A Large-Scale Observational Study of the Causal Effects of a Behavioral Health Nudge

Achille Nazaret[*]
Columbia University
aon2108@columbia.edu

Guillermo Sapiro
Apple
gsapiro@apple.com

Abstract

The Apple Watch encourages users to stand throughout the day by delivering a notification onto the users' wrist if they have been sitting for the first 50 minutes of an hour. This simple behavioral intervention exemplifies the classical definition of *nudge* as a choice architecture that alters behavior without forbidding options or significantly changing economic incentives. In order to estimate from observational data the causal effect of the notification on the user's standing probability throughout the day, we introduce a novel regression discontinuity design for time series data with time-varying treatment. Using over 76 billions minutes of private and anonymous observational standing data from more than 160,000 subjects enrolled in the public Apple Heart and Movement Study from 2019 to 2022, we show that the nudge increases the probability of standing by up to 49.5% across all the studied population. The nudge is similarly effective for participants self-identified as male or female, and it is more effective in older people, increasing the standing probability in people over 75 years old by more than 60%. We also demonstrate that closing Apple Watch Activity Rings, another simple choice architecture that visualizes the participant's daily progress in Move, Exercise, and Stand, correlates with user's response to the intervention; for users who close their activity rings regularly, the standing nudge almost triples their probability of standing. This observational study, which is one of the largest of its kind exploring the causal effects of nudges in the general population, demonstrates the effectiveness of simple behavioral health interventions and introduces a novel application of regression discontinuity design extended here to time-varying treatments.

1 Introduction

For people who are able to stand, prolonged sitting is known to negatively affect their health, e.g., increasing premature mortality risk [17] as well as diabetes and cardiovascular events [20]. Conversely, interrupting prolonged sitting by standing for a few minutes has been shown to limit these risks [8, 10], and even reduce postprandial blood glucose [7]. A key question is therefore how to encourage people to stand more. Here is where the concept of *nudges* enters the equation.

Following the seminal work of Thaler and Sunstein [19], a nudge is any aspect of choice architecture that alters behavior in a predictable way, without forbidding options or significantly changing economic incentives. To count as a mere nudge, the intervention must be easily avoidable. Nudges are not mandates, e.g., putting the fruit at eye level is a nudge but banning junk food is not. Not all nudges are successful, see [6, 16] for positive examples and [2, 14] for negative ones. A major challenge is the design of successful nudges to improve health, and another challenge is to evaluate them with observational studies in the general targeted population (most have been evaluated with experimental design such as A/B testing).

[*] Work performed while at Apple.

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

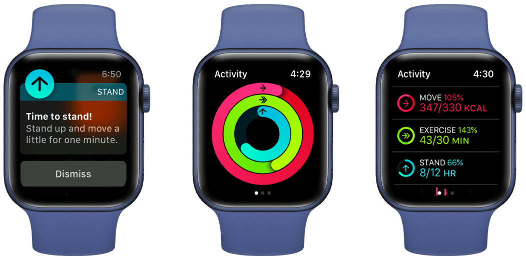

Figure 1: (Left) The standing notification is delivered at 6:50 pm because the user didn't stand from 6:00 pm to 6:50 pm. (Middle) The Apple Watch activity Rings that users are encouraged to *close* each day. (Right) Details of the goals that a user need to reach to close each of the three rings: number of active calorie burned for the Move ring in red, number of exercise minutes for the Exercise ring in green and number of hours containing at least one minute of standing for the Stand ring light blue.

An example of a health nudge is the Apple Watch Standing one, see Figure 1. The *Stand reminder* encourages the user to stand after a long period of inactivity. As described in the Apple Watch settings, the user will "Receive a reminder to stand if [they]'ve been sitting for the first 50 minutes of an hour." The notification will stay on for 10 minutes until the next hour. With the popularity of the Apple Watch, this nudge presents a unique opportunity for studying the causal effects of such behavioral interventions in the general population. In this work, we exploit the Apple Heart and Movement Study [4] to address this. In collaboration with the American Heart Association and Brigham and Women's Hospital, the Apple Heart and Movement Study explores, in a fully private and anonymous fashion, the link between physical activity and heart health. The study is designed to identify factors affecting heart health, mobility, and heart function over time. To illustrate the massive amount of data in this study, the total number of days pooling all users together reaches over 53 million. Using sensors on the Watch, the study records standing data that we bucketed as a binary standing/non-standing variable in 5 minutes intervals, see Figure 2. For our particular study, this data contains over 76 billions minutes of observational standing data from more than 160,000 users collected from 2019 to 2022.

We stress that this is an observational study and not a randomized control trial. As such, it becomes one of the largest study investigating the causal effects of a behavioral health nudge. Being an observational study, we resort to the classical econometrics tool of *regression discontinuity*[2] [3, 5, 13, 15], that we adapt into a novel method for time-series outcomes. A regression discontinuity aims to determine the causal effects of interventions by identifying a threshold above or below which an intervention is assigned. By comparing observations lying closely on either side of the threshold, it is possible to estimate the average treatment effect in environments where randomization is not available. The Apple Watch Standing data is a time series with a *time-varying treatment* (the nudge) with a natural threshold (being idle for at least 50 minutes); in this paper we design a regression discontinuity strategy for time-varying treatment and exploit it to study the causal effects of the Standing nudge. It extends the method of regression discontinuity in time [9] which does not consider time-varying treatments (e.g. nudge, no nudge, no nudge, nudge ...) but only units assigned to a treatment that begins at a particular threshold in time.

2 A regression discontinuity experimental design for time-varying treatment

Standard regression discontinuity design. In a standard regression discontinuity design with binary treatment N and outcome Y [15], the treatment is assigned according to a function g of an additional exogenous variable Z, by $N = g(Z)$. The function g is of the form $z \mapsto g(z) := \mathbb{1}(z \geq \beta)$ for a fixed threshold β. The variable Z is called the forcing variable (or assignment variable) since it determines the treatment assignment. The discontinuity of the function g at $z = \beta$ allows to estimate the causal effect of the treatment N on the outcome Y around $Z = \beta$ using the formula $\lim_{\epsilon \to 0^+} \left(\mathbb{E}\left[Y \mid Z = \beta + \epsilon\right] - \mathbb{E}\left[Y \mid Z = \beta - \epsilon\right]\right)$. A common estimation method is to further assume a parametric model

$$Y = b + a \cdot Z + \tau \cdot N + \epsilon, \tag{1}$$

[2] We also validated our results using synthetic controls [1], this will be reported elsewhere.

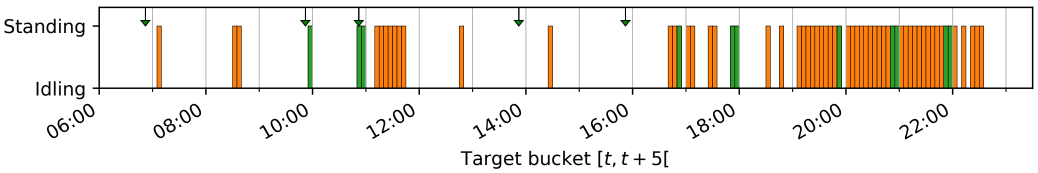

Figure 2: Standing Minutes for a weekday from of one of the author's Health app. Each bar indicates that the user stood during the associated time bucket. The two buckets between 50 and 59 minutes of an hour are drawn in green instead of orange. The top arrows indicate when a nudge notification is triggered. The user ignored the nudges at 13:50 and 15:50, and may have complied to the nudges at 9:55 and 10:50. However, a causal estimator is needed to differentiate if the standing was caused by the nudges or if they were spontaneous standings such as the ones observed at 12:45 and 14:25.

where a, b are the regression coefficients, τ is the causal effect, and ϵ is the random error. Fitting this regression for Z around β then yields an estimation of τ. This design only considers a fixed single treatment N instead of time-varying treatments as in our scenario. It is therefore appealing to extend it for time-varying data, as we do next.

Dynamic treatment strategy for time-varying outcomes. We consider a time-varying outcome variable (Y_t) with a time-varying treatment $(N_t) \in \{0, 1\}$. We assume that (N_t) obeys to a *dynamic treatment strategy* f_t [11, Chap. 19.2]. That is, there exists a function f_t of the previous outcomes that determines each treatment N_t at time t as $N_t = f_t(Y_0, Y_1, ..., Y_{t-1})$. This way, the full history of outcomes becomes a *forcing variable* for the next treatment. For simplicity, and to draw a parallel with the more standard regression discontinuity, we assume that f_t is of the form $f_t(Y_0, ..., Y_{t-1}) = \mathbb{1}(\zeta_t(Y_0, ..., Y_{t-1}) \geq \alpha_t)$ for some function ζ_t and threshold α_t. The function ζ_t projects the multivariate vector of past outcomes into a one-dimensional quantity $\Delta_t := \zeta_t(Y_0, ..., Y_{t-1})$, which is then compared to a threshold α_t to determine the treatment N_t. We are interested in the causal effect of the treatment N_t on the outcome Y_t. In the Apple Standing data in particular, Y_t indicates if the user was standing at time t and N_t if the user was nudged to stand at t. We are therefore interested in the causal effect of the nudge N_t on the standing status Y_t.

Regression discontinuity design for time-varying outcomes. In the time-varying outcome model with dynamic treatment strategy, the treatment is not assigned depending on a single exogenous variable Z, but it is instead assigned depending on the full outcomes history. Nevertheless, if the assignment functions f_t exhibit discontinuities just as the standard function g did, we can estimate the causal effect around the discontinuities with a formula similar to the standard regression discontinuity,

$$\tau_t = \lim_{\epsilon \to 0^+} \mathbb{E}\left[Y_t \mid \Delta_t = \alpha_t + \epsilon\right] - \mathbb{E}\left[Y_t \mid \Delta_t = \alpha_t - \epsilon\right]. \tag{2}$$

In practice, we further assume a parametric linear model adapted from Equation (1) and we fit

$$Y_t = b_t + a_t \cdot \underbrace{\zeta_t(Y_0, ..., Y_{t-1})}_{\Delta_t} + \tau_t \cdot N_t + \epsilon_t \tag{3}$$

around the discontinuity $\Delta_t = \alpha_t$. The coefficients τ_t for each time t estimate the causal effect of N_t on Y_t around $\Delta_t = \alpha_t$ from only observational data. Finally, we define the relative causal effect as $\rho_t = \frac{\tau_t}{\mathbb{E}[Y_t|\Delta_t = \alpha_t] - \tau_t}$, which represents the relative increase in Y_t due to the treatment N_t.

3 Causal effect estimation of the standing nudge

The Stand Minutes data. Users who choose to enroll in the Apple Heart and Movement Study,[3] [4], privately and anonymously share their Stand Minutes data from the Health application. This data indicates if the user is standing or sitting (idling) throughout each day with a resolution of 5 minutes. More precisely, each day is divided into 284 intervals of 5 minutes, referred to as *buckets*, and for

[3] Inclusion criteria: Age > 18 (19 in Alabama/Nebraska and 21 in Puerto Rico); live in the US, own (and not share) an iPhone (with Apple Research App installed) and an Apple Watch; provide informed consent in English.

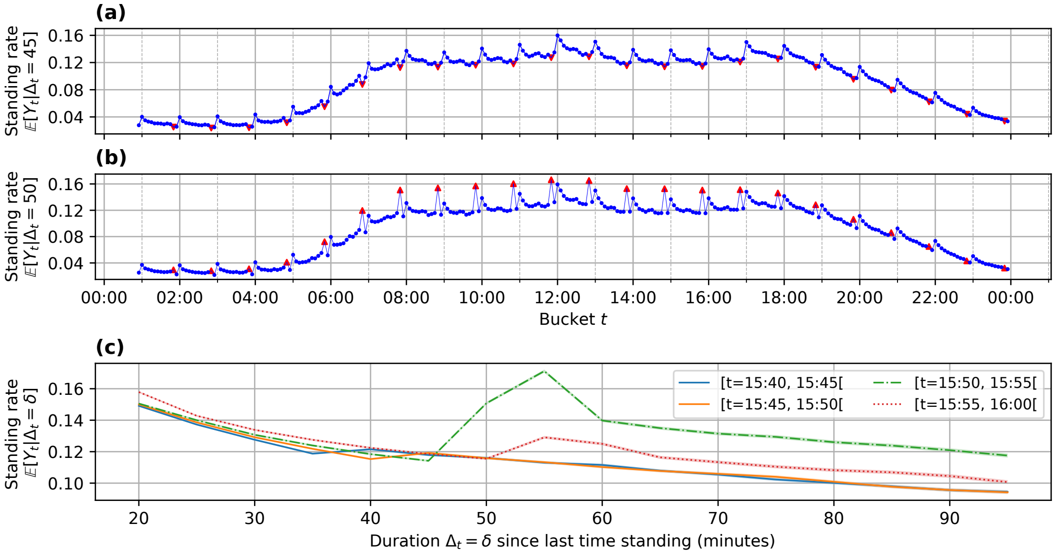

Figure 3: Visualizations of the function $(t, \delta) \mapsto \mathbb{E}\left[Y_t|\Delta_t = \delta\right]$ on corresponding axes. **(a)** Plot of $t \mapsto \mathbb{E}\left[Y_t|\Delta_t = 45\right]$ throughout the day, during which no nudges are triggered because $\Delta_t < 50$. **(b)** Plot of $t \mapsto \mathbb{E}\left[Y_t|\Delta_t = 50\right]$ throughout the day, during which nudges are triggered every hour and 50 minutes because $\Delta_t \geq 50$. In (a,b), red triangles are used for the buckets at each hour and 50 minutes. A discontinuous/jump increase in standing rate is visible from (a) to (b) at each red triangle, corresponding to the nudge activation. **(c)** Function $\delta \mapsto \mathbb{E}\left[Y_t|\Delta_t = \delta\right]$ at four different times t: 15:40 and 15:45 (no nudge), 15:50 (nudge once $\delta \geq 50$), 15:55 (nudge once $\delta \geq 55$). An increase in standing rate happens exactly once the idling δ becomes long enough to trigger the nudge.

each user and bucket, the Stand Minutes data indicates if the user was detected to be standing during the bucket time period. The first bucket of the day is from 00:00:00 to 00:04:59, also written [00:00, 00:05[, and the other buckets follow successively. We refer to buckets by their starting time, such that bucket $[t, t + 5[$ is called bucket t. Figure 2 shows the Stand Minutes profile for a weekday stored in the Health app of one of the authors.

To form the dataset for this large-scale observational study, we pooled the data of 166,034 study participants enrolled between November 14th, 2019, and May 31st, 2022. Days with no standing data were removed. A total of 53,410,838 days of standing data have been accumulated across all participants. It represents a total of 76,911,606,720 minutes and of 15,382,321,344 buckets that are individually labeled *stood* or *idle*. The massive size of this dataset is key for applying a regression discontinuity design, which requires a large quantity of data to eventually form estimators using only data around specific discontinuities.

The time-varying model. For each pair of user u and day d, the binary variable $Y_t^{(u,d)}$ indicates if user u was standing on day d during bucket t. The treatment variable $N_t^{(u,d)}$ is not readily available in the data. Instead, we compute it by replaying the history of standing minutes and apply the rules that are used on the Apple Watch to trigger the nudge:A notification is delivered at 50 minutes of an hour if the subject was idle during the first 50 minutes of the hour. The notification stays on the screen until the end of the hour unless the subject eventually stands or manually dismisses it. This mechanism corresponds to

$$N_t^{(u,d)} = \mathbb{1}\left(\zeta_t\left(Y_{00:00}^{(u,d)}, ..., Y_{t-00:05}^{(u,d)}\right) \geq \alpha_t\right),$$

where ζ_t computes the number of continuous minutes spent idling (not standing) before time t,

$$\Delta_t^{(u,d)} = \zeta_t\left(Y_{00:00}^{(u,d)}, ..., Y_{t-00:05}^{(u,d)}\right) = \max\left\{k \geq 0 \mid (Y_{t-k}^{(u,d)}, Y_{t-k+1}^{(u,d)} ..., Y_{t-00:05}^{(u,d)}) = (0, 0, ..., 0)\right\},$$

and

$$\alpha_t = \begin{cases} +\infty & \text{if } \texttt{t.minutes} \in \{0, 5, ..., 45\}, \\ \texttt{t.minutes} & \text{if } \texttt{t.minutes} \in \{50, 55\}. \end{cases}.$$

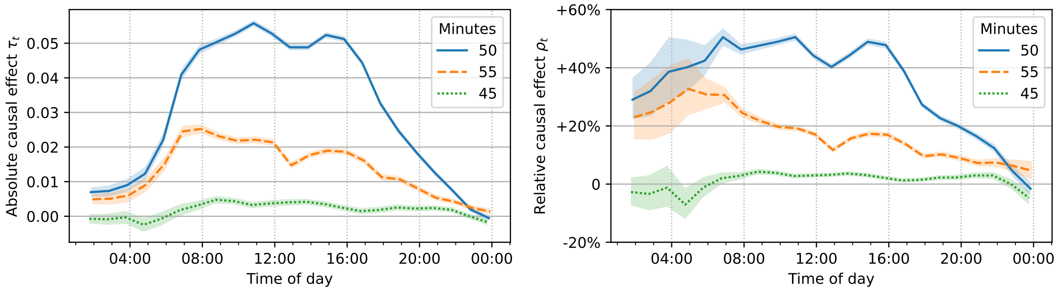

Figure 4: Absolute (left) and relative (right) causal effect of the nudge, estimated over the day at each hour and (intervention) 50, 55, (control) 45 minutes. Error bands show 99% confidence intervals. The relative increase in standing probability at 50 minutes surpasses 40% from 6:00 to 17:00. The notification, which can stay on until 59 minutes, still has an effect at 55 minutes, meanwhile no significant causal effect is detected at 45 minutes.

Figure 3 shows the functions $t \mapsto \mathbb{E}\left[Y_t|\Delta_t = 45\right]$ and $t \mapsto \mathbb{E}\left[Y_t|\Delta_t = 50\right]$ with t varying throughout the day. In Figure 3a, the standing notifications are never triggered because $\Delta_t < 50$. On the contrary, in Figure 3b, the notifications are delivered at 50 minutes of each hour since $\Delta_t \geq 50$. We observe a discontinuous jump/increase in standing probability at each hour and 50 minutes between (a) and (b).

The discontinuity at $\Delta_t = 50$ can be observed in Figure 3c, which shows $\delta \mapsto \mathbb{E}\left[Y_t|\Delta_t = \delta\right]$ for different values of t. For a t that is not at 50 or 55 minutes of an hour (remember that the data is in buckets of 5 minutes), there is no discontinuity, which is expected since no notification nudge are ever triggered before 50 minutes. Instead, for a time t at 50 minutes of an hour (resp. 55), a discontinuity is visible at $\Delta_t = 50$ (resp. $\Delta_t = 55$), which are the thresholds above which $N_t = 1$.

Estimation of the causal effect. We estimate the causal effect τ_t of the notification at each hour and 50 minutes and each hour and 55 minutes, using respectively the discontinuities $\Delta_t = 50$ and $\Delta_t = 55$. We also estimate the causal effect at 45 minutes as a control. We know that there is no notification at 45 minutes and that no causal effect should be discovered. The model from Equation (3) is fitted in Python as a linear model with the library `statsmodels` [18], using the iteratively reweighted least squares (IRLS) method [12]. We restrict the data around the discontinuities $\Delta_t = 50$ or $\Delta_t = 55$ by considering each outcome where $\Delta_t \in [30, 90]$ (Figure 3 suggests that the linear approximation is legitimate on this interval). We also compute the relative effects $\rho_t = \frac{\tau_t}{\mathbb{E}[Y_t|\Delta_t = \alpha_t] - \tau_t}$. These estimations are reported in Figure 4 with 99% confidence intervals. We find a statistically significant positive causal effect of the nudge on the standing probability. The relative increase in standing probability reaches over 40% every hour from 6:00 to 17:00.

In Table 1 we report the relative causal effects of the nudge at 50 minutes, averaged over three periods of the day: the morning, the afternoon, and the evening. We find that the nudge increases the standing rate by 49.5% in the morning and by 43.8% in the afternoon. The relative increase in the evening is around 15.7%. The nudge has a significant impact on the standing rate of the Apple Watch users, especially in the morning and afternoon. These times of day correspond to work hours, which are, for example, moments of prolonged sitting for jobs of the service sector.

4 Controlling the causal effect estimation for population covariates

The Apple Heart and Movement Study contains private and anonymous information about the participants' demographics, which can be used to measure the impact of the nudge on different sub-populations. When the users choose to enroll in the study, they submit their age as well as their biological sex. The age is mandatory to determine eligibility, while the biological sex is optional and can be one of Female, Male or Other. In addition, the Apple Heart and Movement Study contains information about the users' activity level, which we leverage to understand how behavioral differences emerge between people with different engagement with their fitness levels. Tacking into account these sub-groups, we report the relative causal effect between 6:00 and 23:59 in Figure 5 and the relative causal effect grouped by period of day in Table 1.

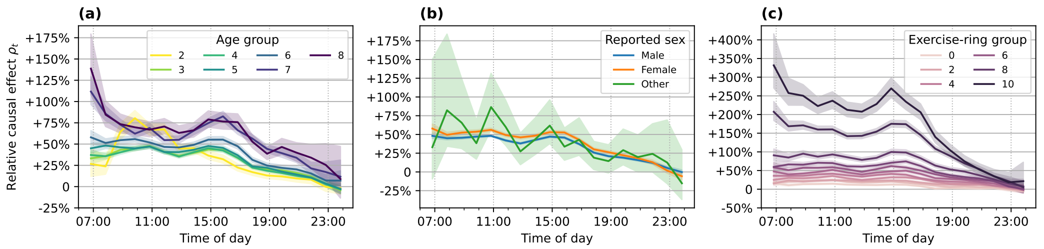

Figure 5: Effects of the Standing nudge for different sub-groups. **(a)** Age effects. Group $n \in \{2, 3, \ldots, 8\}$ contains individuals between $10n - 5$ and $10n + 4$ years old (for $n = 2$ we start at 18 instead of 15). Older individuals are more compliant with the nudge. **(b)** Self-identified biological sex shows no large differences in the effects of the nudge. **(c)** Group n closes the exercise ring between $10n\%$ (included) and $10(n+1)\%$ (excluded) of the days – group 10 closes 100% of the days. The more engaged the participants are in closing the activity rings the more the Standing nudge affects their behavior (we display the Exercise ring, the same patterns are observed for the two other rings).

Controlling for age. We form seven age groups using the users reported age. Group $n \in \{2, 3, 4, 5, 6, 7, 8\}$ contains individuals between $10n - 5$ and $10n + 4$ years old (for $n = 2$ we start at 18 instead of 15). We estimate the causal effect of the nudge on each group separately and report the results in Figure 5a and Table 1. Older individuals are more compliant with the nudge.

Controlling for biological sex. Figure 5b shows that the nudge is similarly effective for participants self-identified as *male*, *female*, or *other*. No statistically significant differences exists between the groups. The group *other* exhibits large confidence intervals because it contains fewer individuals.

Moderating by the engagement with the Apple Rings. We group users by how often they close their Exercise ring (set to a goal of 30 minutes of exercise per day for the majority of users). Individuals in group n achieve their exercise goal between $10n\%$ (included) and $10(n+1)\%$ (excluded) of the days (group 10 completes the goal every single day). Figure 5c shows that participants who close the Exercise ring very often are much more compliant to the Standing nudge. We observe the same behavior with the two other rings.

5 Discussion and concluding remarks

In this paper we extended classical regression discontinuity to time-varying data and applied it to study the causal effects of the Standing nudge in Apple Watch. This works constitutes one of the largest studies on causal effects of behavioral health nudges from observational data, and helped to demonstrate that this simple nudge significantly and positively affects behavior.

Table 1: Quantitative causal effects of the Standing nudge.

		Relative causal effect ρ (in %)		
Sub-population		Morning (6:00-11:59)	Afternoon (12:00-17:59)	Evening (18:00-23:59)
All	–	49.5 $\pm$0.6	43.8 $\pm$0.5	15.7 $\pm$0.6
Age group (years old)	$18 - 34$	46.8 $\pm$1.2	38.0 $\pm$0.9	12.8 $\pm$1.0
	$35 - 54$	45.3 $\pm$0.8	40.3 $\pm$0.7	13.1 $\pm$0.8
	$55 - 84$	63.2 $\pm$1.4	59.9 $\pm$1.2	26.3 $\pm$1.5
Exercise Ring (% of closed days)	$0 - 39$	21.6 $\pm$0.6	17.7 $\pm$0.5	5.7 $\pm$0.6
	$40 - 79$	52.6 $\pm$1.1	49.6 $\pm$0.9	20.8 $\pm$1.1
	$80 - 99$	142.8 $\pm$3.3	140.9 $\pm$2.9	49.7 $\pm$2.7
	100	275.1 $\pm$14.8	276.8 $\pm$14.2	63.4 $\pm$8.4

Acknowledgments

The authors thank Eitan Sapiro-Gheiler from MIT Economics for introducing us to regression discontinuities and discussions on its applications to this work. We also thank Andy Miller, Nick Foti, and Shirley Ren from Apple for valuable discussion and feedback, and the members of the AHMS for their support and constructive comments on this work.

References

[1] A. Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59:391–425, 2021.

[2] S. Agarwal, E. Araral, M. Fan, Y. Qin, and H. Zheng. Water conservation through plumbing and nudging. *Nature Human Behaviour*, pages 1–10, 2022.

[3] J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

[4] Apple. Apple Heart & Movement Study. `https://clinicaltrials.gov/ct2/show/NCT04198194`, 2019. ClinicalTrials.gov Identifier: NCT04198194.

[5] S. Athey and G. W. Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31:3–32, 2017.

[6] S. Berger, A. Kilchenmann, O. Lenz, A. Ockenfels, F. Schlöder, and A. M. Wyss. Large but diminishing effects of climate action nudges under rising costs. *Nature Human Behaviour*, pages 1–5, 2022.

[7] A. J. Buffey, M. P. Herring, C. K. Langley, A. E. Donnelly, and B. P. Carson. The acute effects of interrupting prolonged sitting time in adults with standing and light-intensity walking on biomarkers of cardiometabolic health in adults: A systematic review and meta-analysis. *Sports Medicine*, pages 1–23, 2022.

[8] A. Cooper, S. Sebire, A. Montgomery, T. Peters, D. Sharp, N. Jackson, K. Fitzsimons, C. M. Dayan, and R. Andrews. Sedentary time, breaks in sedentary time and metabolic variables in people with newly diagnosed type 2 diabetes. *Diabetologia*, 55:589–599, 2012.

[9] C. Hausman and D. S. Rapson. Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, 10:533–552, 2018.

[10] G. N. Healy, D. W. Dunstan, J. Salmon, E. Cerin, J. E. Shaw, P. Z. Zimmet, and N. Owen. Breaks in sedentary time: Beneficial associations with metabolic risk. *Diabetes Care*, 31: 661–666, 2008.

[11] M. A. Hernán and J. M. Robins. *Causal Inference: What If.* Chapman & Hall/CRC, 2020.

[12] P. W. Holland and R. E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-Theory and Methods*, 6:813–827, 1977.

[13] G. W. Imbens and T. Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142:615–635, 2008.

[14] A. S. Kristal and A. V. Whillans. What we can learn from five naturalistic field experiments that failed to shift commuter behaviour. *Nature Human Behaviour*, 4:169–176, 2020.

[15] D. S. Lee and T. Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48:281–355, 2010.

[16] K. L. Milkman, M. S. Patel, L. Gandhi, H. N. Graci, D. M. Gromet, H. Ho, J. S. Kay, T. W. Lee, M. Akinola, J. Beshears, et al. A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proceedings of the National Academy of Sciences*, 118:e2101165118, 2021.

[17] N. Owen, G. N. Healy, C. E. Matthews, and D. W. Dunstan. Too much sitting: The population-health science of sedentary behavior. *Exercise and Sport Sciences Reviews*, 38:105, 2010.

[18] S. Seabold and J. Perktold. Statsmodels: Econometric and statistical modeling with python. In *Python in Science Conference*, volume 57, pages 10–25080, 2010.

[19] R. H. Thaler and C. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press, 2008.

[20] E. G. Wilmot, C. L. Edwardson, F. A. Achana, M. J. Davies, T. Gorely, L. J. Gray, K. Khunti, T. Yates, and S. J. Biddle. Sedentary time in adults and the association with diabetes, cardiovascular disease and death: Systematic review and meta-analysis. *Diabetologia*, 55:2895–2905, 2012.