A Tutorial on Matching-based Causal Analysis of Human Behaviors Using Smartphone Sensor Data

GYUWON JUNG, School of Computing, KAIST, Daejeon, South Korea
SANGJUN PARK, School of Computing, KAIST, Daejeon, South Korea
EUN-YEOL MA, Department of Industrial and Systems Engineering, KAIST, Daejeon, South Korea
HEEYOUNG KIM*, Department of Industrial and Systems Engineering, KAIST, Daejeon, South Korea
UICHIN LEE*, School of Computing, KAIST, Daejeon, South Korea

Smartphones can unobtrusively capture human behavior and contextual data such as user interaction and mobility. Thus far, smartphone sensor data have primarily been used to gain behavioral insights through correlation analysis. This paper provides a tutorial on the causal analysis of human behavior using smartphone sensor data by reviewing well-known matching methods. The key steps of the causal inference pipeline employing matching methods are illustrated using a concrete scenario involving the identification of a causal relationship between phone usage and physical activity. Several practical considerations for conducting causal inferences about human behaviors using smartphone sensor data are also discussed.

CCS Concepts: • **Human-centered computing → Empirical studies in ubiquitous and mobile computing**; **Mobile devices**; • **Computing methodologies → Causal reasoning and diagnostics**.

Additional Key Words and Phrases: Smartphone Sensor Data, Causal Inference, Human Behavior, Observational Study

1 INTRODUCTION

We are living in an era of ubiquitous computing, where mobile devices can collect everyday life "human data" anywhere and anytime (e.g., physical activities and phone usage). People interact with their smartphones, which, in turn, enables their seamless integration into smart spaces connected via the Internet of Things (IoT) technologies. Such smart devices can capture users' behaviors and their contexts using multiple types of sensors and data logging features, which can be used to track everyday life activities (e.g., physical movement and interaction with a smartphone) and contexts (e.g., health states, visited places, and ambient light/noise) [19, 48, 74].

In general, human data from smartphones (hereafter, denoted as "smartphone sensor data") can subsequently be converted into features that depict user activities, emotions, and contexts [41, 54]. Such features are also used to perform "digital phenotyping," that is, to uncover behavioral correlates of specific diseases or symptoms (e.g., reduced mobility patterns observed in depression patients) [32, 73]. The common objective of these approaches is to acquire insights into the users' daily lives by learning correlations from data [9] to provide them with

[*] Corresponding author

Authors' addresses: Gyuwon Jung, gwjung@kaist.ac.kr, School of Computing, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, South Korea, 34141; Sangjun Park, sangjun@kaist.ac.kr, School of Computing, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, South Korea, 34141; Eun-Yeol Ma, eyma1127@kaist.ac.kr, Department of Industrial and Systems Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, South Korea, 34141; Heeyoung Kim, heeyoungkim@kaist.ac.kr, Department of Industrial and Systems Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, South Korea, 34141; Uichin Lee, uclee@kaist.ac.kr, School of Computing, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, South Korea, 34141.

ACM 0360-0300/2024/2-ART
https://doi.org/10.1145/3648356

intelligent support, such as monitoring and diagnosing health conditions, predicting symptoms, or providing personalized interventions [15, 65].

Recently, smartphone sensor data have been used to identify the causal relationships among various human behaviors and contexts. For instance, Tsapeli and Musolesi [75] examined the causal effect of place visits on stress levels and found that spending time in places other than home and university had a positive effect on reducing stress. Other studies have investigated the causal relationship between user emotions (e.g., activeness, happiness, and stress) and mobile phone interactions (e.g., notification, mobile app usage, and communication) [53, 62], or the smartphone usage and the incoming notifications [77]. Causal inference using smartphone sensor data is essential in the mobile health field, particularly when designing and evaluating behavior intervention technologies [40]. One way of estimating the effect of behavior intervention technologies is to measure how frequently they are used (i.e., engagement), determine whether the user follows the prescribed intervention (i.e., adherence), and examine the causality with the target symptom [51].

The primary challenge in causal analysis with smartphone sensor data is that we must rely on an "observational study," where data is collected without controlling for any factors [56]. Covariates, including confounding variables (those causally related to both treatment and outcome), can be handled through either an experimental design, such as a random assignment, or a pseudo-experimental design with post-hoc covariate balancing with matching methods, which is the topic illustrated in this tutorial. Here, covariate balancing means that samples for both treated and control groups (e.g., more versus less phone usage) are selected or matched to ensure similar characteristics or covariate distributions (e.g., gender and age composition), preventing potential biases that may arise from non-randomized data.

However, previous studies in the field of mobile computing [53, 75] provide little explanation of how to preprocess the raw smartphone sensor data and balance the covariates within them, which is a prerequisite for estimating causal impact in an observational study. Despite the importance of covariate balancing, researchers usually apply one method, instead of examining reliability using multiple methods, and rarely mention balancing failures, which might sometimes occur in practice. Moreover, factors that should be more carefully considered when inferring causality, particularly due to the unique properties of smartphone sensor data (e.g., very short but frequent interactions), are seldom discussed.

This study provides a comprehensive tutorial on inferring causality based on real-world human data collected by mobile phones. For the sake of illustration, we investigate the representative causality scenario that human behaviors can be explained with smartphone sensor data, specifically, whether mobile app usage causes changes in physical activity. Although previous studies have reported that phone usage is negatively associated with physical activity [3, 28, 45, 46], they relied on self-report based subjective phone usage data, which may fail to capture moment-by-moment interactions with devices or habitual use. In contrast, this work leverages objective human data to explore what everyday data tells about the causal relationship between the two behaviors.

The structure of this paper comprises five sections. First, we introduce how causal inferences can be conducted in observational studies using smartphone sensor data. In particular, this work focuses on "matching," one of the most widely used methods for balancing covariates, in which samples from treated and control groups are matched based on the similarity of their covariates. There are various ways of matching depending on how the similarity is defined (e.g., propensity score matching uses the probability of being treated for given covariates when measuring the similarity). Second, we propose a pipeline for conducting causal analyses based on smartphone sensor data. The process involves four steps: (1) scenario setting, (2) data preprocessing, (3) covariate balancing, and (4) treatment effect estimation. A detailed explanation is given on how to set the scenario and variables of interest, how to process the raw data to extract features about behaviors and contexts, how to balance the covariates to create a set of samples similar to a randomized trial, and how to estimate the effect and evaluate the causality. Third, we offer a case study illustrating the practical application of the proposed analysis pipeline to a real-world dataset obtained from 49 participants over 7 days. We present the inferred causal relationships across various scenarios for each participant, estimated using four different matching methods aimed at balancing covariates. To validate our findings, we applied the same analysis pipeline to an additional, larger dataset gathered from 23 participants over 6 weeks, demonstrating consistent trends in causal relationships across the scenarios. Fourth, we employed causal forest, a machine learning technique for balancing covariates and estimating treatment effect, on the same dataset and conducted cross-validation of our findings obtained through traditional matching methods. Finally, we discuss several practical considerations that researchers might face when utilizing the proposed analysis pipeline on their smartphone sensor data.

The main contributions of this study are as follows:

- We provide a step-by-step tutorial on causal inference based on smartphone sensor data, proposing a causal analysis pipeline and applying it to real-world datasets.
- We make the code and example dataset used in our tutorial available for further studies. The dataset and source code of this study are available at https://bit.ly/3ysuxNV.
- We demonstrate the application of four well-known matching methods to explore the reliability of causal inference in terms of covariate balance and estimated treatment effect.
- We discuss practical considerations that researchers might face when making causal inferences based on smartphone sensor data and propose several suggestions.

2 AN OVERVIEW OF CAUSAL INFERENCE BASED ON OBSERVATIONAL DATA

This section introduces background information on making causal inferences using smartphone sensor data, including (1) an overview of the concept of causal inference in terms of the potential outcomes framework; (2) a description of how covariates can be balanced in observational studies; and (3) a discussion on the “matching” technique, which is the main covariate balancing method utilized throughout this tutorial.

2.1 Causal Inference and Potential Outcomes Framework

One of the key goals of statistical inference is to deduce the causal relationship between variables. For instance, when validating the effect of a new medicine, the researcher’s primary question would be whether taking this medicine causes an improvement in the target symptom, i.e., “*What would be the sole and direct effect of the medicine on the target symptom, compared to when the medicine is not taken?*”

The potential outcomes framework [59], also known as the Rubin causal model, can be used to answer such causal questions by defining an appropriate counterfactual for a given treatment. A potential outcome is the outcome that a subject would have shown if they were to receive the given treatment value. In practice, this approach faces a fundamental limitation in that we can observe only one potential outcome from one subject at a time [31]. In other words, because we cannot create a copy of a subject that is exactly the same as the original in everything except for treatment assignment, it is impossible to measure the treatment effect (i.e., the difference in potential outcomes) at the individual level. Instead, the evaluation needs to be done at the population level by comparing the outcomes between groups of treated and control subjects in which everything but the treatment is identical.

Randomized controlled trials (RCTs) are rigorous experimental studies regarded as the gold standard for estimating causal effects. RCTs generate treatment groups that enable unbiased non-parametric estimation of the direct treatment effect. In typical RCTs with binary treatment, subjects are randomly assigned to either the treated or control group (treatment assigned or not, respectively) to statistically control the effect of the covariates (or confounding variables), which might otherwise alter the treatment assignment, outcome, or both. Because the treatment assignment is random, the distribution of the covariates can be considered the same for the two groups, which then allows the outcomes from each group to be regarded solely due to the treatment given. If the mean difference between the outcomes from the two groups is statistically significant, a causal relationship between the treatment and the outcome can be inferred.

2.2 Covariates in Observational Studies

In many studies, particularly those involving humans, random allocation of subjects is not possible [56]. For example, it can be difficult when an RCT becomes very expensive due to the complexity and multiplicity of test conditions that need to be considered, or because of the long time that it takes for the results of a given treatment to become visible [58]. Previous studies have noted that RCTs may not be appropriate in practice, particularly in the field of health interventions and services due to reasons such as insufficient subjects, lack of generality, and contamination of control groups (e.g., when subjects are unexpectedly exposed to the intervention) [7, 61]. In addition, especially in health-related studies, the random allocation of subjects may pose ethical concerns as noted in [16, 18, 20].

In these circumstances, the data can only be collected through observational studies in which the allocation of the subjects to either the treated or control group cannot be controlled by the researchers. This leads to potential biases (e.g. confounding bias) that are not controlled for and, hence, the outcomes from both groups cannot be directly compared to estimate the treatment effect. Alternatively, we must adjust for covariates that could affect the treatment or outcomes and create biased estimations of the causal effect. There are different strategies for adjusting for such covariates [79, 83], but in this tutorial, we focus on "matching," which is readily applicable and easy to understand.

2.3 Matching for Covariate Balancing

Matching is a straightforward and widely used method for balancing the distribution of covariates between treated and control groups. Rooted in the potential outcomes framework, the primary objective of this method is to create matching pairs of samples that differ in whether a given treatment is administered, but have similar covariate values so that they can be considered counterfactual to each other.

As matching artificially creates a dataset with balanced covariates that mimic randomized trials, to identify the treatment effect, the data used in the matching process should first comply with the following assumptions [39, 57].

- **Stable unit treatment value**: Assigning the treatment to one sample does not affect the potential outcome of another (i.e., no interference between samples), and only one version of the treatment is available.
- **Strong ignorable treatment assignment**: Given a set of pre-treatment covariates, the assignment of the treatment is independent of the potential outcomes, and there is always a positive probability of being treated for every set of covariate values (i.e., the treatment assignment is not deterministic).
- **Consistency**: The potential and observed outcomes under a particular treatment are equal if the same treatment is administered.

As suggested by Stuart [67], different types of matching can be performed depending on detailed criteria such as the matching ratio, matching algorithm, the replacement in matching, setting a caliper, and the way of defining the distance between samples. The matching ratio determines how many samples are used from each group to create the matched pairs. Since there are usually more control samples than treated in observational studies, when matching at a 1:1 ratio, the remaining unmatched control samples are generally discarded. However, additional control samples can be utilized when there are enough of them to increase the matched sample size, although the increased imbalance from the larger distance should be simultaneously considered, or 1:k matching be conducted instead of 1:1 matching.

Matching can be either greedy (also known as "nearest neighbor") or optimal, depending on whether the distance between the remaining unmatched samples or the sum of distances between all pairs of samples is minimized, respectively. In greedy matching, the order in which the samples are matched may affect the resulting covariate balance, as opposed to the optimal approach. Nevertheless, previous literature [24] found that the two algorithms are not significantly different in creating well-balanced groups. To compensate for the insufficient number of controlled samples or to mitigate the effect of the matching order, matching can be performed by allowing the replacement of samples, that is, using the same sample more than once. To increase the quality of covariate balance, we can also adjust the caliper, i.e., the threshold of maximum distance that determines the matched samples.

In addition to these matching criteria, the definition of closeness (i.e., distance) between the samples can generate multiple variations of the matching method. One of them is the *exact matching*, which matches samples only when they have the same value for each covariate. This approach assigns a distance of 0 when samples are identical in all covariates (matching), or infinity otherwise (not matching). However, this is difficult to implement in practice, especially when the number of covariates is large or when they are not discrete but continuous variables. Therefore, relaxed distance metrics have been proposed to search for the most similar rather than perfectly identical pairs of samples.

The *propensity score matching* technique summarizes the set of covariates into one scalar value and determines the closeness based on the absolute difference between sample scores. By definition, the propensity score represents the conditional probability of assigning a treatment given a set of covariates [57]. This score is typically estimated by conducting a logistic regression of treatment (i.e., treated or not treated) on covariates [67], but other methods such as classification trees, random forests, or neural networks could also be utilized in the estimation process [30].

The propensity score can also be applied differently. Full matching [55], a special type of propensity score matching, divides samples into different strata based on the distribution of propensity scores and then matches samples to those within the same stratum. Because particular boundaries are first set based on the propensity scores and then samples are matched within them, the resulting matched samples could have closer propensity scores [26].

In *Mahalanobis distance matching*, the distance between samples is measured by directly using each covariate value rather than using another representative value. The distance between samples calculated in this method is similar to the Euclidean distance, but its formula uses the inverse of the covariance matrix of the covariates [14]. Here, the covariance matrix acts as a scaling factor that transforms the covariate space, enabling the comparison between covariates based on the same unit variance as when measuring the distance between covariates with different scales. Furthermore, as this matrix takes the covariance between two covariates into account, it can reduce the distortion in distance measurements, especially when they are highly correlated.

Coarsened exact matching [35] is a relaxed version of exact matching, which splits each covariate into a certain number of intervals and considers the samples to be equal on the covariate as long as they are within the same interval. Unlike propensity score or Mahalanobis distance matching, which deal with the covariates as a whole when measuring distance, this approach can separately control the imbalance bounding (i.e., the worst allowable balance) of each covariate [34]. As suggested by its name, each variable is coarsened into several groups (e.g., converting the variable "education period (in years)" into groups such as high school, BA, MS, and Ph.D.) and samples are matched within each group. In cases where multiple covariates exist, the combination of these coarsened intervals generates coarsened "bins", and only the samples (from each group) within the same bin will be counted as matched samples.

Moreover, in coarsened exact matching, the degree of balance is determined by setting the number of cutpoints (i.e., points where variables are split during the coarsening). A higher number of cutpoints results in smaller bins, thus making the samples within the same bin almost as identical in covariates as they would be in the exact matching method (i.e., better covariate balance). However, finding pairs of samples within smaller bins becomes harder, which reduces the number of successfully matched samples. Consequently, this method may prune a lot of unmatched samples even if the treated ones are included, so one should carefully consider the balance between increasing the covariate balance and reducing the discarded samples.

Suppose we select sample i and sample j from the treated and control groups, respectively. Also, let $\mathbf{X}_i = \{X_{i1}, X_{i2}, ..., X_{iN}\}$ and $\mathbf{X}_j = \left\{X_{j1}, X_{j2}, ..., X_{jN}\right\}$ denote sets of N chosen covariates for samples i and j, respectively. Then, the distance between samples from the treated and control groups, according to different covariate balancing matching methods, is computed as follows:

- Exact matching

$$dist = \begin{cases} 0 & \text{if } \mathbf{X}_i = \mathbf{X}_j \\ \infty & otherwise \end{cases}$$

- Propensity score matching

$$dist = \left|e_i - e_j\right|$$

  where $e_t = Pr(T_t = 1|\mathbf{X}_t)$ that denotes the estimated propensity score of sample t, and T_t is its treated state either 1 (treated) or 0 (control)

- Mahalanobis distance matching

$$dist = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)^T \mathbf{S}^{-1} (\mathbf{X}_i - \mathbf{X}_j)}$$

  where $\mathbf{S}$ is the covariance matrix of covariates

$$\mathbf{S} = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_N) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_1, X_2) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_N, X_1) & Cov(X_N, X_2) & \cdots & Var(X_N) \end{bmatrix}$$

- Coarsened exact matching

$$dist = \begin{cases} 0 & \text{if } i \in b_u \land j \in b_u \\ \infty & otherwise \end{cases}$$

  where b_u is uth coarsened bin

In practice, researchers can combine several criteria and the distance metric to fine-tune the strategy for matching samples. For instance, one can set the matching as "*1:2 nearest neighbor propensity score matching without replacement using a caliper of 0.25 standard deviations of the propensity score.*"

3 REVIEW OF THE CAUSAL ANALYSIS PIPELINE

In this section, we propose a causal analysis pipeline that examines the existence of causality from human behavior using smartphone sensor data. Overall, this pipeline consists of four main steps (Fig. 1):

- **Scenario setting** which determines the analysis target, defines the treatment and outcome variables, and selects the corresponding *behavior and context* features that should be extracted from the dataset.
- **Data preprocessing** where the chosen features are extracted from the raw smartphone sensor data and aggregated into a dataset characterizing human behavior and contexts within particular time windows.
- **Covariate balancing** that matches the control and treated samples having the most similar covariates thereby balancing the distribution of covariates (as in RCTs).
- **Treatment effect estimation** which involves measuring the average treatment effect and inferring the existence of a causal relationship.

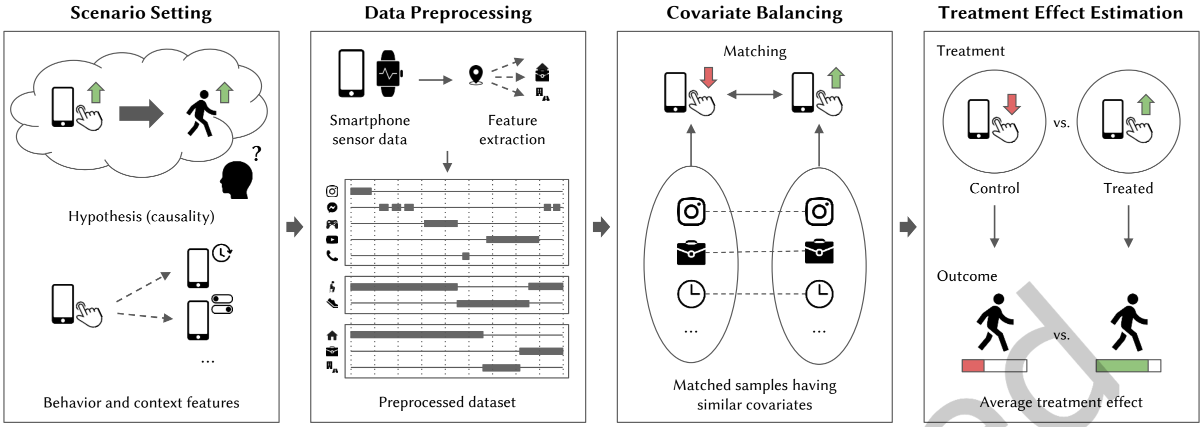


Fig. 1. The causal analysis pipeline proposed in this tutorial, consisting of 4 steps: (1) scenario setting, (2) data preprocessing, (3) covariate balancing, and (4) treatment effect estimation

In the following subsections, we further elaborate on the details of each of these steps. Special emphasis is made on the discussion on data preprocessing, where we describe how to extract the key features from the smartphone sensor data representing the user's interactions with a smartphone, their physical activity, and the surrounding contexts. In addition, in the covariate balancing subsection, we introduce four different matching techniques applied throughout this tutorial, which are widely used and actively discussed in causal studies from various fields.

3.1 Scenario Setting

The scenario setting step determines the target domain and scenario for the causal inference and specifies the corresponding features that should be extracted from the smartphone sensor data. This process begins by setting up a hypothesis of treatment and outcome variables. The hypothesis for the case study presented in this tutorial is that "*changes in mobile app usage cause variations in physical activity level.*" Therefore, the existence of causality will be evaluated by comparing the degree of physical activity in two groups with different levels of mobile app usage.

However, as "mobile app usage" or "physical activity level" may be ambiguous to measure, features are needed to specify these high-level human behaviors. For instance, features such as the "launch count" and "usage time" of apps can be defined to assess the "mobile app usage" level, as in [53]. These features respectively describe how frequently and for how long the user interacts with an app. In addition, mobile app usage can be analyzed by overall or categorical use, rather than by individual apps, to capture the general usage trends and make the results from multiple users comparable even if they use different apps. In the case of physical activity level, the "sedentary time" quantifies how long a user remains still, such as when sitting indoors or riding a vehicle, thus measuring their physical inactivity.

In addition to the treatment and outcome variables mentioned above, context variables, such as location and time, should also be considered because they may affect human behavior and its corresponding causal relationship. For instance, the use of entertainment apps (e.g., games) will be low when users are focused on their tasks at work, where access to these apps is deterred. Similarly, their sedentary time will be longer at night when they are resting back home. Thus, the analysis should link location and temporal information, e.g., by recording the time spent at certain places (e.g., home, work, and others) or activities occurring at different times of day (e.g., morning, afternoon, evening, and night).

3.2 Data Preprocessing

The data preprocessing stage extracts the features defined in the previous step from the smartphone sensor data and constructs a preprocessed dataset that will be directly used for making the causal inference. This step starts by cleaning and organizing the data to prevent invalid data, such as missing values and outliers, from leading to misleading conclusions. In addition, users with too much missing data (e.g., data was not collected for more than 24 hours) are excluded from the analysis, assuming that their data was not appropriately collected.

Subsequently, the features are extracted from the raw smartphone sensor data characterizing human behavior and context. Although data types may vary depending on the purpose of the study, they can divided into two main types: data collected through passive sensors (e.g., GPS or accelerometer) and data generated by active user interactions (e.g., launching mobile apps or making phone calls). Based on data that can be collected using Android APIs, we illustrate below how to extract the key features that are used in this tutorial to evaluate the relationship between smartphone use and physical activity, namely, (1) mobile app usage, (2) sedentary time, (3) location, and (4) time.

3.2.1 *Mobile App Usage.*

Mobile app usage can be inferred based on two usage event types, `MOVE_TO_FOREGROUND` and `MOVE_TO_BACKGROUND`, as illustrated in [2]. Note that after Android API level 29, these events were deprecated and logged as `ACTIVITY_RESUMED` and `ACTIVITY_PAUSED`, respectively. When the user launches an app, it is moved to the foreground (i.e., fully occupying the current screen) and an event of type `MOVE_TO_FOREGROUND` is recorded. In contrast, when the app is closed and no longer in use, it is moved to the background (i.e., moving away from the current screen) and an event of type `MOVE_TO_BACKGROUND` is registered. This implies that the time between these two event types for a given app can be interpreted as “screen time” [66] during which the user is actively interacting with mobile apps.

As in Fig. 2 (a), the launch count and usage time of an app can then be measured based on the number and duration of these screen time events, respectively. However, interactions may occasionally be missed, which results in unmatched pairs of event types, for instance, two successive and repeated `MOVE_TO_FOREGROUND` or `MOVE_TO_BACKGROUND` events. This can be handled by removing one of the repeated event types to complete the pair, assuming that the two events happened but one of them was not logged correctly. Also, since the focus is on measuring the user’s active interactions with mobile apps, other recorded event types, like simple screen-on/off events or incoming notifications, are excluded.

Furthermore, the launch count and usage time of an app can be aggregated to examine how changes in the overall usage of apps belonging to specific categories affect the levels of physical activity. In general, the categories registered in Google Play are used to classify the apps. However, there may be categories of apps that are very rarely launched and simply increase the number of insignificant features. Thus, it is suggested that such rarely used apps are filtered out during data preprocessing or that the number of categories is reduced by grouping similar apps as proposed in [70].

3.2.2 *Sedentary Time.*

Assuming that users bring their smartphones wherever they go, sedentary time can be measured based on the physical activity inferred by the Android system. Android APIs such as Activity Recognition and Activity Transition are useful for identifying the activity type and measuring the user’s sedentary time. These APIs can detect different types of physical activities, including `STILL`, `WALKING`, `RUNNING`, `ON_BICYCLE`, and `IN_VEHICLE`, and the APIs record the start and end times of the given activity using `ACTIVITY_TRANSITION_ENTER` and `ACTIVITY_TRANSITION_EXIT`.

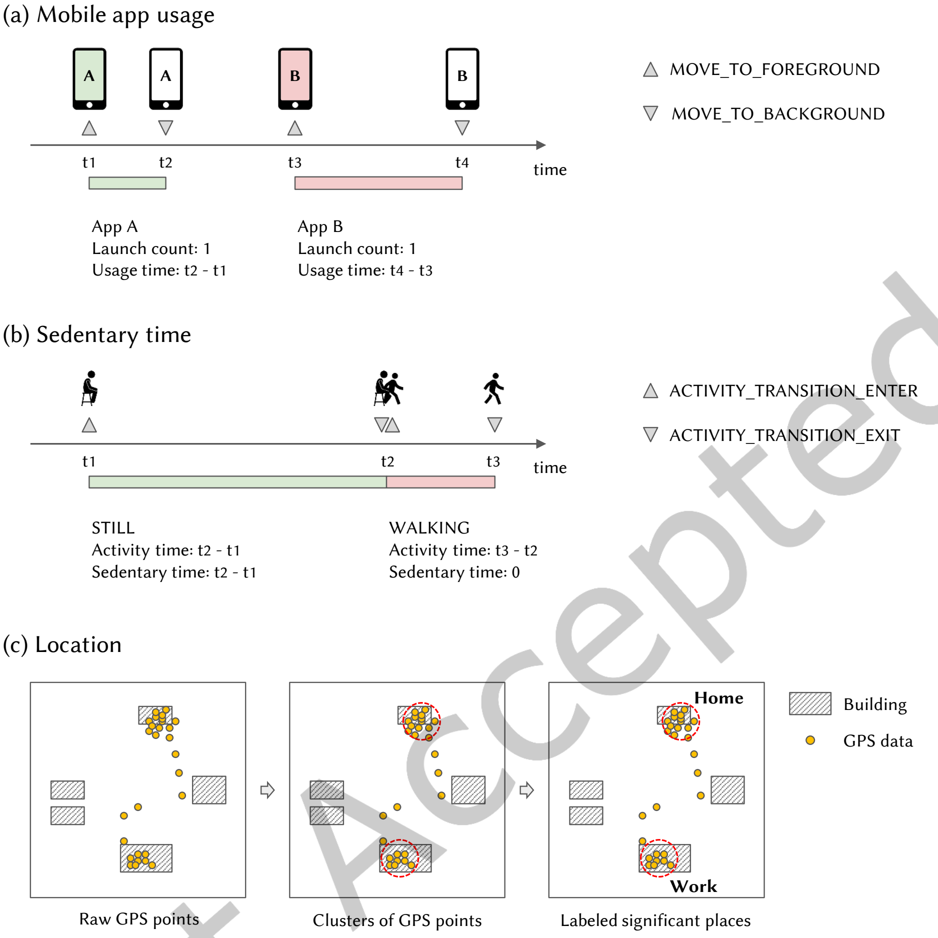


Fig. 2. Extraction of key features from smartphone sensor data: (a) mobile app usage from the usage event types, (b) sedentary time from physical activity types, and (c) location from clustered GPS points which converted into labeled significant places

As in the case of mobile app usage estimation, sedentary time is calculated by aggregating the intervals during which physically inactive events (e.g., STILL or IN_VEHICLE state) are recorded, as illustrated in Fig. 2 (b). For instance, when sitting down in the office, driving a car, or sleeping at home, the APIs log STILL or IN_VEHICLE events, and their duration is counted as sedentary time. Sometimes, when not accurately classified, the physical activity type is recorded as “unknown” and these data should be dropped in the preprocessing step.

3.2.3 Location. The location of a user at a building level and at a particular time can be determined by clustering neighboring GPS data (Fig. 2 (c)). Because GPS data may deviate even if the user is staying at the same place, it is reasonable to cluster close points and regard them as one location rather than count each of them as a different place.

Imputation of missing GPS data is required before creating clusters. Because the Android system collects GPS data only when the user moves more than a certain distance, missing values are possibly due to the user's lack of mobility. Thus, the missing values are imputed with the last location recorded in the dataset.

GPS data are clustered based on the neighboring criteria such as distance (e.g., the maximum radius of the cluster) and timespan (e.g., the duration of stay at a location) [86]. The threshold for each criterion is defined after overviewing the user's activities and contexts; for instance, if the buildings at which the user stays are small or very close, a shorter distance criterion is set to separate the clusters. Also, the clusters generated can be assigned category labels, such as "home", "work", "social venue", and "gym", as in [75], or more abstracted into significant places where people spend most of their time (e.g., home and work) and others [53]. Moreover, because the user's location changes over time, measures such as the time spent at a given location are useful to represent where the user was at a specific instant.

3.2.4 Time. Considering that human behavior follows some patterns over time, mobile app usage and physical activity may correspondingly vary over time. The analysis may be done based on fine-grained time intervals, such as hours, but that would make it difficult to distinguish the user's behavior by time. Instead, time can be labeled with semantic meanings, such as morning, afternoon, evening, and night, that denote timespans where similar events may happen.

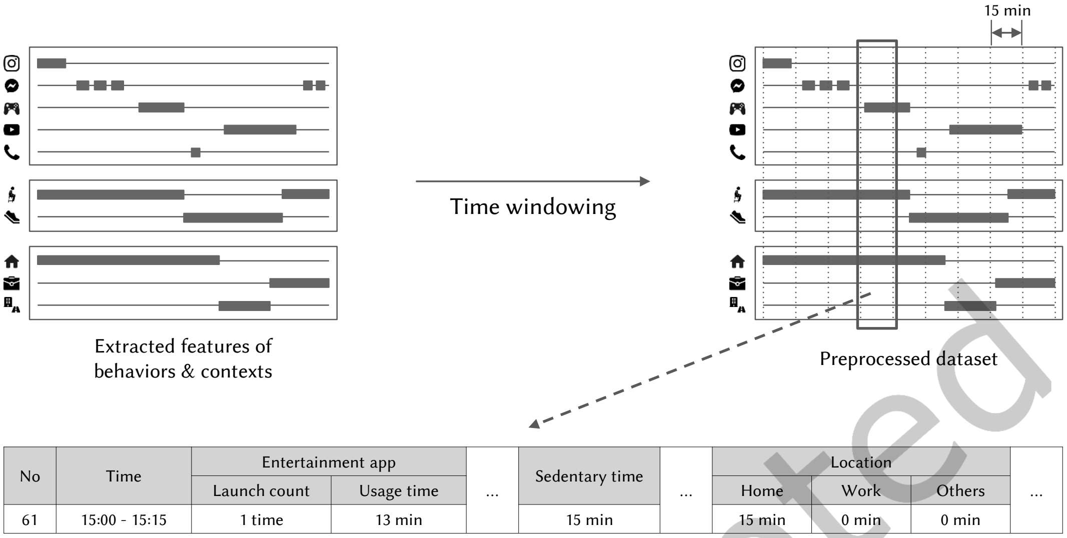
Given that smartphone sensor data records time-series events, the features extracted can be divided into *time windows* to characterize the events that occurred within each interval. This time windowing allows a better capture of human behavior, which is a continuum of events rather than a snapshot. In addition, the samples generated in this process are directly used in conducting the causal inference, for instance, by comparing the physical activity levels between intervals when the mobile app was used frequently and not.

When setting the time window size, a proper size should be chosen to avoid (1) diluting events that rarely or shortly happen if the window becomes too large (e.g., hours and days) and (2) generating too many repeated meaningless values if the window gets too small (e.g., seconds). Throughout this step, researchers can obtain preprocessed datasets where each sample represents the events (i.e., features) that occurred in a given time window. In the example of Fig. 3, the time window is set to 15 minutes and all the extracted features, such as launch count and usage time of entertainment apps or sedentary time during the corresponding window, are arranged as a single sample.

3.3 Covariate Balancing

The objective of the matching process is to artificially create a balanced pseudo-population from the observational data in which the treated and control groups have similar covariate distributions and differ only in whether the treatment has been administered. Because all variables other than treatment application are balanced, it can be concluded that any difference in the outcome of the two groups is solely due to the difference in the treatment.

3.3.1 Treatment Group Assignment. Because the features extracted from the mobile data are typically continuous, it is difficult to assess the causal relationships of interest through matching. Therefore, a simple yet effective method is to binarize the treatment values so that the subjects can be classified as either treated or controlled. For instance, if the researcher selects the "usage time of social apps" as a treatment variable, each sample is assigned to either the treated or control group depending on whether its usage time is larger than a certain threshold, such as the average of all the samples. Although previous studies suggested other ways of dealing with continuous treatment variables [17, 37], binarization is more widely used due to its simplicity. To separate the treated and control groups more clearly, a threshold can be set so that samples whose treatment values are close to the average are excluded, as in [75].



No	Time	Entertainment app		...	Sedentary time	...	Location			...
		Launch count	Usage time				Home	Work	Others	
61	15:00 - 15:15	1 time	13 min		15 min		15 min	0 min	0 min	

Fig. 3. A diagram of the preprocessed dataset after generating samples with 15-minute sized time windows. The table is an example of one sample, which represents the participant's behavior and context that happened in a particular 15-minute window.

3.3.2 Covariate Selection. Next, the covariates that need to be balanced are selected from the dataset. As explained by Zhao et al. [88], not every variable should be considered a covariate and balanced. Instead, variables that affect both the treatment assignment and outcome or solely the outcome should be included in the covariate set.

In general, these variables are chosen based on the domain knowledge or results from prior literature, but studies also proposed an alternative for selecting covariates based on a correlation test between treatment and outcome variables. For example, if the usage time of entertainment apps is significantly correlated with that of social apps (treatment) and sedentary time (outcome), it should be included as a covariate that requires balancing before conducting causal inference. This approach assumes that the existence of a (statistically significant) correlation is a necessary condition for causality, and it may be considered as an alternative when there is a lack of prior knowledge in the domain of interest. However, it should be noted that the correlation test depends on the correlation measure used.

3.3.3 Matching and Balance Checking. After the covariates are determined, matching is conducted to pair samples from the treated and control groups that have similar covariate values and can, thus, act as counterfactuals of each other. In this tutorial, four matching methods based on different distance metrics are applied to reach covariate balance: propensity score matching (1:1 nearest and optimal full), Mahalanobis distance matching, and coarsened exact matching.

Propensity Score Matching. This tutorial demonstrates two representative propensity score-based approaches, namely, 1:1 nearest neighbor and optimal full matching. In 1:1 nearest neighbor matching, which is the simplest and most basic method, each treated sample is paired with the control sample having the most similar propensity score. As a greedy matching, it selects the closest sample among the remaining unmatched ones every time. On the other hand, optimal full matching forms several strata by the propensity score distribution and matches samples within each stratum. When the matching ratio is other than 1:1, weights are given to take the different numbers of samples into account in estimating the treatment effect. Furthermore, this method creates “optimally” matched pairs by minimizing the sum of distances (of matched samples) within each stratum by controlling the number of strata and the assignment of the samples.

After matching, the covariate balance is assessed by the similarity of the covariate distribution in the treated and control groups. For evaluating the balance, several measures can be utilized, such as the standardized mean difference (SMD), variance ratios, and empirical cumulative density functions [22]. Among them, the SMD, which is the most widely used, is calculated for each covariate as the difference in mean covariate between treated and control groups divided by the standard deviation of the covariate among samples. The covariates are regarded as balanced if the absolute SMD for all covariates is smaller than a certain threshold, for instance, 0.1 or 0.25, as suggested in [69].

Mahalanobis Distance Matching. Rather than merging the covariates into a scalar as in propensity score matching, in Mahalanobis distance matching the original values of the covariates are directly used to measure the distance between the samples. As illustrated in the previous section, in Mahalanobis distance the covariance matrix is calculated using all the covariates to uniformly transform the variances of the covariates. Mahalanobis distance matching can be also combined with other criteria such as setting a caliper (i.e., the maximum distance that allows matching), but this tutorial focuses on the basic format of 1:1 nearest neighbor Mahalanobis distance matching. Moreover, the covariate balance is evaluated after matching as in the propensity score matching.

Coarsened Exact Matching. Coarsened exact matching generally creates coarsened bins based on meaningful criteria. However, we propose a simple method for selecting a set of randomly generated cutpoints and finding the optimal one in cases where no clear coarsening criteria exist. This method is required because human behaviors and lifestyles may vary by participants, meaning that a global, common set of cutpoints might not be suitable to balance the covariates of all unique individuals. Therefore, our method begins by randomly generating multiple sets of cutpoints based on a predefined minimum/maximum number of cuts for each covariate to explore and assess the different coarsening strategies. Then, the cutpoint sets in which all the covariates are well balanced after coarsened exact matching, evaluated by the same criteria (i.e., the absolute SMD less than 0.1 or 0.25) as in other methods, are chosen.

Among these balanced cutpoint sets, the one that produces the most paired samples is selected. Specifically, because the number of control samples is generally greater, the number of pairs should be counted based on the matched “treated” samples. This process yields an optimal cutpoint set from multiple candidates, achieving balance in all covariates and maximizing matched treated samples. Considering that the optimal set of cutpoints may vary since it is randomly generated, one can perform matching multiple times and aggregate treatment effect results from each trial.

3.4 Treatment Effect Estimation

As the final step of the causal analysis pipeline, the effect of the treatment on the outcome is estimated based on the samples matched by each matching method. Through matching, the difference between the mean outcome from the treated and control groups is evaluated by an unbiased estimator of the true average treatment effects. When the matching ratio between the treated and control samples is 1:1, the same weights are given to all the samples.

However, methods that create strata, such as the optimal full and coarsened exact matching, may have a matching ratio other than 1:1, in which case the weights are inversely proportional to the number of samples of each group. Suppose that there is one stratum composed of n_t treated and n_c control samples out of a total number of treated and control samples of N_t and N_c. In this case, the assigned weights are 1 for the treated and (n_t/n_c) for the control to adjust the different sample sizes in the stratum. The weights of the control samples are then scaled by (N_t/N_c) so that the total sum of weights from each stratum becomes equal to the number of matched control samples [68].

There are different types of estimands (or values to be estimated via an analysis) for evaluating the treatment effect, such as the average treatment effect in the population (ATE), the average treatment effect in the treated samples (ATT), and the average treatment effect in the remaining matched samples (ATM) [23], which are determined based on the target population and the existence of discarded samples. Among these estimands, the ATT is the most widely used because studies usually focus on the effect of the treatment on treated samples (e.g., the impact of new medicine), and is applicable when each treated sample is matched with either one control sample or multiple weighted control samples. ATE can be applied when all the samples in the original dataset are included (after weighting if the ratio is not 1:1) in the treatment effect estimation, for instance, in optimal full matching. However, if some of the treated samples are pruned after the matching process (e.g., in coarsened exact matching), the only option to estimate the effect is ATM, which takes into account only the successfully matched samples.

The treatment effect [21, 38, 50], can be estimated by performing a linear regression between the outcome and treatment (more strictly, whether treated or not), in which the coefficient of the treatment variable explains the existence and magnitude of causality. If the coefficient is statistically significant (e.g. as determined through a t-test), it can be concluded that causality between the treatment and outcome variables exists. Moreover, the magnitude and sign of the coefficient indicate how intense the change in the outcome is due to the given treatment and whether they are causally related positively or negatively. After all these steps, the estimated treatment effect should be reported along with additional information such as the matching method, the distance measures (e.g., how the propensity score is estimated or how the cutpoints are determined for each covariate), the number of matched/unmatched samples from both the treated and control groups, and the estimand of treatment effect, as suggested in [71].

Unlike previous studies on causal inference among different subjects, the proposed causal analysis pipeline focuses on causal relationships among a single participant's data (i.e., it is subject-level). This is possible because multiple data points are observed for each subject data in smartphone sensor datasets. Therefore, the resulting causal relationship may vary between subjects with different lifestyles. This implies that a direct inter-person comparison of the results may not be appropriate since the differences among people are not controlled. Still, we may observe a general trend that the causal relationship found only in one specific person may exist across the population as well.

4 CASE STUDY

The causal analysis pipeline that we have introduced is now used to illustrate how to conduct the causal inference using a real-world smartphone sensor dataset.

4.1 Dataset

In the case study, we used a smartphone sensor dataset from smartphones collected in a user's daily life context. The data were collected over seven days in 2019 from 74 participants (23 women, average age 23.3 years) in a large university. In the data collection process, a mobile data collection tool was installed on each participant's smartphone, and participants were asked to use their phones as usual. This configuration is typically used in prior smartphone sensor data studies [11, 76, 81].

The data collection software [43] tracked three types of smartphone data as follows (note that all events were logged with a corresponding timestamp in Unix time, allowing us to estimate their duration and frequency):

- **Sensor data** include location information based on GPS data (i.e., latitude, longitude, and altitude) and physical activity states inferred from accelerometer values. Android supported APIs for recognizing physical activities, and our software logged the type of activity and when each activity began and ended with timestamps.

- **Interaction data** detail how a participant interacted with the smartphone; these interactions include generic applications, voice calls, and text messaging. Data for app usage events consisted of the app's name and package name, whether it is a pre-installed app, and the usage event types illustrating whether the app was moved to the foreground or background of the smartphone screen or any other interactions from the app. For voice calls and SMS, the collected data included the encrypted ID of the person on the other end of the line (via a one-way hash function) and the initiator (i.e., incoming or outgoing call/message) and length of the communication (i.e., the call duration and the message length).
- **Device data** collected included the status of the smartphone, such as battery level and temperature, and network status, consisting of data traffic, Wi-Fi connection state, and nearby scanned Wi-Fi access points.

Before the analysis, we excluded participants whose data were inappropriate for causal inference, for instance, if their data were less than 6 days in total, any data were not collected at all, app usage event types were absent, or the GPS data were too sparse to label significant places. Consequently, we included data from 49 participants in this tutorial.

Moreover, we excluded device data from the analysis because they were not directly related to our target scenario, as the causal relationship between mobile app usage and physical activity. Voice calls and SMSs were also not included in this case study, because these events were very sparse in the dataset and would not have a significant effect on the causal relationship. Therefore, in this study, we used smartphone sensor data consisting of GPS, physical activity type (inferred from the accelerometer), and app usage events.

4.2 Method

In this case study, we consider a personalized view of human behaviors and decision-making, known as idiographic perspectives [4], by conducting causal inference on subject levels. Personalized data analyses have been widely used in mobile sensing research [13, 87]. We therefore performed causal inference with each participant's data separately based on the causal analysis pipeline proposed in the previous section. Note that individual results can be aggregated to identify generalizable relationships applicable to all users, known as nomothetic perspectives [6].

First, we examined how the launch count and usage time of the overall app were causally related to sedentary time. We repeated this process for the six app categories including social, entertainment, information, work, system, and health. The causal inference was thus conducted for 686 different test cases in total from 49 participants, two usage metrics (i.e., launch count and usage time), and seven different scenarios (one overall and six categorical usages).

As a tutorial, we first demonstrate how we ran the analysis in detail here, by selecting one particular sample case composing one specific scenario and one participant's data. Then, we show the results from the 686 different cases in the evaluation section concerning their covariate balance and the treatment effect (i.e., causal relationship).

4.2.1 Scenario Setting. We considered a scenario investigating whether one's mobile app usage causes changes in physical activity from the collected smartphone sensor data. Investigation of the causal relationship between the two behaviors could have considerable potential to identify problematic phone usage patterns that might negatively affect health or help design an intervention system that primarily targets such harmful usage patterns. In the case study, we investigated whether the participant's "usage time on social apps" has any causal relationship with their "sedentary time". It is readily apparent in daily life that many people use their smartphones not only when sitting but also while moving around. Among the various categories of apps, this sample case aims to show how the duration of interaction with social apps (i.e., usage time) may causally affect sedentary time, resulting in either longer or shorter sitting duration.

As a running example in this section, we investigated the causal relationship based on data for one participant (ID: P27). The social app usage time can be extracted from the interaction data, referring to the name of the social app and its usage event types (i.e., whether it was opened or closed). In addition, as illustrated in the previous section, we decided to use the recognized physical activity and when it occurred from the sensor data when estimating the participant's sedentary time. Here, we include the launch count and usage time of other categories of apps in the analysis because they may be covariates that affect sedentary time or both social app usage time and sedentary time for P27.

In addition, we included context information such as location (GPS) from the sensor data and time of day, dividing one day into four equal intervals. As in other studies [53, 89], for ease of analysis, we focus on the significant places (i.e., residence and workplace) where people spend most of their time every day and categorize places other than these as "others". Temporal information was used in the form of four time-of-day labels: night (00:00-06:00), morning (06:00-12:00), afternoon (12:00-18:00), and evening (18:00-24:00).

4.2.2 Data Preprocessing. Data preprocessing involves the extraction of features of behavior and context determined in the scenario setting step (Fig. 4).

Mobile App Usage. The usage time of social apps was calculated using the following steps: we labeled each app logged in the dataset with the corresponding category provided by Google Play, and classified them again as social if they were in one of the "social", "communication", and "dating" categories (Table 1).

Next, we created chunks of app usage events using the event type value; one chunk was created for each pair of consecutive `MOVE_TO_FOREGROUND` and `MOVE_TO_BACKGROUND` events, denoting the interval between the participant running and closing of the app. When these pairs were not created well owing to the logging issue

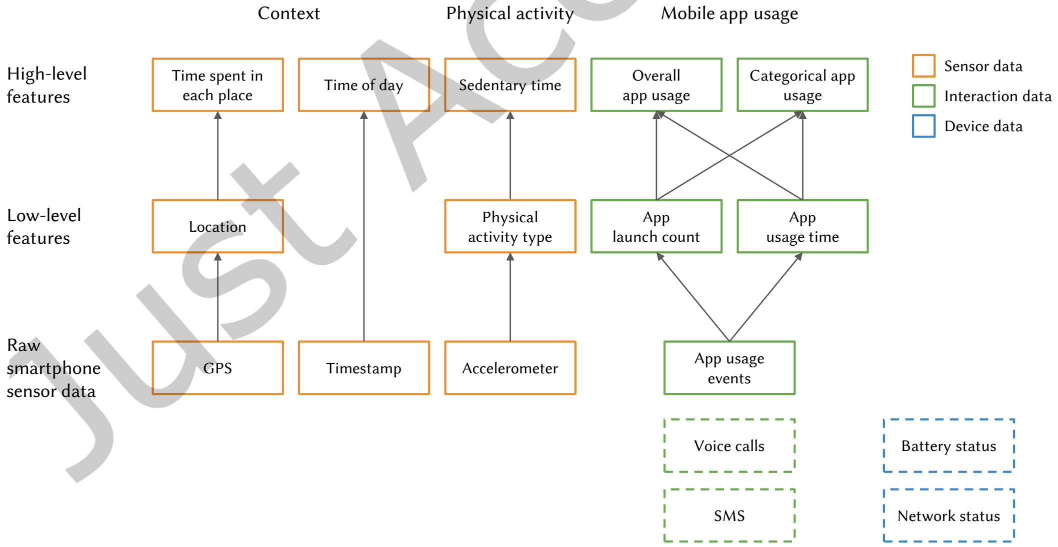


Fig. 4. Extraction of features that illustrate behaviors and contexts of interest from raw smartphone sensor data. The raw data is processed into low-level features, which then are converted into high-level features that will be directly used in the causal inference. Note that the sensor data in dashed boxes are collected in the study but excluded as they are sparse or irrelevant to the analysis.

Table 1. The app categories used in this study (left column) and the original categories from Google Play (right column)

New Category	Categories from Google Play
Social	Communication, Social, Dating
Entertainment	Game, Entertainment, Music and Audio, Cartoon, Video Players, Art and Design, Photography
Information	Lifestyle, Shopping, Travel and Local, Weather, Food and Drink, Map and Navigation, Beauty, Books and Reference, News and Magazine, House and Home
Work	Productivity, Finance, Education, Business
System	Tools, Library and Demo
Health	Health and Fitness, Sports, Medical

(e.g., two successive running or closing app events), one of them was removed. Usage time for social apps was thus estimated by calculating the duration of these chunks. To measure the launch count of social apps, we counted the number of these chunks of interactions. Following the same process, the launch count and usage time of other categories of apps were also extracted by calculating the number and duration, respectively, of chunks.

Sedentary Time. The participant's sedentary time was measured with time intervals, where the activity type was recognized as either `STILL` or `IN_VEHICLE`. Similar to app usage event extractions, we generated chunks of physical activities for each activity type and its start/end time and aggregated them only when the types were physically inactive.

Location. We extracted the participants' location information by creating clusters of GPS data referring to [86] and labeling them as home, work, or others. When forming the clusters, we empirically determined the best clustering criteria as a maximum of 25 m in radius and a minimum duration of 15 min stay at that location. The criteria were sufficient in that they could separate buildings in built-up areas and distinguish instances of motion outside buildings.

After clustering, we obtained 28 different clusters of GPS data, indicating buildings in which the participant stayed for more than 15 min. Of these clusters, the one with the longest stay duration was considered home, and the other top two clusters were labeled work. We found that this participant's home cluster was located in his dormitory inside the university campus, where his GPS data late at night were usually found. In addition, the clusters labeled as "work" were found in the main lecture room and library, which were places that appropriately represented where he worked.

Time. We labeled time information (i.e., time of day) using the four time intervals described earlier. However, we excluded all the events that occurred at night (i.e., time of day assigned to "night") since most of them were composed of records while sleeping, which did not include either mobile app usage or physical activity.

The features extracted above were then cut based on even-sized, 15-minute time windows to create samples describing what happened within each window. Each sample thus included all the information for a particular time, consisting of the app usage, sedentary time, time spent at each location, and time of day. Here, some of the samples include frequent usage of social apps, while others do not, and this difference will be used to categorize the samples into treated and control groups in the next step. In addition, we finalized the data preprocessing by conducting min-max normalization for each feature and transforming the values into decimals between 0 and 1 to adjust the different scales among features.

4.2.3 Covariate Balancing.

Treatment Group Assignment. At the beginning of the matching process, we binarized social app usage time based on its average; we assigned the sample to the treated group if its usage time was above average and to the control group otherwise. This participant's average usage time was 3.53 min (within the 15-min time window), and of 496 samples, 158 were assigned to the treated group, representing longer social app usage than average (Fig. 5). This process led to the addition of a new variable, “treated,” to the dataset to indicate whether the sample was allocated to the treated (treated = 1) or control (treated = 0) group. After assigning samples, the variable “treated” would be used to represent the treatment level in all the following processes instead of the original treatment variable (i.e., social app usage time).

Covariate Selection. Next, we selected covariates among the extracted features (except social app usage time and sedentary time) that needed to be balanced. We utilized correlation tests as described in the causal analysis pipeline, due to the limited prior knowledge about features affecting the treatment and outcome for each participant.

We conducted Kendall’s rank correlation test because the features were not normally distributed and determined them as covariates when they showed significant and moderate correlation with the outcome, which was sedentary time in this example. We set the criteria for the correlation test as follows: (1) the absolute value of the correlation coefficient should be larger than 0.2, and (2) the coefficient should have a p-value less than 0.05, as in [53].

From this test, we identified six covariates: launch count and usage time of each health app and information app and duration at home and “others”, respectively. In addition to this covariate set, we included the time of day separately so that the matching process always considered when those events occurred. Consequently, seven covariates that should be balanced before estimating the treatment effect were selected (Fig. 6).

Matching and Balance Checking. We applied four different matching approaches to balance the covariates: (1) propensity score 1:1 nearest neighbor matching, (2) propensity score optimal full matching, (3) Mahalanobis distance 1:1 nearest neighbor matching, and (4) coarsened exact matching. For matching, we primarily utilized **MatchIt** [30], an R library that supports the creation of matched samples using diverse matching methods. In addition, we used the R library **cem** [33] to randomly generate multiple cutpoints for coarsened exact matching.

For the two propensity score-based methods, we estimated the propensity score using the logistic regression of “treated” on the linear combination of the covariates. The distribution of the propensity score for each 1:1 nearest neighbor matching and optimal full matching is shown in Fig. 7. The former conducted matching between samples with the closest propensity scores and discarded the remaining control samples (i.e., the unmatched

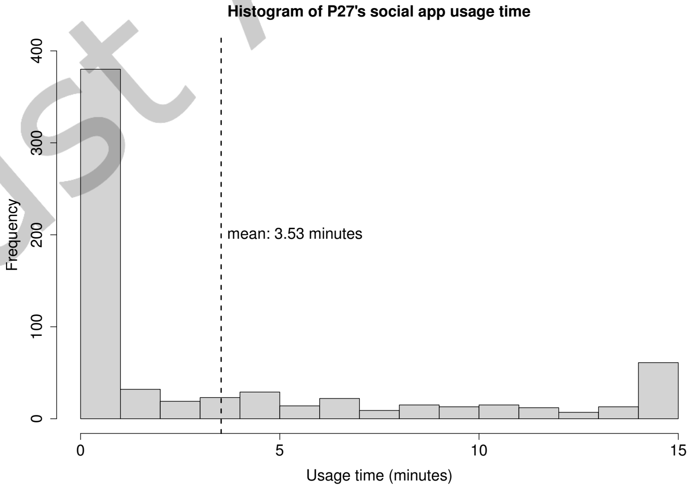


Fig. 5. The distribution of social app usage time of P27 from each sample

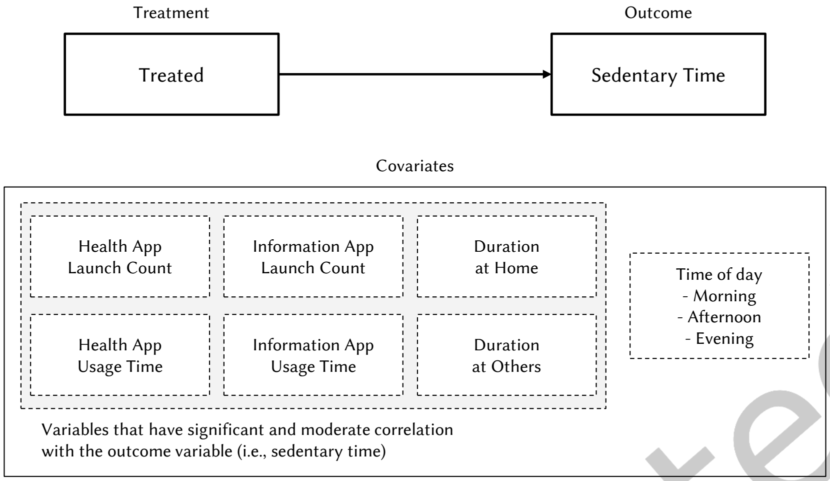
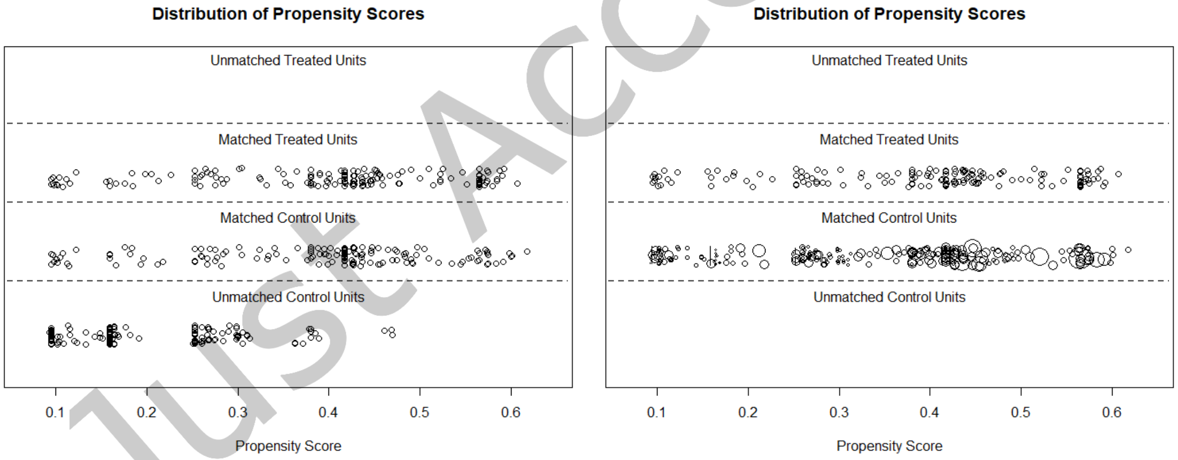


Fig. 6. Sets of features (treatment, outcome, and covariates) that are considered in inferring causality between social app usage time and sedentary time from P27's data. Note that the "treated" state after treatment group assignment is used as the treatment instead of social app usage time, and covariates are selected based on the correlation with the outcome variable (i.e., sedentary time).



(a) Propensity score 1:1 nearest neighbor matching

(b) Propensity score optimal full matching

Fig. 7. The distribution of propensity scores after conducting propensity score-based (a) 1:1 nearest neighbor matching and (b) optimal full matching. Each point denotes one sample, and the size of the point in (b) indicates the relative weights given to each sample.

control units), whereas the latter included all the samples in matching but assigned different weights based on the matching ratio, where the size of each point is proportional to the weight. After matching, we confirmed that

Table 2. The covariate balance result from Mahalanobis distance matching

Covariate	Means (Treated)	Means (Control)	Standardized Mean Difference
Health app launch count	0.1098	0.1044	0.0339
Health app usage time	0.0283	0.0237	0.0494
Information app launch count	0.1425	0.1335	0.0479
Information app usage time	0.0710	0.0709	0.0006
Duration at home	0.3206	0.2802	0.0884
Duration at others	0.6456	0.6965	-0.1126
Time of day (morning)	0.1392	0.1646	-0.0731
Time of day (afternoon)	0.3481	0.3544	-0.0133
Time of day (evening)	0.5127	0.4810	0.0633

all covariates reached a balance in that the maximum absolute SMD (i.e., the worst case in terms of balance) for each case was 0.155 and 0.160, respectively.

For Mahalanobis distance matching, the original values of covariates were used directly without any conversion, as for the propensity score matching. Because we used 1:1 nearest neighbor matching based on the Mahalanobis distance, for each treated sample, we searched for the nearest control sample among the unmatched control samples (i.e., the greedy approach). This method showed a maximum absolute SMD of 0.113 from the duration spent at "others" (Table 2), meaning that all the covariates were well balanced.

Coarsened exact matching randomly generated multiple sets of cutpoints for coarsening, and we chose the optimal set that reached the covariate balance while minimizing the number of treated samples pruned (Fig. 8). The number of cuts differed for each covariate, for instance, the usage time of health apps and duration at home were cut to create four even intervals, whereas the usage time of information apps and the launch count of health apps and information apps produced two intervals. In addition, this set of cutpoints discarded 13 treated samples out of 158 and showed a maximum absolute SMD of 0.0544, implying that all the covariates were well-balanced.

In this specific case, we achieved the covariate balance from all the matching methods we demonstrated. The absolute SMD of all covariates after matching was below 0.25 from all four different approaches we used (Fig 1 in Supplementary Material). We could therefore continue to the next step (estimating the treatment effect) with matched samples from all of these matching methods. This was possible because the distributions of covariates were similar between the treated and control groups even before matching. However, some of the matching approaches might fail to balance the covariates in several cases depending on the distribution of data, and we will discuss these cases in the next section.

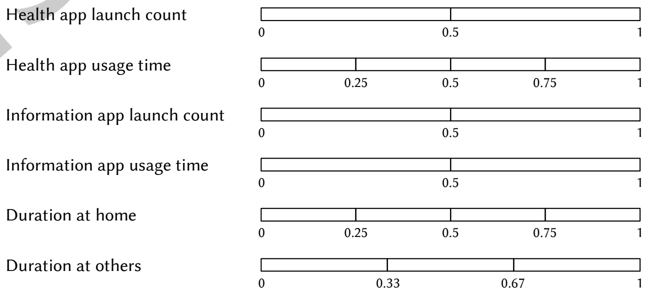


Fig. 8. The optimal set of cutpoints from coarsened exact matching. Time of day was not shown in this figure as it is a categorical variable simply composed of either 0 or 1.

Table 3. The estimated treatment effect from each matching method

Matching Method	Estimate	Std. Error	P-value
Propensity score 1:1 nearest neighbor matching	0.116	0.026	<0.001
Propensity score optimal full matching	0.089	0.027	<0.01
Mahalanobis distance 1:1 nearest neighbor matching	0.068	0.022	<0.01
Coarsened exact matching	0.051	0.023	<0.05

4.2.4 Treatment Effect Estimation. Using the matched samples from each matching method, we calculated the ATT to estimate the effect of social app usage time on sedentary duration. We used linear regression between the two variables, and weights were also included in the estimation if the ratio of matching was not 1:1. We then conducted a t-test on the coefficient of the treatment variable (i.e., "treated") to check causality.

Table 3 shows the results of the estimated treatment effect, where the estimate explains the difference in normalized sedentary time between the treated and control groups and the p-value describes its statistical significance, which confirms the existence of causality. We observe that all p-values are smaller than our threshold (0.05), meaning that causality was found regardless of the matching methods applied in this specific case. As all the estimates have positive numbers despite differences in size, we could infer that P27 sat down longer when using social apps longer.

4.3 Evaluation

As mentioned in the Methods section, we conducted causal inferences for each scenario. Thus, all 49 participants were analyzed using the same causal analysis pipeline. In this section, we summarize the results of the covariate balance and treatment effect estimation from each matching approach applied in this study.

We illustrate how many of the test cases succeeded in reaching covariate balance and how many of these were inferred to have a causal relationship between treatment and outcome variables. Since participant behaviors and contexts differ, we cannot compare the treatment effect directly to conclude the existence of causality in general. Instead, we first describe the results from each causal scenario at the subject level to see whether the participants' data show causality and how they differ across different matching methods. Then, we provide descriptive statistics for the estimated treatment effect and its direction at the group level to see whether such results are generally found across participants or specific to a few of them.

4.3.1 Subject-level Analysis. From the four matching methods, we found differences in the number of test cases where all covariates were successfully balanced, and causality was found from the matched samples after the balancing process (Table 4). The coarsened exact matching achieved a covariate balance for almost all cases because we chose the optimal set of cutpoints that met the balance criteria after randomly generating multiple sets. Other methods resulted in cases in which covariates were not balanced, and we were therefore unable to proceed with the treatment effect estimation. Moreover, each matching method generated different matched

Table 4. The number of test cases that showed covariate balance and the existence of causality from each matching method

		Propensity Score 1:1 Nearest Neighbor	Propensity Score Optimal Full	Mahalanobis Distance 1:1 Nearest Neighbor	Coarsened Exact
Overall	Covariate Balance	40	66	84	98
	Causality Found	24	27	48	55
Categorical	Covariate Balance	149	224	218	511
	Causality Found	29	48	48	157

samples (with different weights) that were used in the treatment effect estimation, resulting in the difference in the causality outcomes.

For overall app usage, there were 98 test cases, composing 49 participants and their launch count and usage time for all apps (two usage metrics). Coarsened exact matching was able to reach the covariate balance for all test cases. Meanwhile, Mahalanobis distance matching showed 84 cases matched, whereas propensity score 1:1 nearest neighbor and optimal full matching succeeded in 66 and 40 cases, respectively. Moreover, 40 cases succeeded in achieving covariate balance from all approaches, and nine cases were balanced based only on coarsened exact matching.

Moreover, we estimated the treatment effect only for cases in which the covariates were balanced. From the four matching methods, on average, 56.1% of the covariate balanced cases had a causal relationship between mobile app usage and sedentary time. Moreover, 21 out of 98 test cases showed causality from all the matching approaches. In the case of categorical app usage, there were a total of 588 test cases from the six categories of apps, two usage metrics, and 49 participants. However, 70 cases could not be used in causal inference because participants either did not use particular app categories at all or used those apps only once; therefore, statistical tests were not applicable. These situations were found mostly for health apps (66 cases), and sometimes for information (three cases) and work (one case) apps. Thus, we used the remaining 518 cases in the covariate balancing and treatment effect estimation process.

From the analyses, coarsened exact matching showed a covariate balance for almost all available cases (511 cases). In the categorical app usage scenario, propensity score 1:1 optimal full matching showed slightly better balancing outcomes than Mahalanobis distance matching. However, the propensity score 1:1 nearest neighbor matching showed the worst balance, as it was in the overall app usage scenario. In addition, 106 out of 518 cases showed a covariate balance from all the matching methods whereas 214 cases were balanced using only coarsened exact matching.

When estimating the treatment effect, we found causality in 30.7% (on average) of balanced cases from the four methods. Moreover, 16 out of 518 test cases were found to have a causal relationship with all the matching methods. They were mostly found in entertainment (9 cases) and social (6 cases) apps, as well as in 1 case for system apps. Note that more details about the causal inference results by scenario and participant are given in the Supplementary Material.

4.3.2 Group-level Analysis. We aggregate each participant's results to identify trends in the causal relationship between mobile app usage and physical activity. Results of covariate balancing and treatment effect estimation of each matching method are shown in Tables 5, 6, 7, and 8. We first present the number of participants achieving covariate balance for the given treatment variable. We then show the number of participants having causality for each case, together with the number of positive and negative causal results. Moreover, the treatment effect was estimated by aggregating the results of participants with causal relationships. Note that the percentage of pruned treated samples is also included in the case of coarsened exact matching, and the estimate used in this case was ATM, not ATT.

Although detailed outcomes vary among the matching methods, overall trends can be identified from the causal relationships. For instance, concerning overall app usage behavior, the launch count showed more negative causalities across the methods, whereas usage time showed the opposite trend. From this result, we may conclude that participants tended to sit shorter when launching the apps more frequently, but longer when using those apps longer.

This result differs from those of prior studies, which usually concluded that greater smartphone usage causes less physical activity. One may assume that this phenomenon is due to the characteristics of app usage behavior. When people use a mobile app for a long time without switching to another, we may imagine that they might be highly focused on interacting with the apps and therefore, staying in one place rather than moving around. In contrast, frequent opening and closing of apps without long use may imply that these behaviors do not require the user's concentration, allowing them to engage in physical activities such as walking.

For categorical app usage, most causalities were found in social and entertainment apps when conducting matching methods, except for coarsened exact matching. Interestingly, the causal relationship between launch count and usage time was reversed for social app usage but was in the same direction for entertainment app usage. We may interpret this result as follows: "*People tend to sit down for shorter durations when they use social apps more frequently and briefly, but when using entertainment apps, they tend to sit for longer durations regardless of the frequency and duration of the interactions.*" Similarly, the existence and direction of causality differ among categories, implying that mobile app usage behaviors in relation to physical activity may vary depending on the category.

We discovered more causal relationships from the more covariate-balanced cases with coarsened exact matching. During this process, we pruned the treated samples to obtain a balance, and the pruning ratio was approximately 19% on average; therefore, we could use most of the treated samples in the treatment effect estimation. This method also revealed a positive causal relationship between entertainment app usage (for both launch count and

Table 5. The results of propensity score 1:1 nearest neighbor matching from total test cases

			Covariate balance	Causality			Treatment effect (ATT)	
				Total	Positive	Negative	Mean	SD
Overall		Launch count	33	21	1	20	-0.093	0.054
		Usage time	7	3	3	0	0.110	0.032
Categorical	Social	Launch count	8	4	2	2	-0.032	0.104
		Usage time	9	5	4	1	0.053	0.082
	Entertainment	Launch count	32	7	7	0	0.122	0.042
		Usage time	23	7	7	0	0.147	0.068
	Information	Launch count	13	0	0	0	N/A	N/A
		Usage time	14	1	0	1	-0.149	N/A
	Work	Launch count	18	1	0	1	-0.137	N/A
		Usage time	15	2	1	1	-0.035	0.117
	System	Launch count	3	1	0	1	-0.136	N/A
		Usage time	8	1	0	1	-0.101	N/A
	Health	Launch count	3	0	0	0	N/A	N/A
		Usage time	3	0	0	0	N/A	N/A

Table 6. The results of propensity score optimal full matching from total test cases

			Covariate balance	Causality			Treatment effect (ATT)	
				Total	Positive	Negative	Mean	SD
Overall		Launch count	40	19	1	18	-0.103	0.056
		Usage time	26	8	7	1	0.098	0.084
Categorical	Social	Launch count	29	10	2	8	-0.070	0.083
		Usage time	16	9	8	1	0.090	0.109
	Entertainment	Launch count	30	7	7	0	0.117	0.044
		Usage time	26	8	8	0	0.145	0.077
	Information	Launch count	24	2	1	1	-0.023	0.169
		Usage time	24	4	0	4	-0.135	0.033
	Work	Launch count	23	2	2	0	0.140	0.008
		Usage time	19	1	1	0	0.088	N/A
	System	Launch count	11	4	1	3	-0.067	0.125
		Usage time	12	1	1	0	0.109	N/A
	Health	Launch count	5	0	0	0	N/A	N/A
		Usage time	5	0	0	0	N/A	N/A

usage time) and sedentary time, but negative causalities were found in all categories other than entertainment, showing that the participants sat down longer only when they interacted with the entertainment apps.

4.4 Case Study with an Additional Dataset

For the generalizability of the matching-based causal analysis pipeline, we conducted additional analysis by collecting another dataset. As there was a lack of open datasets similar to the one in Section 4.1, we decided to collect them using the data collection tool installed on smartphones [43]. Considering that the previous dataset was small and might be limited in sufficiently representing behavior and context, this additional data has been collected over a longer term.

4.4.1 Dataset and Method. We collected the same smartphone sensor data over 6 weeks in 2023 from 23 participants (8 women, average age 21.4 years) in a university. As illustrated in the previous section, the smartphone sensor data collected was composed of GPS, physical activity type, and app usage events to analyze the same causal scenarios. After collecting the data, we followed the proposed causal analysis pipeline to preprocess the data, balance the covariates, and estimate the treatment effect to examine the causal relationship. As in Section 4.3, we estimated the covariate balance and treatment effect of each causal scenario and summarized the results at the subject and group levels.

4.4.2 Evaluation. The treatment effect estimates from the four methods generally align with the original dataset's results in terms of the direction of the causal relationship. Among the 14 causal scenarios (i.e., 2 overall and 12 categorical app usage), we could observe consistent causal relationships from at least 9 cases for each matching method, with a maximum of 13 cases in the case of the coarsened exact matching achieving over 90% agreement. The results varied depending on the matching method, and the coarsened exact matching showed the most similar causality trends between the two datasets. Differences in participants, popular apps, and app interactions may lead to varied causal relationships between the datasets. However, common causality patterns were observed in most scenarios, indicating that those causal relationships could be generally discovered, not limited to a specific dataset.

In addition, more samples of the additional dataset improved the achievement of covariate balance to 91.5% (1,179 cases out of 1,288) when applying the four matching methods. Particularly, most cases from propensity score matching and Mahalanobis distance matching achieved covariate balance, which was impossible in the

Table 7. The results of Mahalanobis distance 1:1 nearest neighbor matching from total test cases

			Covariate balance	**Causality**			**Treatment effect (ATT)**	
				Total	**Positive**	**Negative**	**Mean**	**SD**
Overall		Launch count	43	30	1	29	-0.093	0.047
		Usage time	41	18	10	8	0.013	0.086
Categorical	Social	Launch count	17	6	1	5	-0.068	0.065
		Usage time	13	5	3	2	0.018	0.074
	Entertainment	Launch count	28	9	8	1	0.086	0.066
		Usage time	36	11	11	0	0.140	0.056
	Information	Launch count	19	1	0	1	-0.078	N/A
		Usage time	21	4	1	3	-0.021	0.167
	Work	Launch count	25	4	1	3	-0.056	0.138
		Usage time	24	1	1	0	0.119	N/A
	System	Launch count	9	3	0	3	-0.090	0.041
		Usage time	12	2	0	2	-0.072	0.043
	Health	Launch count	6	0	0	0	N/A	N/A
		Usage time	8	2	0	2	-0.246	0.279

Table 8. The results of coarsened exact matching from total test cases

			Covariate balance	% Pruned Treated	Causality			Treatment effect (ATM)	
					Total	Positive	Negative	Mean	SD
Overall		Launch count	49	2.4%	33	1	32	-0.097	0.048
		Usage time	49	10.1%	22	14	8	0.033	0.100
Categorical	Social	Launch count	49	24.6%	26	4	22	-0.076	0.068
		Usage time	49	24.2%	20	8	12	-0.028	0.092
	Entertainment	Launch count	48	10.5%	7	6	1	0.072	0.080
		Usage time	47	10.5%	11	11	0	0.141	0.062
	Information	Launch count	48	21.4%	15	0	15	-0.189	0.147
		Usage time	46	21.8%	10	0	10	-0.149	0.088
	Work	Launch count	49	18.0%	9	0	9	-0.113	0.043
		Usage time	46	22.7%	6	1	5	-0.073	0.117
	System	Launch count	49	30.9%	24	1	23	-0.088	0.053
		Usage time	49	29.6%	22	4	18	-0.058	0.082
	Health	Launch count	16	25.6%	5	0	5	-0.124	0.040
		Usage time	15	14.1%	2	0	2	-0.148	0.003

previous 7-day dataset. The larger data may increase the possibility of matching similar samples from control and treated groups, as there could be more samples having similar propensity scores or close enough in terms of Mahalanobis distance. More details of causal inference on the additional dataset by matching methods are given in the Supplementary Material.

5 A MACHINE LEARNING ALTERNATIVE FOR MATCHING AND TREATMENT EFFECT ESTIMATION

Recent advances in machine learning have led to the development of machine learning methods for treatment effect estimation as well [25]. Under appropriate assumptions (such as those listed in Section 2.3), these methods take advantage of the strong estimation power of various machine learning models in estimating the treatment effect size. There are several methods that effectively integrate matching within the machine learning framework to accurately estimate the treatment effect, by either conducting matching on a representation space learned through neural networks (e.g., [12, 63, 80]) or using matching to learn a balanced representation space suitable for effect estimation (e.g., [10]).

Among these methods, we briefly introduce one of the widely used machine-learning algorithms for treatment effect estimation, the causal forest [80], and outline how it can be applied to the causal analysis of mobile datasets. The main reasons why we chose the causal forest as our machine learning-based treatment effect estimation algorithm are two-fold. First, the causal forest is similar to matching algorithms in terms of its concept, in which the process of assigning samples to the same leaf node is analogous to matching samples based on the nearest neighbors. Moreover, it inherits the advantages of random forests [8], including the ability to model non-linear representation spaces while maintaining a higher degree of interpretability compared to neural network-based methods.

5.1 Causal Forests

The causal forest algorithm [80] is a treatment effect estimation algorithm based on random forests, designed to estimate the heterogeneous treatment effect in observational studies under the potential outcomes framework. In other words, the causal forest provides estimations for both the conditional average treatment effect (CATE) and the average treatment effect (ATE), which measure the individual-level effect size and group-level effect size, respectively. For simplicity, in consistency with the rest of the paper, we assume binary treatment assignments in our introduction of the causal forest.

Akin to the random forest, which is a collection or an average of multiple decision trees, the causal forest is a collection of causal trees. A causal tree is a tree-based regressor that is grown by recursively partitioning the feature space such that similar samples end up in the same leaf node. If the trees are grown well enough, the samples in the same leaf node will be roughly identically distributed. This enables us to estimate treatment effects as if they were from RCTs by comparing observed outcomes for samples in the same leaf node with different treatments.

5.2 Causal Analysis of Mobile Data Using Causal Forests

5.2.1 Method. As in the matching methods, we estimated the treatment effect of mobile app usage on physical activity. We utilized the R library **grf** [72] to generate causal forests and estimate the treatment effect. Also, we used the preprocessed dataset from Section 4 due to the same scenario setting and data preprocessing steps.

When employing the causal forest, all features except treatment and outcome were considered covariates, which were then used as partitioning criteria (i.e., internal nodes) when generating causal trees. Regarding the causal forest parameters, we configured the forest to consist of 2,000 trees, using 50% of the samples for growing trees and the other 50% for estimating the treatment effect at each leaf node [80]. Also, ATT was estimated to analyze the effect of mobile app usage on physical activity. For each causal scenario, we examined the propensity score of samples and excluded scenarios with extreme scores (near 0 or 1) that might indicate a potential violation of causal inference assumptions.

5.2.2 Evaluation. Overall, the results from the causal forest showed a similar trend to those from matching methods. When aggregating the estimated treatment effect of each participant by causal scenario, the direction of causality was consistent in most cases. For the overall app usage, the launch count had a negative causality with the physical activity whereas the usage time showed the opposite relationship. In the case of categorical app usage, the results from the causal forest followed the major results of the four matching methods.

We also investigated how many single cases (i.e., individual causal scenarios from each person) had the same direction of causality between each matching method and causal forest. Among the cases where ATT could be estimated, we found that around 84% of them revealed consistent causal relationships. From these findings, we could cross-validate the causal relationships inferred from the traditional matching methods, and the high consistency in results may imply that those causalities exist with higher confidence. Note that more details about the analysis results with the causal forest are described in the Supplementary Material.

6 PRACTICAL CONSIDERATIONS FOR CAUSAL INFERENCE USING SMARTPHONE SENSOR DATA

In this section, we review the causal analysis pipeline proposed in this tutorial and discuss practical considerations that researchers should take into account when utilizing this method for smartphone sensor data. The tutorial is composed of two main parts: (1) scenario setting and data preprocessing using smartphone sensor data, and (2) inferring the causal relationship from various scenarios by employing matching methods. We discuss several considerations for each part of the analysis pipeline. Also, we briefly introduce covariate balancing methods other than matching for causal inference.

6.1 Scenario Setting and Data Preprocessing

In the scenario setting and data preprocessing steps of the causal analysis pipeline, we covered topics such as which smartphone sensor data should be collected, how to extract features to represent human behavior and context, and how to set an appropriate size of time window while considering temporal precedence of events. Regarding these steps, there may be challenges in practice and we provide suggestions for dealing with them.

6.1.1 Considerations for Extracting Human Behavioral and Contextual Features. This study mainly dealt with mobile app usage, physical activity, and location, which can be easily tracked and extracted from smartphone sensor data. However, these data are limited in capturing human behaviors and contexts in the real world, such as emotions or activities other than mobile app usage. In that case, researchers may ask people to self-report those factors manually and use them for causal analysis along with automatically tracked data [27]. The experience sampling method (ESM) [49, 78] can be used to collect an individual's state when responding, utilizing the respondent as an additional sensor.

When we extract the features from smartphone sensor data, the granularity (i.e., the degree of detail) of representing human behavior and context should be determined. In this study, we set the granularity of app usage features at the app category level and used them as treatment and covariate variables. However, if the features were too granular (e.g., individual app level), the causal inference could be challenging as the samples of the treated group become sparse and the covariate dimension becomes high. This issue may be more critical if the features are granular compared to the number of samples available. On the other hand, if the features are not granulated (e.g., overall app level), the unique usage of specific apps would be diluted, thereby their causal relationships cannot be found. Therefore, researchers should determine the granularity of features at a level while considering the trade-off between these two opposite cases.

6.1.2 Considerations for Handling Temporal Precedence in Human Behaviors and Contexts. This tutorial infers causality in human behaviors using smartphone sensor data, in which the temporal properties of treatment and outcome may differ from previous studies. As in [67], treatment was typically given at a particular time or period, and the outcome was measured following the end of treatment provision, ensuring a clear temporal separation between the measurements of treatment and outcome. However, for human behaviors such as mobile app usage and physical activity, there are two main differences in the properties of the variables: *multitaskable treatment* and *micro-behavior* characteristics.

Multitaskable Treatment. In everyday life, mobile app usage and physical activity may not be clearly separable based on temporal order. Rather, they can be seen as multitasking events, where treatment and outcome behaviors have a very small temporal gap or occur almost concurrently. This kind of multitaskable treatment (i.e., the treatment that happens almost together with the outcome) can be easily discovered in human behaviors, such as walking while interacting with others through social apps. The temporal precedence between the two behaviors thus becomes less significant and can be relaxed in this tutorial's scenario, unlike prior studies in other domains.

Micro-behavior. The everyday event, such as mobile app usage, can be viewed as a micro-behavior, which occurs very briefly and frequently. It can be easily seen that people run mobile apps several times, even in a short time, to check new incoming messages or retrieve information. Because of this micro-behavioral nature, we aggregated multiple app usage events that occurred within a particular time window and collectively extracted features (i.e., launch count and usage time) rather than dealing with each event as a distinctive sample for the causal inference.

In this circumstance, we may suppose that there exists limited interaction (or dependency) between mobile app usage events that occur within different time windows. For example, it may not be reasonable to determine that one's current social app usage or walking behavior is directly affected by a few seconds of interaction with a social app that happened 15 min before. Rather, other events may influence current behaviors to a greater extent, such as external cues (e.g., incoming notifications) given immediately before app usage or events that are close to physical activity.

However, if a smaller time window is set to analyze the behavior at a more micro-level, one may need to consider the interactions between events from adjacent time windows. Particularly, if the treatment behavior has a long duration and cannot be seen as a micro-behavior (e.g., emotions that can be continued for several hours, as in [53]), behaviors in prior time windows should also be included in the analysis when balancing covariates. In addition, there may exist a seasonality in human behavior that repeats over time, such as staying at the workplace in the morning or resting at home at night. Therefore, although there was limited dependency between the time windows, we suggest including "the time of day," a macro-level context variable, in the covariate to balance such seasonality.

6.2 Covariate Balancing and Treatment Effect Estimation

After generating samples from the two previous steps, we described how to examine causalities in the covariate balancing and treatment effect estimation steps. They explained how to assign samples to treated and control groups, balance covariates from high-dimensional smartphone sensor data, apply different matching methods, and interpret the estimated treatment effect. In this section, we discuss several points to be considered and provide our suggestions.

6.2.1 Considerations for Conducting Causal Inference with Continuous Variables as Treatments. In this tutorial, many features representing human behavior and context are continuous variables, such as sedentary state duration, app launch counts, and time spent at places. This differs from causal inference in other domains where treatment is binary (i.e., treated or not treated). We mimicked this by assigning samples to the control or treated group based on the average treatment level. As Lu et al. [52] noted, for samples with continuous treatment variables, stratification into multiple groups by treatment levels and matching may help ensure treatment level differences between matched samples.

Still, the approach of splitting samples with a continuous treatment variable into two groups can be practical when considering the simplicity of the analysis process and result interpretation. In mobile app usage, usage distribution was positively skewed so that only samples with near zero usage were assigned to the control group, resulting in a reasonable separation in the treatment level between the two groups. However, when the treatment level distribution is concentrated around the mean, dividing based solely on the average may not yield significant differences between the control and treated groups in terms of treatment level. As shown in [75], this issue can be addressed by pruning samples with treatment levels close to their mean and making the treatment difference between the two groups more evident.

6.2.2 Considerations for Selecting and Balancing Covariates from Various Features. Previous studies [60] recommended that researchers should include as many covariates as possible to minimize bias in estimating the treatment effects. However, including all features extracted from multi-modal and high-dimensional smartphone sensor data (e.g., hundreds of features) may not be suitable because covariate balancing becomes more difficult to achieve.

Therefore, there are several practical strategies for selecting key covariate features and focusing on their balance to mitigate this issue. Basically, in many cases, there is existing domain knowledge that helps us narrow down features (e.g., by following mental health diagnosis guidelines [82]). Furthermore, researchers may conduct correlation tests on the features as illustrated in this tutorial and include them if they are significantly correlated with treatment and outcome. In addition, the double selection procedure proposed by Belloni et al. [5] can be utilized for reducing the dimension of covariates, where covariates are selected based on the regularization technique such as LASSO.

6.2.3 Considerations for Employing Multiple Methods for Balancing Covariates. In this tutorial, we demonstrated four different matching methods and found that the set of matched samples and achievement of covariate balance varied depending on the applied methods. Among the methods, we noticed that the coarsened exact matching was advantageous in balancing all covariates since it creates coarsened bins for each covariate separately and matches samples belonging to the same bin. The other methods such as propensity score matching or Mahalanobis distance matching explore the closest samples in terms of one transformed distance metric, which can succeed in balancing the joint distribution but may fail to balance each individual covarate [34].

This is more critical when the covariate distribution of control and treated groups barely overlaps. Since the coarsened exact matching matches the samples within the same bin, the matching mainly happens for the overlapped area and the covariate balance is achievable. On the contrary, the other methods will match the samples even if they are outside the overlapping region for the covariate, indicating that the covariate's distribution would remain separated and the balance quality may not be improved even after matching. Nevertheless, researchers should consider that the coarsened exact matching may discard many samples (especially, samples from the treated group) as it only takes the bins containing samples from both groups, implying that the remaining samples may not represent the original population well.

Furthermore, we suggest conducting diverse methods and reporting their results as the estimated treatment effect can also vary depending on the covariate balancing methods. Each balancing method may generate different matched samples having particular weights by matching criteria (e.g., distance metric, matching ratio, or sample pruning), resulting in a difference in the estimated treatment effect. Therefore, it is important to investigate whether similar results are observed from those methods for the robustness of the causal inference results. In reporting the estimated treatment effect, researchers should also report information about the balancing methods used and the samples included after balancing (e.g., the number of samples for each group and the improvements in covariate balance) [71].

6.2.4 Considerations for Interpreting and Utilizing the Causal Inference Results. When investigating causal relationships using smartphone sensor data, we should note between-individual variations. As shown earlier, the causal relationships can vary among individuals even in the same scenario. This variation is primarily due to individual differences in mobile app usage and daily routines. To examine whether a particular causal relationship is generalizable, researchers may further compare the causal results of people with similar characteristics (like a cohort analysis) and see if there are any common patterns of causalities. Also, it would be meaningful to cluster people having similar causal relationships and investigate whether they share common characteristics to figure out the reasons for the pattern.

One of the fundamental limitations of this study is that there is no ground truth for result validation. We can only assess result consistency through comparisons across various matching methods or different people. This limitation is inherent in any studies investigating causal relationships based on observational datasets [62, 75, 77]. To verify the causal relationships, we need to conduct controlled experiments where treatment can be randomly assigned (e.g., doing particular behaviors). Particularly, if the experiment is targeting a single subject (i.e., investigating causality from an individual), we may utilize methods such as mirco-randomized trials or self-experimentations as in [44, 47].

6.3 Alternative Methods for Causal Inference

In this study, we consider matching the primary method for causal inference because of its simplicity and applicability. Despite its widespread use, causal inference via matching relies on creating approximately balanced datasets, which may lead to inexact causal conclusions [36]. In matching, the estimation is made for the matched population, which may not be representative of the entire population. Furthermore, the matching estimators used to quantify the treatment effect size may be biased and require adjustments [1]. Therefore, in some cases, other methods might be better suited for causal inference, such as adjustment via weighting instead of matching, or using well-designed regression models [29].

Furthermore, researchers may choose recent machine learning methods for treatment effect estimation or use models that directly handle raw temporal data [25]. In particular, representation learning techniques could be leveraged to generate balanced samples of treated and control groups [42]. There are several ways such as counterfactual regression [64], local similarity preserved individual treatment effect estimation [85], and feature selection representation matching [10]. Yao et al. [84] introduces more alternative methods of covariate balancing in causal inference.

7 CONCLUSION

This tutorial introduces a causal inference pipeline using smartphone sensor data for human behavior understanding that is useful for digital health service design and evaluation. This pipeline was used to investigate the causality between mobile app usage and physical activity. The pipeline consists of (1) setting a scenario and corresponding variables, (2) preprocessing the data to extract features, (3) balancing the covariates to generate comparable (treated and control) groups, and (4) estimating the treatment effect to validate the causal relationship. For reliability reasons, we employed four well-known matching methods to evaluate the differences in covariate balancing results and the corresponding estimated treatment effect. Furthermore, we validated our findings by (1) applying the same analysis pipeline to an additional dataset and (2) employing another method based on the machine learning technique.

In this tutorial, we focused on a few primary features when inferring causal relationships between mobile app usage and physical activity to simplify the analysis process. However, those relationships could be thoroughly analyzed by including additional, complex human behavior and contexts such as social settings, emotional states, or activities other than mobile app usage. Also, we may include data from other sources such as wearables, smart speakers, and IoT devices to capture diverse behavioral features. Moreover, the proposed analysis pipeline could be further improved by referring to the practical considerations we offered. Along with the growing interest in digital health through smart devices, we expect this tutorial to support researchers in mobile computing in investigating causality using the data in practice.

ACKNOWLEDGMENTS

This research was supported by the National Research Foundation (NRF) funded by the Korean government (MSIT) (2021M3A9E4080780, 2022M3J6A1063021), and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1A6A3A13072027).

REFERENCES

[1] Alberto Abadie and Guido W Imbens. 2006. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica* 74, 1 (2006), 235–267.

[2] Android Developers. 2023. UsageEvents.Event. https://developer.android.com/reference/android/app/usage/UsageEvents.Event (accessed: 2024-02-01).

[3] Jacob E Barkley and Andrew Lepp. 2016. Mobile Phone Use Among College Students Is a Sedentary Leisure Behavior Which May Interfere With Exercise. *Computers in Human Behavior* 56 (2016), 29–33.

[4] David H Barlow and Matthew K Nock. 2009. Why Can't We Be More Idiographic in Our Research? *Perspectives on Psychological Science* 4, 1 (2009), 19–21.

[5] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. 2014. High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives* 28, 2 (2014), 29–50.

[6] Adriene M Beltz, Aidan GC Wright, Briana N Sprague, and Peter CM Molenaar. 2016. Bridging the Nomothetic and Idiographic Approaches to the Analysis of Clinical Data. *Assessment* 23, 4 (2016), 447–458.

[7] Nick Black. 1996. Why We Need Observational Studies to Evaluate the Effectiveness of Health Care. *Bmj* 312, 7040 (1996), 1215–1218.

[8] Leo Breiman. 2001. Random Forests. *Machine Learning* 45 (2001), 5–32.

[9] Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. 2022. Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications. *Information Fusion* 81 (2022), 59–83.

[10] Zhixuan Chu, Stephen L Rathbun, and Sheng Li. 2020. Matching in Selective and Balanced Representation Space for Treatment Effects Estimation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 205–214.

[11] Seungeun Chung, Chi Yoon Jeong, Jeong Mook Lim, Jiyoun Lim, Kyoung Ju Noh, Gague Kim, and Hyuntae Jeong. 2022. Real-World Multimodal Lifelog Dataset for Human Behavior Study. *ETRI Journal* 44, 3 (2022), 426–437.

[12] Oscar Clivio, Fabian Falck, Brieuc Lehmann, George Deligiannidis, and Chris Holmes. 2022. Neural Score Matching for High-Dimensional Causal Inference. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 151)*. PMLR, New York, New York, USA, 7076–7110.

[13] Marios Constantinides, Jonas Busk, Aleksandar Matic, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E Bardram. 2018. Personalized Versus Generic Mood Prediction Models in Bipolar Disorder. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. Association for Computing Machinery, New York, NY, USA, 1700–1707.

[14] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. 2000. The Mahalanobis Distance. *Chemometrics and Intelligent Laboratory Systems* 50, 1 (2000), 1–18.

[15] René A de Wijk and Lucas PJJ Noldus. 2021. Implicit and Explicit Measures of Food Emotions. In *Emotion Measurement (Second Edition)*. Woodhead Publishing, Cambridge, United Kingdom, 169–196.

[16] Angus Deaton and Nancy Cartwright. 2018. Understanding and Misunderstanding Randomized Controlled Trials. *Social Science & Medicine* 210 (2018), 2–21.

[17] Christian Fong, Chad Hazlett, and Kosuke Imai. 2018. Covariate Balancing Propensity Score for a Continuous Treatment: Application to the Efficacy of Political Advertisements. *The Annals of Applied Statistics* 12, 1 (2018), 156–177.

[18] Thomas R Frieden. 2017. Evidence for Health Decision Making—Beyond Randomized, Controlled Trials. *New England Journal of Medicine* 377, 5 (2017), 465–475.

[19] Tasha Glenn and Scott Monteith. 2014. New Measures of Mental State and Behavior Based on Data Collected From Sensors, Smartphones, and the Internet. *Current Psychiatry Reports* 16, 12 (2014), 1–10.

[20] Cory E Goldstein, Charles Weijer, Jamie C Brehaut, Dean A Fergusson, Jeremy M Grimshaw, Austin R Horn, and Monica Taljaard. 2018. Ethical Issues in Pragmatic Randomized Controlled Trials: A Review of the Recent Literature Identifies Gaps in Ethical Argumentation. *BMC Medical Ethics* 19, 1 (2018), 1–10.

[21] Robin Gomila. 2021. Logistic or Linear? Estimating Causal Effects of Experimental Treatments on Binary Outcomes Using Regression Analysis. *Journal of Experimental Psychology: General* 150, 4 (2021), 700.

[22] Noah Greifer. 2024. Covariate Balance Tables and Plots: A Guide to the cobalt Package. https://cran.r-project.org/web/packages/cobalt/vignettes/cobalt.html (accessed: 2024-02-01).

[23] Noah Greifer and Elizabeth A. Stuart. 2023. Choosing the Causal Estimand for Propensity Score Analysis of Observational Studies. arXiv:2106.10577 [stat.ME]

[24] Xing Sam Gu and Paul R Rosenbaum. 1993. Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. *Journal of Computational and Graphical Statistics* 2, 4 (1993), 405–420.

[25] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. 2020. A Survey of Learning Causality With Data: Problems and Methods. *ACM Computing Surveys (CSUR)* 53, 4 (2020), 1–37.

[26] Ben B Hansen. 2004. Full Matching in an Observational Study of Coaching for the SAT. *J. Amer. Statist. Assoc.* 99, 467 (2004), 609–618.

[27] Gabriella M Harari, Nicholas D Lane, Rui Wang, Benjamin S Crosier, Andrew T Campbell, and Samuel D Gosling. 2016. Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges. *Perspectives on Psychological Science* 11, 6 (2016), 838–854.

[28] Severin Haug, Raquel Paz Castro, Min Kwon, Andreas Filler, Tobias Kowatsch, and Michael P Schaub. 2015. Smartphone Use and Smartphone Addiction Among Young People in Switzerland. *Journal of Behavioral Addictions* 4, 4 (2015), 299–307.

[29] Miguel A. Hernán and James M. Robins. 2020. *Causal Inference: What If.* Chapman & Hall/CRC, Boca Raton.

[30] Daniel Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2011. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* 42, 8 (2011), 1–28.

[31] Paul W Holland. 1986. Statistics and Causal Inference. *J. Amer. Statist. Assoc.* 81, 396 (1986), 945–960.

[32] Kit Huckvale, Svetha Venkatesh, and Helen Christensen. 2019. Toward Clinical Digital Phenotyping: A Timely Opportunity to Consider Purpose, Quality, and Safety. *NPJ digital medicine* 2, 1 (2019), 1–11.

[33] Stefano Iacus, Gary King, and Giuseppe Porro. 2009. cem: Software for Coarsened Exact Matching. *Journal of Statistical Software* 30, 9 (2009), 1–27.

[34] Stefano M Iacus, Gary King, and Giuseppe Porro. 2011. Multivariate Matching Methods That Are Monotonic Imbalance Bounding. *J. Amer. Statist. Assoc.* 106, 493 (2011), 345–361.

[35] Stefano M Iacus, Gary King, and Giuseppe Porro. 2012. Causal Inference Without Balance Checking: Coarsened Exact Matching. *Political Analysis* 20, 1 (2012), 1–24.

[36] Stefano M Iacus, Gary King, and Giuseppe Porro. 2019. A Theory of Statistical Inference for Matching Methods in Causal Research. *Political Analysis* 27, 1 (2019), 46–68.

[37] Kosuke Imai and David A Van Dyk. 2004. Causal Inference With General Treatment Regimes: Generalizing the Propensity Score. *J. Amer. Statist. Assoc.* 99, 467 (2004), 854–866.

[38] Guido W. Imbens. 2004. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics* 86, 1 (2004), 4–29.

[39] Guido W Imbens and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge University Press, Cambridge, United Kingdom.

[40] OT Inan, P Tenaerts, SA Prindiville, HR Reynolds, DS Dizon, K Cooper-Arnold, M Turakhia, MJ Pletcher, KL Preston, HM Krumholz, BM Marlin, KD Mandl, P Klasnja, B Spring, E Iturriaga, R Campo, P Desvigne-Nickens, Y Rosenberg, SR Steinhubl, and RM Califf. 2020. Digitizing Clinical Trials. *NPJ digital medicine* 3, 1 (2020), 1–7.

[41] Thomas R Insel. 2017. Digital Phenotyping: Technology for a New Science of Behavior. *JAMA* 318, 13 (2017), 1215–1216.

[42] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning Representations for Counterfactual Inference. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48).* PMLR, New York, New York, USA, 3020–3029.

[43] Soowon Kang, Woohyeok Choi, Cheul Young Park, Narae Cha, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Heepyung Kim, Yong Jeong, and Uichin Lee. 2023. K-EmoPhone: A Mobile and Wearable Dataset with In-Situ Emotion, Stress, and Attention Labels. *Scientific Data* 10, 1 (2023), 351.

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

[45] Erica L Kenney and Steven L Gortmaker. 2017. United States Adolescents' Television, Computer, Videogame, Smartphone, and Tablet Use: Associations With Sugary Drinks, Sleep, Physical Activity, and Obesity. *The Journal of Pediatrics* 182 (2017), 144–149.

[46] Sung-Eun Kim, Jin-Woo Kim, and Yong-Seok Jee. 2015. Relationship Between Smartphone Addiction and Physical Activity in Chinese International Students in Korea. *Journal of Behavioral Addictions* 4, 3 (2015), 200–205.

[47] Predrag Klasnja, Shawna Smith, Nicholas J Seewald, Andy Lee, Kelly Hall, Brook Luers, Eric B Hekler, and Susan A Murphy. 2019. Efficacy of Contextually Tailored Suggestions for Physical Activity: A Micro-Randomized Optimization Trial of HeartSteps. *Annals of Behavioral Medicine* 53, 6 (2019), 573–582.

[48] Lampros C Kourtis, Oliver B Regele, Justin M Wright, and Graham B Jones. 2019. Digital Biomarkers for Alzheimer's Disease: the Mobile/Wearable Devices Opportunity. *NPJ digital medicine* 2, 1 (2019), 1–9.

[49] Reed Larson and Mihaly Csikszentmihalyi. 2014. *The Experience Sampling Method.* Springer Netherlands, Dordrecht, 21–34.

[50] Finbarr P Leacy and Elizabeth A Stuart. 2014. On the Joint Use of Propensity and Prognostic Scores in Estimation of the Average Treatment Effect on the Treated: A Simulation Study. *Statistics in Medicine* 33, 20 (2014), 3488–3508.

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

[52] Bo Lu, Robert Greevy, Xinyi Xu, and Cole Beck. 2011. Optimal Nonbipartite Matching and Its Statistical Applications. *The American Statistician* 65, 1 (2011), 21–30.

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

[54] David C Mohr, Mi Zhang, and Stephen M Schueller. 2017. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual review of clinical psychology* 13 (2017), 23.

[55] Paul R Rosenbaum. 1991. A Characterization of Optimal Designs for Observational Studies. *Journal of the Royal Statistical Society: Series B (Methodological)* 53, 3 (1991), 597–610.

[56] Paul R Rosenbaum. 2005. *Observational Study.* John Wiley & Sons, Ltd, Chichester, West Sussex, United Kingdom, Chapter 3, 1451–1462.

[57] Paul R Rosenbaum and Donald B Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70, 1 (1983), 41–55.

[58] Donald B Rubin. 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66, 5 (1974), 688.

[59] Donald B Rubin. 2005. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

[60] Donald B Rubin and Neal Thomas. 1996. Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics* 52, 1 (1996), 249–264.

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

[62] Zhanna Sarsenbayeva, Gabriele Marini, Niels van Berkel, Chu Luo, Weiwei Jiang, Kangning Yang, Greg Wadley, Tilman Dingler, Vassilis Kostakos, and Jorge Goncalves. 2020. Does Smartphone Use Drive Our Emotions or Vice Versa? A Causal Analysis. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–15.

[63] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. 2019. Perfect Match: A Simple Method for Learning Representations for Counterfactual Inference With Neural Networks. arXiv:1810.00656 [cs.LG]

[64] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating Individual Treatment Effect: Generalization Bounds and Algorithms. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*. PMLR, New York, New York, USA, 3076–3085.

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

[66] Neza Stiglic and Russell M Viner. 2019. Effects of Screentime on the Health and Well-Being of Children and Adolescents: A Systematic Review of Reviews. *BMJ open* 9, 1 (2019), e023191.

[67] Elizabeth A Stuart. 2010. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 25, 1 (2010), 1.

[68] Elizabeth A Stuart and Kerry M Green. 2008. Using Full Matching to Estimate Causal Effects in Nonexperimental Studies: Examining the Relationship Between Adolescent Marijuana Use and Adult Outcomes. *Developmental Psychology* 44, 2 (2008), 395.

[69] Elizabeth A Stuart, Brian K Lee, and Finbarr P Leacy. 2013. Prognostic Score–Based Balance Measures Can Be a Useful Diagnostic for Propensity Score Methods in Comparative Effectiveness Research. *Journal of Clinical Epidemiology* 66, 8 (2013), S84–S90.

[70] Thomas Stütz, Thomas Kowar, Michael Kager, Martin Tiefengrabner, Markus Stuppner, Jens Blechert, Frank H Wilhelm, and Simon Ginzinger. 2015. Smartphone Based Stress Prediction. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer International Publishing, Cham, 240–251.

[71] Felix J Thoemmes and Eun Sook Kim. 2011. A Systematic Review of Propensity Score Methods in the Social Sciences. *Multivariate Behavioral Research* 46, 1 (2011), 90–118.

[72] Julie Tibshirani, Susan Athey, Rina Friedberg, Vitor Hadad, David Hirshberg, Luke Miner, Erik Sverdrup, Stefan Wager, and Marvin Wright. 2023. grf: Generalized Random Forests. https://cran.r-project.org/web/packages/grf/index.html (accessed: 2024-02-01).

[73] John Torous, Mathew V Kiang, Jeanette Lorme, and Jukka-Pekka Onnela. 2016. New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *JMIR Mental Health* 3, 2 (2016), e16.

[74] Alina Trifan, Maryse Oliveira, and José Luís Oliveira. 2019. Passive Sensing of Health Outcomes Through Smartphones: Systematic Review of Current Solutions and Possible Limitations. *JMIR Mhealth Uhealth* 7, 8 (2019), e12649.

[75] Fani Tsapeli and Mirco Musolesi. 2015. Investigating Causality in Human Behavior From Smartphone Sensor Data: A Quasi-Experimental Approach. *EPJ Data Science* 4, 1 (2015), 24.

[76] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. 2017. Recognizing Detailed Human Context in the Wild From Smartphones and Smartwatches. *IEEE Pervasive Computing* 16, 4 (2017), 62–74.

[77] Niels van Berkel, Simon Dennis, Michael Zyphur, Jinjing Li, Andrew Heathcote, and Vassilis Kostakos. 2021. Modeling Interaction as a Complex System. *Human–Computer Interaction* 36, 4 (2021), 279–305.

[78] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The Experience Sampling Method on Mobile Devices. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 1–40.

[79] Tyler J VanderWeele. 2019. Principles of Confounder Selection. *European Journal of Epidemiology* 34, 3 (2019), 211–219.

[80] Stefan Wager and Susan Athey. 2018. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.

[81] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery, New York, NY, USA, 3–14.

[82] Rui Wang, Weichen Wang, Alex DaSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–26.

[83] Janine Witte and Vanessa Didelez. 2019. Covariate Selection Strategies for Causal Inference: Classification and Comparison. *Biometrical Journal* 61, 5 (2019), 1270–1289.

[84] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A Survey on Causal Inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 5 (2021), 1–46.

[85] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2018. Representation Learning for Treatment Effect Estimation from Observational Data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Curran

Associates, Inc., Red Hook, NY, USA, 2638–2648.

[86] Yang Ye, Yu Zheng, Yukun Chen, Jianhua Feng, and Xing Xie. 2009. Mining Individual Life Pattern Based on Location History. In *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*. IEEE, IEEE Computer Society, USA, 1–10.

[87] Fengpeng Yuan, Xianyi Gao, and Janne Lindqvist. 2017. How Busy Are You? Predicting the Interruptibility Intensity of Mobile Users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 5346–5360.

[88] Qin-Yu Zhao, Jing-Chao Luo, Ying Su, Yi-Jie Zhang, Guo-Wei Tu, and Zhe Luo. 2021. Propensity Score Matching With R: Conventional Methods and New Features. *Annals of Translational Medicine* 9, 9 (2021).

[89] Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, and Loren Terveen. 2007. Discovering Personally Meaningful Places: An Interactive Clustering Approach. *ACM Transactions on Information Systems (TOIS)* 25, 3 (2007), 12–es.