

Causal Analysis of Self-tracked Time Series Data Using a Counterfactual Framework for N-of-1 Trials*

Eric J. Daza

Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, CA, USA

Keywords

Causal inference, counterfactual, n-of-1 trial, single subject, time series

Summary

Background: Many of an individual's historically recorded personal measurements vary over time, thereby forming a time series (e.g., wearable-device data, self-tracked fitness or nutrition measurements, regularly monitored clinical events or chronic conditions). Statistical analyses of such n-of-1 (i.e., single-subject) observational studies (N10Ss) can be used to discover possible cause-effect relationships to then self-test in an n-of-1 randomized trial (N1RT). However, a principled way of determining how and when to interpret an N10S association as a causal effect (e.g., as if randomization had occurred) is needed.

Objectives: Our goal in this paper is to help bridge the methodological gap between risk-factor discovery and N1RT testing by introducing a basic counterfactual framework for N10S design and personalized causal analysis.

Methods and Results: We introduce and characterize what we call the average period

treatment effect (APTE), i.e., the estimand of interest in an N1RT, and build an analytical framework around it that can accommodate autocorrelation and time trends in the outcome, effect carryover from previous treatment periods, and slow onset or decay of the effect. The APTE is loosely defined as a contrast (e.g., difference, ratio) of averages of potential outcomes the individual can theoretically experience under different treatment levels during a given treatment period. To illustrate the utility of our framework for APTE discovery and estimation, two common causal inference methods are specified within the N10S context. We then apply the framework and methods to search for estimable and interpretable APTEs using six years of the author's self-tracked weight and exercise data, and report both the preliminary findings and the challenges we faced in conducting N10S causal discovery.

Conclusions: Causal analysis of an individual's time series data can be facilitated by an N1RT counterfactual framework. However, for inference to be valid, the veracity of certain key assumptions must be assessed critically, and the hypothesized causal models must be interpretable and meaningful.

1. Introduction

Celia wants to know if and how exercise affects her body weight. She's recorded her weight and physical activity (e.g., step count) over the past couple of years. She looks at her data, and asks, "Is there evidence that changing my average level of physical activity and maintaining it at that level for a given period of time would affect my weight? If so, how?"

This example illustrates one of many personal research questions this paper may help answer by introducing a basic framework for personalized causal analysis. While particular techniques for causal discovery and effect estimation will later be presented and applied to the author's own health data, we will remain agnostic to the actual techniques chosen. Rather, these methods will be used to demonstrate how our framework allows the analyst to state causal assumptions precisely, thereby strengthening analytical decisions and conclusions in single-subject research.

Clinical or biomedical research conducted on one subject or individual is often called a single-subject, single-case, or *n-of-1* study, and an individual who undertakes an n-of-1 study on herself is said to self-track her own data. Such studies have been described as idiographic (i.e., population-of-one) in the psychological literature, in contrast to a nomothetic (i.e., population-of-many) study that characterizes a group of individuals [1]. N-of-1 studies are used in a variety of fields, including clinical trials and biomedical research [2–7]. Guidance on n-of-1 trial implementation and analysis has been codified by various investigators [8–11], and by the U.S. Department of Health and Human Services Agency for Healthcare Research

Correspondence to:

Eric J. Daza
Stanford Prevention Research Center
Stanford University School of Medicine
Medical School Office Building, X3C16
1265 Welch Road, Mail Code 5411
Stanford, CA 94305–5411
USA
E-mail: ericjdaza@stanford.edu

Methods Inf Med 2018; 57(Open 1): e10–e21

<https://doi.org/10.3414/ME16-02-0044>

received: November 20, 2016

accepted: August 16, 2017

Funding

This work was supported by the National Institutes of Health (NIH) grant 2T32HL007034–41. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIH.

* Supplementary material published on our website <https://doi.org/10.3414/ME16-02-0044>

and Quality (AHRQ) [12]. N-of-1 trials have even been offered as a clinical service in Australia, Canada, and the United States [13, 14]. In the field of mobile health (mHealth), Chen *et al* (2012) [15] proposed that mobile or wearable devices may help facilitate implementation of n-of-1 trials. Barr *et al* (2015) [16] are currently running a randomized controlled trial (RCT) to assess the feasibility and effectiveness of helping chronic-pain patients and their clinicians conduct n-of-1 trials using a smartphone app. Both AHRQ and a recent Nature article have even included n-of-1 trials under “personalized medicine” [12, 17].

While statistical methods for causal inference have largely been developed for n-of-1 randomized trials (N1RTs), to date there are few if any such methods for n-of-1 observational studies (N1OSs). We define an N1OS as a non-randomized single-subject study with the two-part goal of discovering both causal effects and possible N1RT treatments for subsequent testing of putative effects. Toward this end, we propose that the randomization-based approach of the Neyman-Rubin-Holland counterfactual framework [18–20] can be used to analyze self-tracked N1OS time series. In this literature, marginal structural models and the time-varying g -formula have been extensively developed for analyzing time-varying effects in longitudinal health data [21–23]. However, these two methods are used to conduct inference on average effects over a target population of many individuals, and hence may not readily apply to N1OSs. Randomized study designs more closely related to N1RTs (that may therefore be better suited to developing N1OS methods) include micro-randomization trials (MRTs) [24] and sequential multiple assignment randomized trials (SMARTs) [25], which are commonly used to develop just-in-time adaptive interventions (JITAIs) [26]. While these approaches focus on optimizing personalized treatments by finding the best set of treatment rules (i.e., rather than treatments) applicable to each individual, they still rely on averaging over a set of such individualized treatment regimes. (A MRT or SMART might be understood as a series of N1RTs [12].)

Causal inference methods that use only a single unit’s time series data do provide some direction. Aalen and Frigessi (2007) [27] and Aalen *et al* (2012) [28] proposed a mechanism-focused approach, rather than a counterfactual one. White and Kennedy (2009) [29] demonstrated equivalences between Granger and structural causality under a key assumption of conditional exogeneity, and derived useful methods for causal analysis of time series. White and Lu (2010) [30] drew formal connections between Granger and counterfactual-based causality, and Lu *et al* (2017) [31] showed how these concepts applied to the setting of cross-sectional and longitudinal data analysis. A good survey of relevant causal inference time series concepts can be found in Eichler (2012) [32], Eichler and Didelez (2012) [33], and Eichler (2013) [34], who connect the theory behind these ideas to those of various causal graphing systems. Unfortunately, almost all of these developments focus on econometric or financial applications with no direct analogue to the health settings of N1RTs.

The goal of this paper is to help bridge the methodological gap between risk-factor discovery and N1RT testing by introducing a basic counterfactual framework for N1OS design and analysis. The rest of this paper is organized as follows. In Section 2, we briefly review the counterfactual framework. We then define an idiographic causal estimand called the average period treatment effect in Section 3, present a framework for its specification and analysis, and specify two common estimation methods within this framework. In Section 4, we estimate average period treatment effects relevant to the author’s weight and physical activity data using our framework. We conclude in Section 5 with a brief discussion, and we propose a basic procedure for performing n-of-1 causal discovery. Further notes and derivations are provided in an online Appendix. All analyses were conducted in R version 3.3.1.

2. Counterfactual-based Causal Inference

Throughout this article, we use the following notation. Random variables and fixed

values are written in upper-case and lower-case, respectively. Let $p(A = a)$ denote the probability mass or density of random variable A at a , with shorthand $p(a)$. Let $\{(A)\}$ denote a stochastic process; i.e., a time series of random variables. For any index j , let $\{(j)\}$ denote a sequence. For any random variable B , let $B|A$ denote the event B conditional on A , with shorthand $B|a$ for $B|A = a$. Let $B \perp A$ denote statistical independence of B and A .

Suppose we have a scalar-valued function of random variables, as are specified in structural equation models [35]. Let the left side (i.e., area to the left of the equal sign) consist of an outcome or output variable, and let the right side consist of three components: predictors, a completely random zero-mean error or disturbance term ξ that is independent of all predictors, and a function relating these two sets of input variables to the outcome, with the error term suppressed in the function notation unless needed for conceptual clarification. Suppose this function is constrained such that all inputs must temporally occur before the outcome. We define a data-generating process (DGP) to be such a time-constrained function (e.g., the univariate structural equations in White and Lu, 2010 [30]), and call the structural equation expression of a DGP a data-generating model (DGM).

Counterfactuals have been broadly defined in terms of different types of interventions [36, 32]. We take a statistics-based approach, and consider the types of effects identifiable from randomized interventions (i.e., randomly selected, assigned, or otherwise manipulated) [18–20]. Such causal effects are defined in terms of statistical associations between a treatment (or intervention) and an outcome if the treatment mechanism is ignorable (i.e., hypothetical effects at different treatment levels remain unchanged regardless of actual treatment assignment). For example, ignorability is implied if the treatment is randomized. In this paper, we only consider the case of ignorability implied by randomization, and henceforth write “treatment” in place of “randomized treatment”. We define an exposure to be a measured phenomenon that may be considered a treatment; i.e., all treatments are exposures, while the converse does not hold in general. Henceforth,

we use “causal effect”, “treatment effect”, and “effect” interchangeably.

Let $Y = g^Y(X, U)$ denote the DGP of observed outcomes, where X represents an exposure, and U represents the set of all other (possibly unobserved) outcome predictors. For a given individual, consider a hypothetical value of Y under exposure level a and predictor values U if X and U are independent. We formalize this concept by defining the *counterfactual* (i.e., counterfactual outcome) of Y corresponding to $X = a$ and U as

$$Y^a(U) = g_a^Y(U, X \perp U),$$

where a represents a fixed value that is not a predictor. Under *causal consistency* (CC), the observed and counterfactual outcomes under $X = a$ are identical; i.e.,

$$g^Y(X = a, U, \xi) = g_a^Y(U, X \perp U, \xi).$$

The term “counterfactual” is used because if $X = a$ is in fact observed, then observation of Y^a for any $a' \neq a$ is “counter to fact” (i.e., Y^a cannot be observed under CC). A counterfactual is also called a potential outcome because it is a potentially observable outcome resulting from an exposure.

Each individual i has the counterfactual $Y_i^a(u_i) = g_a^Y(u_i, X \perp U, \xi_i)$ at $U = u_i$. A contrast between $E_{\xi_i}(Y_i^a(u_i))$ and $E_{\xi_i}(Y_i^a(u_i))$, where $E(\cdot)$ denotes the expectation function, is called an individual treatment effect (ITE). This is the desired estimand of counterfactual-based causal inference (hereafter, causal inference). Unfortunately, an ITE is generally not identifiable because for any individual, we cannot simultaneously observe both $Y_i^a(u_i)$ and $Y_i^a(u_i)$ (i.e., the *fundamental problem of causal inference* [20]), much less estimate their expectations. Now let $E(Y^a(u))$ represent the average or mean counterfactual corresponding to a taken over the population of individuals with $U = u$, often conceptualized as the expected outcome if everyone in such a population is randomized to treatment a . While not directly observable due to the fundamental problem, this quantity may be consistently estimated if U is either fully observed, or partly observed and treatment is randomized (see ► Appendix equations (1) and (3)). Hence, comparisons or contrasts of $E(Y^a(u))$ and

$E(Y^{a'}(u))$ may be of interest.

Many authors introduce the counterfactual as Y^a , with attendant contrasts between $E(Y^a)$ and $E(Y^{a'})$ called average treatment effects (ATEs). These are often the primary estimands of interest because in nomothetic studies with randomized interventions, all other outcome predictors U need not be observed in order to consistently estimate $E(Y^a)$. (see paragraph below on estimation). In particular,

$$Y^a = E_U(Y^a(U)|X = a, X \perp U) = E_U(Y^a(U)|X \perp U)$$

averages over all other true outcome predictors, as does $E(Y^a)$ by implication. This approach is particularly useful when there is little heterogeneity in the treatment effects across, for example, settings, contexts, groups, or individuals, that can be formalized using U . (Dependence of the effects on U is often the main interest in the literature on heterogeneous treatment effects, which focuses on conditional ATEs [37].) If $E(Y^a) \neq E(Y^{a'})$ for some $a' \neq a$, then X is said to have an ATE on Y , and we will call X a causal predictor (i.e., cause) of Y . For example, if treatment is randomized as $X = 0, 1$, then possible ATEs include $E(Y^1) - E(Y^0)$ and $E(Y^1)/E(Y^0)$. If all DGP predictors are causal, we will call this DGP a causal process, and its corresponding DGM a causal model.

Estimation of any of these quantities, however, requires observed (i.e., not counterfactual) outcomes. Let $R = 1$ denote the implementation of randomization to $X = a$, and let $R = 0$ denote the absence of randomization (i.e., corresponding with the ecological, natural, or otherwise undisturbed state of X). Suppose the outcome DGP might vary depending on whether or not X is randomized, denoted as $Y = g^Y(X, U, R)$. Then the same outcome will be generated whether or not X is randomized if $g^Y(X, U, R, \xi) = g^Y(X, U, \xi)$. We will refer to this equivalence as *data-generation invariance* (DGI) because it describes invariance of the DGP to randomization status. Importantly, note that if DGI holds, then $p(y|x, u, r) = p(y|x, u)$, while the converse is not true in general. If DGI and CC both hold, then $E(Y^a)$ can be identified using observed outcomes if $R = 1$ (see ► Appendix equations (1)-(2)). If $R = 0$ (as in Section 3.3), then $p(u|r) = p(u)$ is also needed to

identify $E(Y^a)$. We will call this last condition *distributional invariance* (DI); i.e., all other outcome predictors U are independent of the randomization status of X . (In the ► Appendix, we relate our conceptual approach and assumptions to the standard ITE-based statistics concepts of causal consistency and conditional ignorability/exchangeability.)

All observations in an NIOS belong to a single individual, and in this sense constitute a single context. Hence, in beginning to develop counterfactual theory for single-subject causal analysis, we will focus on some individualized quantity analogous to the ATE; one that averages over other outcome predictors specific to that individual throughout the timespan of her self-tracked observations. Future methods can and should be developed to fashion conditional ATEs that more properly account for the varied sub-contexts within an individual's own experiences (e.g., seasonality). Finally, note that while we can rely on randomization to enable estimation of an ATE, we generally consider corresponding DGMs that are fit in practice to be, at best, approximations to the hypothesized true causal mechanism (i.e., the true and unknown processes by which a cause produces an effect).

3. Average Period Treatment Effect

In this section, we define an average treatment effect for analytical use in both randomized and non-randomized idiographic settings, and introduce a framework for specifying and analyzing this average effect. Two common estimation methods are specified within this framework, and stationarization is briefly illustrated as a way to model confounding. We rely on formalisms similar to the general dynamic structural equations of White and Kennedy (2009) [29] and White and Lu (2010) [30], Section 22.5 in Eichler (2012) [32], and Section 5 of Eichler (2013) [34]. Throughout, we assume that DGI and CC hold.

3.1 Definition

Let $\{(X, Y)\}$ represent a stochastic process. The standard NIRT is a randomized cross-over design used to assess an ATE of X on Y . However, methods for conducting inference regarding ATEs are almost exclusively nomothetic. In particular, researchers generally wish to draw inference on the mean counterfactual taken over a population of individuals, as mentioned previously. Because there is only one individual in an n-of-1 study, the ATE definition in this idiographic setting needs to be modified. The definition that follows is motivated by standard NIRT concepts [12].

In the basic NIRT, a two-level treatment X is randomized at each time period t , defined as a set of measurement time points. Let $t(j)$ denote a time point within period t for $j = 1, \dots, m_t$. Treatment level is randomized per period only at $t(1)$, and is otherwise kept constant; i.e., randomized assignment $X_{t(1)} = a$ implies $X_{t(j)} = a$ for $j \in (2, \dots, m_t)$ if $m_t > 1$. We will call a treatment administered in a period consisting of only one time point (i.e., $m_t = 1$) a *point treatment*, and write t instead of $t(1)$ in such cases; otherwise, a treatment may be called a *period treatment* for clarification. Consider the simple case of a point treatment, where Y_{t+1} has a time-invariant association with X_t and other predictors W_t , there is no autocorrelation or time trend in $\{(W)\}$ (e.g., a white noise process, which is a strictly stationary time series), and $\{X_t, W_t\} \perp Y_{t+1}$ for all $t' < t$. Suppose each outcome Y_{t+1} is independent of and identically distributed with all other outcomes, conditional on $\{X_t, W_t\}$ (where this relationship is constant over time). Hence, the outcome DGP is $Y_{t+1} = g^Y(X_t, W_t)$. There is no autocorrelation in $\{(Y)\}$, and because $p(y_{t+1})$, $p(x_t)$, and $p(w_t)$ are constant over time, no time trend exists in $\{(Y, X, W)\}$.

Since we are interested in the effect of X_t on Y_{t+1} , it is reasonable to think of the pair $\{X_t, Y_{t+1}\}$ as an idiographic unit of observation. Let Y_{t+1}^a represent the counterfactual of Y_{t+1} corresponding to $X_t = a$. We define a *period treatment effect* (PTE) to be a contrast between Y_{t+1}^a and $Y_{t+1}^{a'}$ for $a \neq a'$, and call a contrast of $E(Y_{t+1}^a)$ and $E(Y_{t+1}^{a'})$ an *average period treatment effect* (APTE). The APTE is the estimand of in-

terest in an NIRT. This mean counterfactual represents the expected outcome if the individual is randomized to treatment a at t , but not over all time points, as would be directly analogous to the interpretation of an ATE mean counterfactual (i.e., taken over all individuals). This is an important distinction, because randomization to a at all time points may violate the DI assumption, which is a key condition needed for identification of an APTE in the presence of confounding, as discussed in Section 3.3.

In our simple case, there is no carryover of effects from any past periods. There is no slow onset/activation of the APTE (e.g., due to delayed uptake of the treatment), and neither is there any slow decay/deactivation. Both $\{(X)\}$ and $\{(Y)\}$ are strictly stationary processes integrated of order 0 [38], thus permitting straightforward estimation of the APTE.

3.2 N-of-1 Counterfactual Framework

We present the following framework for specifying an APTE that allows for autocorrelation or a time trend in the outcomes, or carryover or slow onset/decay of the effect. Suppose observations or measurements occur at evenly spaced time points indexed by j . For any random variable B , let $\bar{B}_j = (B_j, B_{j-1}, B_{j-2}, \dots)$ denote the history of B at $j+1$. Let Y and X denote the outcome and treatment of interest, respectively, where X is a categorical variable. Suppose $Y_{j+1} = g_j^Y(\bar{X}_j, \bar{V}_j)$ in general, where \bar{V}_j represents all other predictors of Y_{j+1} . Likewise, suppose in general that $X_j = g_{j-1}^X(\bar{V}_{j-1}, \bar{X}_{j-1}, \bar{Z}_{j-1})$, where \bar{Z}_{j-1} represents all other predictors of X_j and $\bar{V}_j \cap \bar{Z}_{j-1} = \emptyset$.

We first distinguish between a treatment and an exposure. If $R_{j-1} = 1$, then X_j has no predictors by definition. We denote this mechanistic relationship by re-specifying the DGP as $X_j = g_{j-1}^X(\bar{V}_{j-1}, \bar{X}_{j-1}, \bar{Z}_{j-1}, R_{j-1})$; in particular, $g_{j-1}^X(\bar{V}_{j-1}, \bar{X}_{j-1}, \bar{Z}_{j-1}, R_{j-1} = 1, \xi_{j-1}) = g_{j-1}^X(R_{j-1} = 1, \xi_{j-1})$. For example, suppose randomization to either treatment or control occurs at every time point; i.e., $X_j = 1, 0$, respectively. Then one reasonable DGM is $X_j = I(\xi_{j-1} > \Pr(\bar{X}_j = \|\bar{V}_{j-1}, \bar{X}_{j-1}, \bar{Z}_{j-1}, R_{j-1}\|))$, where $I(b) = 1$ if expression b is true and $I(b) = 0$ otherwise, and ξ_{j-1} is uniformly distributed between 0 and 1.

Treatment periods are constructed as follows. Partition $\{(j)\}$ into $\{(t)\}$ such that $t = (t(1), \dots, t(m_t))$, where treatments in period t are observed at each point $t(j)$; i.e., $\{(t)\}$ is a structured time series. Let $X_{t(j)}$ denote the categorically defined treatment at time point $j = 1, \dots, m_t$ in period t . Randomization for period t can be implemented at $\{R_{t(j-1)} : j \in (1, \dots, m_t)\}$. The last outcome for period t occurs at $t(m_t+1) \equiv t+1(1)$, and the outcomes for treatment period t are $\{Y_{t(j+1)} : j \in (1, \dots, m_t)\}$. Our general formulation permits randomization of multiple treatments within a period; e.g., (a_1, \dots, a_m) could represent a dynamic treatment regime [39] in a JITAI, MRT, or SMART. However, we will only consider the standard NIRT case where only the first treatment is randomized, and then held constant for the rest of the period; i.e., $R_{t(0)} = 1$ and $R_{t(j-1)} = 0$ for $j \in (2, \dots, m_t)$, and implies $X_{t(j)} = a$ for $j \in (1, \dots, m_t)$. In this way, an NIRT might be a type of cluster-randomized trial in which a period constitutes a cluster, or perhaps a kind of non-adaptive MRT with period treatments.

Suppose each treatment effect is bounded, and may stabilize or destabilize over time after treatment introduction. We define the association (e.g., coefficient in a linearized model) of an outcome with a predictor as *stable* if their associations at $\{t(j), t'(j)\}$ are identical for any pair of periods $\{t, t' \neq t\}$ at any j . If all outcome-predictor associations are stable, then $g_{t(j)}^Y(\cdot) = g_{t'(j)}^Y(\cdot)$ for exactly equal input values at $\{t(j), t'(j)\}$, and we write $g_j^Y(\cdot)$ instead. We define the association of an outcome with a predictor as *period-stable* if their associations at $\{t(j), t'(j)\}$ are identical for any pair of points $\{j, j' \neq j\}$ at any t . If all stable outcome-predictor associations are period-stable, then $g_j^Y(\cdot) = g_{j'}^Y(\cdot)$ for exactly equal input values at $\{j, j'\}$, and we write $g^Y(\cdot)$ instead. Henceforth, we only consider stable associations for simplicity.

Suppose we have randomized period treatments (i.e., $R_{t(0)} = 1$ and $R_{t(j-1)} = 0$ for $j \in (2, \dots, m_t)$) with period-stable associations. In the rest of this section, we assume all distributional statements are conditioned on $R_{t(j-1)}$, and therefore suppress this notation. For any random variable B , let $\bar{B}_{t(j)} = (B_{t(j)}, B_{t(j-1)}, \dots, B_{t(1)}, B_{t-1(m_{t-1})}, B_{t-1(m_{t-1}-1)}, \dots)$. Suppose the outcome DGP is

$Y_{t(j+1)} = g^y(X_{t(j)}, \bar{V}_{t(j)})$ where there is no autocorrelation or time trend in $\{(V)\}$. Hence, there is no carryover or slow onset/decay, and no autocorrelation or time trend in either $\{(X)\}$ or $\{(Y)\}$. The mean counterfactual corresponding to $X_{t(j)} = a$ is therefore $E(Y_{t(j+1)}^a) = E(Y_{t(j+1)} | X_{t(j)} = a)$. (See ▶ Appendix equations (1)-(3) for derivations of this and all remaining mean counterfactual expressions stated in this section.) Let $\alpha_{t(k+1)}(a', a)$ denote a contrast function of $E(Y_{t(k+1)}^a)$ and $E(Y_{t(k+1)}^a)$, where $a' \neq a$. We now redefine the APTE as a function of some pre-specified subset $\{\alpha_{t(k+1)}(a', a) : k \in \mathbf{k}\}$, where $\mathbf{k} \subseteq \{1, \dots, m_t\}$. In the current simple case, $\alpha_{t(j+1)}(a', a) = \alpha_t(a', a)$ for $j \in (1, \dots, m_t)$ because all associations are period-stable, and $\alpha_t(a', a) = \alpha(a', a)$ for all t because all associations are stable. Hence, $\mathbf{k} = (1, \dots, m_t)$ might be specified, along with $apte(a', a) = \alpha(a', a)$ for any $\{a', a\}$.

Suppose autocorrelation in $\{(Y)\}$ is also present, such that $Y_{t(j+1)} = g^y(X_{t(j)}, \bar{V}_{t(j)}, \bar{V}_{t(j)})$. Note that $\{(X)\}$ Granger-causes $\{(Y)\}$, a related but distinct causal concept; i.e., $\{(X)\}$ can Granger-cause $\{(Y)\}$ even if $\{(X)\}$ is not a randomized-treatment series [40–42]. A model for $E(Y_{t(j+1)} | X_{t(j)}, \bar{V}_{t(j)})$ can be specified and used to estimate an APTE specified with $E(Y_{t(j+1)}^a | \bar{V}_{t(j)}) = E(Y_{t(j+1)} | X_{t(j)} = a, \bar{V}_{t(j)})$. Note that specifying an APTE with $E(Y_{t(j+1)}^a) - E_{\bar{V}_{t(j)}}\{E(Y_{t(j+1)} | X_{t(j)} = a, \bar{V}_{t(j)})\}$ is not straightforward because $R_{t(j-1)} = 1$ only at $j = 1$; we will see how to handle cases in which $R_{t(j-1)} = 0$ in Section 3.3.

Suppose further that there is a time trend in $\{(Y)\}$. The same DGP applies, but $\{(Y)\}$ is no longer stationary, which is required for consistent estimation of model parameters. One option is to define this trend to be a function of some predictors of $Y_{t(j+1)}$ (see Section 3.5 and ▶ Appendix equation (3)), and model $E(Y_{t(j+1)} | X_{t(j)}, \bar{V}_{t(j)}, \bar{V}_{t(j)})$ in order to estimate an APTE specified with $E(Y_{t(j+1)}^a | \bar{V}_{t(j)}, \bar{V}_{t(j)}) = E(Y_{t(j+1)} | X_{t(j)} = a, \bar{V}_{t(j)}, \bar{V}_{t(j)})$. Explicit modeling might be avoided by using a randomization scheme that balances the treatments across periods (e.g., a randomized-block design limiting the viable block permutations, where a block is defined as a set of consecutive periods). However, even this approach tacitly assumes some general structure to the trend (e.g., linear, quadratic) in order to determine how balance can best be achieved.

Now suppose carryover is present from $\ell \in \mathbb{N}$ lagged effects, such that $Y_{t(j+1)} = g^y(X_{t(j)}, \bar{X}_{t-\ell:j-1}, \bar{V}_{t(j)}, \bar{V}_{t(j)})$ where $\bar{X}_{t-\ell:j-1} = (X_{t(j)}, X_{t(j-1)}, \dots, X_{t(j-\ell)})$. (Since all elements of $\bar{X}_{t-\ell:j-1}$ are randomized, carryover is a type of causal interference [43] in that a given period's potential outcomes are a function of possible treatment levels in both the current and past periods.) The conditional mean counterfactual corresponding to $X_{t(j)} = a$ is therefore

$$E(Y_{t(j+1)}^a | \bar{X}_{t-\ell:j-1}, \bar{V}_{t(j)}, \bar{V}_{t(j)}) = E(Y_{t(j+1)} | X_{t(j)} = a, \bar{X}_{t-\ell:j-1}, \bar{V}_{t(j)}, \bar{V}_{t(j)}).$$

The DGM of $g^y(X_{t(j)}, \bar{X}_{t-\ell:j-1}, \bar{V}_{t(j)}, \bar{V}_{t(j)})$ that needs to be specified and fit is usually unknown in practice, unfortunately, but washouts may be used to avoid having to fully specify this DGM.

A washout period can be defined in order to eliminate carryover. In a designed-washout approach, treatment is not administered during the washout period, which would then be excluded from the main analysis to estimate the APTE. However, not administering treatment is itself a treatment. Let a_0 denote such a washout treatment (henceforth, washout). Note that the control treatment and washout need not be identical, as in the case of an active control; nor must the washout equal an exposure that occurs naturally (i.e., outside of a randomized trial). A designed-washout approach is a type of randomized-block design in which at least one washout period is assigned immediately following the treatment period (i.e., the block is at least two periods long). Suppose enough washouts are assigned to cover all lags; i.e. $\bar{X}_{t-\ell:j-1} = a_0$, where a_0 is a $1 \times \ell$ vector with every element equal to a_0 . Also suppose that the washouts are assigned properly, such that $p(Y_{t(j+1)} | X_{t(j)} = a, \bar{X}_{t-\ell:j-1} = a_0, \bar{V}_{t(j)}, \bar{V}_{t(j)}) = p(Y_{t(j+1)} | X_{t(j)} = a, \bar{V}_{t(j)}, \bar{V}_{t(j)})$. Hence, a model for $E(Y_{t(j+1)} | X_{t(j)}, \bar{V}_{t(j)}, \bar{V}_{t(j)})$ can be specified and used to estimate an APTE specified with $E(Y_{t(j+1)}^a | \bar{V}_{t(j)}, \bar{V}_{t(j)}) = E(Y_{t(j+1)} | X_{t(j)} = a, \bar{V}_{t(j)}, \bar{V}_{t(j)})$. (Note that this additionally requires observing all other predictors $\bar{V}_{t(j)}$.) For example, if $\ell = 1$ and $m_t = 1$ for all t , then it can be shown that washouts are properly assigned if a washout period always follows a treatment period, and vice versa.

In an analytic-washout approach, each period consists of multiple measurements, and the washout subperiod is defined as a set of measurements occurring early in the

period (i.e., the set of early measurements with effects from previous time periods). In conducting the main APTE analysis, this approach involves not collecting, excluding, down-weighting, or otherwise reducing the influence of washout subperiod observations on estimation. Analytic washouts may be applied if $Y_{t(k+1)} = g^y(X_{t(k)}, \bar{V}_{t(k)}, \bar{V}_{t(k)})$ for any $k \in \mathbf{k}_{\text{postwash}}$, where $\mathbf{k}_{\text{postwash}} \subseteq \{1, \dots, m_t\}$ denotes the post-washout subperiod; i.e., the later measurements of a period, when there are no lingering carryover effects. In such cases, an APTE specified with $E(Y_{t(k+1)}^a | \bar{V}_{t(k)}, \bar{V}_{t(k)}) = E(Y_{t(k+1)} | X_{t(k)} = a, \bar{V}_{t(k)}, \bar{V}_{t(k)})$ can be estimated. Because all associations are period-stable, $\alpha_{t(k+1)}(a', a) = \alpha_t(a', a)$ for $k \in \mathbf{k}_{\text{postwash}}$, and because all associations are stable, $\alpha_t(a', a) = \alpha(a', a)$ for all t . The APTE might then be specified as $apte(a', a) = \alpha(a', a)$.

Finally, suppose slow onset or decay may be present, such that stable effects may no longer be period-stable, implying $Y_{t(j+1)} = g^y(X_{t(j)}, \bar{X}_{t-\ell:j-1}, \bar{V}_{t(j)}, \bar{V}_{t(j)})$. Suppose subperiod-stable effects are present for a subset $\mathbf{k}_{\text{stable}} \subseteq \{1, \dots, m_t\}$ (i.e., the stable subperiod), such that $Y_{t(k+1)} = g^y(X_{t(k)}, \bar{X}_{t-\ell:j-1}, \bar{V}_{t(k)}, \bar{V}_{t(k)})$ for any $k \in \mathbf{k}_{\text{stable}}$; i.e., the middle or later measurements of a period, when an effect is fully activated and stable, before it begins deactivating (if applicable). In such cases, we will refer to the subperiod intervals before and after the stable subperiod as the stabilization and destabilization subperiods, respectively. An APTE specified with $E(Y_{t(k+1)}^a)$ can be estimated by specifying a model for $E(Y_{t(k+1)} | X_{t(k)} = a, \bar{V}_{t(k)}, \bar{V}_{t(k)})$ for $k \in \mathbf{k}_{\text{postwash}} \cap \mathbf{k}_{\text{stable}}$. Because all other associations are period-stable, $\alpha_{t(k+1)}(a', a) = \alpha_t(a', a)$ for $k \in \mathbf{k}_{\text{postwash}} \cap \mathbf{k}_{\text{stable}}$, and because all associations are stable, $\alpha_t(a', a) = \alpha(a', a)$ for all t . The APTE might then be specified as $apte(a', a) = \alpha(a', a)$. Note that the stabilization subperiod may be equivalent to the washout subperiod if the current and previous treatment levels differ. For example, this could be the case for a binary treatment consisting of mutually exclusive treatment levels; e.g., administration or removal of one active treatment that does not destabilize.

3.3 Confounding

Suppose we now have period exposures instead of treatments; i.e., $R_{t(j-1)} = 0$ for $j \in (1, \dots, m_t)$ at all t . Suppose there is no autocorrelation or time trend in $\{(Z)\}$, and $X_{t(j)} = g^x(\bar{W}_{t-1(m_t)}^x, R_{t(j-1)} = 0) = g^x(\bar{W}_{t-1(m_t)}^x, R_{t(0)} = 0)$, where $\bar{W}_{t-1(m_t)}^x \subseteq \{\bar{X}_{t-1(j-1)}, \bar{Y}_{t-1(m_t)}, \bar{Z}_{t-1(m_t)}\}$ and $\bar{V}_{t(j)} \cap \bar{Z}_{t-1(m_t)} \neq \emptyset$ in general. Let $\bar{W}_{t(j)}^y \subseteq \{\bar{X}_{t-1(j-1)}, \bar{Y}_{t(j)}, \bar{V}_{t(j)}\}$, and let $\bar{C}_{t(j)} \neq \emptyset$ represent $\bar{W}_{t(j)}^y \cap \bar{W}_{t-1(m_t)}^x$.

If a variable B is a causal predictor of both $X_{t(j)}$ and $Y_{t(j+1)}$, we say that B confounds the relationship between $X_{t(j)}$ and $Y_{t(j+1)}$. Suppose every element of $\bar{C}_{t(j)}$ is a confounder. This assumption may be too strong to defend in practice, but can be relaxed using the rules of d-separation [44] (specifically, to avoid “M-bias”), a topic beyond the scope of this paper. Hence

$$E(Y_{t(j+1)} | X_{t(j)} = a, R_{t(j-1)} = 0) = E_{\bar{W}_{t(j)}^y} \left\{ E(Y_{t(j+1)} | X_{t(j)} = a, \bar{W}_{t(j)}^y, R_{t(j-1)} = 0) | X_{t(j)} = a, R_{t(j-1)} = 0 \right\} \neq E(Y_{t(j+1)}^a),$$

in general. However, if

$$Y_{t(j+1)} = g_y^y(X_{t(j)}, \bar{W}_{t(j)}^y, R_{t(j-1)}, \xi_{t(j)}) = g_y^y(X_{t(j)}, \bar{W}_{t(j)}^y, \xi_{t(j)}), \quad \text{DGI}$$

$$p(\bar{w}_{t(j)}^y | r_{t(j-1)}) = p(\bar{w}_{t(j)}^y), \quad \text{DI}$$

then

$$E(Y_{t(j+1)}^a) = E_{\bar{W}_{t(j)}^y} \left\{ E(Y_{t(j+1)} | X_{t(j)} = a, \bar{W}_{t(j)}^y, R_{t(j-1)} = 0) | R_{t(j-1)} = 0 \right\}, \quad (1)$$

which is identifiable from observed data if the inner expectation DGM is known (see ► Appendix equation (5)).

We will jointly refer to DGI and DI as *invariance to randomization* (hereafter, *invariance*), a concept akin to that of “distributional stability”; i.e., the joint probability distribution of predictors, outcomes, and covariates is invariant to the predictor’s intervention regime (e.g., observational vs. randomized) [45–47; 32]. Invariance is a powerful condition because if it holds, then an APTE specified with $E(Y_{t(j+1)}^a)$ can be estimated in the absence of randomization. Hence, in discussing the strength of causal inference, it is crucial for the analyst to acknowledge that she is making an assumption that invariance holds, and to assess and report the veracity of this assumption.

Identifiability of an APTE specified with $E(Y_{t(j+1)}^a)$ if $R_{t(j-1)} = 0$ also relies on the positivity condition that $p(x_{t(j)}, \bar{w}_{t(j)}^y) > 0$ for all $x_{t(j)}$ and $\bar{w}_{t(j)}^y$; i.e., all possible combinations of $X_{t(j)}$ and $\bar{w}_{t(j)}^y$ are theoretically observable. (Note that positivity is implied if $R_{t(j-1)} = 1$.) We implicitly assumed that this condition holds in deriving (1), and its importance will become particularly apparent in the IPW formula of Section 3.4.

In general, post-washout and stable sub-periods are not properly assigned (let alone specified a priori) in non-randomized settings. Instead, k_{postwash} and k_{stable} can be assumed, and an APTE specified with $E(Y_{t(k+1)}^a)$ can be estimated by specifying a model for $E(Y_{t(k+1)} | X_{t(k)} = a, \bar{W}_{t(k)}^y, R_{t(j-1)} = 0)$ for $k \in k_{\text{postwash}} \cap k_{\text{stable}}$. As in the randomized case, because all other associations are period-stable, and all associations are stable, the APTE might then be specified as $\alpha_{t(k+1)}(a', a) = \alpha_t(a', a) = \text{apte}(a', a)$. The analyst could then vary the assumed values of k_{postwash} and k_{stable} , and characterize how the estimated APTE changes.

3.4 Estimation

The following two causal inference methods are commonly used to estimate an ATE in the presence of confounding, assuming positivity and invariance hold. Here, we specify them for an APTE. If the DGM for $Y_{t(j+1)} = g^y(X_{t(j)}, \bar{W}_{t(j)}^y)$, also called an *outcome model*, is correctly specified, then (1) can be estimated directly. This is known as the g-formula method [48, 49], and in the epidemiological literature is also called direct standardization [23, 50], or stratification elsewhere [51, 52]. The key insight is that the outer expectation is taken over $\bar{W}_{t(j)}^y$, not over $(\bar{W}_{t(j)}^y | X_{t(j)} = a)$ as is required by $E(Y_{t(j+1)} | X_{t(j)} = a, R_{t(j-1)} = 0)$.

For a binary-valued $X_{t(j)}$, another strategy is to instead argue that the functional form of $\Pr(X_{t(j)} = 1 | \bar{W}_{t-1(m_t)}^x, R_{t(j-1)} = 0)$, also called the *propensity model*, is correctly specified; e.g.,

$$g^x(\bar{W}_{t-1(m_t)}^x, R_{t(j-1)} = 0, \xi_{t-1(m_t)}^x) = I(\xi_{t-1(m_t)}^x > \Pr(X_{t(j)} = 1 | \bar{W}_{t-1(m_t)}^x, R_{t(j-1)} = 0))$$

where $\xi_{t-1(m_t)}^x$ is uniformly distributed between 0 and 1. An APTE specified with

$$E(Y_{t(j+1)}^a) = E \left\{ \frac{I(X_{t(j)} = a) Y_{t(j+1)}}{\Pr(X_{t(j)} = a | \bar{W}_{t-1(m_t)}^x, R_{t(j-1)} = 0)} | R_{t(j-1)} = 0 \right\}$$

IPW formula can then be estimated (see ► Appendix for derivation). This is known as the method of inverse probability weights (IPWs), which uses the reciprocal (i.e., inverse) of the conditional probability of X . The conditional probability that $X_{t(j)} = 1$ is also known as the propensity score because it reflects the propensity of receiving exposure $a = 1$ [53]. Consistent estimation of APTEs specified with $E(Y_{t(j+1)}^a)$ is often performed using a Horvitz-Thompson ratio estimator (see ► Appendix). (Many common matching methods also use the propensity score as a way to balance covariate values between exposure levels in order to estimate putative treatment effects; e.g., by selecting subsamples of the original sample.)

Note that the g-formula method does not require specification of the propensity model, while the IPW method does not require specification of the outcome model. An advanced technique called the augmented IPW or doubly robust estimation method is useful for gaining statistical efficiency if specifications of both the propensity and outcome models may be reasonably asserted as true, with only one of the two required to be correctly specified to ensure consistent estimation of model parameters.

3.5 Stationarization

Following Lu and Zeger (2007) [54], in this section we argue that stationarization may be understood as a way to model confounding. Estimation of the mean counterfactual requires both the predictor and outcome time series to be weak- or wide-sense stationary (WSS) processes [38]. If a time series is not WSS, the methods of taking first differences (or pre-whitening) or detrending are commonly employed. If the outcomes are continuous, then both are special cases of the same general expression that can itself be used to specify a model for confounding. Throughout this section, we assume we have period exposures, and suppress the $R_{t(j-1)} = 0$ notation.

Suppose Y is continuous. Let $\Delta_{t(j+1)}^y = Y_{t(j+1)} - H_{t(j)}^y$ and $\Delta_{t(j)}^x = X_{t(j)} - H_{t(0)}^x = X_{t(1)} - H_{t(0)}^x$, where

$H_{t(j)}^Y = h_{t(j)}^Y(\bar{w}_{t(j)}^Y)$ and $H_{t(j)}^X = h_{t(j)}^X(\bar{w}_{t(j)}^X)$ are scalar-valued functions. Suppose

$$Y_{t(j+1)} = H_{t(j)}^Y + \bar{\Delta}_{t(j)}^Y \beta^Y + \bar{\Delta}_{t(j)}^X \beta^X + \xi_{t(j)}, \quad (2)$$

where $\bar{\Delta}_{t(j)}^X = (\Delta_{t(j)}^X, \Delta_{t(j-1)}^X, \dots, \Delta_{t(j-l)}^X)$, β^Y and β^X are conformable coefficient vectors, and $\xi_{t(j)}$ is completely random zero-mean error with finite variance. Noting that $\bar{\Delta}_{t(j)}^X \beta^X = \beta_0^X X_{t(j)} - \beta_0^X H_{t(0)}^X + \bar{\Delta}_{t(j-1)}^X (\beta_1^X, \dots, \beta_l^X)'$, we can define the mean counterfactual corresponding to $X_{t(j)} = a$ as

$$E(Y_{t(j+1)}^a) = E_{H_{t(0)}^X, \bar{\Delta}_{t(j)}^X, H_{t(j)}^X} \{E(Y_{t(j+1)}^a | X_{t(j)} = a, H_{t(0)}^X, \bar{\Delta}_{t(j)}^X, H_{t(j)}^X)\} \\ = E(H_{t(j)}^Y) + E(\bar{\Delta}_{t(j)}^Y) \beta^Y + \beta_0^X a - \beta_0^X E(H_{t(0)}^X) + E(\bar{\Delta}_{t(j-1)}^X) (\beta_1^X, \dots, \beta_l^X)$$

Now suppose $\{(\Delta^Y, \Delta^X)\}$ is a marginally WSS process. For example, this would hold for $\{(\Delta^Y)\}$ if either the first-differenced or de-trended process is WSS: $H_{t(j)}^Y = Y_{t(j)}$ in the former case, while $H_{t(j)}^Y = E(Y_{t(j+1)} | \bar{w}_{t(j)}^Y)$ in the latter case (for example), such that $E(\Delta_{t(j+1)}^Y) = \mu$ (i.e., is constant). If $Y_{t(j+1)}$ has no predictors, then $H_{t(j)}^Y$ is either a constant or completely random with a constant mean (e.g., white noise). Rewriting (2) as $\Delta_{t(j+1)}^Y = \bar{\Delta}_{t(j)}^Y \beta^Y + \bar{\Delta}_{t(j)}^X \beta^X + \xi_{t(j)}$, we see that consistent estimation of β^Y and β^X is straightforward if $|\beta^X| < \mathbf{1}$ where $\mathbf{1}$ is a vector of ones.

4. Empirical Application

A dataset of the author's body weight and physical activity (PA) spanning six years was analyzed. Following Partridge *et al* (2016) [55], we hypothesized that a change in PA regimen causes a change in weight. Outcomes, exposures, and treatment periods were first defined or specified. DGMs for the g-formula and IPW methods were then defined, adjusted for stationarity, and used to estimate and interpret putative APTEs. All hypothesis tests were performed at the 0.05 significance level unless stated otherwise. Throughout this section, we assume we have period exposures, and therefore suppress the $R_{t(j-1)} = 0$ notation.

4.1 Definitions and Specifications

The raw outcome was defined as per-day average body weight, and the constructed

outcome was defined as the average centered body weight (ACBW) per week. Centered body weight was defined as the difference in body weight in kilograms from the empirical average body weight taken over all six years. The raw exposure was defined as engaging in PA on a given day, where PA was defined as some combination of cardiovascular or resistance training (e.g., running, swimming, cycling, rock climbing, weight lifting, push-ups, pull-ups). Following common definitions of one-week PA summary variables (e.g., minutes/week, steps/week, days/week) [56–59], the constructed exposure was defined as the proportion of days per week when any PA was reported, among days when body weight was reported (i.e., non-missing). The resulting constructed time series consisted of 290–293 time points, depending on the specifications below.

The NIOS treatment was specified as a regular pattern of PA spanning one or more weeks. However, among eight highly relevant RCTs in a systematic review by Schoeppe *et al* (2016) [59] with PA or weight as outcomes, studies designed to detect empirically apparent (i.e., statistically significant) effects of interventions spanned six to 14 weeks. Likewise, many of the relevant study periods in a systematic evaluation by Afshin *et al* (2016) [58] were at least two, four, or six weeks long. Hence, we did not expect to find plausibly stable effects for periods shorter than about six weeks. To identify possible treatment period lengths, we conducted changepoint analysis on the constructed-exposure series. Changepoint analysis detects where the mean of an otherwise stationary series changes over time, thereby partitioning the series into a sequence of segments of varying length [60]. Once segments were identified, they were considered to be periods with fixed lengths $\{(m_t)\}$ (i.e., considered as a priori, pre-specified periods of an NIRT). For each segment, PA intensity was defined as “high” if the PA segment mean was greater than 5/7 (i.e., indicating more than 5 days of PA in a week); otherwise, PA was dened as “low”. The treatment level for each one-week-long segment was assigned to the previous segment's treatment level, and the segment identifier was likewise changed to that of the previous segment,

thus ensuring each segment was at least $m_t = 2$ weeks long.

The analysis outcome was defined as weekly ACBW. We assumed each treatment might require time to stabilize, but did not thereafter destabilize until a change in level. Because both treatment levels were mutually exclusive, the washout subperiod was assumed to be a subset of the stabilization subperiod for any period. We assumed that effect stability of either high or low PA in any period was reached by some stability point, i.e., k_0 . To model carryover or stabilization in periods shorter than k_0 , dummy variables corresponding to each observed $k < k_0$ would be included in the model as a function of previous treatment and period length (because, e.g., the effect of low PA might already be stable if preceded by a sufficiently long low-PA period). In addition, we posited that the current outcome might depend on the previous outcome, and that this dependence varies by treatment level. Hence, we specified the g-formula outcome model as

$$Y_{t(j+1)} = \beta_0 + \begin{cases} \{1 - X_{t(j)}\} i \\ \{X_{t-1(j)} + (1 - X_{t-1(j)}) I(m_{t-1} < k_0 - 1)\} i \\ \sum_{k=1}^{k_0-1} \beta_{0k} I(j = k) + \\ \beta_1 Y_{t(j)} + \\ \beta_2 + \\ \{(1 - X_{t-1(j)}) + X_{t-1(j)} I(m_{t-1} < k_0 - 1)\} i \\ \sum_{k=1}^{k_0-1} \beta_{2k} I(j = k) + \\ \beta_3 Y_{t(j)} \\ X_{t(j)} + \\ \xi_{t(j)} \end{cases} \quad i$$

where $\mathbf{k}_{\text{stable}} = \{k : k_0 \leq k \leq m_t\}$. The stable low-PA average baseline effect (ABE) (i.e., baseline average ACBW during weeks of low PA) and APTE were specified as $abe = E(Y_{t(k+1)}^0) = \beta_0 + \beta_1 E(Y_{t(k)})$ and $apte = E(Y_{t(k+1)}^1) - E(Y_{t(k+1)}^0) = \beta_2 + \beta_3 E(Y_{t(k)})$ for $k \in \mathbf{k}_{\text{stable}}$, respectively, assuming $E(Y_{t(j)} | j < k_0) = E(Y_{t(j)} | j \geq k_0)$ holds (see ► Appendix for derivation). Hence, $\bar{w}_{t(j)}^Y = Y_{t(j)}$, and the corresponding estimators are denoted as \widehat{abe} and \widehat{apte} . The IPW method would be applied in a secondary analysis, with its propensity model specified based on our experiences fitting the outcome model.

Stability of both the APTE and ABE would be assessed and reported. We rea-

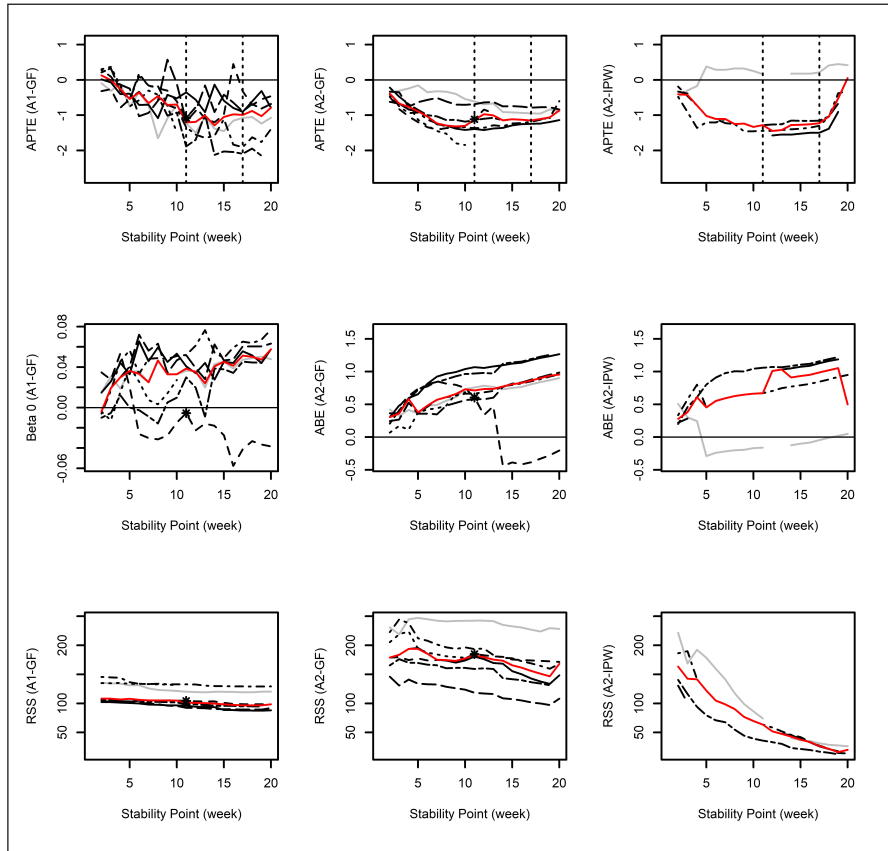


Fig. 1 Trends in valid (i.e., using stationary series) estimates and residual sums of squares (RSSs) across values of k_0 for start days 1 to 7. (In each graph, different line types indicate different start days, and the red line indicates the median value. In the top row, the dotted lines demarcate an interval with a possibly stable APTE. In the left and center columns, the black asterisk indicates start day 3 at stability point $k_0 = 11$, which was chosen for illustration in ► Figures 2 and 3.)

soned that if k_0 equals the true stability point, denoted k_0^* , then mean effects estimated for $k > k_0$ should vary around the true stable APTE regardless of raw series start day (i.e., the first day of both raw outcome end exposure series used to define each constructed series). Hence, we would first vary the value of k_0 from 2 to the length of the second-longest PA segment. The corresponding values of \widehat{abe} and \widehat{apte} would be graphed as a function of k_0 . We would also assess the robustness of our stability findings by varying the start day for each set of analyses. Because the constructed variables were defined using seven days of data, it was reasonable to vary the raw series start day from 1 to 7; we would also examine the findings for start days 8 to 14.

Missing constructed outcomes and exposures would be imputed as follows. To simplify the demonstration of our meth-

ods, we assumed constructed variables were missing completely at random (MCAR). (Data were likely to be at least missing at random [61, 62], and more refined analyses should examine the sensitivity of results to such missingness assumptions; these are beyond the scope of this paper.) Missing values would be linearly interpolated using na.interpolation(), and Gaussian noise added to the imputed constructed outcomes using the empirical means and standard deviations of their non-missing counterparts.

4.2 Post Hoc Analyses

The Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) unit-root tests were performed using `adf.test()` and `kpss.test()`, respectively, to assess stationarity. Stationarity tests of

$\{(Y_{t(j)})\}$ indicated that this series was likely not stationary across most start days and values of k_0 . However, letting $\Delta_{t(j+1)}^y = Y_{t(j+1)} - Y_{t(j)}$ represent the change in outcome from the previous outcome (i.e., first difference), these tests indicated that $\{(\Delta_{t(j+1)}^y)\}$ might have been stationary in most cases. The ADF and KPSS tests were also used to assess stationarity of $\{(X_{t-1(1)})\}$ in each case. Hence, we instead specified the following two analyses.

In Analysis 1, the g-formula was used to model the change in outcome from the previous outcome as

$$\begin{aligned}
 Y_{t(j+1)}^y &= \beta_0 + \\
 &\quad \left\{1 - X_{t(j)}\right\} i \\
 &\quad \left\{X_{t-1(1)} + (1 - X_{t-1(1)}) I(m_{t-1} < k_0 - 1)\right\} i \\
 &\quad \sum_{k=1}^{k_0-1} \beta_{0k} I(j = k) + \\
 &\quad \beta_1 Y_{t(j)}^y + \\
 &\quad \left\{(1 - X_{t-1(1)}) + X_{t-1(1)} I(m_{t-1} < k_0 - 1)\right\} i \\
 &\quad \sum_{k=1}^{k_0-1} \beta_{2k} I(j = k) \\
 &\quad X_{t(j)} + \\
 &\quad \xi_{t(j)}
 \end{aligned} \tag{A1-GF}$$

for $k_0 > 1$ (i.e., to allow at least one week for a treatment effect to occur), where k_{stable} was specified as before. We specified the corresponding APTE of interest as

$$\begin{aligned}
 apte_1 &= \sum_{k=1}^{k_0-1} E(\Delta_{t(k+1)}^y | X_{t-1(1)} = 0, k < k_0) - \\
 &\quad (k_0 - 1) E(\Delta_{t(j+1)}^y | X_{t-1(1)} = 0, j \geq k_0) = \\
 &\quad \sum_{k=1}^{k_0-1} \beta_{2k}
 \end{aligned}$$

because this quantity represents the total mean change in the outcome attributable to high PA (after a period of low PA) before the high-PA mean effect stabilizes if $E(\Delta_{t(k)}^y | X_{t-1(1)} = 0, k < k_0) = E(\Delta_{t(j)}^y | X_{t-1(1)} = 0, j < k_0)$ for all $k < k_0$, and if $E(\Delta_{t(j)}^y | X_{t-1(1)} = 0, j < k_0) = E(\Delta_{t(j)}^y | X_{t-1(1)} = 0, j \geq k_0)$ (see ► Appendix for derivation); hence, $\widehat{w}_{t(j)}^{\Delta^y} = \Delta_{t(j)}^y$. In particular, $\beta_0 = 0$ if $k_0 = k_0^*$, so for Analysis 1 we examined trends in $\widehat{\beta}_0$ across different values of k_0 to assess the stability assumption.

In Analysis 2, we modeled the change in outcome from the previous period's last outcome or the average of its stable out-

comes, whichever occurred last; i.e., $\Delta_{t(j+1)} = Y_{t(j+1)} - Y_{t-1}^*$ where

$$Y_{t-1}^* = I(m_{t-1} \leq k_0) Y_{t-1(m_{t-1})} + I(m_{t-1} > k_0) (m_{t-1} - k_0 + 1)^{-1} \sum_{k=k_0}^{m_{t-1}} Y_{t-1(k+1)}.$$

We posited that $\Delta_{t(j+1)}$ might depend on $\Delta_{t-1}^y = Y_{t-1}^* - Y_{t-2}^*$ (which the ADF and KPSS tests indicated may have been stationary in most cases), and that this dependence varies by treatment level. The Analysis-2 outcome model was specified as

$$\begin{aligned} \epsilon_{(j+1)} = & \gamma_0 + \\ & \left\{ 1 - X_{t(j)} \right\} i \\ & \left\{ X_{t-1(1)} + (1 - X_{t-1(1)}) I(m_{t-1} < k_0 - 1) \right\} i \\ & \sum_{k=1}^{k_0-1} \gamma_{0k} I(j = k) + \\ & \gamma_1^y Y_{t-1}^* + \\ & \gamma_2 + \\ & \left\{ (1 - X_{t-1(1)}) + X_{t-1(1)} I(m_{t-1} < k_0 - 1) \right\} i \\ & \sum_{k=1}^{k_0-1} \gamma_{2k} I(j = k) + \\ & \gamma_3^y Y_{t-1}^* \\ & X_{t(j)} + \\ & \epsilon_{t(j)} \end{aligned}$$

A2-GF

where k_{stable} was specified as before. The ABE (redefined as the baseline average change in ACBW from Y_{t-1}^*) and APTE were specified as

$abe_2 = E(\Delta_{t(k+1)}^0) = \gamma_0 + \gamma_1 E(\Delta_{t-1}^y)$ and $apte_2 = E(\Delta_{t(k+1)}^1) - E(\Delta_{t(k+1)}^0) = \gamma_2 + \gamma_3 E(\Delta_{t-1}^y)$ for $k \in \mathbf{k}_{\text{stable}}$, assuming $E(\Delta_{t-1}^y | j < k_0) = E(\Delta_{t-1}^y | j \geq k_0)$ holds (see ► Appendix for derivation); hence, $\bar{w}_{t(j)}^\Delta = \Delta_{t-1}^y$.

A propensity model for Analysis 2 (A2-IPW) was then specified as follows. We posited that treatment assignment at the start of the current period may have depended on treatment level during the preceding period, and on changes in outcomes over the past one or two periods, defined as $\Delta_{t-1}^y = I(m_{t-1} > 1)(Y_{t-1}^* - Y_{t-2}^*) + I(m_{t-1} = 1)\Delta_{t-2}^y$ where

$$Y_{t-1}^* = I(m_{t-1} - 1 \leq k_0) Y_{t-1(m_{t-1})} + I(m_{t-1} - 1 > k_0) (m_{t-1} - k_0)^{-1} \sum_{k=k_0}^{m_{t-1}-1} Y_{t-1(k+1)}.$$

The treatment propensity model was therefore specified as

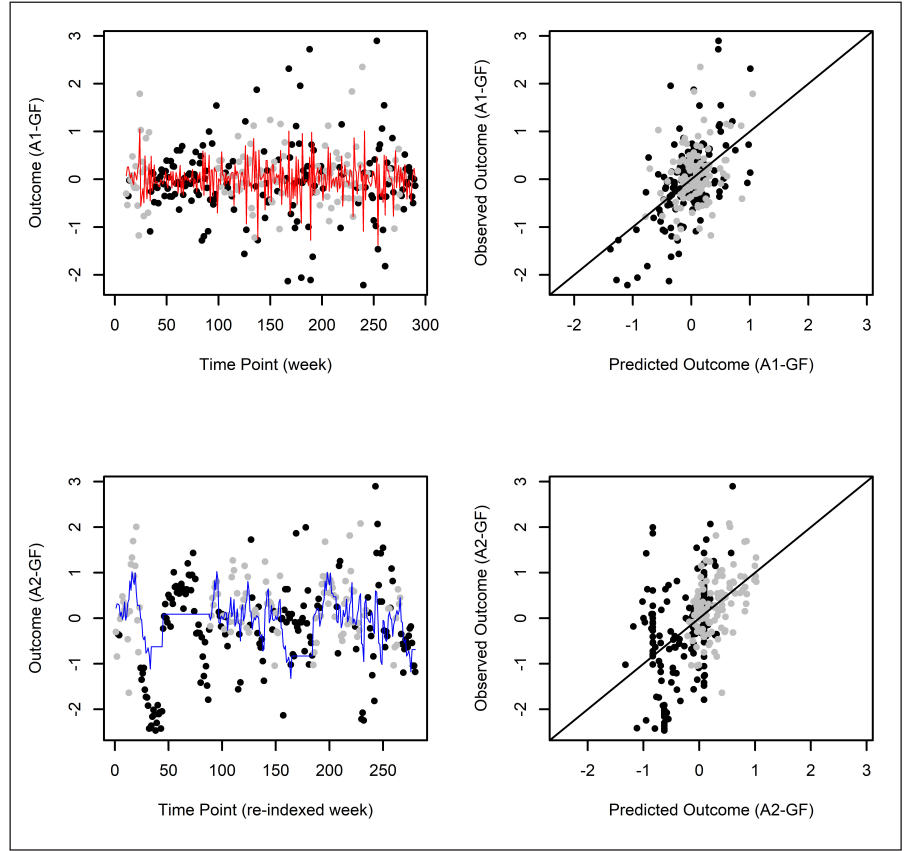


Fig. 2 Analysis outcomes: Observed and predicted outcomes using the g-formula, for start day 3 at stability point $k_0 = 11$. (In each graph, high and low physical-activity analysis outcomes are plotted as black and gray circles, respectively. In the left column, the red and blue lines indicate predicted values for Analyses 1 and 2, respectively. In the right column, observed versus predicted analysis outcomes are plotted.)

$$\begin{aligned} \text{logit}\left(\Pr\left(X_{t(j)} = 1 \mid \bar{w}_{t-1(m_{t-1})}^X\right)\right) = \\ \alpha_0 + \alpha_1 X_{t-1(1)} + \alpha_2 \Delta_{t-1}^{y*} \end{aligned} \quad \text{A2-IPW}$$

where $\bar{w}_{t-1(m_{t-1})}^X \subseteq \{X_{t-1(1)}, \Delta_{t-1}^{y*}\}$. The ABE and APTE were likewise specified as $abe_2 = E(\Delta_{t(k+1)}^0)$ and $apte_2 = E(\Delta_{t(k+1)}^1) - E(\Delta_{t(k+1)}^0)$ for $k \in \mathbf{k}_{\text{stable}}$, but were instead estimated using the IPW formula in Section 3.4. Note, however, that this standard IPW formula cannot be used to flexibly model unstable APTEs; hence, we fit this model using only observations in $\mathbf{k}_{\text{stable}}$. This limitation (coupled with not having known k_0^* , or if it even existed) greatly reduced the IPW method's immediate utility. We nonetheless were able to further assess our stability assumptions by comparing the IPW and g-formula APTE estimates (as noted below).

4.3 Results and Interpretation

► Figure 1 illustrates the trends in valid estimates across values of the assumed stability point k_0 for start days 1 to 7. Findings were considered valid only in cases where in the exposure $\{(X_{t(1)})\}$ and the relevant predictor (i.e., $\Delta_{t(j)}^y$, Δ_{t-1}^y , or Δ_{t-1}^{y*}) were both deemed sufficiently stationary, as indicated by at least one of the two test results corresponding to each series. The median trends (i.e., across all valid findings at each k_0) are also plotted. Only values for $k_0 \leq 20$ are shown, as estimated effects seem to have been either sparse or very noisy past $k_0 = 20$. For Analysis 1 (left panel), the APTE may have been stable between 11 and 17 weeks of treatment (indicated by the median, plotted in red between the dotted vertical lines); however, $\hat{\beta}_0$ did not seem to systematically stabilize to zero. For

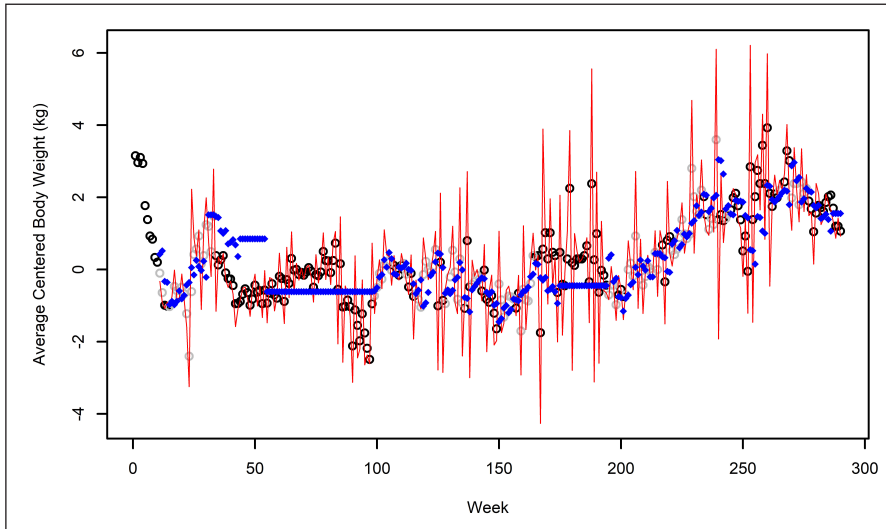


Fig. 3 Average centered body weight: Observed and predicted outcomes using the g-formula, for start day 3 at stability point $k_0 = 11$. (The black and gray circles indicate analysis outcomes corresponding to high and low physical activity, respectively, and the red line and blue asterisks indicate predicted values for Analyses 1 and 2, respectively.)

Analysis 2 (center and right panels), the g-formula APTE may have been stable also between 11 and 17 weeks of treatment; however, the ABE seemed to generally increase. These findings were reflected in the corresponding IPW plots. A sensitivity analysis using the raw series for start days 8–14 produced somewhat qualitatively similar results (see ► Appendix).

For illustration, we report and interpret the set of findings for start day 3 at $k_0 = 11$, marked with black asterisks in ► Figure 1. The valid Analysis-1 estimates were $\widehat{ap}_{te_1} = -1.08\text{kg}$ and $\widehat{\beta}_0 = -0.01\text{kg}$, with a residual sum of squares (RSS) of 104.33. The corresponding estimates for Analysis 2 were $\widehat{ap}_{te_2} = -1.12\text{kg}$ and $\widehat{abe}_2 = 0.60\text{kg}$, with a RSS of 184.91. Our Analysis-1 results meant that 11 weeks of high PA after a period of low PA may have reduced ACBW by about 1.08kg on average, where the estimated APTE may have been stable between 11 and 17 weeks of high PA. Our Analysis-2 results meant that 11 weeks of high PA may have reduced ACBW by about 1.12kg on average, while low PA may have increased ACBW by about 0.60kg on average; the estimated APTE may have been stable between 11 and 17 weeks. Both sets of findings qualitatively resembled those in Naimark *et al* (2015) [57] in association (though not necessarily cau-

sation): After 14 weeks, the intervention group (i.e., who used a health-promoting app) increased their PA by 63 minutes per week on average, while control subjects decreased theirs by an average of 30 minutes. Intervention subjects concurrently lost an average of 1.44kg, while control subjects lost an average of 0.13kg. Our observed and predicted outcomes are plotted in ► Figure 2. The top panel shows that the Analysis-1 predictions modestly fit the analysis outcomes, while the quality of fit of Analysis-2 predictions in the bottom panel was somewhat mixed. In particular, because Analysis 2 assumed APTE stability after 11 weeks, it failed to capture trends in analysis outcomes during the high-PA interval roughly between time points 20 and 90. This can likewise be seen roughly between weeks 40 and 100 in ► Figure 3.

As a sensitivity analysis, we assessed the analytical impact of raw-variable missing data. The median proportions of missing values for weekly ACBW and proportion of PA days across all values of k_0 and all 14 start days were 0.134 (range: 0.127 to 0.158) and 0.052 (0.041 to 0.055), respectively. For each analysis outcome and predictor, we weighted each analysis outcome by the total proportion of days without missing values out of all possible days that could have been used in its derivation.

Bigger analytic weights corresponded to fewer missing raw values (i.e., analytic weight of 1 for no missing values, less than 1 otherwise). The resulting weighted generalized linear model (weighted-GLM) analyses produced similar findings to the unweighted analyses (see ► Appendix).

A few immediate modeling improvements could be made in a future study by noting the following limitations. Our self-tracked data did not include reliable measurements of dietary factors, which likely confounded the exposure-outcome relationships in both analyses. We also did not investigate reverse causality (i.e., the effect of theoretically manipulable weight on PA propensity, e.g., through examining dietary patterns as causes), which might help disentangle possible causal feedback structures between ACBW and PA intensity. Factors such as aging may also play a role in modifying APTEs over time (e.g., by inducing a time trend in an APTE itself). While we did not address moderation or mediation of APTEs, our framework does allow formal specification of such contextual influences. Finally, the impact of noise on the first-difference outcome in Analysis 1 could be characterized using simulations in order to assess its impact on both the analysis procedure and APTE estimates.

5. Discussion

We showed how a counterfactual framework based on n-of-1 randomized trials can be used to specify and estimate causal effects using observational n-of-1 time series data. Our framework is modular: It allows for nesting such that each time point $t(j)$ can itself be specified as a set of sub-points, thus permitting finer-grained specification of causal relationships. The framework might also accommodate traditional RCTs or series-of-N1RTs (i.e., by adding the subscript i to index study participants), as well as help formalize or model more complex causal mechanisms at different scales (using, e.g., hidden Markov models, control theory and dynamical systems models), or mechanisms that better account for context (e.g., sufficient-component causes [63–65]). In the future, we will formalize this framework in terms of causal

graphs in order to ease conceptualization of causal structures.

In an NIOS, a priori definitions and specifications that would be used in an NIRT may not be reasonable, feasible, or even possible. While the analytical goal in an NIRT is to estimate a posited causal effect, the goal in an NIOS is instead to discover which effects to posit in the first place (i.e., causal discovery). We distill Section 4 into the following general six-step NIOS procedure, which might be used to encourage discovery of estimable stable APTEs that can address the original research hypotheses by eschewing stationarization of the treatment. 1. Use relevant research (both idiographic and nomothetic) to generally define outcomes and treatments. 2. Specify treatments, and search for candidate sets of treatment periods. 3. Specify outcomes and stable subperiods. 4. Specify APTEs and models. Assess whether or not invariance could hold. 5. Conduct main analyses. Assess stationarity. 6. Conduct sensitivity analyses, and address missing data. These steps can be repeated as the exploratory study evolves, and relevant analytical developments should be reported. The specifications that yield the best fit and interpretability might be highlighted as yielding the most conclusive findings. Our Section 4 analyses also highlighted the commonly encountered analytical tradeoff between APTE estimability and interpretability.

Statistical learning methods can be used to strengthen the search and modeling components in the above procedure. These include the search for sensible treatment periods or stable subperiods (e.g., through time series clustering), and the search for the outcome and propensity models that fit the data well. In particular, cross-validation and predictive modeling may be quite well-suited to finding the best-fitting DGMs for the g-formula, IPW, or doubly robust estimation methods. This sort of “causal predictive modeling” [66–69] would incorporate principles of statistical estimation and inference, causal modeling, and statistical learning. A number of investigators have taken a similar approach towards such causal discovery (i.e., what Gelman and Imbens, 2013 [67], call reverse causal inference), in particular van der Laan *et al*

(2009) [68], Austin (2012) [69], Athey and Imbens (2015) [37], and Spirtes and Zhang (2016) [66].

We are excited to see how related efforts may likewise help advance idiographic causal discovery in the fields of personalized health and medicine. Still, in pursuing this line of inquiry, it should be kept in mind that “for causal inference, issues of design are of utmost importance; a lot more is needed than just an algorithm” [70]. Rubin (2008) [71] sums it up nicely: “For objective causal inference, design trumps analysis.”

Acknowledgment

Special thanks go to Dr. Michael Baiocchi and Dr. Alex Keil for reviewing my manuscript and providing critical feedback and guidance. I thank Dr. Theodore Walls (at the University of Rhode Island) for introducing me to the concepts of idiographic and nomothetic studies. Thank you to the 2016 Quantified Self Public Health Symposium and Health Data Exploration Network Meeting (where I met Dr. Walls) for their inspiration and direction in the growing field of quantified n-of-1 studies. Thank you to my family and friends for their constant support. I dedicate this paper to Professor E. Michael Foster, a seminal mentor in my study of causal inference, former health econometrician in the field of maternal and child health, fellow Tar Heel, and well-missed friend. To Filipinos, Filipino-Americans, and others who are underrepresented or unacknowledged in science, technology, engineering, and math (STEM), and in academia: Kaya natin 'to! To you, the reader: Know yourself, help others, and find meaning in all things.

References

1. Ponterotto JG. Qualitative research in counseling psychology: A primer on research paradigms and philosophy of science. *Journal of Counseling Psychology* 2005; 52(2): 126–136.
2. Guyatt G, Sackett D, Taylor DW, Ghong J, Roberts R, Pugsley S. Determining optimal therapy – randomized trials in individual patients. *New England Journal of Medicine* 1986; 314(14): 889–892.
3. 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. *Annals of Internal Medicine* 1990; 112(4): 293–299.
4. Backman CL, Harris SR. Case studies, single-subject research, and n of 1 randomized trials: Comparisons and contrasts. *American Journal of Physical Medicine & Rehabilitation* 1999; 78(2): 170–176.
5. Gabler NB, Duan N, Vohra S, Kravitz RL. N-of-1 trials in the medical literature: A systematic review. *Medical Care* 2011; 49(8): 761–768.
6. Lillie EO, Patay B, Diamant J, Issell B, Topol EJ, Schork NJ. The n-of-1 clinical trial: The ultimate strategy for individualizing medicine? *Personalized Medicine* 2011; 8(2): 161–173.
7. Duan N, Kravitz RL, Schmid CH. Single-patient (n-of-1) trials: A pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *Journal of Clinical Epidemiology* 2013; 66(8): S21–S28.
8. Naughton F, Johnston D. A starter kit for undertaking n-of-1 trials. *European Health Psychologist* 2014; 16(5): 196–205.
9. Nikles J, Mitchell G. *The Essential Guide to N-of-1 Trials in Health*. Springer; 2015.
10. Shamseer L, Sampson M, Bukutu C, Schmid CH, Nikles J, Tate R, et al. CONSORT extension for reporting N-of-1 trials (CENT) 2015: Explanation and elaboration. *Journal of Clinical Epidemiology* 2016; 76: 18–46.
11. 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. *Journal of Clinical Epidemiology* 2016; 76: 9–17.
12. Kravitz R, Duan N, the DeCIde Methods Center N-of-1 Guidance Panel (Duan N, Eslick I, Gabler N, Kaplan H, et al. Design and Implementation of N-of-1 Trials: A User's Guide. AHRQ Publication No. 13(14)-EHC122-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2014. <http://www.effectivehealthcare.ahrq.gov/N-1-Trials.cfm>.
13. Nikles CJ, Mitchell GK, Del Mar CB, Clavarino A, McNairn N. An n-of-1 trial service in clinical practice: Testing the effectiveness of stimulants for attention-deficit/hyperactivity disorder. *Pediatrics* 2006; 117(6): 2040–2046.
14. Kravitz RL, Duan N, Niedzinski EJ, Hay MC, Subramanian SK, Weisner TS. What ever happened to n-of-1 trials? Insiders' perspectives and a look to the future. *Milbank Quarterly* 2008; 86(4): 533–555.
15. Chen C, Haddad D, Selsky J, Homan JE, Kravitz RL, Estrin DE, et al. Making sense of mobile health data: An open architecture to improve individual- and population-level health. *Journal of Medical Internet Research* 2012; 14(4): e112.
16. Barr C, Marois M, Sim I, Schmid CH, Wilsey B, Ward D, et al. The PREEMPT study-evaluating smartphone-assisted n-of-1 trials in patients with chronic pain: Study protocol for a randomized controlled trial. *Trials* 2015; 16(1).
17. Schork NJ. Personalized medicine: Time for one-person trials. *Nature* 2015; 520(7549): 609–611.
18. Neyman J. On the application of probability theory to agricultural experiments. *Essay on principles*. Section 9. *Statistical Science*. 1923, tr 1990;5(4):465–480. Translated and edited by D.M. Dabrowska and T.P. Speed from the Polish original, which appeared in *Roczniki Nauk Rolniczych*

- Tom X (1923) 1–51 (*Annals of Agricultural Sciences*).
19. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; 66(5): 688–701.
 20. Holland PW. Statistics and causal inference. *Journal of the American Statistical Association* 1986; 81(396): 945–960.
 21. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. Springer 2000; 95–133.
 22. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; 11: 550–560.
 23. Robins JM, Hernan MA. Estimation of the causal effects of time-varying exposures. *Longitudinal Data Analysis* 2009; 553–599.
 24. Klasnja P, Hekler EB, Shiman S, Boruvka A, Almirall D, Tewari A, et al. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology* 2015; 34(S): 1220–1228.
 25. Murphy SA. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* 2005; 24(10): 1455–1481.
 26. Nahum-Shani I, Hekler EB, Spruijt-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychology* 2015; 34(S): 1209–1219.
 27. Aalen OO, Frigessi A. What can statistics contribute to a causal understanding? *Scandinavian Journal of Statistics* 2007; 34(1): 155–168.
 28. Aalen OO, Roysland K, Gran JM, Ledergerber B. Causality, mediation and time: A dynamic viewpoint. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2012; 175(4): 831–861.
 29. White H, Kennedy P, Bates White L. Retrospective estimation of causal effects through time. *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F Hendry* 2009; 59–87.
 30. White H, Lu X. Granger causality and dynamic structural systems. *Journal of Financial Econometrics* 2010; 8(2): 193–243.
 31. Lu X, Su L, White H. Granger causality and structural causality in cross-section and panel data. *Econometric Theory* 2017; 33(2): 263–291.
 32. Eichler M. Causal inference in time series analysis. *Causality: Statistical Perspectives and Applications* 2012; 327–354.
 33. Eichler M, Didelez V. Causal reasoning in graphical time series models. *arXiv preprint arXiv:12065246*. 2012; 109–116.
 34. Eichler M. Causal inference with multiple time series: Principles and problems. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 2013; 371(1997): 20110613.
 35. Bollen KA. *Structural Equations with Latent Variables*. Wiley-Interscience; 1989.
 36. Collins JD, Hall EJ, Paul L. *Causation and Counterfactuals*. MIT Press; 2004.
 37. Athey S, Imbens GW. Machine learning methods for estimating heterogeneous causal effects. *arXiv:150401132v1 [statML]* 5 Apr 2015. 2015; 1–24.
 38. Hayashi F. *Econometrics*. Princeton University Press; 2000.
 39. Murphy SA. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2003; 65(2): 331–355.
 40. Granger CW. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* 1969; 424–438.
 41. Granger CW. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control* 1980; 2: 329–352.
 42. Granger CW. Some recent development in a concept of causality. *Journal of Econometrics* 1988; 39(1): 199–211.
 43. Rosenbaum PR. Interference between units in randomized experiments. *Journal of the American Statistical Association* 2007; 102(477): 191–200.
 44. Pearl J. *Causality*. 2nd ed. Cambridge University Press, USA; 2009.
 45. Dawid AP. Influence diagrams for causal modeling and inference. *International Statistical Review* 2002; 70(2): 161–189.
 46. Dawid AP, Didelez V. Identifying the consequences of dynamic treatment strategies. *Research Report*; 2005.
 47. Dawid AP, Didelez V. Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys* 2010; 4: 184–231.
 48. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling* 1986; 7(9–12): 1393–1512.
 49. Pearl J, Robins J. Probabilistic evaluation of sequential plans from causal models with hidden variables. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc 1995; 444–453.
 50. Hernan MA, Robins JM. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health* 2006; 60(7): 578–586.
 51. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* 2004; 23(19): 2937–2960.
 52. Morgan SL, Winship C. *Counterfactuals and Causal Inference*. Cambridge University Press; 2014.
 53. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70(1): 41–55.
 54. Lu Y, Zeger SL. On the equivalence of case-cross-over and time series methods in environmental epidemiology. *Biostatistics* 2006; 8(2): 337–344.
 55. Partridge SR, McGeechan K, Bauman A, Phongsavan P, Allman-Farinelli M. Improved eating behaviours mediate weight gain prevention of young adults: Moderation and mediation results of a randomised controlled trial of TXT2BFiT, mHealth program. *International Journal of Behavioral Nutrition and Physical Activity* 2016; 13(1): 44.
 56. Trinh L, Wilson R, Williams HM, Sum AJ, Naylor PJ. Physicians promoting physical activity using pedometers and community partnerships: a real world trial. *British Journal of Sports Medicine* 2012; 46(4): 284–290.
 57. Naimark JS, Madar Z, Shahar DR. The impact of a Web-based app (eBalance) in promoting healthy lifestyles: Randomized controlled trial. *Journal of Medical Internet Research* 2015; 17(3).
 58. Afshin A, Babalola D, Mclean M, Yu Z, Ma W, Chen CY, et al. Information technology and lifestyle: A systematic evaluation of internet and mobile interventions for improving diet, physical activity, obesity, tobacco, and alcohol use. *Journal of the American Heart Association* 2016; 5(9): e003058.
 59. Schoeppe S, Alley S, Van Lippevelde W, Bray NA, Williams SL, Duncan MJ, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: A systematic review. *International Journal of Behavioral Nutrition and Physical Activity* 2016; 13(1): 127.
 60. Killick R, Eckley I. changepoint: An R package for changepoint analysis. *Journal of Statistical Software* 2014; 58(3): 1–19.
 61. Rubin DB. Inference and missing data. *Biometrika* 1976; 63(3): 581–592.
 62. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley & Sons; 2014.
 63. Rothman KJ. Causes. *American Journal of Epidemiology* 1976; 104(6): 587–592.
 64. VanderWeele TJ, Hernan MA. From counterfactuals to sufficient component causes and vice versa. *European Journal of Epidemiology* 2006; 21(12): 855–858.
 65. VanderWeele TJ, Robins JM. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *American Journal of Epidemiology* 2007; 166(9): 1096–1104.
 66. Spirtes P, Zhang K; Springer. *Causal discovery and inference: Concepts and recent methodological advances* 2016; 3(1): 3.
 67. Gelman A, Imbens G. Why ask why? Forward causal inference and reverse causal questions. *National Bureau of Economic Research* 2013.
 68. van der Laan MJ, Rose S, Gruber S. *Readings in targeted maximum likelihood estimation* 2009.
 69. Austin PC. Using ensemble-based methods for directly estimating causal effects: An investigation of tree-based G-computation. *Multivariate Behavioral Research* 2012; 47(1): 115–135.
 70. Sekhon JS. Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science* 2009; 12: 487–508.
 71. Rubin DB. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics* 2008; 808–840.