What possibly affects nighttime heart rate? Conclusions from N-of-1 observational data

Almeida Matias, Igor Alexandre; Daza, Eric J.; Wac, Katarzyna

How to cite

ALMEIDA MATIAS, Igor Alexandre, DAZA, Eric J., WAC, Katarzyna. What possibly affects nighttime heart rate? Conclusions from N-of-1 observational data. In: Digital health, 2022, vol. 8, p. 53 p. doi: 10.1177/20552076221120725

This publication URL: https://archive-ouverte.unige.ch/unige:163106

Publication DOI: 10.1177/20552076221120725

DIGITAL HEALTH

What Possibly Affects Nighttime Heart Rate? Conclusions from N-of-1 Observational Data

Journal:	*Digital Health*
Manuscript ID	DHJ-22-0330.R2
Manuscript Type:	Original Research
Date Submitted by the Author:	n/a
Complete List of Authors:	Matias, Igor; University of Geneva, Quality of Life Technologies Lab Daza, Eric; Evidation Health Wac, Katarzyna; University of Geneva, Quality of Life Technologies Lab
Keywords:	auto experimentation, causal inference, endogeneity, longitudinal, nighttime heart rate, n-of-1 trial, resting heart rate, self-reporting; stress, Wearables < Personalised medicine
Abstract:	Background: Heart rate (HR), especially at nighttime, is an important biomarker for cardiovascular health. It is known to be influenced by overall physical fitness, as well as daily life physical or psychological stressors like exercise, insufficient sleep, excess alcohol, certain foods, socialization, or air travel causing physiological arousal of the body. However, the exact mechanisms by which these stressors affect the nighttime HR are unclear and may be highly idiographic (i.e., individual-specific). A single-case or "n-of-1" observational study (N1OS) is useful in exploring such suggested effects by examining each subject's exposure to both stressors and baseline conditions, thereby characterizing suggested effects specific to that individual. Objective: Our objective was to test and generate individual-specific N1OS hypotheses of the suggested effects of daily life stressors on nighttime HR. As an N1OS, this study provides conclusions for each participant, thus not requiring a representative population. Methods: We studied three healthy, nonathlete individuals, collecting the data for up to four years. Additionally, we evaluated model-twin randomization (MoTR), a novel Monte Carlo method facilitating the discovery of personalized interventions on stressors in daily life. Results: We found that physical activity can increase the nighttime heart rate amplitude, whereas there were no strong conclusions about its suggested effect on total sleep time. Self-reported states such as exercise, yoga, and stress were associated with increased (for the first two) and decreased (last one) average nighttime heart rate. Conclusions: This study implemented the MoTR method evaluating the suggested effects of daily stressors on nighttime heart rate, sleep time, and physical activity in an individualized way: via the N-of-1 approach. A Python implementation of MoTR is freely available.

Original Research

What Possibly Affects Nighttime Heart Rate? Conclusions from N-of-1 Observational Data

Igor Matias[1,*], Eric J. Daza[2], Katarzyna Wac[1]

[1]Quality of Life Technologies Lab, University of Geneva, 1205 Geneva, Switzerland
[2]Evidation Health, San Mateo, CA. 94401, United States of America

[*]Corresponding Author:
Igor Matias, Quality of Life Technologies Lab, University of Geneva
Route de Drize 7, Battelle A
CH-1227 Carouge
Switzerland

Email: igor.matias@unige.ch
Phone: +41 223 790 234
Twitter: IgorMatias3

Abstract

Background: *Heart rate* (HR), especially at nighttime, is an important biomarker for cardiovascular health. It is known to be influenced by overall physical fitness, as well as daily life physical or psychological stressors like exercise, insufficient sleep, excess alcohol, certain foods, socialization, or air travel causing physiological arousal of the body. However, the exact mechanisms by which these stressors affect the nighttime HR are unclear and may be highly idiographic (i.e., individual-specific). A single-case or *"n-of-1" observational study* (N1OS) is useful in exploring such suggested effects by examining each subject's exposure to both stressors and baseline conditions, thereby characterizing suggested effects specific to that individual.

Objective: Our objective was to test and generate individual-specific N1OS hypotheses of the suggested effects of daily life stressors on nighttime HR. As an N1OS, this study provides conclusions for each participant, thus not requiring a representative population.

Methods: We studied three healthy, nonathlete individuals, collecting the data for up to four years. Additionally, we evaluated *model-twin randomization* (MoTR), a novel Monte Carlo method facilitating the discovery of personalized interventions on stressors in daily life.

Results: We found that physical activity can increase the nighttime heart rate amplitude, whereas there were no strong conclusions about its suggested effect on total sleep time. Self-reported states such as exercise, yoga, and stress were associated with increased (for the first ~~three~~two) and decreased (last one) average nighttime heart rate.

Conclusions: This study implemented the MoTR method evaluating the suggested effects of daily stressors on nighttime heart rate, sleep time, and physical activity in an individualized way: via the N-of-1 approach. A Python implementation of MoTR is freely available.

Keywords: auto experimentation; causal inference; endogeneity; longitudinal; nighttime heart rate; n-of-1 trial; resting heart rate; self-reporting; stress; wearables.

INTRODUCTION

Background

The emergence and ubiquitous availability of personal miniaturized technologies, including self-tracking mobile and wearable devices, enable continuous, longitudinal data collection and facilitate "self-knowledge through numbers," fulfilling the vision put forward by the "quantified-self" founders ((1,2)). Motivated individuals leverage these technologies, as well as self-reporting tools to track their behaviors, including those related to physical activity, sleep, alcohol consumption, foods, presence of psychological stress, air travels, or more ((3)). Additionally, these technologies enable capturing certain physiological signals like body *temperature* (temp),

respiration rate (RR), *heart rate* (HR), *heart rate variability* (HRV), or *galvanic skin response* (GSR) corresponding to the physical or psychological state of the individual ((3,4)).

Individuals can track a single behavior at its simplest, and use their self-tracking data for self-experimentation, changing it in the desired direction, like walking more steps or sleeping enough. However, these technologies can also enable more complex interventions and, if paired with disciplined scientific approaches to data analysis, they can provide more robust personalized insights ((5,6)). They are also able to help detect or even predict health issues by the mean of more advanced measurements like an *electrocardiogram* (ECG) ((7)). When combining wearable ECG signals with artificial intelligence algorithms, illness prediction is possible ((8)), transforming these ubiquitous and accessible devices into a powerful source of self-information.

This study employs an *n-of-1 observational study* (N1OS) design and integrates data from two different technological touchpoints: a consumer-grade behavior and physiology tracking device; and an electronic self-reporting tool. We use the data to characterize nonathlete individuals and test our main hypotheses on the correlation of daily stressors with *nighttime HR*, an important health concern in the context of cardiovascular health. The nighttime HR is specifically defined as a nighttime resting heart rate when the body returns to a baseline, and no daily-life stressors are present ((9)). We will sometimes use the term "correlation" interchangeably with the broader and more statistically accurate term "association" for ease of understanding. However, note that the statistical definition of "correlation" is narrower than is commonly meant; i.e., a non-linear statistical association or dependence is not a statistical correlation.

Additionally, we evaluate the analytic impact of *model-twin randomization* (MoTR) ((10)) on our inferences and conclusions. MoTR ("motor") is a new causal inference method that artificially emulates an n-of-1 randomized trial (i.e., the gold standard due to randomization) from the N1OS dataset. It does so by first modeling the outcome of interest as a function of the exposure of interest, along with an individual's assumed recurring confounders (i.e., daily observed variables thought to influence or affect both the exposure and the outcome). MoTR then randomly shuffles (i.e., permutes) the exposures, which were originally only observed, thereby simulating an n-of-1 randomized trial. This allows us to infer more accurately a suggested effect of daily stressors beyond just correlation.

Note that this study is not a case report, an observational study of a single participant. Unlike a case report, which has limited internal validity ((11)), our study uses MoTR to improve the veracity of findings of possible causal effects. In this way, an N1OS enables the discovery of findings for a given individual that is hard to achieve with standard group-based observational study designs ((12)) — and MoTR adjusts these findings to suggest possible interventions. These causal inference methods also facilitate subsequent design and testing of the suggested effects in an n-of-1 randomized trial of these discovered effects.

The operational objective of this paper is to establish the feasibility of the N1OS design augmented with MoTR for generating and evaluating hypotheses about the idiographic (i.e., individual-specific) recurring average effect of an exposure (e.g.,

daily stressors) on the self-tracked outcome of each participant (e.g., nighttime HR). The analogous nomothetic (i.e., group-level) effects in randomized controlled trials (RCTs) are called "*average causal effects*" or "*average treatment effects*" (ATEs).
We chose daily life stressors like physical activity, insufficient sleep, excess of alcohol consumption, certain foods, presence of psychological stress, air travels as the exposure variables because they have a profound and acute effect on several aspects of health in a short, as well as long-term, especially when repeated and are behaviors that may be commonly tracked on current consumer devices or via a minimum self-reporting efforts. The nighttime HR is our selected health biomarker because it is affected by daily stressors in nonathletes and is also an important outcome measure of cardiovascular health ((13–16)).
This hypothesis exploration is based on the relevant literature on the importance of managing daily life stressors for short-term and long-term health outcomes. The intentional choice of nonathletic individuals was with an eye for preventing a chronic disease involving the cardiovascular system. The additional objective was to evaluate the MoTR method for generating and testing such idiographic hypotheses, potentially facilitating personalized management of stressors in daily life. As a result, we demonstrate an observational study design and analysis plan to contribute to and help guide rigorous self-tracking and n-of-1 study designs.

N-of-1 Study Designs: Experimental and Observational

An n-of-1 study, also known as a single-subject or single-case study, is a scientific study focused on a single individual. Such studies are used to better understand the individual-specific effect of an intervention or association of exposure with an outcome, e.g., how behavioral changes in a specific person causally affect or associate with daily-life stressors and nighttime HR the following night. There are two types of n-of-1 studies.
An n-of-1 randomized trial (N1RT) is a randomized crossover design in which a participant acts as their baseline (i.e., control), and is randomly assigned to an active treatment over multiple treatment periods. For example, a participant may be randomized to the sequence of treatment period denoted ABBA, where A and B represent an active and baseline treatment, respectively ((12,17)). The N1RT design requires experimentation (i.e., comparing the intervention results with a baseline), and has been increasingly used in clinical trials and biomedical research. A related study design, the *single-case experimental design* (SCED), has been extensively applied in psychology and education ((5,18–21)).
The target quantity that can be inferred using an N1RT is a recurring average effect that (22) calls an "*average period treatment effect*" (APTE). In an N1RT, a period is defined as a recurring time interval during which a treatment or intervention is randomly assigned. Treatment levels are not required to change every period, but they must be randomized. For example, for intervention levels A and B each randomized twice with equal probability, the sequences ABBA, AABB, and BABA all have equal probability. The APTE is the n-of-1 analog to the population ATE of RCTs (defined above).
An *n-of-1 observational study* (N1OS) is a scientific study design involving a single individual without any structured randomization, akin to ecological or

epidemiological studies. While an N1RT generally carries more internal validity (i.e., cause-effect relationship) than an N1OS, the latter generally carries more external or ecological validity (i.e., generalizability to real-world situations) than an N1RT. The would-be "intervention" and "baseline" are derived from real-world data, without randomization of the individual to the condition ((23)).

In an N1OS, a period is defined as a recurring time interval during which an exposure is observed (e.g., a day after sleeping for a certain amount of time). As with N1RT treatments, exposure levels are not required to change every period; however, they are generally not randomized, being observed as they naturally occur. "Exposure" is the epidemiological term for the would-be randomized treatment in a corresponding N1RT design; i.e., we wish to infer a reasonable or plausible causal effect (i.e., APTE) of the exposure on the outcome that we would otherwise more definitively infer by randomizing the exposure.

N1OS designs offer several opportunities for health psychology and behavioral medicine; they can be used to describe changes in naturally occurring phenomena (e.g., behaviors) over time. They can also enable testing the hypotheses related to the relationships between variables, such as those specified in behavioral theories ((24–27)). Furthermore, N1OS can be used to design highly personalized, data-driven interventions based on the unique predictive relationships identified at the individual level ((28)).

Note that while causal inference methods for observational data like MoTR can actually help estimate possible treatment effects (which prediction/correlation methods generally cannot), none of these methods — not even MoTR — can definitively estimate an average causal effect. This is because estimating an APTE in an N1OS requires measuring all (or at least the strongest) treatment-effect confounders, and also correctly modeling the relationship between the exposure, outcome, and all confounders. These two strong assumptions highlight that "there is no free lunch" in trying to infer a causal effect without randomization. To convincingly estimate an APTE, the exposure must be manipulated or randomized; otherwise, these two crucial assumptions must hold — assumptions that generally cannot be tested simply by updating or trying new models using the same dataset (i.e., without manipulating the exposure directly to produce a new dataset).

At best, we can and will assume that we have observed the strongest (not necessarily all) confounders, and that our chosen model is reasonably correct (i.e., correctly approximates the "true" causal model). This relaxed assumption allows for bias, but is much more realistic. With our observational data, the most we can do is hope that our set of confounders is complete enough, and that our models are correct enough, to keep the true, unknown bias in estimating the APTE small.

Relationship to Longitudinal Studies

N1OS are related to longitudinal studies that use common statistical approaches like mixed- or random-effects modeling or generalized estimating equations. However, these two study designs differ fundamentally in their analytic goals concerning levels of inference.

In a longitudinal study, the analytic goal is to infer the average trend overtime over a group of participants, i.e., it is a nomothetic goal. However, repeated measurements

for each study participant may induce within-individual autocorrelation that reduces the overall information on the group-level trend. This may increase the variance of the trend estimator, which is a "nuisance" to reaching the goal; hence the common need to deal with statistical nuisance parameters" in longitudinal studies. In an N1OS, the analysis goal is to infer a recurring average association over a set of repeated measurements within one participant; i.e., it is an idiographic goal. A group-level association may be useful as a starting point to help specify individual-level a priori hypotheses or as a starting value in some iterative analytic approaches. However, the entire approach of conducting an N1OS assumes that the within-individual average meaningfully differs from the group average — even, perhaps, for a large group of similar individuals.

Hence, N1OS a priori hypotheses ideally rely on the participant's own experiences, opinions and beliefs, and their past self-tracked data if available. That is its core principle. If group-based findings in the scientific literature are deemed useful for structuring the idiographic a priori hypotheses, these can and should also be used. However, an N1OS by design privileges the participant's own prior beliefs above any group-based findings — some of which may inform those prior beliefs. The process of each participant designing their own a priori hypotheses resembles prior elicitation in Bayesian modeling, but where the study participant is the "domain expert" of their N1OS.

Daily-Life Stressors: Exercise, Sleep, Alcohol and Food, Psychological Stress, Aircraft Travels, and Nighttime HR

The growing body of research indicates the importance of HR (including nighttime HR) as a prognostic factor and potential therapeutic target in populations at large ((29)). The resting HR shows a clear circadian rhythm, being substantially higher during waking hours, but the variations are relatively small, between 10±6 beats/min ((30)). Additionally, HR also changes with posture, being some 3 beats/min higher in the sitting compared with the supine position ((31)).

In this work, the nighttime HR considered is specifically defined as HR while sleeping, when there are no daily life stressors ((9)). Research shows that although it may be difficult to define an optimal HR for a given individual, it seems desirable to maintain low nighttime HR ((29)). High nighttime HR has direct detrimental effects on the progression of *cardiovascular diseases* (CVD) ((32)). Studies have specifically found a continuous increase in the risk of CVD with nighttime HR above 60 beats/min ((33–35)), which is very important, especially given the increasing prevalence of CVDs leading to premature deaths ((36,37)). When considering a desirable or optimal HR for an individual, demographic and measurement factors must also be considered. Namely, HR has been reported to decrease with age ((38)), although this has not been seen in all studies, and HR is higher in women than in men ((39)).

The nighttime HR is influenced by physical stressors experienced in the preceding day, like exercise or mental stress. On the one hand, research shows that higher overall activity level and athletic capacity leads to lower heart rate ((40,41)). A systematic meta-review by Reimers et al. shows that especially endurance training, yoga, and strength training conducted at least 2 times a week for at least 4 weeks

have shown decreases in heart rate ((42)). However, with acute exercise exertion like multiple hours of running or biking, the resulting nighttime HR is shown to be higher ((13,43)). Additionally, a systematic meta-review by Kredlow et al. shows that the overall activity levels and acute exercise have small beneficial effects on sleep duration ((44)). Additionally, nighttime HR is influenced by insufficient and variable time of sleep ((14)), excess alcohol, certain foods (e.g., greasy) ((15)), presence of psychological stress ((15,45–47)), or air travels decreasing oxygen saturation in the blood ((16)).

When considering the nighttime HR and factors influencing it, it is also important to understand their minimally important differences in their values, which may be considered as relevant from the clinical perspective. Concerning HR itself, the research focused mostly on evaluating changes in resting heart rate (measured when awake and calm) in longitudinal observational studies. Chen et al. ((48)) show that an increase of 1 BPM in 10 years was associated with a 3% higher risk for all-cause death, 1% higher risk for CVD, and 2% higher risk for coronary heart disease. An increase of 5 BPM is associated with a higher risk of cardiovascular disease, heart failure, and overall all-cause mortality ((49)). A further increase of 10 BPM relates to an increased probability of mortality ((50)), while a resting HR above 60 BPM increases this risk almost exponentially ((29)). Summarizing these findings, we consider minimally important differences in nighttime HR as 1 BPM from the clinical relevance perspective.

The research results are as follows when considering the minimally important differences in *total sleep time* (TST). Overall, it is recommended for an adult to sleep between 7-9 *hours* (h) a day ((51–53)). Furthermore, it is important to notice that most of the sleep-related variables in longitudinal observational studies are self-reported ((54)). Sleeping less than 5 hours relates to an increase in the risk of chronic illness ((55)) while sleeping less than 6 h compared to 7-8 h was correlated with more fat accumulation ((56)). For patients with knee osteoarthritis, a cut-off point of 382 min (6.5h) of TST has been found important for their health; more sleep corresponded to better disease management ((57)). A decrease of total sleep time of 23 min or about 0.6 min per year for 36 years of follow-up has been assigned to effects of aging ((58)). Patients with poorer overall health were found to sleep 39-46 min less than comparable healthy populations ((59)). Summarizing these findings, we consider minimally important differences in TST as 23 min from the clinical relevance perspective.

When considering the minimally important differences in steps and step length, the research results are as follows. A value of 121 steps/day has been indicated as a result of a minimally important change in physical intervention studies — RCTs evaluated via a systematic literature review ((60)). An increase in 226 daily steps has been found for one of the intervention groups (RCT) in a study involving CVD patients ((61)). Similarly, in an RCT with *chronic obstructive pulmonary disease* (COPD) patients, an improvement of 427 steps or deterioration of 456 steps a day has been found clinically significant for their health outcomes ((62)). In a similar study that value was 600 steps ((63)).

Overall research results show that walking an additional 1000 steps per day can help to achieve better health outcomes in cancer patients ((64)), in fibromyalgia

patients ((65)), and lower the risk of all-cause mortality in the general population ((66)). For every increase of 2000 steps per day in the general population, the risk of chronic illness is decreased across multiple health outcomes ((67)). As for the stride length, there exist fewer research results linking it to the health outcomes in longitudinal studies, likely because of the challenge in its instrumentation to measure it accurately in daily life environments. Boyer et al. focused on the in-lab assessment of millimeter changes in stride length in the context of an assessment of the impact of injuries on patients' mobility ((68)). Hannik et al. assessed stride length with 1-centimeter accuracy in the context of geriatric care ((69)), while Rampp et al. achieved 1.5 cm accuracy in a similar research context ((70)). From the clinical relevance perspective, we summarize these findings and consider minimally important differences in steps per day as 121, and stride length as 1 cm.

Organization of the Document

The remainder of our paper is organized as follows. We present our study design, the devices used and their accuracy, our hypotheses, and our analysis plan in the Materials and Methods section, particularly how to collect and analyze data across devices and time within the individuals contributing to this N1OS. We report our main findings in the Results section, along with findings that can better inform a future N1OS or even N1RT designs in the same context of managing stressors and health outcomes. We summarize our findings in the Discussion section and reflect on our findings and experiences in this study in the Conclusions section, indicating the potential future work areas.

MATERIALS AND METHODS

The section below presents the resources and methods applied while conducting the study. Hence, in the subsection "Study Design," we present the description and organization of the different data types used, as well as general statistical principles and overall modeling approach. The subsection "Participants and Collected Data Summary" provides information about the three participants (IM, EJD, KW) and the mean values for all the exposure values. In the subsection "Accuracy of Sleep Duration, Steps, Distance, and Heart Rate Monitored with Fitbit and Apple Watch," we discuss the validity of the wearable devices used to collect data. Finally, in the subsection "Statistical Analysis Plan," we present our a priori hypotheses and the statistical planning of this research.

Study Design

Exploratory N1OS Study Goals and Approach

This is an exploratory N1OS. This is not a confirmatory study, which has the goal of replicating fairly well-known relationships between well-defined variables (i.e., testing/confirming discovered or formerly reproduced scientific hypotheses). Instead, the goal of this study is to characterize largely unknown relationships between variables that are not yet well-defined in the scientific literature; i.e., its goal is to suggest scientific hypotheses to be tested or be confirmed in future studies.

With respect to our hypotheses, (71) investigated a number of these that we relied on in forming our a priori hypotheses in Subsection "A Priori Hypotheses". True to our exploratory goals, these are broad in scope and do not specify exact quantities, but rather directions (e.g., increasing or introducing X causes Y to decrease). Rather, we created our a priori hypotheses based on both the findings of (71) and our own experiences and reflections.

This approach reflects the N1OS core principle mentioned in Section "Relationship to Longitudinal Studies". The participant is also the study's "principal investigator" and domain expert — the domain being their own past health history and experiences. Other information (e.g., the scientific literature) only serves to supplement their own understanding of this domain, how to create a priori hypotheses, and how to assess exploratory hypotheses.

Estimating Credibility and True Quantity Discernibility

In this study, we depart from common statistical practice in one important way that we hope improves our scientific communication. The term significant" is largely misunderstood as meaning "scientifically, clinically, or practically important". Statistical significance is unrelated to scientific significance but has been ubiquitously misunderstood as meaning "significant". This well-documented and long-standing phenomenon is called the "significance fallacy" ((72–75)), a key contributor to the replication crisis in biomedicine and psychology. Hence, leading statistical authorities have strongly recommended abandoning the phrase "statistical significance" entirely ((76–79)), necessitating a search for another phrase to describe the amount of statistical evidence in research findings.

Instead, we will proceed as follows. We will continue to describe a *P* value in terms of its "statistical significance". However, we will describe its corresponding estimate in terms of "statistical discernibility". For example, if an estimated effect of 2 has a *P* value of .001, we might describe the estimate as being statistically discernible for the true, unknown effect (or simply say the estimate is "discernible"). That is, there is sufficient statistical evidence that 2 is a statistically valid estimate of the true, unknown value. If that estimate of 2 had a *P* value of .83, we might say 2 is not statistically discernible as the true effect. (80) is an example of a publication that successfully used this lexical strategy.

Our hope in taking this approach is to avoid committing the common error of making scientifically unsupported claims (i.e., based on statistical qualities of an estimate, rather than on the size and direction of the estimate itself). For example, we might incorrectly claim that "there was a significant effect of getting more sleep on step count the next day, i.e., more sleep causes an increase of 2 steps (p=.001)", when in fact the true finding is, "there was a discernible effect of getting more sleep on step count the next day; i.e., more sleep causes a credible increase of 2 steps (p=.001), but this small increase may not be practically significant."

Modeling Approach

For each participant, we fit Granger models ((81)) over each participant's analysis period (i.e., time frame of available data). These are the time series linear models fit by (71), combinations of which might together comprise a vector autoregression.

"Granger" refers to so-called "Granger causality", which by causal inference definition is in fact only an association/correlation/prediction of one time series with another time series — not a causal effect of one time series on another, as we are attempting to estimate in this study. Predictors include lagged values of both the *dependent variable* (DV) and *independent variables* (IDVs). Our own DVs and IDVs resemble theirs, as detailed below. We will not fit any generalized additive models like they did, as they found that these did not perform notably better than their linear models.

We also included calendar-based control variables in our models, following the examples in (71) Table 1 (e.g., weekend indicator). To enforce temporal order needed to conduct causal inference, we made sure all model DVs occurred after their IDVs and control variables, i.e., generally no overlap in time is allowed between any model predictor and its corresponding outcome.

Participants and Collected Data Definition

The participants of this study were all its three authors, Igor Matias (IM), Eric J. Daza (EJD), and Katarzyna Wac (KW). The data were collected via self-reports and personal wearables used by all three authors (IM, EJD, KW) for different periods. Seventeen types of data are organized into three main categories: calendar-based (CB) control variables; self-reported (SR); and wearable-measured (WM). Table 1 illustrates their splitting and main characteristics. Two main categories of hypotheses to test were defined according to the time frame of the available data:

- Type A hypotheses - included data from all three individuals (IM, EJD, KW) from 14 August 2020 until 8 January 2021 (148 days per person), with only WM data for the first two (IM and EJD), and with CB, SR, and WM for KW.
- Type B hypotheses - included data only from one participant (KW) from 13 February 2017 until 13 August 2020 (1278 days), including CB, SR, and WM.

The CB data type is defined as follows. *Weekend*, a binary variable, with "1" for Saturday or Sunday and "0" for any other day of the week. *Year*, a discrete variable between "0" and "4", representing the years of 2017, 2018, 2019, 2020, and 2021, respectively. *Month*, a discrete variable between "0" and "11" for every month of the year, chronologically ordered from January to December. *Season*, a discrete variable between "0" and "3", representing "Summer", "Autumn", "Winter", and "Spring", respectively, according to the astronomical seasons.

For the WM category, we defined five variables as follows. *Total sleep time* (TST) is a continuous variable for the total seconds of sleep during the main nighttime sleep period, excluding naps. *Steps per awake time* (SAT) is defined as a daily average, calculated as the incremental steps that day divided by the seconds between the waking time and going to bed that night (akin to average daily walking speed). We used SAT instead of total steps per day, as the number of steps is dependent on the awake time each day. *Step length* is defined as a daily average (measured in meters), calculated as the total distance logged that day divided by the total number of steps. *Nighttime HR* is calculated as the average HR during that night's sleep. *Difference HR* (DIF-HR) is defined as the difference between the maximum and the minimum heart rate registered during sleep that night after, and it helps characterize the maximum range in nighttime HR.

The third category, SR, included seven binary variables defined as follows. *Socializing* is defined as "1" when socializing in the evening (which in most cases implied eating-out, hence consumption of non-routine foods and potentially of moderate amounts of alcoholic drinks) and "0" otherwise. *Yoga* is defined as "1" when practicing yoga during the afternoon/evening and "0" when not. *Exercise* with "1" when any acute physical exercise was practiced during the day (e.g., gym session, running, or long biking). *Fasting* defined between "1" and "0" whether the participant fasted (since the dinner a day before, for a full day) or not, respectively. *Tired/Sick/Stress* is defined as "1" when having a tiring day, feeling sick, or experiencing high-stress levels during the day. *Holiday* (for Type A) or *Vacations* (for Type B) is positive when going on vacations or having a non-working day, such as a weekend. *Short air travel* is defined as "1" when traveling by air within the same continent during the daytime and "0" when not traveling or traveling for longer periods or nighttime. The SR variables were collected daily, using manual annotation of personal notes.

Table 1. Types of data used in the study. CB stands for "calendar-based," WM for "wearable-measured," and SR for "self-reported" control variables.

Variable	CB	WM	SR	Used for hyp. type(s)	Type	Units/values
Weekend	X			A and B	Binary	0 or 1
Year	X			B	Discrete	0 to 4
Month	X			A and B	Discrete	0 to 11
Season	X			A and B	Discrete	0 to 3
TST		X		A	Continuous	Seconds
SAT		X		A	Continuous	Steps per second
Step length		X		A	Continuous	Meters
Nighttime HR		X		A and B	Continuous	Beats per minute
DIF-HR		X		A and B	Continuous	Beats per minute
Socializing			X	B	Binary	0 or 1
Yoga			X	B	Binary	0 or 1
Exercise			X	B	Binary	0 or 1
Fasting			X	B	Binary	0 or 1
Tired/Sick/Stress			X	B	Binary	0 or 1
Holiday			X	A	Binary	0 or 1
Vacations			X	B	Binary	0 or 1
Short air travel			X	B	Binary	0 or 1

Participants and Collected Data Summary

As defined above, the participants of this study were all its three authors: IM; EJD; KW. On the last day of the experiment (8 January 2021), IM was a 24-year-old male with a normal *body mass index* (BMI), EJD was a 41-year-old male with a normal BMI, and KW was a 41-year-old female with a normal BMI. All the participants were

healthy (i.e., no unusual medical history), not experiencing any notable work- or family-related stresses, nor disturbances or abnormalities in walking, sleeping, or in any cardiovascular aspects.

For the Type A hypotheses' time frame, the mean value of SAT for IM, EJD, and KW, was 0.07 ± 0.03, 0.10 ± 0.05, and 0.22 ± 0.09 steps per second awake, respectively. In the same way, the mean value of TST for IM, EJD, and KW was 25854.55 ± 3530.30 (7 hours, 10 minutes, 54.55 seconds ± 58 minutes, 50.30 seconds), 28683.21 ± 5127.09 (7 hours, 58 minutes, 03.21 seconds ± 1 hour, 25 minutes, 27.09 seconds), and 30358.39 ± 3232.53 (8 hours, 25 minutes, 58.39 seconds ± 53 minutes, 52.53 seconds) respectively.

For both A and B hypotheses' time frames (1426 days in total), the numbers of positive days for socializing were 369 (25.88%), 68 for yoga (4.77%), 125 for exercise (8.77%), 22 for fasting (1.54%), 150 for tired/sick/stress (10.52%), 280 for holiday (A) and vacations (B) combined (19.64%), and 117 for a short air travel (8.21%).

Accuracy of Sleep Duration, Steps, Distance, and Heart Rate Monitored with Fitbit and Apple Watch

All *wearable measured* (WM) data were collected using a *Fitbit Charge 2*™ (FC2), *Charge 3*™ (FC3), *Charge 4*™ (FC4) (Fitbit, Inc., San Francisco, CA, USA), and an *Apple Watch* (AW) Series 5™ (Apple Computer, Inc., Cupertino, CA, USA). All of them connect via Bluetooth™ to a smartphone, the last one (AW) only fully compatible with iOS™ devices. Within the Type A hypothesis' time frame, IM used an AW Series 5, EJD an FB3, and KW used an FC4. As for Type B's time frame, KW used an FB2 until 17 April 2020, changing to FB4 afterward. Although all devices can measure all the WM data this study needed, these devices use different sensors and components. It is therefore important to discuss the accuracy of each one of them.

As for sleep, because this study did not consider the sleep stages, we will only evaluate the accuracy of sleep total time assessment for the used devices. As validated by (82), FC2 overestimated TST by 9 minutes when compared with polysomnography ($p < .05$). In the same way, (83) compared FC3 and found an inverse conclusion, with an underestimation of TST of about 11 minutes. For FC4, studies of evaluation were not found. Last, AW (no version specified by the literature) overestimated TST by 4.65 minutes, as tested by (84).

As for the steps, when evaluating FC2 against an *ActiGraph GT3X*™, found an overestimation of 2451.3 ± 2085.4 steps per day by using the average over seven days of comparison, 32.2 ± 40.7% above the comparison measurement, with a correlation of $r = 0.58$, $p = .02$. By performing a 24 minutes exercise, at different speeds, (85) concluded about an error of 1.07 steps for the AW compared to the manual count obtained from video recordings, with a total error of 0.034% and a correlation of $r = 0.96$, $p < .001$.

The evaluation results for HR are as follows. To validate the HR measured with a FC2 and AW Series 3, (86) compared both to a gold standard electrocardiograph and found that both devices slightly underestimated HR across 24 hours. While sleeping, FC2 showed a *mean absolute error* (MAE) of 2.15 *beats per minute* (BPM) and *mean absolute percentage error* (MAPE) of 3.36%, where AW Series 3 had a MAE of 1.96

BPM (MAPE of 3.12%). (87) compared FC3 to other well-known wearable devices such as *Polar H10*™, and documented an underestimation of HR by 7 BPM by a FC3, although this study did not follow the same gold standard approach as the first one. Finally, when evaluating the measurement of distance traveled during the day, (88) compared a Fitbit Charge(FC) device with others available at the time, placing FC among the best with a MAPE lower than 5.6%. (89) followed a similar approach and compared AW Series 4 with other brand's devices, documenting that the overall MAPE <5% ranges from 0.9% to 4.1% only.

Statistical Analysis Plan

A Priori Hypotheses

We investigated a total of eleven a priori hypotheses of two types (A and B, defined in Subsection "Participants and Collected Data Definition") and divided them into three groups. We tested an association between an exposure (i.e., IDV) and an outcome (i.e., DV) for each hypothesis. The exposures are SAT, TST, socializing, yoga, exercise, fasting, tired/sick/stress, vacations, and short air travel. The outcomes are TST, step length, DIF-HR, and nighttime HR.

All outcomes were log-transformed and treated as continuous variables. All exposures were treated as binary variables indicating the presence (enumerated as 1) relative to the absence (enumerated as 0) of the exposure or as having a high (1) versus low (0) exposure value. We dichotomized continuous exposures in keeping with the traditional *Neyman-Holland-Rubin potential-outcomes approach* that compares average outcomes between only two treatment levels ((90–92)). We assigned a threshold for each exposure per participant to separate their high and low values. These thresholds were set as the observed per-participant mean value of the exposure over each participant's entire analysis period (see Subsection "Participants and Collected Data Definition" for values).

After dividing all the hypotheses into two types (A and B), we assigned them to three different groups (Steps-TST, Diff-HR, Nighttime HR). The first group, *Steps-TST*, included two hypotheses (H1 and H2) with TST and step length as outcomes. The second group, *Diff-HR*, included two other hypotheses (H3 and H4) in which the outcome was the DIF-HR. The third group, *Nighttime HR*, included the remaining seven hypotheses (H5, H6, H7, H8, H9, H10, and H11) having nighttime HR as the outcome.

Table 2 illustrates the splitting of the hypotheses across the two types (A, B) and three groups (Steps-TST, Diff-HR, Nighttime HR). In this paper, the Steps-TST hypotheses were more conjectural in nature (i.e., stemming from curiosity); in contrast, we had stronger prior beliefs about our remaining nine nighttime heart rate hypotheses (i.e., Diff-HR and Nighttime HR).

We specified the two Steps-TST hypotheses as follows. *Hypothesis 1* (H1) was that an increase in SAT was associated with an average increase in TST the next night. *Hypothesis 2* (H2) was that an increase in TST was associated with a longer average step length the day after.

We specified the two Diff-HR hypotheses as follows. *Hypothesis 3* (H3) was that an increase in SAT was associated with an average decrease in DIF-HR. *Hypothesis 4*

(H4) was that socializing was associated with an average increase in DIF-HR afterward. As in Subsection "Participants and Collected Data Definition", we define *DIF-HR* afterward as the difference between the highest and the lowest HR during the sleep period after the socializing event. For example, if the socializing refers to a social event in the evening of the day, the DIF-HR refers to the night that same day (recall this is defined while sleeping), after the evening ends.

We specified the seven Nighttime HR hypotheses as follows. *Hypothesis 5* (H5) was that an increase in SAT was associated with an average decrease in nighttime HR the following night. We expected to have different levels of association for H3 and H5 (outcome change being DIF-HR for the first and Nighttime HR for the other). *Hypotheses 6-8* (H6 to H8) were that yoga, exercise, and fasting were all associated with an average decrease in nighttime HR the night after. *Hypothesis 9* (H9) was that a tiring day, sickness, or high-stress levels during the day were collectively associated with an average increase in nighttime HR. *Hypothesis 10* (H10) was that going on vacation was associated with an average increase in nighttime HR, as the possible sources of stress of vacations were distinct from those on non-vacation stressful days (i.e., due to different physical activities, different sleep hours, sleeping in a different bed, alcohol intake, traveling, among others). *Hypothesis 11* (H11) was that short air travel would not be associated with a meaningful average change in nighttime HR.

Table 2. A priori hypotheses. TST stands for "total sleep time," SAT stands for "steps per awake time," and TSS stands for "tired/sick/stress."

Hyp.	Exposure	Exp. change	Outcome	Out. Change	Type	Group
H1	SAT	Increase	TST	Increase	A	Steps-TST
H2	TST	Increase	Step length	Increase	A	Steps-TST
H3	SAT	Increase	DIF-HR	Decrease	A	Diff-HR
H4	Socializing	Presence	DIF-HR	Increase	B	Diff-HR
H5	SAT	Increase	Nighttime HR	Decrease	A	Nighttime HR
H6	Yoga	Presence	Nighttime HR	Decrease	B	Nighttime HR
H7	Exercise	Presence	Nighttime HR	Decrease	B	Nighttime HR
H8	Fasting	Presence	Nighttime HR	Decrease	B	Nighttime HR
H9	TSS	Presence	Nighttime HR	Increase	B	Nighttime HR
H10	Vacations	Presence	Nighttime HR	Increase	B	Nighttime HR
H11	Short air travel	Presence	Nighttime HR	None	B	Nighttime HR

We also included variables in each model to account for suggested effect modification by the CB variables. These are specified as interaction terms between an IDV and each CB variable included in a model. We did so in case the average daily effect of an IDV on a DV might vary based on a CB variable. For example, in H1, suppose the effect of taking more steps increases average TST the following night in the Summer than in the Fall for IM. This might be because IM has fewer scheduled

early workdays in the Summer, allowing him to sleep longer after an active day with many steps, or even due to better weather conditions during Summer.

Causal Hypotheses via Model-Twin Randomization (MoTR)

Thus far, all hypotheses have been assumptions of statistical association or correlation, not causation. In this paper, we went further and employed the MoTR method to simulate an N1RT after adjusting for other assumed confounders. MoTR allows us to change these hypotheses of association to hypothesized effects that can be statistically tested.

MoTR is a Monte Carlo approach to estimating the APTE that works as follows. It takes as its input a model fit to a dataset, randomly shuffles the exposures (IDVs previously dichotomized as in Subsection "A Priori Hypotheses"), and then sequentially predicts the outcome (DV) for all time points (or "periods" in APTE parlance) in the study period. The average outcomes under high and low exposures are compared, yielding an APTE estimate with a *P* value (and, thereby, confidence interval). Because many random sequences of exposures are possible given the longitudinal datasets, MoTR repeats this procedure many times by randomly shuffling exposures differently each time. This creates multiple Monte Carlo simulation runs. The final mean APTE estimate and *P* value were reported once the APTE estimate stabilized (after a minimum of 1000 runs), or at run 10000 (to set a computational time limit on the MoTR algorithm), whichever occurred first. (See the Supplementary Materials and Formulas for details on these convergence criteria.)

Exploratory, Testing, and Confirmatory Phases

In general, we conducted four types of analyses. *Exploratory A* and *Exploratory B* analyses were first conducted as this was the main goal of our paper. We then proceeded to the *Testing* and *Confirmatory* phases. We conducted a few Confirmatory analyses based on loosely defined a priori hypotheses. This was done to demonstrate how to apply MoTR in a confirmatory study. These types were separated according to their input and main goal, as represented in Table 3 and Table 4.

We define *lag* as the number of days preceding the exposure day, including the exposure day, i.e., the day for which the DV was obtained for each hypothesis. The lag, therefore, defines the number of days for which the data has been acquired for DV. For example, using a lag of 2 days means the hypothesis considered data from the DV on the study day (t) and each of the two days before (t-1 and t-2), thus enabling the analysis of the variation of the DV before the exposure. There are 240 hypotheses for Exploratory A, 140 for Exploratory B, 10 hypotheses for the Testing phase, and 24 for the Confirmatory phase. These are defined as follows:

- *Exploratory A Phase* - this phase was intended to explore the impact on analytic results of bigger lags (from 1 to 10 days), as well as the changes in DV produced by controlling for interactions within IDVs. It was applied to all Type A hypotheses (H1, H2, H3, H5) one at a time.
- *Exploratory B Phase* - similarly to the Exploratory A, this phase aimed at exploring the impact on analytic results of lags bigger than one day (lag from 1 to 10 days), but on a longer period of days and only from one participant

(KW), non-overlapping with the respective participant's data on Type A hypotheses. It was applied to all Type B hypotheses (H4, H6, H7, H8, H9, H10, H11) one at a time.

- *Testing Phase* - the goal of this phase was to assess the accuracy of using a model fitted on 1278 days of KW's data (Type B time frame) to predict the 148 days after (Type A time frame). Using the machine learning model fitted for KW's data on the Exploratory B phase (one repetition for each of the Type B hypotheses), we predicted the DV values for each study day on the Type A's data.
- *Confirmatory Phase* - this phase aimed to assess the suggested effect (or not) of each hypothesis' IDV on the DV without using any lag bigger than 1 day (i.e., lag 1 only) nor using interactions within the IDVs as controls. It was applied to all Type A hypotheses (H1, H2, H3, H5) one at a time.

The Testing phase included several tasks as follows. (1) Using the model fit with data from part B of KW's data during Exploratory B, we predicted the DV values of part A and added Gaussian noise to each prediction using the standard deviation (SD) of the model residuals from part B. (2) We estimated the association of the IDV with the DV, or naive effect estimate," by comparing the means of the noisy predicted DV values between high/low or present/absent exposure (and its t test *P* value) using the first method; hence, we refer to this as the "naive method." (3) We also assessed the fit of that same model on KW's data from part A, using the *mean squared error* (MSE) of its residuals. (4) Then we predicted the DV values (with noise, as before) of part A using the MoTR method, now calculating the hypothetical suggested effect of the IDV on the DV (and its t test *P* value).

In the end, we compared both the fit of the model on observed data in parts B and A measured as their MSEs and the difference between the naive effect estimate of the IDV in part A with its hypothetical suggested effect estimated using MoTR. We chose to compare model fit between parts B and A using the MSE rather than R-squared because this metric expresses the same qualitative information as the R-squared in how well the model explains random variation in the DV. However, the MSE also preserves the original scale of the DV, such that it conveys this added information that is masked when calculating the R-squared.

The dataset processing, programming language libraries used, and the original code used to deploy the MoTR method are described in the Supplementary Materials and Formulas at the end of this paper.

The data flow between hypotheses and phases is represented in Figure 1.

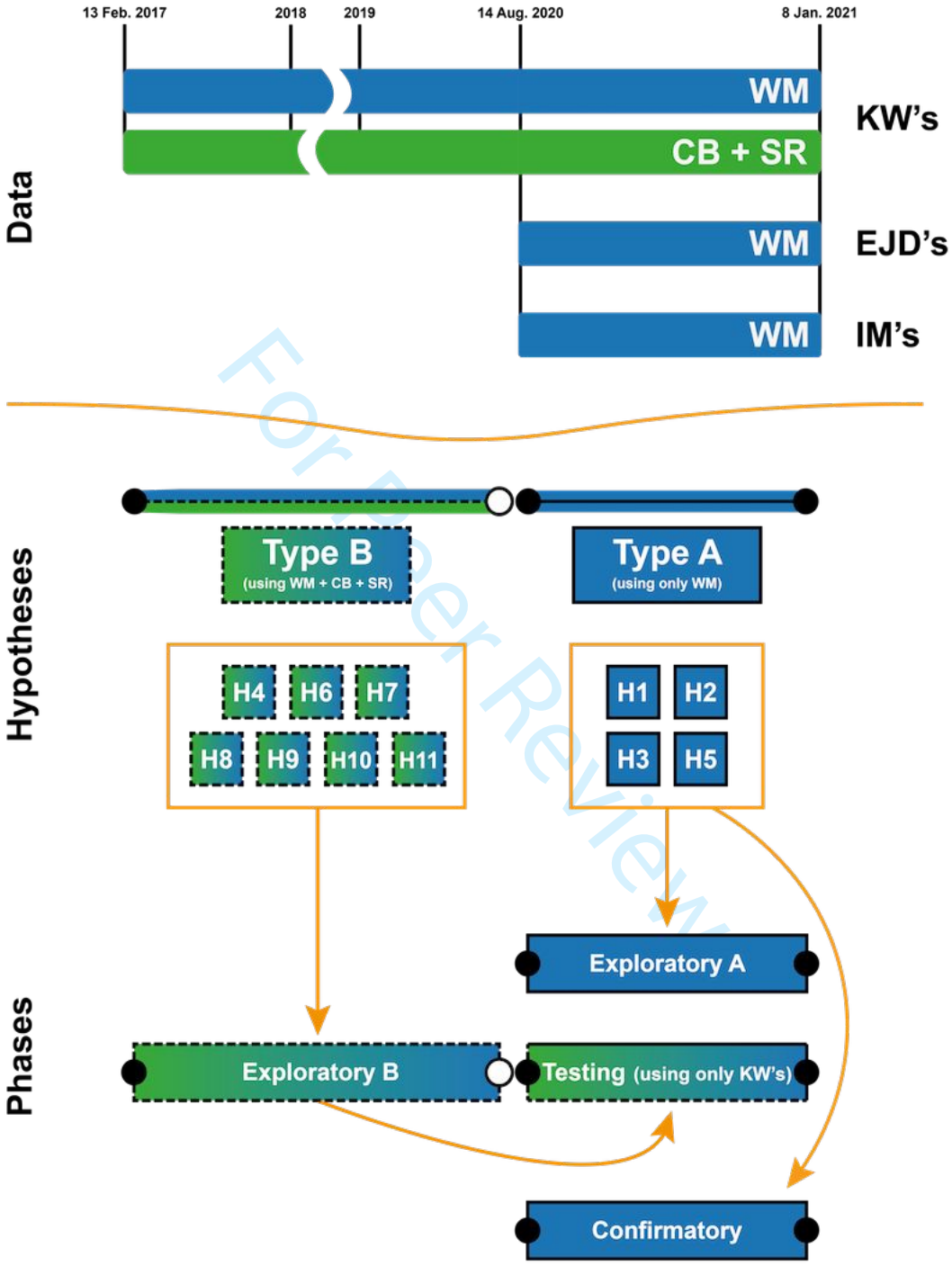

Figure 1. Representation of the data flow between the different hypotheses and phases of the methods. The orange arrows represent the hypotheses pool used in

each phase (e.g., the Testing phase only considered the hypotheses selected by Exploratory B from the entire pool). (WM = wearable-measured, CB = calendar-based, SR = self-reported, KW = Katarzyna Wac, EJD = Eric J. Daza, IM = Igor Matias).

Table 3. Exploratory and Confirmatory phases planning.

Exploratory A							
	Controlling for	”Weekend,” ”Holiday,” ”Month,” interactions IDV*”Weekend,” IDV*”Holiday,” and IDV*”Month” (A-MONTH)			”Weekend,” ”Holiday,” ”Season,” interactions IDV*”Weekend,” IDV*”Holiday,” and IDV*”Season” (A-SEASON)		
	Days of lag	1 to 10 days	1 to 10 days	1 to 10 days	1 to 10 days	1 to 10 days	1 to 10 days
	Participants	KW	EJD	IM	KW	EJD	IM
	Hypotheses	H1, 2, 3, 5	H1, 2, 3, 5	H1, 2, 3, 5	H1, 2, 3, 5	H1, 2, 3, 5	H1, 2, 3, 5
	Total hyp.	40	40	40	40	40	40
		120			120		
		240					
Exploratory B							
	Controlling for	”Weekend,” ”Vacations,” ”Year,” ”Month,” interactions IDV*”Weekend,” IDV*”Vacations,” IDV*”Year,” and IDV*”Month” (B-MONTH)			“Weekend,” ” Vacations,” ”Year,” ”Season,” interactions IDV*”Weekend,” IDV*”Vacations,” IDV*”Year,” and IDV*”Season” (B-SEASON)		
	Days of lag	1 to 10 days			1 to 10 days		
	Participants	KW			KW		
	Hypotheses	H4, 6, 7, 8, 9, 10, 11			H4, 6, 7, 8, 9, 10, 11		
	Total hyp.	70			70		
		140					
Confirmatory							
	Controlling for	”Weekend,” ”Holiday,” and ”Month” (C-MONTH)			”Weekend,” ”Holiday,” and ”Season” (C-SEASON)		
	Days of lag	1	1	1	1	1	1
	Participants	KW	EJD	IM	KW	EJD	IM
	Hypotheses	H1, 2, 3, 5	H1, 2, 3, 5	H1, 2, 3, 5	H1, 2, 3, 5	H1, 2, 3, 5	H1, 2, 3, 5
	Total hyp.	4	4	4	4	4	4
		12			12		
		24					

Table 4. Testing phase planning. Refer to the Subsection “Results Selection Criteria” for details on the two criteria used.

Testing											
	Criteria	MDE						MBL			
	Controlling for	B-MONTH			B-SEASON			B-MONTH		B-SEASON	
	Days of lag	6	10	10	6	10	4	3	4	3	3
	Participants	KW	KW	KW	KW	KW	KW	KW	KW	KW	KW
	Hypothesis	H4	H6	H9	H4	H6	H9	H4	H9	H4	H9
	Total hyp.	3			3			2		2	
		6						4			
		10									

RESULTS

Missing Data Imputation

Of all the 148 total days of data per person used in Type A hypotheses, TST, SAT, sleep HR, and DIF-HR was missing on 2 days (1.35%) for KW's data and 4 days (2.70%) for IM's data, with no missing data for EJD. From the 1278 days of data used in Type B hypotheses, TST was missing on 28 days (2.19%), SAT was missing on 46 days (3.60%), and sleep HR and DIF-HR were missing on 30 days (2.35%). We considered the data to be missing at random (93,94). The missing data were imputed using linear interpolation for the missing values only, keeping the original values that were not missing in the interval.

Results Selection Criteria

Although we calculated results for all lags (i.e., 1 to 10) for all the hypotheses in Table 2, only the models with the most interesting results are presented in detail and discussed. Each Exploratory A and B hypothesis model included only one of 10 possible lags, chosen using the following criteria. (The Confirmatory hypotheses had only one model each, with a lag of one day; hence, we report all their results, and no results selection was done for the Confirmatory hypotheses.)

We used the following two criteria to select models with the most interesting results.

- *Most discernible effect* (MDE) - We selected the model with the lowest *P* value after applying the MoTR method.
- *MDE, best-fitting model, and largest confounding influence* (MBL) - We selected the model that jointly met three criteria: the lowest *P* value after the MoTR method was applied (MDE); the smallest value of each model's Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and F statistic *P* value; and the largest value of the confounding influence, defined as the absolute difference between the mean differences in outcomes under the two different exposure levels (e.g., low and high), before and after applying MoTR. Please refer to the Supplementary Materials and Formulas section for results and discussion using this criterion.

MDE Criterion

The MDE criterion selects the model that produces the most statistically discernible suggested effect among all candidate models. The selected model has the greatest statistical evidence (i.e., lowest t test *P* value) that the there is a true mean difference in its predicted noisy outcomes between the two IDV levels (e.g., low/high exposure), after following the MoTR procedure to randomize the IDV. Recall that randomizing the IDV makes this mean difference an estimate of a suggested causal effect of the IDV on the DV—not just an association or correlation between IDV and outcome.

Note that the model selected using the MDE criterion does not necessarily produce the largest estimated suggested effect. The correct interpretation is that the selected model has the most statistical evidence for the existence of an effect of the IDV on the DV, which may be large or small. However, it can only ever be a suggested effect; the model may or may not resemble the true data-generating model needed to calculate

the true, unknown effect (if any). Recall that this is the main limiting assumption of the MoTR method, and indeed, of all models used in causal inference for observational studies (wherein the exposure or IDV was never randomized).

Procedure

After the first round of model selection, we still had 16 results for the Exploratory A phase (i.e., eight for MDE and eight for MBL), and 28 for Exploratory B (i.e., 14 for MDE and 14 for MBL). Thus, we performed a second round of selection of results.
In this second round, for each hypothesis, we selected the model with both $P \leq .05$ and an estimated suggested effect higher than the minimally important difference in the outcome from the clinical perspective (defined in Subsection "Daily-Life Stressors: Exercise, Sleep, Alcohol and Food, Psychological Stress, Aircraft Travels, and Nighttime HR") and at the same time higher than device's error as defined in Subsection "Accuracy of Sleep Duration, Steps, Distance, and Heart Rate Monitored with Fitbit and Apple Watch". Therefore, we consider a suggested effect of at least 11 minutes (660 seconds) for TST, at least 0.035 (KW and EJD) or 0.039 (IM) meters for step length (i.e., 5% of the mean from all three individuals), and at least 2 BPM for heart rate data. These inclusion criteria are summarized by Table 5.
Following the second round's selection above described, applying the MDE criterion resulted in the inclusion of results for H1 and H3 (only from IM's data) for the Exploratory A phase, and results for H4, H6, H7, and H9 for Exploratory B (KW's data). MBL resulted in the inclusion of results for H1 and H3 (only from IM's data) for the Exploratory A phase and results for H4 and H9 for Exploratory B (KW's data).

Table 5. Results inclusion criteria: minimally important difference in exposure/outcome considered. TST stands for total sleep time, and HR stands for heart rate.

DV	Minimum effect of IDV on the DV
TST	660 seconds (11 minutes)
Step length — KW	0.035 meters
Step length — EJD	0.035 meters
Step length — IM	0.039 meters
HR	2 BPM

Exploratory Phases Results Using MDE Criterion

Table 6. Selected results from Exploratory A and Exploratory B phases using MDE criterion. Time values are in the format "minutes:seconds."

Exploratory A					
	H1	IM's data	Controlling for	A-MONTH	A-SEASON
			Lag	2 days	1 day
			IDV effect on DV	+ 21:26.901	-18:21.077

			t test *P* value	.026	.027
	H3	IM's data	Controlling for	A-MONTH	A-SEASON
			Lag	1 day	6 days
			IDV effect on DV	+ 5:61 BPM	+ 2.53 BPM
			t test *P* value	< .001	.057
Exploratory B					
	H4	KW's data	Controlling for	B-MONTH	B-SEASON
			Lag	6 days	6 days
			IDV effect on DV	+ 2.84 BPM	+ 4.99 BPM
			t test *P* value	.004	.002
	H6	KW's data	Controlling for	B-MONTH	B-SEASON
			Lag	10 days	10 days
			IDV effect on DV	+ 4.58 BPM	+ 10.90 BPM
			t test *P* value	.051	.020
	H7	KW's data	Controlling for	B-MONTH	B-SEASON
			Lag	2 days	1 day
			IDV effect on DV	+ 3.36 BPM	+ 2.523 BPM
			t test *P* value	.051	.016
	H9	KW's data	Controlling for	B-MONTH	B-SEASON
			Lag	10 days	9 days
			IDV effect on DV	- 4.00 BPM	- 6.63 BPM
			t test *P* value	.001	.001

Exploratory A - Suggested Effect of SAT on TST (H1)

For H1, only IM's data yielded results with a t test *P* value below .05 and a suggested effect greater than 660 seconds (11 minutes). KW's data led to *P* values higher than .10 and EJD's higher than .23, thus being discarded.

Following the selection process previously described Table 6 presents the metrics of the lag with the lowest t test *P* value for each type of control. When controlling for A-MONTH, the lowest t test *P* value was .026 for 2 days of lag, measuring a suggested effect of 1286.901 seconds more (+ 21 minutes and 26.901 seconds) on TST when IM's number of steps per awake time was higher than his daily mean. When controlling for A-SEASON, the lowest t test *P* value was .027, very similar to the one controlling for A-MONTH, for 1 day instead of 2, for which the suggested effect has the opposite meaning, that is, when IM's steps per awake time value are above mean it results in a decreased TST of - 1101.077 seconds (- 18 minutes and 21.077 seconds), instead of a positive suggested effect as before.

Because of this inverse suggested effect, while controlling whether for month or season (the only difference between A-MONTH and A-SEASON), Table 7 presents the direct comparison of the two selected results with its correspondence (labeled as "not selected") on the other control type, that is, the results using the same lag size. Although the correspondent lags do not surpass the minimum 660 seconds for being plausible, we can still confirm that the positive/negative suggested effect stays the

same when controlling for A-MONTH (always positive) or A-SEASON (always negative). These results will be further discussed in the later sections of this article.

Table 7. Comparison of the selected results for H1 (Exploratory A), using MDE criterion, with its correspondents (same days of lag) on the other control type. Time values are in the format "minutes:seconds".

H1	IM's data	Controlling for	A-MONTH	A-SEASON
		Lag	2 days	2 days (not selected)
		IDV effect on DV	+ 21:26.901	- 6:26.480
		t test *P* value	.026	.037
	IM's data	Controlling for	A-MONTH	A-SEASON
		Lag	1 day (not selected)	1 day
		IDV effect on DV	+ 6:34.288	- 18:21.077
		t test *P* value	.048	.027

Exploratory A: Suggested Effect of SAT on DIF-HR (H3)

Like for H1, for H3, only IM's data yielded results with a t test *P* value below .05, this time with a suggested effect higher than 2 BPM. KW's data led to *P* values higher than .07 and EJD's higher than .20, thus being discarded.

Like the last stated hypothesis, Table 6 presents the metrics of the lag with the lowest t test *P* value for each type of control. Although the t test *P* value when controlling for A-SEASON is slightly above .05, we still consider it. Thus, whether controlling for A-MONTH or A-SEASON, when IM's steps per awake time value are above the daily mean, the difference between the highest and the lowest HR during sleep increases (5.61 BPM controlling for A-MONTH with 1 day of lag, 2.53 BPM controlling for A-SEASON with 6 days of lag).

Exploratory B: Suggested Effect of Socializing on DIF-HR (H4)

While for the Exploratory A phase, we screened all the results from all three data sources (KW, EJD, and IM), the only data used for the Exploratory B phase came from KW, as previously described in this article.

For the first selected results, that is, for the suggested effect of socializing on nighttime HR, the lowest t test *P* value was obtained when considering 6 days of lag for both control types. The suggested effect was positive in both controls when KW's data reported the existence of socializing, increasing the nighttime HR after by 2.84 BPM (controlling for B-MONTH) and 4.99 BPM (controlling for B-SEASON), as shown by Table 6.

Exploratory B: Suggested Effect of Yoga on Nighttime HR (H6)

Like for H4, for the hypothesis of yoga affecting the nighttime HR (H6), the results showed a positive suggested effect with 10 days of lag on both control types, as detailed by Table 6.When controlling for B-MONTH, yoga exercise affects the HR during sleep after in + 4.58 BPM and + 10.90 BPM when controlled for B-MONTH and B-SEASON, respectively. Like for H3 above, we considered the value when controlling for B-MONTH even with a *P* value slightly above .05.

Exploratory B: Suggested Effect of Exercise on Nighttime HR (H7)

As in Table 6, the suggested effect of exercise on nighttime HR was positive as the suggested effect of yoga. While controlling for B-MONTH, we found a positive suggested effect of 3.36 BPM with 2 days of lag. Controlling for B-SEASON allowed us to reveal a possible positive effect of 2.52 BPM with 1 day of lag.

Exploratory B: Suggested Effect of Tired/Sick/Stress on Nighttime HR (H9)

Inversely to socializing, yoga, and exercise, the presence of a tired/sick/stress state during the day of KW revealed a negative suggested effect on the average nighttime HR, with 10 days of lag and 9 days of lag while controlling for B-MONTH and B-SEASON, respectively, as shown by Table 6. The strongest suggested effect was found while controlling for B-SEASON with - 6.63 BPM of change, compared with - 4.00 BPM when controlling for the other type.

Table 8. Comparison of the a priori hypotheses and the results obtained using the MDE criterion. For H1 there were different results when controlling for A-MONTH (increase) and A-SEASON (decrease). TST stands for "total sleep time," SAT stands for "steps per awake time," and TSS stands for "tired/sick/stress."

	A priori				Results with MDE	
Hyp. / participant	Exposure	Exp. change	Outcome	Out. change	Out. change	Result
H1/IM	SAT	Increase	TST	Increase	Inc./Dec.	Inconclusive
H3/IM	SAT	Increase	DIF-HR	Decrease	Increase	Not supported
H4/KW	Socializing	Presence	DIF-HR	Increase	Increase	Supported
H6/KW	Yoga	Presence	Nighttime HR	Decrease	Increase	Not supported
H7/KW	Exercise	Presence	Nighttime HR	Decrease	Increase	Not supported
H9/KW	TSS	Presence	Nighttime HR	Increase	Decrease	Not supported

Table 8 presents a comparison between the a priori hypotheses from both phases Exploratory A and B and the results obtained following the MDE criterion.

Testing Phase for the Results Using MDE Criterion

Table 9. Testing phase's results for the hypotheses selected using the MDE criterion. MSE stands for "mean squared error."

H4	KW's data	Controlling for	B-MONTH	B-SEASON
		Lag	6 days	6 days
		R^2 in B	0.312	0.290
		MSE in B	0.001	0.001
		IDV effect on DV	0.12 BPM	0.12 BPM

		(naïve method)		
		t test *P* value (naïve method)	.885	.885
		R^2 in A	- 0.143	- 0.078
		R^2 in B - R^2 in A	0.455	0.368
		MSE in A	0.001	0.001
		MSE in B - MSE in A	0.000	0.001
		IDV effect on DV (MoTR)	- 0.43 BPM	- 0.72 BPM
		t test *P* value (MoTR)	.221	.224
H6	KW's data	Controlling for	B-MONTH	B-SEASON
		Lag	10 days	10 days
		R^2 in B	0.291	0.281
		MSE in B	0.001	0.001
		IDV effect on DV (naïve method)	- 1.23 BPM	- 1.23 BPM
		t test *P* value (naïve method)	.082	.082
		R^2 in A	- 0.364	- 0.598
		R^2 in B - R^2 in A	0.655	0.879
		MSE in A	0.001	0.001
		MSE in B - MSE in A	0.000	0.000
		IDV effect on DV (MoTR)	- 1.50 BPM	- 2.27 BPM
		t test *P* value (MoTR)	.670	.660
H9	KW's data	Controlling for	B-MONTH	B-SEASON
		Lag	10 days	4 days
		R^2 in B	0.302	0.283
		MSE in B	0.001	0.002
		IDV effect on DV (naïve method)	- 1.35 BPM	- 1.31 BPM
		t test *P* value (naïve method)	.134	.141
		R^2 in A	0.014	- 0.090
		R^2 in B - R^2 in A	0.288	0.373
		MSE in A	0.001	0.001
		MSE in B - MSE in A	0.001	0.001
		IDV effect on DV (MoTR)	- 1.31 BPM	- 0.89 BPM
		t test *P* value (MoTR)	.576	.422

As described in the Subsection "Exploratory, Testing, and Confirmatory Phases", the Testing phase was intended to assess the accuracy of the models fitted with part B of KW's data (from Type B time frame) for predicting data from part A (from Type A time frame). Because the models used were selected according to the two criteria used (MDE and MBL), this first subsection presents the results only for the models obtained from the results chosen using MDE (H4, H6, H7, and H9). Table 9 shows the metrics for

the hypotheses H4, H6, and H9. The testing phase could not be applied to H7 because KW's data did not include any positive values for Exercise in part A of the data.

Suggested Effect of Socializing on DIF-HR (H4)

The difference between the MSE of the model in B and the MSE in A is almost null for both control types (B-MONTH and B-SEASON). However, the suggested effect of IDV and t test *P* value between the naive method and the MoTR method requires additional attention. The t test *P* value is notably smaller (approximately four times) when using the MoTR method, even though it is always above .05. When measuring the suggested effect of Socializing on DIF-HR, the biggest value is obtained when controlling for B-SEASON and using the MoTR method (-0.72 BPM). However, none of the calculated suggested effects is above the minimum suggested effect defined in Table 5, and we obtain a positive impact when using the naive method. In contrast, it is negative if we consider the MoTR method.

Suggested Effect of Yoga on Nighttime HR (H6)

The suggested effect of yoga on nighttime HR is negative whether we consider the naive method or the MoTR method, always being less than the minimum suggested effect defined in Table 5 except when controlling for B-SEASON and using the MoTR method. The model fitted has an MSE of 0.001 in part A and part B, being this fitting difference virtually non-existent. However, the t test *P* value is always higher than .05, making all the results not statistically discernible — the naive method gave lower *P* values than the MoTR.

Suggested Effect of Tired/Sick/Stress on Nighttime HR (H9)

Like the H6, the suggested effect of tired/sick/stress on nighttime HR was always below the minimum suggested effect defined in Table 5, being approximately the same between the naive method and the MoTR method when controlling for B-MONTH. The suggested effect was smaller when controlling for B-SEASON while using the MoTR method, although the lag differed between the two control types. The MSE differed in approximately 0.001 between the two processes.

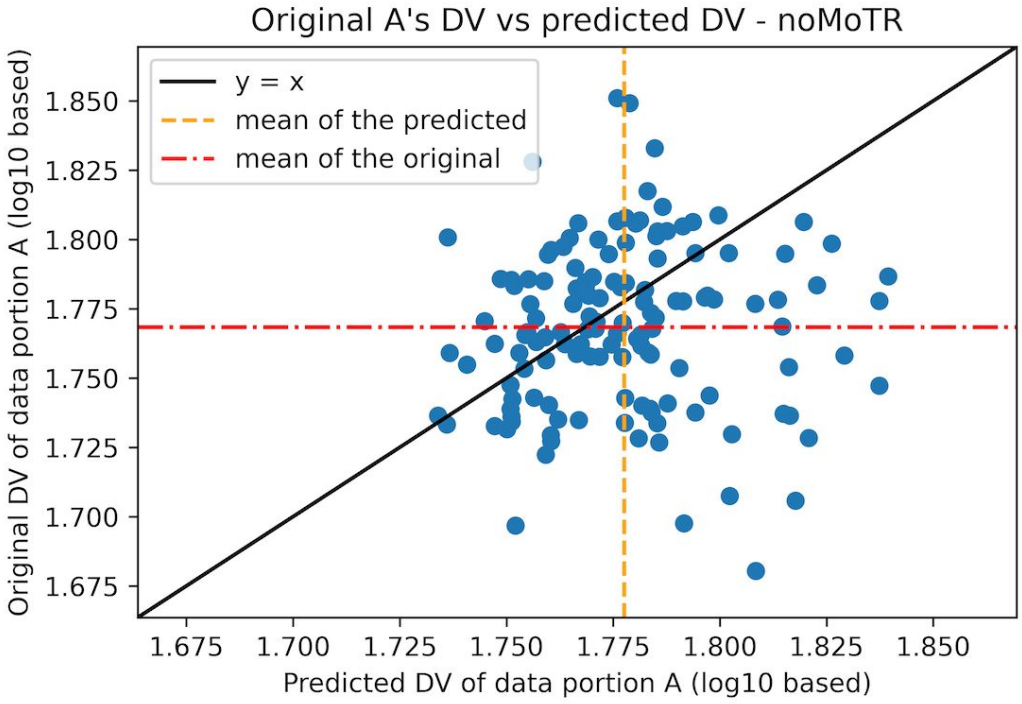

Figure 2. Scatter plot comparing the observed (original) with the predicted DV values for H6, controlling for B-SEASON with 10 days of lag. Result selected using MDE criterion.

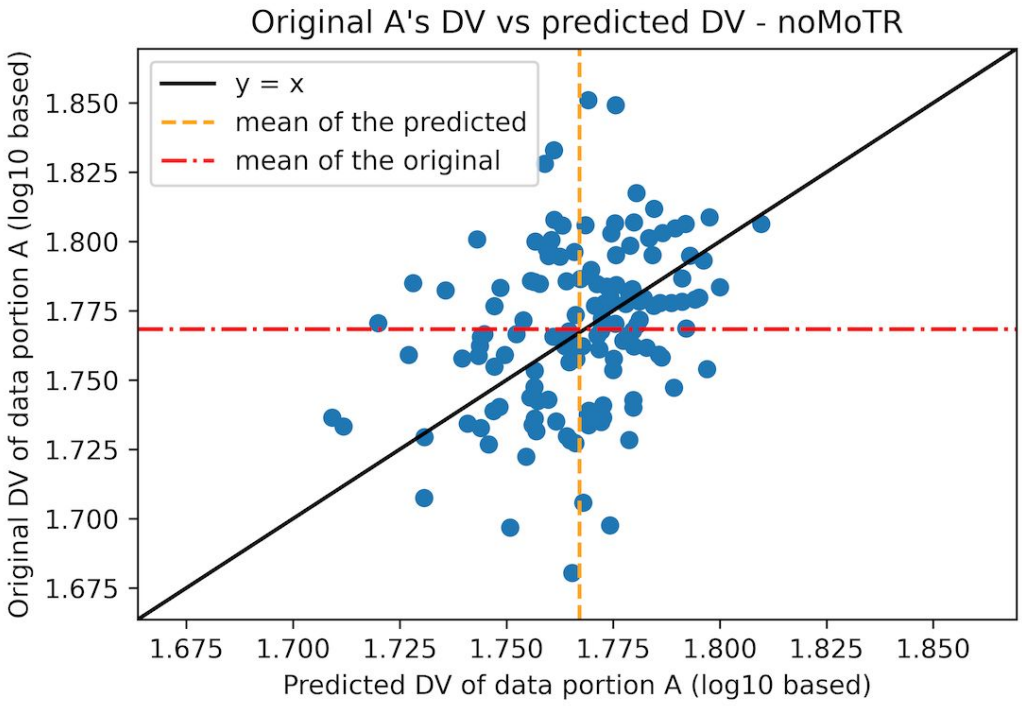

Figure 3. Scatter plot comparing the observed (original) with the predicted DV values for H9, controlling for B-MONTH with 10 days of lag. Result selected using MDE criterion.

In Table 9, note that the R-squared values for part B are positive for both H9 controlling for B-MONTH, and for H6 controlling for B-SEASON. This makes sense because the outcomes are predicted using the model fit to the same data in each case. However, while the R-squared value is positive for part A for H9 controlling for B-MONTH, the R-squared is negative for part A for H6 controlling for B-SEASON. This is because the R-squared formula (see Supplementary Materials and Formulas) relies on the empirical overall mean of the observed outcomes in the target dataset in which predicted outcomes are calculated.

To elaborate, predicted outcomes are created using a model fit to a dataset with a certain empirical mean (e.g., part B). If the empirical mean of the new target dataset (e.g., part A) differs from the original dataset's empirical mean, then the R-squared value calculated using predictions calculated using the original model can be negative. The mean of the predicted values will resemble that of the original dataset's outcomes, while the new dataset's overall mean outcome will not. Such a difference in empirical means between the target dataset's predicted and observed values is shown between the two example cases mentioned above in

Figure 2 and

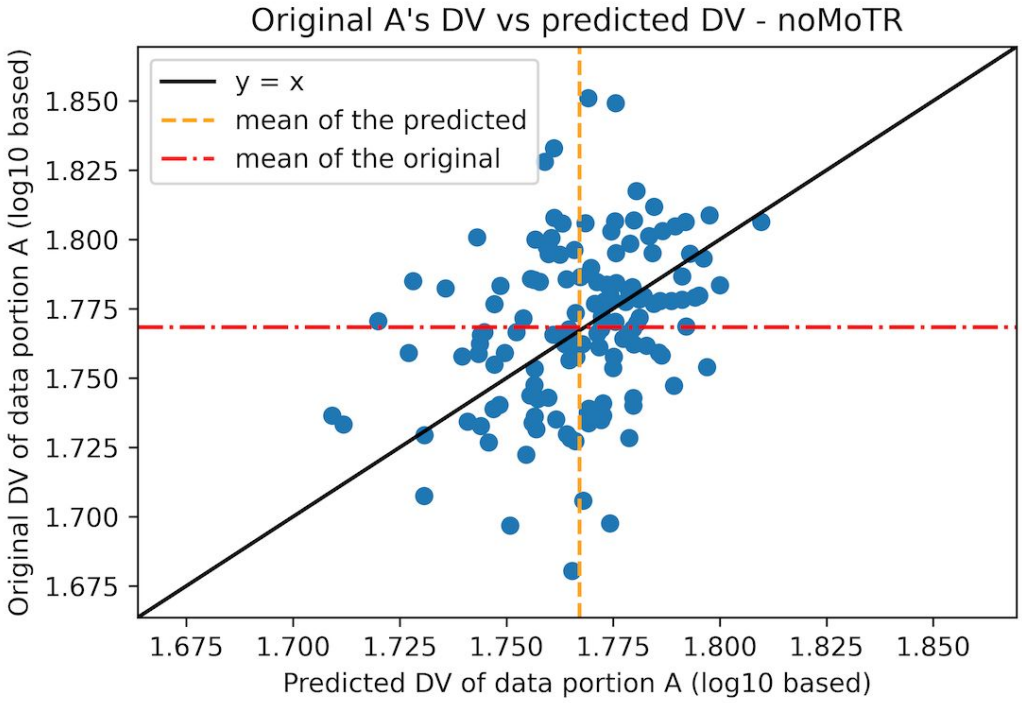

Figure 3.

Confirmatory Phase Results

Table 10. Confirmatory phase's results for all the studied hypotheses. Time values are in the format "minutes:seconds."

H1	KW's data	Controlling for	C-MONTH	C-SEASON
		IDV effect on DV	- 13:35.953	- 8:54.645
		t test *P* value	.142	.153
	EJD's data	Controlling for	C-MONTH	C-SEASON
		IDV effect on DV	+ 00:22.307	- 00:19.555
		t test *P* value	.512	.538
	IM's data	Controlling for	C-MONTH	C-SEASON
		IDV effect on DV	- 23:24.859	- 16:25.250
		t test *P* value	.027	.022
H2	KW's data	Controlling for	C-MONTH	C-SEASON
		IDV effect on DV	+ 0.014 meters	+ 0.006 meters
		t test *P* value	.014	.683
	EJD's data	Controlling for	C-MONTH	C-SEASON
		IDV effect on DV	- 0.002 meters	- 0.002 meters
		t test *P* value	.543	.655
	IM's data	Controlling for	C-MONTH	C-SEASON
		IDV effect on DV	- 0.003 meters	- 0.006 meters
		t test *P* value	.442	.460

H3	KW's data	Controlling for	C-MONTH	C-SEASON
		IDV effect on DV	- 517.69 BPM	+ 9.20 BPM
		t test *P* value	.116	.169
	EJD's data	Controlling for	C-MONTH	C-SEASON
		IDV effect on DV	+ 1.92 BPM	- 0.59 BPM
		t test *P* value	.472	.482
	IM's data	Controlling for	C-MONTH	C-SEASON
		IDV effect on DV	+ 1.66 BPM	+ 0.72 BPM
		t test *P* value	.290	.128
H5	KW's data	Controlling for	C-MONTH	C-SEASON
		IDV effect on DV	- 2.23 BPM	+ 2.99 BPM
		t test *P* value	.126	.168
	EJD's data	Controlling for	C-MONTH	C-SEASON
		IDV effect on DV	+ 0.26 BPM	+ 1.75 BPM
		t test *P* value	.558	.520
	IM's data	Controlling for	C-MONTH	C-SEASON
		IDV effect on DV	- 0.09 BPM	- 0.06 BPM
		t test *P* value	.017	.017

As shown in Table 10, most of the results obtained during the Confirmatory phase are not statistically discernible at the .05 level of statistical significance. Hence, the data we have cannot infer any conclusions about 19 of these 24 hypotheses reliably. Nevertheless, if we apply the same filtering criteria defined in subsection "Results Selection Criteria", we can only use one group of results. That is the H1 using IM's data, showing a negative suggested effect of the SAT (when greater than daily average) in the TST of - 1404.86 seconds (- 23 minutes and 24.86 seconds) and - 985.25 seconds (- 16 minutes and 25.25 seconds), while controlling for C-MONTH and C-SEASON, respectively.

We also focus on the result obtained for H3 with KW's data while controlling for C-MONTH, as the suggested effect obtained with the MoTR method is impossible to obtain in real life (- 517.69 BPM). The most plausible explanation for this result is that we used noise simulation for every MoTR implementation, with a value randomly generated out of a normal distribution, with the mean being the mean value of the data with which the model was fitted previously. By doing so, this specific simulation occasionally generated too many outliers or one extra distant outlier that increased the noise inputted in one or more simulations. It is important to mention that this simulation was repeated during the results-making process, so we could be sure this was not caused by any code or compiler error — this was possible because MoTR implementation has a fixed initial seed for all the randomly generated values.

DISCUSSION

Discussion of the Results Using MDE Criterion

From all the a priori hypotheses from the Exploratory A phase, only the results of two were included following the inclusion criteria defined in Subsection "Results Selection

Criteria", and both results were based on the IM's data (H1 and H3). One of these hypotheses, H3, was not supported. The other, H1, was supported when controlling for A-MONTH but not supported when controlling for A-SEASON. That may tell us that the month to which the data is referring influences the suggested effect of SAT on the TST. A possible explanation for that is that the data used by this type of hypothesis (A) includes a total of six months (August to January) and a total of three seasons (Summer, Autumn, and Winter), for which Summer and Winter are only represented by approximately one and two months, respectively. In contrast, Autumn is fully represented with three months of data. If we assume the month influences the outcome being studied (TST), then we might be looking at a version of Simpson's Paradox ((95)), in which the effect size of every month is lost when we combine them into seasons, thus explaining why we get a positive suggested effect when controlling for all the months, but we get a negative suggested effect when we control for the season, that is, a combined version of the months. According to this paradox, the correct result is given when controlling A-MONTH. That is, the hypothesis can be supported. Additionally, this observed difference might be caused by different distributions of physical activity—steps—throughout the daytime, which may depend on the season or weather conditions, for example.

From the Exploratory B phase, only H4 was supported, showing that KW's socializing events could positively have affected the DIF-HR. The difference between the maximum and minimum heart rate registered during the following sleep period increased. The biggest suggested effect was obtained with six days of lag, that is, considering the history of the past six days. That means that the KW's DIF-HR values are dependent on the days before. A possible cause of that is that socialization happened at a weekend day (6th day), with a stressful proceeding week, in which DIF-HR was slightly higher than usual mean DIF-HR; the socialization event resulted in the following DIF-HR notably higher than usual. In the case of days with usual DIF-HR, followed by one day of socialization, the resulting nighttime DIF-HR may not be notably higher than usual; the body metabolizes" well socialization exposure. All the other hypotheses (H6, H7, and H9) were not supported, for which the MoTR method always revealed the contrary change on the outcome variables compared to the initial hypotheses. Although, if we compare the results using MoTR with the suggested effect direction shown by the original data (cf. Table 11), H7 and H9 are supported by the original data and not by MoTR. A possible cause is that the MoTR method meticulously evaluates the causation between the IDV and DV variables, while the original data can give false conclusions about that causation.

When looking at the results selection criteria, there might have been some results being discarded erroneously because the minimum suggested effect definitions (cf. Table 5) were considered as the highest possible device error according to the literature. Because multiple devices were used to collect the data, there is a possibility that some of the data originating from a more accurate device still got rejected according to the minimum value inclusion thresholds defined in this study.

Table 11. Comparison between the outcome change previewed by the a priori hypotheses, obtained by the results selected using the MDE criterion, and the change observed using the naive method for H4, H6, H7, and H9 (note: only applies to

participant KW). Refer to Table 8 for details on the Hypotheses. For all the hypotheses, the naive method resulted in the same suggested effect direction, whether controlling for B-MONTH or B-SEASON.

	A priori	Results with MDE	Results with naïve method
Hypothesis	Outcome change	Outcome change	Outcome change
H4	Increase	Increase	Decrease
H6	Decrease	Increase	Increase almost always (decrease with 10 days of lag)
H7	Decrease	Increase	Decrease
H9	Increase	Decrease	Increase

The main goal of the Testing phase was to test if the model used in MoTR could be used in future data, that is, data in which the model would have never been trained/fitted. To this analysis, we assessed two results: (1) the model fitting; (2) the model estimates for the new data. Firstly, as shown in Table 9, the R-squared in A values were almost always negative except for H9 controlling for B-MONTH. That means that the estimates of the model for the new data (part A) were shifted from the actual original new data (original values of part A), thus revealing that even with the model being accurate in estimating the new data values, adjustments must be made to obtain them in the right range of values. The smallest differences between the R-squared in B and the R-squared in A are for H9 (controlling for B-MONTH and B-SEASON) and H4 (controlling for B-SEASON), all having a difference below 0.400. These three results were selected based on an arbitrary decision of including half of the total amount of results, although this can be shifted to suit the researcher's needs. Second, to evaluate the model estimates for the new data, we should compare the direction of the suggested effect measured by MoTR in the new data with the result shown in Table 8. Picking only the H4 and H9 selected as previously detailed, the IDV suggested effect on DV for H4 was a decreasing of the DV, that is, the contrary direction of the suggested effect estimated by the model for the type B data (in which it was trained). A note should be made about the direction of the suggested effect calculated using the naive method for the H4, which increases the DV identically to the results in data of type B. For H9, both the naive and the MoTR methods estimated a negative suggested effect of IDV, that is, a decreasing of the DV, consistent with the results in type B data. However, all the *P* values of H4 and H9 testing phase results were higher than .05, making those results not statistically discernible.

From the fourth and last phase of results, the Confirmatory, only the H2 and H5 were supported, although for different participants (H2 for KW and H5 for IM) and never with a discernible result, that is, always with a suggested effect estimated below the minimum suggested effect inclusion criteria (cf. Table 5 - 0.035 meters for KW's step length and 2 BPM for all heart rate values).

Because this is an observational study and not an interventional one, no causality can be strongly concluded but only hypothesized, using the presented MoTR method to do so.

Potential Unobserved Confounding Variables

Additionally, there is a possible explanation for not obtaining any statistically discernible result for KW's nor EJD's data on the Exploratory A phase. Because the used data of type A was collected from 14 August 2020 until 8 January 2021 (cf. Subsection "Participants and Collected Data Definition"), the lockdown due to the COVID-19 pandemic may have been a confounder that this study did not account for. Three types of behavior occurred with the three participants: KW had a relatively active daily life, walking to work every day as usual; EJD was in lockdown for approximately half of the data collection time; IM was in lockdown during almost all the days of data collection. That may help justify why EJD's data did not show the needed consistency for the results to be statistically discernible. As for KW's data, although she did not stop the normal daily physical activity involved in commuting to work every day, there may have been context changes influencing the measurements made during that period, and for which this study did not account for, like for example changes in social interaction, travels, or in exercise patterns.

Finally, we acknowledge some possible confounders that this study did not account for when applying the methods. The country where KW lived in might have impacted her data, as in study period B (February 2017 – August 2020) she was moving every few months between Denmark, Switzerland, and the United States of America (mostly during Summer), that have had influenced her overall lifestyle patterns, the sleep and steps taken per day and exercising, but also patterns of nutritional intake that influences the metabolism and hence the HR patterns. For EJD's and IM's data, it is possible that external factors might have influenced the collected data in addition to lockdown, namely alcohol intake, late meals, and visual and psychological stimulation (e.g., watching movies, working until late, mobile devices used before or in bed) that were not measured and can influence HR levels and the TST.

Applicability of the Method

The N-of-1 Observational Study method presented here provides a new tool to interpret self-collected data and correlate it with daily-life stressors. Specifically, the MoTR method is presented as a new tool to assess potential causality using intensive longitudinal data ((96–98)).

One use case of the MoTR method is to help develop N1RTs for diagnosis or intervention/treatment. After applying it to data, researchers can use this method to select findings with the largest statistically discernible differences from the naive. That will indicate possible confounding variables that can change decisions on a future intervention. Finally, MoTR can also help discuss a possible intervention plan by analyzing the highest potential causality between intensive longitudinal data and the outcome expected to be affected during a study.

Thus, this novel method is intended to be used with one's data. However, the best practice would be to (a) use occasional self-screening mental health questionnaires and (b) work with a health expert to analyze the data and conclude about them.

Conclusions and Future Work Areas

Self-tracking devices are nowadays very common and used for a multitude of purposes. Most of those are used to track sleep, heart rate, and physical activity data

only to enable a generic self-perception of one's daily behaviors and state of the body. They are a ubiquitous, simple, and useful tool to conclude about the individual's behavior patterns. Additionally, if used longitudinally, they can also enable the acquisition of the datasets that can further help understand how certain behaviors and external factors (such as stressors) impact the physiology and functioning of the body. Therefore, this study implemented the MoTR method evaluating the suggested effects of daily stressors on nighttime heart rate, sleep time, and physical activity in an individualized way: via the N-of-1 approach. For one of the three participants (IM), we found that physical activity can increase the nighttime heart rate amplitude, whereas there are no strong conclusions about its suggested effect on total sleep time. For one of the other participants (KW), socializing, yoga, and self-reported exercise ~~were associated with~~may have increased the nighttime heart rate. On contrary, being tires/sick/stressed (a self-reported state) may have decreased the nighttime heart rate~~, which decreased when the participant self-reported being tired/sick/stressed~~.
Our study had the following limitations. The interval of collection of data of type A might have been too short to accurately evaluate the suggested effect of the selected IDVs (approximately only five months long, under changing seasons and variable COVID-19 conditions). There were only self-reported and wearable-collected data daily, thus losing any detail that an intra-day sampling might have provided. The data regarding physical activity—steps—had no detail of the moment of the day. Thus, the results in this paper only focus on the possible effect of the aggregate total number of steps rather than the steps taken during, for example, the morning, afternoon, or night periods. This might also have caused the differences between the two groups of control data when estimating the effect of SAT on TST. The self-reported (except stress) data were coded by the user (KW) at the moment of this study deployment (early/mid-2021), possibly containing a bias based on unclear calendar notes/events for 2017-2020. Specifically, when collecting stress data, KW did not account for its use in this study, using it asmomentary week-to-week management of health and work/life balance. Thus, a minimal bias is also expected in this data. The lockdown during the COVID-19 pandemic might have influenced the measured suggested effects of the IDV's, especially for one of the three participants (IM) who was in an almost full lockdown during the collection period for data of type A. For example, this lockdown might have interfered with the participant's sleep-wake regimen before its start.
Additionally, many other confounders likely influenced the variables measured in both types of data (self-reported and wearable-collected).
Future studies shall focus on collecting, analyzing, and modeling intraday data for the hypotheses stated during this study. A longer data collection period would be beneficial too. We will also assess multilevel models for this same approach. Data collected before the COVID-19 lockdown could help understand this paper's suggested effects by comparing the conclusions before and after that moment. Finally, a randomized control trial should also be conducted to test the causality of variables suggested by the methods presented here.
Additionally, in a future study, rather than using the AIC, BIC, and F statistic to perform model selection in order to meet the MBL criterion (cf. Subsection "Results Selection Criteria" and SUPPLEMENTARY MATERIALS AND FORMULAS), we may instead apply k-fold cross-validation. For example, we might conduct leave-one-out cross-validation

by first calculating each model's *predicted residual sum-of-squares* (PRESS) statistic (99), and then selecting the model with the highest PRESS statistic. We would then fit this model on all training data and use its estimated coefficients to predict values in any new (i.e., test) dataset.

DECLARATIONS

Conflicting Interests

EJD works for as a Lead Biostatistician in Data Science. The other authors declare no conflicts of interest.

Funding

KW efforts were supported by internal grants from the University of Copenhagen (Denmark) and University of Geneva (Switzerland), as well as the H2020 WellCo Project (769765), AAL GUARDIAN project (AAL-2019-6-120-CP), Data+ AI@CARE, swissuniversities AGE-INT project, and COST action ReMO (CA19117). IM efforts were supported by internal grants from the University of Geneva (Switzerland).

Ethical approval

All data analyzed were collected using personal wearable devices from three participants. The three participants are the three authors of this research, who consented to data usage for this research's purposes before its analysis. The research done by the Quality of Life Technologies Lab has been approved by the "La Commission D'Ethique de La Recherche de L'Universite de Geneve" (CUREG).

Guarantor

IM

Contributorship

Conceptualization, I.M., E.J.D. and K.W.; Data curation, I.M. and E.J.D.; Formal analysis, I.M., E.J.D. and K.W.; Funding acquisition, E.J.D. and K.W.; Investigation, I.M., E.J.D. and K.W.; Methodology, I.M. and E.J.D.; Project Administration, I.M., E.J.D. and K.W.; Resources, I.M., E.J.D. and K.W.; Software, I.M. and E.J.D.; Supervision, I.M., E.J.D. and K.W.; Validation, I.M., E.J.D. and K.W.; Visualization, I.M. and E.J.D.; Writing, I.M., E.J.D. and K.W. All authors have read and agreed to the published version of the manuscript.

Acknowledgment

We thank the Reviewers for taking the time and effort necessary to review the manuscript. We sincerely appreciate all valuable comments and suggestions, which helped us to improve the manuscript's quality.

REFERENCES

1. Gary Wolf. Know Thyself : Tracking Every Facet of Life , from Sleep to Mood to Pain ,. Wired Mag 365 [Internet]. 2013 [cited 2021 Jun 28];40–2. Available from: https://www.wired.com/2009/06/lbnp-knowthyself/?currentPage=all
2. Wolf G. The Data-Driven Life - The New York Times. New York Times [Internet]. 2010 [cited 2021 Jun 28];17:1–18. Available from: https://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html?pagewanted=all&_r=0
3. Wac K. From Quantified Self to Quality of Life. In 2018. p. 83–108.
4. Wac K, Tsiourti C. Ambulatory assessment of affect: Survey of sensor systems for monitoring of autonomic nervous systems activation in emotion. IEEE Trans Affect Comput. 2014 Jul 1;5(3):251–72.
5. Guyatt GH, Keller JL, Jaeschke R, Rosenbloom D, Adachi JD, Newhouse MT. The n-of-1 randomized controlled trial: Clinical usefulness. Our three-year experience. Ann Intern Med. 1990;112(4):293–9.
6. McKinley WO, Johns JS, Musgrove JJ. Clinical presentations, medical complications, and functional outcomes of individuals with gunshot wound-induced spinal cord injury. Am J Phys Med Rehabil [Internet]. 1999 Mar [cited 2021 Jul 1];78(2):102–7. Available from: https://pubmed.ncbi.nlm.nih.gov/10088582/
7. Athavale Y, Krishnan S. Biosignal monitoring using wearables: Observations and opportunities. Vol. 38, Biomedical Signal Processing and Control. Elsevier; 2017. p. 22–33.
8. Matias I, Garcia N, Pirbhulal S, Felizardo V, Pombo N, Zacarias H, et al. Prediction of Atrial Fibrillation using artificial intelligence on Electrocardiograms: A systematic review. Vol. 39, Computer Science Review. Elsevier; 2021. p. 100334.
9. Johansen CD, Olsen RH, Pedersen LR, Kumarathurai P, Mouridsen MR, Binici Z, et al. Resting, night-time, and 24 h heart rate as markers of cardiovascular risk in middle-aged and elderly men and women with no apparent heart disease. Eur Heart J [Internet]. 2013 Jun 14 [cited 2021 Jun 28];34(23):1732–9. Available from: https://academic.oup.com/eurheartj/article/34/23/1732/425158
10. Daza EJ, Schneider L. Model-Twin Randomization (MoTR): A Monte Carlo Method for Estimating the Within-Individual Average Treatment Effect Using Wearable Sensors. Prepr Prog. 2022;
11. Margolis A, Giuliano C. Making the switch: From case studies to N-of-1 trials. Epilepsy Behav Reports [Internet]. 2019 [cited 2021 Dec 13];12:100336. Available from: /pmc/articles/PMC6854058/
12. Nikles J, Mitchell G. The Essential Guide to N-of-1 Trials in Health. The Essential Guide to N-of-1 Trials in Health. 2015.
13. Myllymäki T, Kyröläinen H, Savolainen K, Hokka L, Jakonen R, Juuti T, et al. Effects of vigorous late-night exercise on sleep quality and cardiac autonomic activity. J Sleep Res [Internet]. 2011 Mar 1 [cited 2021 Jul 1];20(1 PART II):146–53. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2869.2010.00874.x
14. Faust L, Feldman K, Mattingly SM, Hachen D, V. Chawla N. Deviations from

normal bedtimes are associated with short-term increases in resting heart rate. npj Digit Med [Internet]. 2020 Dec 1 [cited 2021 Jul 1];3(1):1–9. Available from: https://doi.org/10.1038/s41746-020-0250-6

15. Valentini M, Parati G. Variables Influencing Heart Rate. Prog Cardiovasc Dis. 2009 Jul 1;52(1):11–9.
16. Li X, Dunn J, Salins D, Zhou G, Zhou W, Schüssler-Fiorenza Rose SM, et al. Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information. PLoS Biol [Internet]. 2017 Jan 12 [cited 2021 Jul 1];15(1):2001402. Available from: https://www.va.gov/
17. Kravitz RL, Duan N, DEcIDE Methods Centre. Design and Implementation of N-of-1 Trials: A User's Guide | Effective Health Care [Internet]. AHRQ Publication No. 13(14)-EHC122-EF. Rockville, MD: Agency for Healthcare Research and Quality. 2014 [cited 2021 Mar 20]. Available from: https://effectivehealthcare.ahrq.gov/products/n-1-trials/research-2014-5
18. Guyatt G, Sackett D, Taylor DW, Ghong J, Roberts R, Pugsley S. Determining Optimal Therapy — Randomized Trials in Individual Patients. N Engl J Med [Internet]. 1986 Apr 3 [cited 2021 Mar 20];314(14):889–92. Available from: http://www.nejm.org/doi/abs/10.1056/NEJM198604033141406
19. Lillie EO, Patay B, Diamant J, Issell B, Topol EJ, Schork NJ. The n-of-1 clinical trial: The ultimate strategy for individualizing medicine? Per Med [Internet]. 2011 Mar [cited 2021 Mar 20];8(2):161–73. Available from: https://pubmed.ncbi.nlm.nih.gov/21695041/
20. Chen C, Haddad D, Selsky J, Hoffman JE, Kravitz RL, Estrin DE, et al. Making sense of mobile health data: An open architecture to improve individual- and population-level health. J Med Internet Res [Internet]. 2012 [cited 2021 Mar 20];14(4). Available from: https://pubmed.ncbi.nlm.nih.gov/22875563/
21. Vohra S, Shamseer L, Sampson M, Bukutu C, Schmid CH, Tate R, et al. CONSORT extension for reporting N-of-1 trials (CENT) 2015 Statement. BMJ [Internet]. 2015 May 14 [cited 2021 Mar 20];350:h1738. Available from: http://www.bmj.com/
22. Daza EJ. Causal Analysis of Self-tracked Time Series Data Using a Counterfactual Framework for N-of-1 Trials. Methods Inf Med [Internet]. 2018 [cited 2021 Nov 22];57:e10–21. Available from: https://doi.org/10.3414/ME16-02-0044
23. McDonald S, Quinn F, Vieira R, O'Brien N, White M, Johnston DW, et al. The state of the art and future opportunities for using longitudinal n-of-1 methods in health behaviour research: a systematic literature overview. Health Psychol Rev [Internet]. 2017 Oct 2 [cited 2021 Jun 28];11(4):307–23. Available from: https://pubmed.ncbi.nlm.nih.gov/28406349/
24. Hobbs N, Dixon D, Johnston M, Howie K. Can the theory of planned behaviour predict the physical activity behaviour of individuals? Psychol Heal. 2013;28(3):234–49.
25. Johnston DW, Johnston M. Useful theories should apply to individuals. Br J Health Psychol. 2013 Sep;18(3):469–73.
26. Sutton C. Developing and evaluating complex interventions. Matern Child Nutr [Internet]. 2014 [cited 2021 Jun 28];10(2):163–5. Available from: www.mrc.ac.uk/complexinterventionsguidance

27. Quinn F, Johnston M, Johnston DW. Testing an integrated behavioural and biomedical model of disability in N-of-1 studies with chronic pain. Psychol Heal. 2013 Dec;28(12):1391–406.
28. McDonald S, Vieira R, Godfrey A, O'Brien N, White M, Sniehotta FF. Changes in physical activity during the retirement transition: A series of novel n-of-1 natural experiments. Int J Behav Nutr Phys Act [Internet]. 2017 Dec 8 [cited 2021 Jun 28];14(1). Available from: https://pubmed.ncbi.nlm.nih.gov/29221449/
29. Fox K, Borer JS, Camm AJ, Danchin N, Ferrari R, Lopez Sendon JL, et al. Resting Heart Rate in Cardiovascular Disease. J Am Coll Cardiol. 2007 Aug 28;50(9):823–30.
30. Nakagawa M, Iwao T, Ishida S, Yonemochi H, Fujino T, Saikawa T, et al. Circadian rhythm of the signal averaged electrocardiogram and its relation to heart rate variability in healthy subjects. Heart [Internet]. 1998 [cited 2022 Jan 13];79(5):493–6. Available from: https://pubmed.ncbi.nlm.nih.gov/9659198/
31. Kristal-Boneh E, Harari G, Weinstein Y, Green MS. Factors affecting differences in supine, sitting, and standing heart rate: The Israeli CORDIS study. Aviat Sp Environ Med. 1995;66(8):775–9.
32. Kovar D, Cannon CP, Bentley JH, Charlesworth A, Rogers WJ. Does Initial and Delayed Heart Rate Predict Mortality in Patients with Acute Coronary Syndromes? Clin Cardiol [Internet]. 2004 [cited 2021 Jun 28];27(2):80–6. Available from: https://pubmed.ncbi.nlm.nih.gov/14979625/
33. Palatini P, Thijs L, Staessen JA, Fagard RH, Bulpitt CJ, Clement DL, et al. Predictive value of clinic and ambulatory heart rate for mortality in elderly subjects with systolic hypertension. Arch Intern Med [Internet]. 2002 Nov 15 [cited 2021 Jun 28];162(20):2313–21. Available from: https://pubmed.ncbi.nlm.nih.gov/12418945/
34. Jouven X, Empana J-P, Schwartz PJ, Desnos M, Courbon D, Ducimetière P. Heart-Rate Profile during Exercise as a Predictor of Sudden Death. N Engl J Med [Internet]. 2005 May 12 [cited 2021 Jun 28];352(19):1951–8. Available from: https://pubmed.ncbi.nlm.nih.gov/15888695/
35. Diaz A, Bourassa MG, Guertin MC, Tardif JC. Long-term prognostic value of resting heart rate in patients with suspected or proven coronary artery disease. Eur Heart J [Internet]. 2005 May [cited 2021 Jun 28];26(10):967–74. Available from: https://pubmed.ncbi.nlm.nih.gov/15774493/
36. Naghavi M, Abajobir AA, Abbafati C, Abbas KM, Abd-Allah F, Abera SF, et al. Global, regional, and national age-sex specifc mortality for 264 causes of death, 1980-2016: A systematic analysis for the Global Burden of Disease Study 2016. Lancet [Internet]. 2017 Sep 16 [cited 2021 Jun 28];390(10100):1151–210. Available from: https://vizhub.healthdata.org/
37. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual Causes of Death in the United States, 2000. J Am Med Assoc [Internet]. 2004 Mar 10 [cited 2021 Jun 28];291(10):1238–45. Available from: https://pubmed.ncbi.nlm.nih.gov/15010446/
38. Palatini P, Benetos A, Julius S. Impact of increased heart rate on clinical outcomes in hypertension: Implications for antihypertensive drug therapy.

Drugs [Internet]. 2006 Sep 17 [cited 2022 Jan 5];66(2):133–44. Available from: https://link.springer.com/article/10.2165/00003495-200666020-00001

39. Bonnemeier H, Wiegand UKH, Brandes A, Kluge N, Katus HA, Richardt G, et al. Circadian profile of cardiac autonomic nervous modulation in healthy subjects: Differing effects of aging and gender on heart rate variability. J Cardiovasc Electrophysiol [Internet]. 2003 Aug 1 [cited 2022 Jan 5];14(8):791–9. Available from: https://onlinelibrary.wiley.com/doi/full/10.1046/j.1540-8167.2003.03078.x
40. Chapman JH. Profound sinus bradycardia in the athletic heart syndrome. J Sports Med Phys Fitness. 1982;22(1):45–8.
41. Amagasa S, Kamada M, Sasai H, Fukushima N, Kikuchi H, Lee IM, et al. How well iphones measure steps in free-living conditions: Cross-sectional validation study. JMIR mHealth uHealth [Internet]. 2019 Jan 1 [cited 2022 Jan 5];7(1). Available from: https://pubmed.ncbi.nlm.nih.gov/30626569/
42. Reimers AK, Knapp G, Reimers CD. Effects of Exercise on the Resting Heart Rate: A Systematic Review and Meta-Analysis of Interventional Studies. J Clin Med [Internet]. 2018 Dec 1 [cited 2022 Jan 5];7(12). Available from: /pmc/articles/PMC6306777/
43. O'Toole ML. Gender differences in the cardiovascular response to exercise. Cardiovasc Clin [Internet]. 1989 [cited 2022 Jan 5];19(3):17–33. Available from: https://pubmed.ncbi.nlm.nih.gov/2644030/
44. Kredlow MA, Capozzoli MC, Hearon BA, Calkins AW, Otto MW. The effects of physical activity on sleep: a meta-analytic review. J Behav Med [Internet]. 2015 Jun 1 [cited 2022 Jan 5];38(3):427–49. Available from: https://pubmed.ncbi.nlm.nih.gov/25596964/
45. Vrijkotte TGM, Van Doornen LJP, De Geus EJC. Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability. Hypertension [Internet]. 2000 [cited 2022 Jan 5];35(4):880–6. Available from: https://www.ahajournals.org/doi/abs/10.1161/01.HYP.35.4.880
46. Schnall PL, Schwartz JE, Landsbergis PA, Warren K, Pickering TG. Relation between job strain, alcohol, and ambulatory blood pressure. Hypertension [Internet]. 1992 [cited 2022 Jan 5];19(5):488–94. Available from: https://pubmed.ncbi.nlm.nih.gov/1568768/
47. Schnall PL, Schwartz JE, Landsbergis PA, Warren K, Pickering TG. A longitudinal study of job strain and ambulatory blood pressure: Results from a three-year follow-up. Psychosom Med [Internet]. 1998 [cited 2022 Jan 5];60(6):697–706. Available from: https://pubmed.ncbi.nlm.nih.gov/9847028/
48. Chen XJ, Barywani SB, Hansson PO, Östgärd Thunström E, Rosengren A, Ergatoudes C, et al. Impact of changes in heart rate with age on all-cause death and cardiovascular events in 50-year-old men from the general population. Open Hear [Internet]. 2019 Mar 1 [cited 2022 Jan 5];6(1):e000856. Available from: https://openheart.bmj.com/content/6/1/e000856
49. Vazir A, Claggett B, Jhund P, Castagno D, Skali H, Yusuf S, et al. Prognostic importance of temporal changes in resting heart rate in heart failure patients: An analysis of the CHARM program. Eur Heart J [Internet]. 2015 Mar 14 [cited 2022 Jan 5];36(11):669–75. Available from:

https://academic.oup.com/eurheartj/article/36/11/669/492036
50. Münzel T, Hahad O, Gori T, Hollmann S, Arnold N, Prochaska JH, et al. Heart rate, mortality, and the relation with clinical and subclinical cardiovascular diseases: results from the Gutenberg Health Study. Clin Res Cardiol [Internet]. 2019 Dec 1 [cited 2022 Jan 5];108(12):1313–23. Available from: https://link.springer.com/article/10.1007/s00392-019-01466-2
51. Hirshkowitz M, Whiton K, Albert SM, Alessi C, Bruni O, DonCarlos L, et al. National sleep foundation's sleep time duration recommendations: Methodology and results summary. Sleep Heal [Internet]. 2015 Mar 1 [cited 2022 Jan 5];1(1):40–3. Available from: https://pubmed.ncbi.nlm.nih.gov/29073412/
52. Paruthi S, Brooks LJ, D'Ambrosio C, Hall WA, Kotagal S, Lloyd RM, et al. Recommended amount of sleep for pediatric populations: A consensus statement of the American Academy of Sleep Medicine. J Clin Sleep Med [Internet]. 2016 [cited 2022 Jan 5];12(6):785–6. Available from: /pmc/articles/PMC4877308/
53. Watson NF, Badr MS, Belenky G, Bliwise DL, Buxton OM, Buysse D, et al. Recommended amount of sleep for a healthy adult: A joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society. In: Sleep [Internet]. Oxford University Press; 2015 [cited 2022 Jan 5]. p. 843–4. Available from: /pmc/articles/PMC4434546/
54. Ibáñez V, Silva J, Cauli O. A survey on sleep assessment methods. PeerJ [Internet]. 2018 May 25 [cited 2022 Jan 5];2018(5):e4849. Available from: https://peerj.com/articles/4849
55. Perfect MM, Beebe DW, Levine-Donnerstein D, Frye SS, Bluez GP, Quan SF. The development of a clinically relevant sleep modification protocol for youth with type 1 diabetes. Clin Pract Pediatr Psychol [Internet]. 2016 [cited 2022 Jan 5];4(2):227–40. Available from: https://psycnet.apa.org/journals/cpp/4/2/227
56. Chaput JP, Bouchard C, Tremblay A. Change in sleep duration and visceral fat accumulation over 6 years in adults. Obesity [Internet]. 2014 May 1 [cited 2022 Jan 5];22(5):E9–12. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/oby.20701
57. Salwen JK, Smith MT, Finan PH. Mid-treatment sleep duration predicts clinically significant knee osteoarthritis pain reduction at 6 months: Effects from a behavioral sleep medicine clinical trial. Sleep [Internet]. 2017 [cited 2022 Jan 5];40(2). Available from: /pmc/articles/PMC6251549/
58. Hublin C, Haasio L, Kaprio J. Changes in self-reported sleep duration with age - a 36-year longitudinal study of Finnish adults. BMC Public Health [Internet]. 2020 Sep 9 [cited 2022 Jan 5];20(1):1–8. Available from: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-020-09376-z
59. Cepeda MS, Stang P, Blacketer C, Kent JM, Wittenberg GM. Clinical relevance of sleep duration: Results from a cross-sectional analysis using NHANES. J Clin Sleep Med [Internet]. 2016 [cited 2022 Jan 5];12(6):813–9. Available from: https://jcsm.aasm.org/doi/abs/10.5664/jcsm.5876
60. Chaudhry UAR, Wahlich C, Fortescue R, Cook DG, Knightly R, Harris T. The effects of step-count monitoring interventions on physical activity: Systematic

review and meta-analysis of community-based randomised controlled trials in adults. Int J Behav Nutr Phys Act [Internet]. 2020 Oct 9 [cited 2022 Jan 5];17(1):1–16. Available from: https://ijbnpa.biomedcentral.com/articles/10.1186/s12966-020-01020-8

61. Shcherbina A, Hershman SG, Lazzeroni L, King AC, O'Sullivan JW, Hekler E, et al. The effect of digital physical activity interventions on daily step count: a randomised controlled crossover substudy of the MyHeart Counts Cardiovascular Health Study. Lancet Digit Heal [Internet]. 2019 Nov 1 [cited 2022 Jan 5];1(7):e344–52. Available from: http://www.thelancet.com/article/S2589750019301293/fulltext
62. Polgar O, Patel S, Walsh JA, Barker RE, Clarke SF, Man WD-C, et al. Minimal clinically important difference for daily pedometer step count in COPD. ERJ Open Res [Internet]. 2021 Jan 1 [cited 2022 Jan 5];7(1):00823–2020. Available from: https://openres.ersjournals.com/content/7/1/00823-2020
63. Demeyer H, Burtin C, Hornikx M, Camillo CA, Van Remoortel H, Langer D, et al. The minimal important difference in physical activity in patients with COPD. PLoS One [Internet]. 2016 Apr 1 [cited 2022 Jan 5];11(4). Available from: /pmc/articles/PMC4849755/
64. Gresham G, Hendifar AE, Spiegel B, Neeman E, Tuli R, Rimel BJ, et al. Wearable activity monitors to assess performance status and predict clinical outcomes in advanced cancer patients. npj Digit Med [Internet]. 2018 Jul 5 [cited 2022 Jan 5];1(1):1–8. Available from: https://www.nature.com/articles/s41746-018-0032-6
65. Kaleth AS, Slaven JE, Ang DC. Does Increasing Steps Per Day Predict Improvement in Physical Function and Pain Interference in Adults With Fibromyalgia? Arthritis Care Res (Hoboken) [Internet]. 2014 Dec 1 [cited 2022 Jan 5];66(12):1887–94. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/acr.22398
66. Hall KS, Hyde ET, Bassett DR, Carlson SA, Carnethon MR, Ekelund U, et al. Systematic review of the prospective association of daily step counts with risk of mortality, cardiovascular disease, and dysglycemia. Int J Behav Nutr Phys Act [Internet]. 2020 Jun 20 [cited 2022 Jan 5];17(1):1–14. Available from: https://ijbnpa.biomedcentral.com/articles/10.1186/s12966-020-00978-9
67. Kraus WE, Janz KF, Powell KE, Campbell WW, Jakicic JM, Troiano RP, et al. Daily Step Counts for Measuring Physical Activity Exposure and Its Relation to Health. Med Sci Sports Exerc [Internet]. 2019 [cited 2022 Jan 5];51(6):1206–12. Available from: http://links.lww.
68. Boyer ER, Derrick TR. Select Injury-Related Variables Are Affected by Stride Length and Foot Strike Style during Running. Am J Sports Med [Internet]. 2015 Sep 3 [cited 2022 Jan 5];43(9):2310–7. Available from: https://journals.sagepub.com/doi/full/10.1177/0363546515592837
69. Hannink J, Kautz T, Pasluosta CF, Barth J, Schulein S, Gabmann KG, et al. Mobile Stride Length Estimation with Deep Convolutional Neural Networks. IEEE J Biomed Heal Informatics. 2018 Mar 1;22(2):354–62.
70. Rampp A, Barth J, Schülein S, Gaßmann KG, Klucken J, Eskofier BM. Inertial Sensor-Based Stride Parameter Calculation From Gait Sequences in Geriatric

Patients. IEEE Trans Biomed Eng. 2015 Apr 1;62(4):1089–97.
71. Chevance G, Baretta D, Romain AJ, Godino J, Bernard P. Day-to-day associations between sleep and physical activity: a set of person-specific analyses in adults with overweight and obesity. 2021 [cited 2021 Jun 28]; Available from: https://osf.io/preprints/sportrxiv/nfjqv/
72. Kühberger A, Fritz A, Lermer E, Scherndl T. The significance fallacy in inferential statistics Psychology. BMC Res Notes [Internet]. 2015 [cited 2022 Mar 14];8(1). Available from: https://pubmed.ncbi.nlm.nih.gov/25888971/
73. Silva-Ayçaguer LC, Surez-Gil P, Fernndez-Somoano A. The null hypothesis significance test in health sciences research (1995-2006): Statistical analysis and interpretation. BMC Med Res Methodol [Internet]. 2010 May 19 [cited 2022 Mar 14];10(1):1–9. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-10-44
74. Barnett ML, Mathisen A. Tyranny of the p-Value: The Conflict between Statistical Significance and Common Sense [Internet]. Vol. 76, Journal of Dental Research. J Dent Res; 1997 [cited 2022 Mar 14]. p. 534–6. Available from: https://pubmed.ncbi.nlm.nih.gov/9042074/
75. Gelman A, Stern H. The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. Am Stat [Internet]. 2006 [cited 2022 Mar 14];60(November). Available from: www.ics.uci.edu/
76. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance [Internet]. Vol. 567, Nature. Nature Publishing Group; 2019 [cited 2022 Mar 14]. p. 305–7. Available from: https://www.nature.com/articles/d41586-019-00857-9
77. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon Statistical Significance. Am Stat [Internet]. 2019 Mar 29 [cited 2022 Mar 14];73(sup1):235–45. Available from: https://www.tandfonline.com/doi/abs/10.1080/00031305.2018.1527253
78. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p < 0.05" [Internet]. Vol. 73, American Statistician. Taylor & Francis; 2019 [cited 2022 Mar 14]. p. 1–19. Available from: https://www.tandfonline.com/doi/abs/10.1080/00031305.2019.1583913
79. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose [Internet]. Vol. 70, American Statistician. American Statistical Association; 2016 [cited 2022 Mar 14]. p. 129–33. Available from: https://www.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108
80. Daza EJ, Wac K, Oppezzo M. Effects of sleep deprivation on blood glucose, food cravings, and affect in a non-diabetic: An n-of-1 randomized pilot study. Healthc [Internet]. 2020 Dec 25 [cited 2022 Mar 14];8(1):6. Available from: https://www.mdpi.com/2227-9032/8/1/6/htm
81. Granger CWJ. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica. 1969 Aug;37(3):424.
82. de Zambotti M, Goldstone A, Claudatos S, Colrain IM, Baker FC. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. Chronobiol Int [Internet]. 2018 Apr 3 [cited 2021 Jun 25];35(4):465–76.

Available from: https://pubmed.ncbi.nlm.nih.gov/29235907/

83. Menghini L, Yuksel D, Goldstone A, Baker FC, de Zambotti M. Performance of Fitbit Charge 3 against polysomnography in measuring sleep in adolescent boys and girls. Chronobiol Int [Internet]. 2021 [cited 2021 Jun 25];38(7):1010–22. Available from: https://www.tandfonline.com/action/journalInformation?journalCode=icbi20
84. Roomkham S, Hittle M, Lovell D, Perrin D. Can we use the Apple Watch to measure sleep reliably? J Sleep Res. 2018 Oct;27:e153_12766.
85. Veerabhadrappa P, Moran MD, Renninger MD, Rhudy MB, Dreisbach SB, Gift KM. Tracking Steps on Apple Watch at Different Walking Speeds. J Gen Intern Med [Internet]. 2018 Jun 1 [cited 2021 Jun 25];33(6):795–6. Available from: https://doi.org/10.1371/journal.pone.0154420
86. Nelson BW, Allen NB. Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: Intraindividual validation study. JMIR mHealth uHealth [Internet]. 2019 Mar 1 [cited 2021 Jun 30];7(3):e10828. Available from: https://mhealth.jmir.org/2019/3/e10828
87. Muggeridge DJ, Hickson K, Davies AV, Giggins OM, Megson IL, Gorely T, et al. Measurement of heart rate using the polar OH1 and fitbit charge 3 wearable devices in healthy adults during light, moderate, vigorous, and sprint-based exercise: Validation study. JMIR mHealth uHealth [Internet]. 2021 Mar 1 [cited 2021 Jun 25];9(3):e25313. Available from: https://mhealth.jmir.org/2021/3/e25313
88. Wahl Y, Düking P, Droszez A, Wahl P, Mester J. Criterion-validity of commercially available physical activity tracker to estimate step count, covered distance and energy expenditure during sports conditions. Front Physiol [Internet]. 2017 Sep 22 [cited 2021 Jun 25];8(SEP):725. Available from: www.frontiersin.org
89. Gilgen-Ammann R, Schweizer T, Wyss T. Accuracy of distance recordings in eight positioning-enabled sport watches: Instrument validation study. JMIR mHealth uHealth [Internet]. 2020 Jun 1 [cited 2021 Jun 25];8(6):e17118. Available from: https://mhealth.jmir.org/2020/6/e17118
90. Splawa-Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat Sci [Internet]. 1990 Nov 1 [cited 2022 Jan 5];5(4):465–72. Available from: https://projecteuclid.org/journals/statistical-science/volume-5/issue-4/On-the-Application-of-Probability-Theory-to-Agricultural-Experiments-Essay/10.1214/ss/1177012031.full
91. Holland PW. Statistics and Causal Inference. J Am Stat Assoc. 1986 Dec;81(396):945.
92. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688–701.
93. Little RJA, Rubin DB. Statistical analysis with missing data. Statistical Analysis with Missing Data. Wiley; 2014. 1–381 p.
94. Rubin DB. Inference and Missing Data. Biometrika. 1976 Dec;63(3):581.
95. Simpson EH. The Interpretation of Interaction in Contingency Tables. J R Stat Soc Ser B [Internet]. 1951 [cited 2021 Dec 14];13(2):238–41. Available from: https://about.jstor.org/terms

96. Walls TA, Schafer JL. Models for Intensive Longitudinal Data. Models for Intensive Longitudinal Data. Oxford University Press; 2012. 1–310 p.
97. Tan X, Shiyko MP, Li R, Li Y, Dierker L. A time-varying effect model for intensive longitudinal data. Psychol Methods [Internet]. 2012 Mar [cited 2021 Dec 14];17(1):61–77. Available from: /pmc/articles/PMC3288551/
98. Timms KP, Martin CA, Rivera DE, Hekler EB, Riley W. Leveraging intensive longitudinal data to better understand health behaviors. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014. Institute of Electrical and Electronics Engineers Inc.; 2014. p. 6888–91.
99. Allen DM. The Relationship between Variable Selection and Data Agumentation and a Method for Prediction. Technometrics. 1974 Feb;16(1):125.
100. Rossum G Van, Drake FL. Python Reference Manual. October [Internet]. 2006 [cited 2021 Dec 15];22:9117–29. Available from: http://www.cs.cmu.edu/afs/cs.cmu.edu/project/gwydion-1/OldFiles/OldFiles/python/Doc/ref.ps
101. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature [Internet]. 2020 Sep 16 [cited 2021 Dec 15];585(7825):357–62. Available from: https://www.nature.com/articles/s41586-020-2649-2
102. McKinney W. Data Structures for Statistical Computing in Python. In: Proceedings of the 9th Python in Science Conference. 2010. p. 56–61.
103. Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos Pedregosa, Varoquaux, Gramfort et al. Matthieu Perrot. J Mach Learn Res [Internet]. 2011 [cited 2021 Dec 15];12:2825–30. Available from: http://scikit-learn.sourceforge.net.
104. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007 May 1;9(3):90–5.
105. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods [Internet]. 2020 Feb 3 [cited 2021 Dec 15];17(3):261–72. Available from: https://www.nature.com/articles/s41592-019-0686-2

SUPPLEMENTARY MATERIALS AND FORMULAS

The dataset generated and analyzed during this study is available upon reasonable request from the *Yareta* repository, the research data repository of Geneva's Higher Education Institutions.
The organization of the dataset was done using *Microsoft Office Excel* software.
The analysis code in *Python* language (100) is available at https://gitfront.io/r/IgorMatias3/E34WEsuwBnrR/MoTR-python/ (to be replaced by a public repository when article accepted for publication).
The following *Python* language packages were used, aside from its standard libraries. *NumPy* ((101)), *Pandas* ((102)), *Scikit-Learn* ((103)), *Matplotlib* ((104)), and *SciPy* ((105)).

Stopping Rule

Let j index each Monte Carlo simulation run corresponding to a randomly shuffled exposure vector. Let μ represent the true, unknown APTE, and let Y_j represent the statistically consistent (i.e., unbiased with larger and larger samples) APTE estimate we calculate at run j from our data, with true sample-to-sample variance $\sigma^2 = V(Y)$ and standard deviation σ.
Consider the model and the data used to fit the model as fixed. Specifically, suppose Y_j randomly varies only due to the random shuffling, conditional on the data and corresponding model parameters or metrics. Then the set of Y_j values are identically and independently distributed.
Let $\overline{Y}_k = \frac{1}{k}\sum_{j=1}^{k} Y_j$ represent the estimator for μ defined as the empirical mean of all the Y_j values for runs $j = 1$ through k, with true variance $V(\overline{Y}_k) = \frac{1}{k^2}\sum_{j=1}^{k} V(Y_j) = \frac{1}{k}V(Y_j) = \frac{\sigma^2}{k}$. Hence, the standard error of the estimator is $SE(\overline{Y}_k) = \sqrt{V(\overline{Y}_k)} = \frac{\sigma}{\sqrt{k}}$. Because we do not know σ, the true standard deviation of Y, we will estimate it using the sample standard deviation of the k APTE estimates, denoted $\hat{\sigma}_k$ with corresponding estimated standard error $\widehat{SE}(\overline{Y}_k) = \frac{\hat{\sigma}_k}{\sqrt{k}}$.
We defined the stopping rule using the empirical quantity $acv_k \widehat{SE}(\overline{Y}_k)/|\overline{Y}_k|$.For a sufficiently long time series, Slutsky's theorem can be used to show that acv_k approaches the true, unknown absolute coefficient of variation $(\frac{\sigma}{\sqrt{k}})/|\mu|$ — a quantity that standardizes the between-sample variation in APTEs based on the APTE size. We used the absolute value of μ because we wished to make standard comparisons based on APTE size, not direction.
Furthermore, acv_k approaches zero as k increases because the numerator $\frac{\sigma}{\sqrt{k}}$ approaches zero. Hence, this quantity could potentially stabilize within finite computing time as the number of MoTR simulation runs increased, provided the true variance σ^2 was finite. It could also be applied across a range of APTE sizes and variances thanks to its standardizing property. We chose it for this reason, and arbitrarily defined $acv_k < 0.5$ as indicating stability of APTE estimates by run k.

The final stopping rule had two conditions or criteria. We stopped the MoTR runs when one of the following two conditions were met: (1) $k \geq 1000$ and $acv_k < 0.5$, or (2) $k = 10000$ (to set a computational time limit on the algorithm).

R-squared Calculation

Let Y_i denote observed outcome $i = 1,...,n$, and $\hat{Y}_i$ denote the corresponding predicted outcome. Let $\overline{Y}_n = \sum_{i=1}^{n} Y_i$ denote the empirical overall mean outcome of the target dataset. The R-squared formula is $R^2 = 1 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)/\sum_{i=1}^{n}(Y_i - \overline{Y}_n)$.
To calculate the R-squared values, the following Python code was used.

*r_squared = 1 - sum((y-yhat)**2) / sum((y-numpy.mean(y))**2)*

Here, *y* and *yhat* represent Y_i and $\hat{Y}_i$, respectively. *numpy.mean* is the empirical mean function from the *Numpy* library.

Results Selection Using the MBL Criterion

Results Using the MBL Criterion

The MBL criterion selects the model that has the most statistical evidence for a suggested effect, fits the original observed data best, and estimates a suggested effect that differs most from its corresponding naive effect estimate. It is a way to distinguish correlation from causation by quantifying findings of correlation with those of causation, comparing them, and selecting the model with the greatest difference between the two.
The MBL criterion itself consists of three sub-criteria, including the MDE criterion for identifying the model with the most statistically supported evidence of a *suggested effect* (a finding that can be interpreted in terms of causation). However, the MDE procedure does not directly compare this effect estimate with the naive effect estimate simply calculated by comparing the mean observed outcomes between the two IDV levels (a finding that can only be interpreted in terms of correlation/association). Hence, the MBL criterion requires that two other sub-criteria be met when selecting a final model. The first is the "best-fitting model" sub-criterion, the "B" in MBL. Of all candidate models, the model that meets this sub-criterion explains the most variation in the outcome, measured using three goodness-of-fit metrics: AIC, BIC, and the omnibus F test *P* value.
The second is the "largest confounding influence" sub-criterion, the "L" in MBL. Of all candidate models, the model that meets this sub-criterion shows the largest difference in size between the naive effect estimate and the MoTR effect estimate. That is, of all candidate models, it has the largest absolute difference between the mean differences in outcome under low and high exposure, before and after applying MoTR.
Note that the magnitude of the three quantities used in each sub-criterion can vary greatly across both criteria and models. Hence, to be able to simultaneously compare all three values for each model with the three values of any other model, we needed to first standardize all three quantities. To do so, for each quantity, we divided the raw

value (e.g., MDE *P* value) by the difference between the minimum and maximum raw values of all candidate models. This ensured that all three sub-criteria used standardized values ranging from 0 to 1.

Exploratory Phases Results Using MBL Criterion

Table 12. Selected results from Exploratory A and Exploratory B phases using MBL criterion. Time values are in the format "minutes:seconds."

Exploratory A					
	H1	IM's data	Controlling for	A-MONTH	A-SEASON
			Lag	3 days	1 day
			IDV effect on DV	+ 17:47.047	-18:21.077
			t test *P* value	.034	.027
	H3	IM's data	Controlling for	A-MONTH	A-SEASON
			Lag	1 day	2 days
			IDV effect on DV	+ 5:61 BPM	+ 1.02 BPM
			t test *P* value	< .001	.093
Exploratory B					
	H4	KW's data	Controlling for	B-MONTH	B-SEASON
			Lag	3 days	3 days
			IDV effect on DV	+ 2.65 BPM	+ 5.64 BPM
			t test *P* value	.009	.002
	H9	KW's data	Controlling for	B-MONTH	B-SEASON
			Lag	4 days	3 days
			IDV effect on DV	- 4.17 BPM	- 6.79 BPM
			t test *P* value	.002	.001

Exploratory A: Suggested Effect of SAT on TST (H1)

Like the MDE criterion results, only IM's data provided statistically discernible results for the H1.

As presented by Table 12, the suggested effect of SAT above the daily average on the TST was positive when controlling for A-MONTH and negative when controlling for A-SEASON. These selected results show the same as using the MDE criterion, although the selected lag is different when controlling for A-MONTH.

Like for the MDE criterion, Table 13 shows the comparison between the same-lag results for the results above, as the two lag levels are not the same. Like for the first-used criteria, the inverse suggested effect of sleep depending on the control used is still present when considering another number of days as lag.

Table 13. Comparison of the selected results for H1 (Exploratory A), using MBL criterion, with its correspondents (same days of lag) on the other control type. Time values are in the format "minutes:seconds".

H1	IM's data	Controlling for	A-MONTH	A-SEASON

		Lag	3 days	3 days (not selected)
		IDV effect on DV	+ 17:47.047	- 4.58.967
		t test *P* value	.034	.054
	IM's data	Controlling for	A-MONTH	A-SEASON
		Lag	1 day (not selected)	1 day
		IDV effect on DV	+ 6:34.288	- 18:21.077
		t test *P* value	.048	.027

Exploratory A: Suggested Effect of SAT on DIF-HR (H3)

For the third hypothesis, Table 12 shows the selected lags according to the MBL criterion. Compared to the other used criteria, only when controlling for A-SEASON we find a different result, this time being the suggested effect approximately double and with a higher *P* value, which makes this selection not statistically discernible.

Exploratory B: Suggested Effect of Socializing on DIF-HR (H4)

The selected results for H4 differ from those obtained with the MDE criterion in the number of days for lag (6 days for the first criteria and 3 for this). Nevertheless, the suggested effect is approximately the same across both criteria. As presented by Table 12 the suggested effect measured is slightly less and higher when controlling both for B-MONTH and B-SEASON, respectively.

Exploratory B: Suggested Effect of Tired/Sick/Stress on nighttime HR (H9)

For the final selection of results according to the MBL criterion, hypothesis H9, the suggested effect of a state of tired/sick/stress on nighttime HR is negative like observed on the results from the first criteria.
The suggested effect increased in both controlling cases, although the MBL criterion led us to 4 and 3 days of lag instead of 10 and 9 like MDE, respectively, for B-MONTH and B-SEASON controls.

Table 14. Comparison of the a priori hypotheses and the results obtained using the MBL criterion. For H1 there were different results when controlling for A-MONTH (increase) and A-SEASON (decrease). TST stands for "total sleep time," SAT stands for "steps per awake time," and TSS stands for "tired/sick/stress."

	A priori				Results with MBL	
Hyp. / participant	Exposure	Exp. change	Outcome	Out. change	Out. change	Result
H1/IM	SAT	Increase	TST	Increase	Inc./Dec.	Inconclusive
H3/IM	SAT	Increase	DIF-HR	Decrease	Increase	Not supported
H4/KW	Socializing	Presence	DIF-HR	Increase	Increase	Supported
H9/KW	TSS	Presence	Nighttime HR	Increase	Decrease	Not supported

Table 14 presents a comparison between the a priori hypotheses from both phases Exploratory A and B and the results obtained following the MDE criterion.

Testing Phase for the Results Using MBL Criterion

Table 15. Testing phase's results for the hypotheses selected using the MBL criterion. MSE stands for "mean squared error."

H4	KW's data	Controlling for	B-MONTH	B-SEASON
		Lag	3 days	3 days
		R^2 in B	0.309	0.284
		MSE in B	0.002	0.002
		IDV effect on DV (naïve method)	0.14 BPM	0.14 BPM
		t test *P* value (naïve method)	.872	.872
		R^2 in A	- 0.171	- 0.127
		R^2 in B - R^2 in A	0.480	0.411
		MSE in A	0.001	0.001
		MSE in B - MSE in A	0.001	0.001
		IDV effect on DV (MoTR)	- 0.53 BPM	- 0.64 BPM
		t test *P* value (MoTR)	.370	.366
H9	KW's data	Controlling for	B-MONTH	B-SEASON
		Lag	4 days	3 days
		R^2 in B	0.297	0.260
		MSE in B	0.002	0.002
		IDV effect on DV (naïve method)	- 1.31 BPM	- 1.32 BPM
		t test *P* value (naïve method)	.141	.138
		R^2 in A	- 0.111	- 0.086
		R^2 in B - R^2 in A	0.408	0.346
		MSE in A	0.001	0.001
		MSE in B - MSE in A	0.001	0.001
		IDV effect on DV (MoTR)	- 1.41 BPM	- 0.67 BPM
		t test *P* value (MoTR)	.409	.380

This second subsection presents the results only for the models obtained from the results chosen using MBL (H4 and H9). Table 15 shows the metrics for both hypotheses.

Suggested Effect of Socializing on DIF-HR (H4)

When considering MDE selection criteria, the results of using the models selected using MBL show an almost null difference between the MSE in B and A. The t test *P* value is also notably smaller when applying the MoTR method (approximately 2.5 times) and always above .05. The suggested effect of the IDV is also positive when

applying the naive approach and negative when using the MoTR. None of the calculated suggested effects is higher than the minimum suggested effect defined in Table 5.

Suggested Effect of Tired/Sick/Stress on Nighttime HR (H9)

The fitting of the model in parts B and A considering H9 is almost similar between the naive method and the MoTR, like it was shown for the selected results using the MDE criterion. Similarly, the suggested effect of IDV is also negative and very close to the calculated for the chosen models with the MDE criterion. The t test *P* value is always higher than .05 and lower when applying the naive method, like the results selected with the other criteria.

Discussion of the Results Using MBL Criterion

From all the a priori hypotheses from the Exploratory A phase, only the results of two were included following the inclusion criteria defined in Subsection "Results Selection Criteria", and both results were based on the IM's data (H1 and H3), just like for the MDE criterion. As for MDE, one of these hypotheses, H3, was not supported, and the other, H1, was supported when controlling for A-MONTH but not supported when controlling for A-SEASON. The possible meaning and explanation for this are the same as for the first criterion and are described in the previous subsection. According to that explanation, the correct result is given when controlling A-MONTH, thus the hypothesis can be supported.

From the Exploratory B phase, only H4 and H9 were included following the MBL criterion. From those, only H4 was supported, showing that KW's socializing events could positively have affected the DIF-HR. The difference between the maximum and minimum heart rate registered during the following sleep period increased. The biggest suggested effect was obtained with three days of lag that is, considering the history of the past three days (compared to six days of lag when using the MDE criterion of selection). That means that the KW's DIF-HR values are dependent on the days before. A possible cause of that would be the same as for the selected results using the other criterion (MDE). The other hypothesis (H9) was not supported, for which the MoTR method revealed the contrary change in the outcome variables compared to the initial hypotheses. Although, if we compare the results using MoTR with the suggested effect direction shown by the original data (cf. Table 16), H9 is supported by the original data and not by MoTR. A possible cause is that the MoTR method meticulously evaluates the causation between the IDV and DV variables, while the original data can give false conclusions about that causation. This also occurred with the results using the MDE criterion for H9.

Table 16. Comparison between the outcome change previewed by the a priori hypotheses, obtained by the results selected using the MBL criterion, and the change observed using the naive method for H4 and H9 (note: only applies to participant KW). Refer to Table 14 for details on the Hypotheses. The naive method resulted in the same suggested effect direction, whether controlling for B-MONTH or B-SEASON.

	A priori	Results with MDE	Results with naïve method

Hypothesis	Outcome change	Outcome change	Outcome change
H4	Increase	Increase	Decrease
H9	Increase	Decrease	Increase

As for the results of the Testing phase selected using the MDE criterion, we assessed two results: (1) the model fitting; (2) the model estimates for the new data. Firstly, as shown in Table 15, the R-squared in A values were always negative. That means that the estimates of the model for the new data (part A) were shifted from the actual original new data (original values of part A), thus revealing that even with the model being accurate in estimating the new data values, adjustments must be made to obtain them in the right range of values. The smallest difference between the R-squared in B and the R-squared in A is for H9 controlling for B-SEASON, with a difference below 0.400. This threshold (0.400) was selected to match the same used when analyzing the results selected using the MDE criterion. Second, to evaluate the model estimates for the new data, we should compare the direction of the suggested effect measured by MoTR in the new data with the result shown in Table 8. Picking only H9 selected as previously detailed, both the naive and the MoTR methods estimated a negative suggested effect of IDV, that is, a decrease of the DV, consistent with the results in type B data. However, all the *P* values of H9 testing phase results were higher than .05, making those results not statistically discernible.

As previously stated in Subsection “Discussion of the Results Using MDE Criterion”, this is an observational study and not an interventional one, no causality can be strongly concluded but only hypothesized, using the presented MoTR method to do so.