0 Feature Enhancements. Bluetooth® Technology Website. 2018. Available from: https://www.bluetooth.com/bluetooth-resources/bluetooth-5-go-faster-go-further/ [accessed Mar 25, 2024]

120. Brown VA. An Introduction to Linear Mixed-Effects Modeling in R. Advances in Methods and Practices in Psychological Science SAGE Publications Inc; 2021 Jan 1;4(1):2515245920960351.

121. Ekstrom D, Quade D, Golden RN. Statistical analysis of repeated measures in psychiatric research. Arch Gen Psychiatry 1990 Aug;47(8):770–772. PMID:2378548

122. Gueorguieva R, Krystal JH. Move over ANOVA: progress in analyzing repeated-measures data and its reflection in papers published in the Archives of General Psychiatry. Arch Gen Psychiatry 2004 Mar;61(3):310–317. PMID:14993119

123. Harris T, Hilbe JM, Hardin JW. Modeling count data with generalized distributions. Stata J SAGE Publications; 2014 Sep;14(3):562–579.

124. Hilbe JM. Modeling Count Data. Cambridge, England: Cambridge University Press; 2014. ISBN:9781107611252

125. Garay AM, Hashimoto EM, Ortega EMM, Lachos VH. On estimation and influence diagnostics for zero-inflated negative binomial regression models. Comput Stat Data Anal 2011 Mar 1;55(3):1304–1318.

126. Lambert D. Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. Technometrics Taylor & Francis; 1992 Feb 1;34(1):1–14.

127. Brooks S, Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. Statistician JSTOR; 1996;45(2):266.

128. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016. ISBN:9780262337373

129. Cybenko G. Approximation by superpositions of a sigmoidal function. Math Control Signals Systems 1989 Dec 1;2(4):303–314.

130. Katz G, Barrett C, Dill DL, Julian K, Kochenderfer MJ. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. Computer Aided Verification Springer International Publishing; 2017. p. 97–117.

131. Rumelhart DE, Hinton GE, Williams RJ. Learning Internal Representations by Error Propagation, Parallel Distributed Processing, Explorations in the Microstructure of Cognition, ed. DE Rumelhart and J. McClelland. Vol. 1. 1986. Biometrika.

132. Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez JE, Stoica I. Tune: A Research Platform for Distributed Model Selection and Training. arXiv [csLG]. 2018. Available from: http://arxiv.org/abs/1807.05118

133. Ray Tune: Hyperparameter Tuning — Ray 2.10.0. Available from: https://docs.ray.io/en/latest/tune/index.html [accessed Apr 1, 2024]

134. Li L, Jamieson K, Rostamizadeh A, Gonina E, Hardt M, Recht B, Talwalkar A. A System for Massively Parallel Hyperparameter Tuning. arXiv [csLG]. 2018. Available from: http://arxiv.org/abs/1810.05934

135. Wu J, Chen X-Y, Zhang H, Xiong L-D, Lei H, Deng S-H. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb. Dianzi Keji Daxue Xuebao 2019 Mar 1;17(1):26–40.

136. Korinek EV, Phatak SS, Martin CA, Freigoun MT, Rivera DE, Adams MA, Klasnja P, Buman MP, Hekler EB. Adaptive step goals and rewards: a longitudinal growth model of daily steps for a smartphone-based walking intervention. J Behav Med 2018 Feb;41(1):74–86. PMID:28918547

137. Klasnja P, Smith S, Seewald NJ, Lee A, Hall K, Luers B, Hekler EB, Murphy SA. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. Ann Behav Med Oxford University Press US; 2019;53(6):573–582.

138. Rozet A, Kronish IM, Schwartz JE, Davidson KW. Using machine learning to derive just-in-time and personalized predictors of stress: Observational study bridging the gap between nomothetic and ideographic approaches. J Med Internet Res JMIR Publications Inc.; 2019 Apr 26;21(4):e12910.

139. Clague J, Bernstein L. Physical activity and cancer. Curr Oncol Rep 2012;14(6):550–558. PMID:22945451

140. Bianchini F, Kaaks R, Vainio H. Weight control and physical activity in cancer prevention. Obes Rev 2002;3(1):5–8. PMID:12119660

141. U.S. Department of Health and Human Services, Promotion, National Center for Chronic Disease Prevention & Health. Behavioral risk factor surveillance system physical activity trends data nationwide. 2011. Available from: http://apps.nccd.cdc.gov/PASurveillance/DemoCompareResultV.asp?State=0&Cat=1&Year=2007&CI=on&Go=GO#result

142. Bauman A, Bull F, Chey T, Craig CL, Ainsworth BE, Sallis JF, Bowles HR, Hagstromer M, Sjostrom M, Pratt M, Grp IPS. The International Prevalence Study on Physical Activity: results from 20 countries. International Journal of Behavioral Nutrition and Physical Activity 2009;6:11.

143. Matthews CE. Physical activity in the United States measured by accelerometer: Comment. Medicine and Science in Sports and Exercise 2008;40(6):1188–1188.

144. Sugiyama T, Healy GN, Dunstan DW, Salmon J, Owen N. Joint associations of multiple leisure-time sedentary behaviours and physical activity with obesity in Australian adults. International Journal of Behavioral Nutrition and Physical Activity 2008;5:6.

145. Troiano RP, Berrigan D, Dodd KW, Masse LC, Tilert T, McDowell M. Physical Activity in the United States measured by accelerometer. Medicine & Science in Sports & Exercise 2008;40(1):181–188.

146. Warburton DE, Nicol CW, Bredin SS. Health benefits of physical activity: the evidence. CMAJ: Canadian Medical Association Journal 2006;174(6):801–809.

147. Woolf K, Reese CE, Mason MP, Beaird LC, Tudor-Locke C, Vaughan LA. Physical activity is associated with risk factors for chronic disease across adult women's life cycle. Journal of the American Dietetic Association 2008;108(6):948–959.

148. Eakin E, Reeves M, Winkler E, Lawler S, Owen N. Maintenance of Physical Activity and Dietary Change Following a Telephone-Delivered Intervention. Health Psychology 2010;29(6):566–573.

149. Fjeldsoe B, Neuhaus M, Winkler E, Eakin E. Systematic Review of Maintenance of Behavior Change Following Physical Activity and Dietary Interventions. Health Psychology 2011;30(1):99–109.

150. King AC, Hekler EB, Castro CM, Buman MP, Marcus BH, Friedman RH, Napolitano MA. Exercise advice by humans versus computers: Maintenance effects at 18 months. Health Psychology 2014;33(2):192.

151. Lü L, Medo M, Yeung CH, Zhang Y-C, Zhang Z-K, Zhou T. Recommender systems. Phys Rep Elsevier; 2012 Oct 1;519(1):1–49.

152. Hekler EB, Klasnja P, Chevance G, Golaszewski NM, Lewis D, Sim I. Why we need a small data paradigm. BMC Med 2019 Jul 17;17(1):133. PMID:31311528

153. Hekler E, Tiro JA, Hunter CM, Nebeker C. Precision Health: The Role of the Social and Behavioral Sciences in Advancing the Vision. Ann Behav Med academic.oup.com; 2020 Nov 1;54(11):805–826. PMID:32338719

154. Kreuter MW, Wray RJ. Tailored and targeted health communication: strategies for enhancing information relevance. Am J Health Behav 2003;27 Suppl 3:S227-32. PMID:14672383

155. Prochaska JO, Butterworth S, Redding CA, Burden V, Perrin N, Leo M, Flaherty-Robb M, Prochaska JM. Initial efficacy of MI, TTM tailoring and HRI's with multiple behaviors for employee health promotion. Prev Med 2008;46(3):226–231. PMID:18155287

156. Martín CA, Rivera DE, Hekler EB, Riley WT, Buman MP, Adams MA, Magann AB. Development of a Control-Oriented Model of Social Cognitive Theory for Optimized

mHealth Behavioral Interventions. IEEE Trans Control Syst Technol 2020 Mar;28(2):331–346. PMID:33746479

157. Spruijt-Metz D, Marlin BM, Pavel M, Rivera DE, Hekler E, De La Torre S, El Mistiri M, Golaszweski NM, Li C, Braga De Braganca R, Tung K, Kha R, Klasnja P. Advancing Behavioral Intervention and Theory Development for Mobile Health: The HeartSteps II Protocol. Int J Environ Res Public Health 2022;19(4). PMID:35206455

158. Abraham C, Michie S. A taxonomy of behavior change techniques used in interventions. Health Psychology 2008;27(3):379.

159. Michie S, Ashford S, Sniehotta FF, Dombrowski SU, Bishop A, French DP. A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: the CALO-RE taxonomy. Psychology & health 2011;26(11):1479–1498.

160. Achtziger A, Gollwitzer PM, Sheeran P. Implementation Intentions and Shielding Goal Striving From Unwanted Thoughts and Feelings. Personality and Social Psychology Bulletin 2008;34(3):381–393.

161. Hagger MS, Luszczynska A. Implementation Intention and Action Planning Interventions in Health Contexts: State of the Research and Proposals for the Way Forward. Applied Psychology: Health and Well-Being 2014;6(1):1–47.

162. Armstrong T, Bull F. Development of the World Health Organization Global Physical Activity Questionnaire (GPAQ). J Public Health Springer; 2006 Apr 1;14(2):66–70.

163. Adams R. Revised Physical Activity Readiness Questionnaire. Can Fam Physician 1999 Apr;45:992, 995, 1004–5. PMID:10216799

164. Rivera DE, Braun MW, Mittelmann HD. Constrained Multisine Inputs for Plant-friendly Identification of Chemical Processes. IFAC Proceedings Volumes Elsevier; 2002 Jan 1;35(1):425–430.

165. Martin CA, Deshpande S, Hekler EB, Rivera DE. A system identification approach for improving behavioral interventions based on Social Cognitive Theory. 2015 American Control Conference (ACC) IEEE; 2015. p. 5878–5883.

166. Martín CA, Rivera DE, Hekler EB. An identification test monitoring procedure for MIMO systems based on statistical uncertainty estimation. 2015 54th IEEE Conference on Decision and Control (CDC) 2015. p. 2719–2724.

167. El Mistiri M, Rivera DE, Klasnja P, Park J, Hekler E. Enhanced Social Cognitive Theory Dynamic Modeling and Simulation Towards Improving the Estimation of " Just-In-Time " States. 2022 American Control Conference (ACC) 2022. p. 468–473.

168. Martin CA, Rivera DE, Riley WT, Hekler EB, Buman MP, Adams MA, King AC. A dynamical systems model of Social Cognitive Theory. 2014 American Control Conference IEEE; 2014. doi: 10.1109/acc.2014.6859463

169. Gotzian L. Modeling the decreasing intervention effect in digital health: a computational model to predict the response for a walking intervention. 2023. doi: 10.31219/osf.io/6v7d5

170. Gosling SD, Rentfrow PJ, Swann WB Jr. A very brief measure of the Big-Five personality domains. J Res Pers Elsevier; 2003;37(6):504–528.

171. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav JSTOR; 1983 Dec;24(4):385–396. PMID:6668417

172. Horne JA, Ostberg O. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. Int J Chronobiol 1976;4(2):97–110. PMID:1027738

173. Marcus BH, Eaton CA, Rossi JS, Harlow LL. Self-efficacy, decision-making, and stages of change: An integrative model of physical Exercise1. J Appl Soc Psychol Wiley; 1994 Mar;24(6):489–508.

174. Balbim GM, Marques IG, Marquez DX, Patel D, Sharp LK, Kitsiou S, Nyenhuis SM. Using Fitbit as an mHealth Intervention Tool to Promote Physical Activity: Potential Challenges and Solutions. JMIR Mhealth Uhealth 2021 Mar 1;9(3):e25289. PMID:33646135

175. Climate Data Online. Climate Data Online: Web Services Documentation. Available from: https://www.ncdc.noaa.gov/cdo-web/webservices/v2 [accessed May 18, 2023]

176. Ljung L. System Identification: Theory for the User. Prentice-Hall; 1987. ISBN:9780138816407

177. Stenman A. Model on demand: Algorithms, analysis and applications. Department of Electrical Engineering, Linköping University; 1999.

178. Kha RT, Rivera DE, Klasnja P, Hekler E. Model Personalization in Behavioral Interventions using Model-on-Demand Estimation and Discrete Simultaneous Perturbation Stochastic Approximation. Proc Am Control Conf 2022 Jun;2022:671–676. PMID:36340266

179. Moher D. The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomized Trials. JAMA 2001 Apr 18;285(15):1987.

180. HeartSteps Github Repository. Available from: https://github.com/kpwhri/heartsteps [accessed May 20, 2023]

181. Abernethy A, Adams L, Barrett M, Bechtel C, Brennan P, Butte A, Faulkner J, Fontaine E, Friedhoff S, Halamka J, Howell M, Johnson K, Long P, McGraw D, Miller R, Lee P, Perlin J, Rucker D, Sandy L, Savage L, Stump L, Tang P, Topol E, Tuckson R, Valdes K. The Promise of Digital Health: Then, Now, and the Future. NAM Perspect 2022 Jun 27;2022. PMID:36177208

182. Duffy A, Christie GJ, Moreno S. The Challenges Toward Real-world Implementation of Digital Health Design Approaches: Narrative Review. JMIR Hum Factors 2022 Sep 9;9(3):e35693. PMID:36083628

183. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-A call for innovative evidence generation approaches. NPJ Digit Med 2020 Aug 27;3:110. PMID:32904379

184. Xiong S, Lu H, Peoples N, Duman EK, Najarro A, Ni Z, Gong E, Yin R, Ostbye T, Palileo-Villanueva LM, Doma R, Kafle S, Tian M, Yan LL. Digital health interventions for non-communicable disease management in primary health care in low-and middle-income countries. NPJ Digit Med 2023 Feb 1;6(1):12. PMID:36725977

185. El Mistiri M, Rivera DE, Klasnja P, Park J, Hekler E. Model Predictive Control Strategies for Optimized mHealth Interventions for Physical Activity. 2022 American Control Conference (ACC) 2022. p. 1392–1397.

186. Django Web Framework. Available from: https://www.djangoproject.com/ [accessed Mar 25, 2024]

187. Burnham KP, Anderson DR, Huyvaert KP. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav Ecol Sociobiol 2011 Jan 1;65(1):23–35.

188. Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP. Why do we still use stepwise modelling in ecology and behaviour? J Anim Ecol 2006 Sep;75(5):1182–1189. PMID:16922854

189. Steidl RJ. Model Selection, Hypothesis Testing, and Risks of Condemning Analytical Tools. Wildfire The Wildlife Society; 2006 Dec;70(6):1497–1498.

190. Johnson JB, Omland KS. Model selection in ecology and evolution. Trends Ecol Evol 2004 Feb;19(2):101–108. PMID:16701236

191. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner P-C. Rank-Normalization, Folding, and Localization: An Improved Rˆ for Assessing Convergence of MCMC (with Discussion). ba International Society for Bayesian Analysis; 2021 Jun;16(2):667–718.

192. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv [csLG]. 2014. Available from: http://arxiv.org/abs/1412.6980

193. Nielsen A. Practical Time Series Analysis: Prediction with Statistics and Machine Learning. 1st ed. O'Reilly; 2019. ISBN:9781492041658

194. Fushiki T. Estimation of prediction error by using K-fold cross-validation. Stat Comput 2011 Apr 10;21(2):137–146.

195. Diebold FX, Mariano RS. Comparing Predictive Accuracy. J Bus Econ Stat Taylor & Francis; 2002 Jan 1;20(1):134–144.

196. Diebold FX. Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. J Bus Econ Stat Taylor & Francis; 2015 Jan 2;33(1):1–1.

197. Chevance G, Baretta D, Golaszewski N, Takemoto M, Shrestha S, Jain S, Rivera DE, Klasnja P, Hekler E. Goal setting and achievement for walking: A series of N-of-1 digital interventions. Health Psychol 2021 Jan;40(1):30–39. PMID:33252961

198. Figueroa CA, Deliu N, Chakraborty B, Modiri A, Xu J, Aggarwal J, Jay Williams J, Lyles C, Aguilera A. Daily Motivational Text Messages to Promote Physical Activity in University Students: Results From a Microrandomized Trial. Ann Behav Med 2022 Feb 11;56(2):212–218. PMID:33871015

199. Klasnja P, Rosenberg DE, Zhou J, Anau J, Gupta A, Arterburn DE. A quality-improvement optimization pilot of BariFit, a mobile health intervention to promote physical activity after bariatric surgery. Transl Behav Med 2021 Mar 16;11(2):530–539. PMID:32421187

200. Saponaro M, Vemuri A, Dominick G, Decker K. Contextualization and individualization for just-in-time adaptive interventions to reduce sedentary behavior. Proceedings of the Conference on Health, Inference, and Learning New York, NY, USA: Association for Computing Machinery; 2021. p. 246–256.

201. Vandelanotte C, Trost S, Hodgetts D, Imam T, Rashid M, To QG, Maher C. Increasing physical activity using an just-in-time adaptive digital assistant supported by machine learning: A novel approach for hyper-personalised mHealth interventions. J Biomed Inform 2023 Aug;144:104435. PMID:37394024

202. Künzler F. Assessing and Predicting States of Receptivity for Physical Activity Interventions. research-collection.ethz.ch; 2020. Available from: https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/455509/1/PhD_Thesis_A5_upload_v2.pdf [accessed May 2, 2024]

203. De Bra P. Challenges in User Modeling and Personalization. IEEE Intell Syst IEEE; September/October 2017;32(5):76–80.

204. Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial Intelligence and the Implementation Challenge. J Med Internet Res 2019 Jul 10;21(7):e13659. PMID:31293245

205. Colin Cameron A, Trivedi PK. Regression Analysis of Count Data. Cambridge University Press; 2013. ISBN:9781107717794

206. Hernandez-Lobato JM, Adams R. Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks. Bach F, Blei D, editors. Lille, France: PMLR; 07--09 Jul 2015;37:1861–1869.

207. Kononenko I. Bayesian neural networks. Biol Cybern 1989 Sep 1;61(5):361–370.

208. Marcot BG, Penman TD. Advances in Bayesian network modelling: Integration of modelling technologies. Environmental Modelling & Software 2019 Jan 1;111:386–393.

209. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. arXiv [statML]. 2012. Available from: https://proceedings.neurips.cc/paper_files/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf [accessed Apr 11, 2023]

210. Kim J-M, Ha ID. Deep learning-based residual control chart for count data. Qual Eng Taylor & Francis; 2022 Jul 3;34(3):370–381.

211. Granat MH. Event-based analysis of free-living behaviour. Physiol Meas 2012 Nov;33(11):1785–1800. PMID:23110873

212. Sokas D, Paliakaitė B, Rapalis A, Marozas V, Bailón R, Petrėnas A. Detection of Walk Tests in Free-Living Activities Using a Wrist-Worn Device. Front Physiol 2021 Aug 12;12:706545. PMID:34456748

213. Granat M, Clarke C, Holdsworth R, Stansfield B, Dall P. Quantifying the cadence of free-living walking using event-based analysis. Gait Posture 2015 Jun;42(1):85–90. PMID:25953505

214. Grant PM, Granat MH, Thow MK, Maclaren WM. Analyzing free-living physical activity of older adults in different environments using body-worn activity monitors. J Aging Phys Act 2010 Apr;18(2):171–184. PMID:20440029

215. Toth LP, Park S, Springer CM, Feyerabend MD, Steeves JA, Bassett DR. Video-Recorded Validation of Wearable Step Counters under Free-living Conditions. Med Sci Sports Exerc 2018 Jun;50(6):1315–1322. PMID:29381649

216. Thune I, Furberg AS. Physical activity and cancer risk: Dose-response and cancer, all sites and site-specific. Med Sci Sports Exerc 2001;33(6 SUPPL.). PMID:11427781

217. Blackwell DL, Clarke TC. State variation in meeting the 2008 federal guidelines for both aerobic and muscle-strengthening activities through leisure-time physical activity among adults aged 18-64: United States, 2010-2015. Natl Health Stat Report 2018;2018(112). PMID:30248007

218. Paganini S, Terhorst Y, Sander LB, Catic S, Balci S, Küchler A-M, Schultchen D, Plaumann K, Sturmbauer S, Krämer LV, Lin J, Wurst R, Pryss R, Baumeister H, Messner E-M. Quality of Physical Activity Apps: Systematic Search in App Stores and Content Analysis. JMIR mHealth and uHealth 2021 Jun 9;9(6):e22587.

219. Riley WT, Rivera DE, Atienza AA, Nilsen W, Allison SM, Mermelstein R. Health behavior models in the age of mobile interventions: are our theories up to the task? Transl Behav Med Oxford Academic; 2011 Mar 1;1(1):53–71. PMID:21796270

220. Norman GJ, Zabinski MF, Adams MA, Rosenberg DE, Yaroch AL, Atienza AA. A Review of eHealth Interventions for Physical Activity and Dietary Behavior Change. Am J Prev Med Elsevier; 2007 Oct 1;33(4):336-345.e16. PMID:17888860

221. Bandura A. Health promotion by social cognitive means. Health Educ Behav SAGE Publications; 2004 Apr 30;31(2):143–164. PMID:15090118

222. Locke EA, Latham GP. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. Am Psychol American Psychological Association Inc.; 2002;57(9):705–717. PMID:12237980

223. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, Murphy SA. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. Ann Behav Med 2018 May 18;52(6):446–462.

224. Strecher VJ, Seijts GH, Kok GJ, Latham GP, Glasgow R, Devellis B, Meertens RM, Bulger DW. Goal Setting as a Strategy for Health Behavior Change. Health Educ Behav Sage PublicationsSage CA: Thousand Oaks, CA; 1995 Sep 4;22(2):190–200. PMID:7622387

225. Folkman S, Moskowitz JT. Coping: Pitfalls and Promise. Annu Rev Psychol 2004;55(1):745–774.

226. Ben-Zeev D, Brenner CJ, Begale M, Duffecy J, Mohr DC, Mueser KT. Feasibility, Acceptability, and Preliminary Efficacy of a Smartphone Intervention for Schizophrenia. Schizophr Bull 2014 Nov 1;40(6):1244–1253.

227. Gustafson DH, McTavish FM, Chih M-Y, Atwood AK, Johnson RA, Boyle MG, Levy MS, Driscoll H, Chisholm SM, Dillenburg L, Isham A, Shah D. A Smartphone Application to Support Recovery From Alcoholism. JAMA Psychiatry 2014 May 1;71(5):566.

228. van Dantzig S, Geleijnse G, van Halteren AT. Toward a persuasive mobile application to reduce sedentary behavior. Pers Ubiquit Comput 2013 Aug 12;17(6):1237–1246.

229. Mair JL, Hayes LD, Campbell AK, Buchan DS, Easton C, Sculthorpe N. A Personalized Smartphone-Delivered Just-in-time Adaptive Intervention (JitaBug) to Increase Physical Activity in Older Adults: Mixed Methods Feasibility Study. JMIR Formative Research 2022 Apr 7;6(4):e34662.

230. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Cent Rep 2019 Jan 21;49(1):15–21.

231. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med 2015 Jan 6;162(1):55–63.

232. Zongben Xu, Mingwei Dai, Deyu Meng. Fast and Efficient Strategies for Model Selection of Gaussian Support Vector Machine. IEEE Trans Syst Man Cybern B Cybern 2009 Oct;39(5):1292–1307.

233. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016. p. 785–794.

234. Hinton G. Deep learning—a technology with the potential to transform health care. JAMA American Medical Association; 2018;320(11):1101–1102.

235. Ho TK. Random decision forests. Proceedings of 3rd international conference on document analysis and recognition IEEE; 1995. p. 278–282.

236. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: Machine learning in Python. the Journal of machine Learning research JMLR. org; 2011;12:2825–2830.

237. Keras. Available from: https://keras.io/ [accessed Mar 24, 2024]

238. XGBoost. Available from: https://github.com/dmlc/xgboost

239. Sci-Keras. Available from: https://github.com/adriangb/scikeras

240. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: Review of methods and applications. Expert Syst Appl 2017 May;73:220–239.

241. Yap BW, Rani KA, Rahman HAA, Fong S, Khairudin Z, Abdullah NN. An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. 2014. p. 13–22.

242. Rodríguez P, Bautista MA, Gonzàlez J, Escalera S. Beyond one-hot encoding: Lower dimensional target embedding. Image Vis Comput 2018 Jul;75:21–31.

243. Gaier A, Ha D. Weight agnostic neural networks. Adv Neural Inf Process Syst 2019;32.

244. Liashchynskyi P, Liashchynskyi P. Grid search, random search, genetic algorithm: A big comparison for NAS. arXiv preprint arXiv:191206059 2019;

245. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS One Public Library of Science San Francisco, CA USA; 2017;12(6):e0177678.

246. All of Us Research Program Investigators, Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, Jenkins G, Dishman E. The "All of Us" Research Program. N Engl J Med 2019 Aug 15;381(7):668–676. PMID:31412182

247. Derry JMJ, Mangravite LM, Suver C, Furia MD, Henderson D, Schildwachter X, Bot B, Izant J, Sieberts SK, Kellen MR, Friend SH. Developing predictive molecular maps of human disease through community-based modeling. Nat Genet 2012 Feb 27;44(2):127–130.

248. Krawczyk B, Schaefer G, Woźniak M. A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification. Artif Intell Med 2015 Nov;65(3):219–227.

249. Jiang J, Liu X, Zhang K, Long E, Wang L, Li W, Liu L, Wang S, Zhu M, Cui J, Liu Z, Lin Z, Li X, Chen J, Cao Q, Li J, Wu X, Wang D, Wang J, Lin H. Automatic diagnosis of imbalanced ophthalmic images using a cost-sensitive deep convolutional neural network. Biomed Eng Online 2017 Dec 21;16(1):132.

250. Quiroz JC, Feng Y-Z, Cheng Z-Y, Rezazadegan D, Chen P-K, Lin Q-T, Qian L, Liu X-F, Berkovsky S, Coiera E, Song L, Qiu X, Liu S, Cai X-R. Development and Validation of a Machine Learning Approach for Automated Severity Assessment of COVID-19 Based on Clinical and Imaging Data: Retrospective Study. JMIR Medical Informatics 2021 Feb 11;9(2):e24572.

251. Novaković JD, Veljović A, Ilić SS, Papić Ž, Milica T. Evaluation of classification models in machine learning. Theory and Applications of Mathematics & Computer Science 2017;7(1):39–46.

252. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020 Dec 2;21(1):6.

253. Lundberg I, Johnson R, Stewart BM. What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. Am Sociol Rev 2021 Jun 1;86(3):532–565.

254. Stanley KO, Miikkulainen R. Evolving neural networks through augmenting topologies. Evol Comput MIT Press; 2002;10(2):99–127.