American Economic Review: Papers & Proceedings 2017, 107(5): 481–485
https://doi.org/10.1257/aer.p20171085

Wearable Technologies and Health Behaviors: New Data and New Methods to Understand Population Health[†]

By Benjamin Handel and Jonathan Kolstad*

In 2014, US health care spending was $3 trillion ($9,523 per person) or 17.5 percent of GDP, more than any other nation. Yet US life expectancy, and many other measures of population health, were worse than OECD counterparts (OECD 2016). Poor diet, lack of exercise, and limited medication adherence, amongst other behaviors, could affect this gap and are seen as critical to improving health outcomes and to reducing the cost of health care. Up to one-third of all deaths in the United States result from a condition that can be modified by changes in behavior (Loewenstein, Brennan, and Volpp 2007). It is crucial for economic and policy analysis of health care to better understand why such behaviors persist and how they might be changed.

One potential explanation is that the private returns to behavior improvement are small relative to social cost. For example, moral hazard that arises when someone does not bear the cost of their illness may undermine their willingness to make up-front investments in health-improving behaviors. However, in general in the context of US health care, the private returns for individuals to changing health behavior—increased longevity and improved health—are large. Take the case of individuals who have had a heart attack: Jackevicius, Mamdani, and Tu (2002) show that among patients suffering an acute myocardial infarction (AMI), the share taking cholesterol lowering medication declines to 50 percent within one year. This is despite the fact that financial costs are small and health costs are both salient and substantial: these medications dramatically lower rates of future heart attacks, other cardiac disease, and mortality.

Improving health behaviors at the population level must, therefore, address not only the externalities of health behaviors but the internalities driving a gap between the true private optimal health-improving actions for an individual and the effort undertaken in practice (Baicker, Mullainathan, and Schwartzstein 2015). One approach to close this gap is to help individuals overcome barriers by aiding them in (i) identifying the health impacts of different behaviors; (ii) setting up a plan to improve behaviors; and (iii) monitoring and supporting follow-through on the plan for behavior change.

Small-scale experiments have demonstrated that individuals who plan to improve behaviors and stick to that plan see better outcomes (Milkman et al. 2011). Whether these findings can be scaled up is less clear. One way to accomplish this is through *wearable* and other personal IT tools that enable individuals to understand and make a personal plan to change their health behaviors (e.g., exercise, eating, sleep). We study this question empirically in the context of a randomized intervention at a large employer.

I. Data and Setting

In order to study the impact of wearable technology and its components, we implemented a large-scale randomized control trial between October 2015 and May 2016. The firm provided a subsidy to employees to purchase a wearable wristband beginning in October 2016. The subsidy was large (75 percent of the cost of the approximately $200 device), encouraging widespread adoption by a broad cross section of the

* Handel: Economics Department, UC Berkeley, 521 Evans Hall, Berkeley, CA 94720, and NBER (e-mail: handel@berkeley.edu); Kolstad: Haas School, UC Berkeley, 2200 Piedmont Ave., Berkeley, CA, 94720, and NBER (e-mail: jkolstad@berkeley.edu). We thank Eva Lyubich and Aaron Kaye for outstanding research assistance and Microsoft Research for support of the project. We also thank Amy Finkelstein, Paul Gertler, Justin Sydnor, and Reed Walker for helpful comments. All errors are our own.

[†] Go to https://doi.org/10.1257/aer.p20171085 to visit the article page for additional materials and author disclosure statement(s).

employee population. The wristband included 11 different sensors that allowed a user to, amongst other things, measure exercise activities (e.g., steps, cycling), skin conductivity, UV light, heart rate, and sleep. In addition, the wearable could be linked with a user's phone to receive e-mails, texts and, importantly for our purposes, reminders and updates to help adherence to their personal sleep and exercise plans.

The primary treatment in our study is access to a web-based tool that allows users to upload their data from the wearable, assess their performance, and to select and customize plans to improve sleep and exercise. For example, a plan to improve sleep would allow a user to establish a target time to be in bed each night as well as a target wake-up time. The user can also establish a reminder for times after which she should not drink caffeine and a time after which she should not use electronics (the light from which can disrupt sleep). After setting up this plan, she could then receive reminders on her wearable about when to change behavior over the course of each day. Similarly, an individual wanting to improve exercise could establish daily step goals, receive updates on progress over the course of the day as well as reminders that he needs to improve performance.

A large portion of employees at the partner firm were given the option to participate in the study. Subsidies of $150 were offered to 20,211 individuals; 17,276 individuals ultimately purchased a wearable and, of that group, 14,911 connected to the website to be included in the study. Seventy-five percent of this final study group were randomized into the treatment condition: they were invited to enroll in a wearable-based plan to improve sleep and exercise. The remaining 3,600 individuals remained in the control group through the entire sample period, ending in May 2016. Among the group who were invited to join a plan, 27 percent joined a plan to improve sleep and 20 percent enrolled in a plan to improve exercise. Our primary analysis period is between January 15 and May 15, 2016.

The control group was able to log into the website and to load their data but they were not able to monitor performance and use planning tools to establish a program to improve sleep or exercise. The distinction between the groups, therefore, is the planning aspects of the experience.

We obtain micro-level data from the wearable technology for all individuals in the study throughout the study period. Aggregating these data, we are able to monitor daily exercise, sleep, and use/wearing of the wristband. In addition, we administered a survey to gather data on (i) knowledge about their own health behaviors; (ii) knowledge about the wearable and associated planning tools; (iii) health state and medical conditions; and (iv) some basic family demographic measures.

II. Evidence on Population Sleep

Prior to our analysis of the impact of planning on sleep and exercise behavior, we present some simple summaries of population-level sleep behavior. Little research in economics has focused on the role of sleep in key outcomes of interest. One potential barrier to this kind of analysis is the ability to capture and assess data on sleep. Wearables have the potential to overcome this barrier, for sleep behavior as well as other measures of population health.

To assess sleep in the population, we estimate a simple model that includes individual level fixed effects and daily fixed effects. These results are presented in Figure 1. Panel A presents the distribution of individual fixed effects for sleep hours. The main mass is at zero relative to the population mean of 6.3 hours per night. We see a wide range in average sleep with skewness in the negative direction. It appears only a small share of the population is consistently sleeping the 7–9 hours recommended by the National Sleep Foundation. Interestingly, we find a small number of employees appear to have very little sleep consistently. Panel B presents within-individual variation in sleep, captured by the distribution of day fixed effects from the same regression. We see a large share of days with no significantly different effect. However we also find a mass above suggesting a consistent pattern within individuals in which they increase their average nightly sleep by approximately 25 minutes. The effect primarily captures the increases in sleep in the population on weekend days.

III. Results

A. *Average Impact*

We first compare the average outcome for the treatment to the control group during the study period. The difference between the groups is the

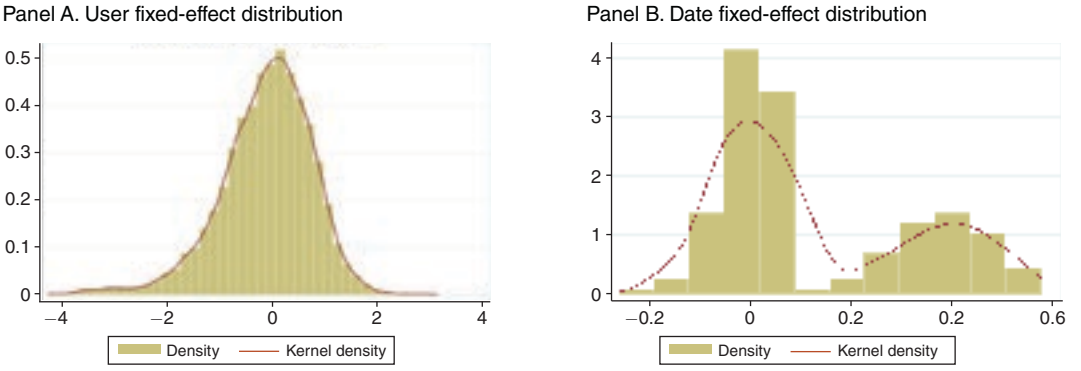

FIGURE 1. DISTRIBUTION OF INDIVIDUAL AND DAY FIXED EFFECTS FOR SLEEP HOURS

Intent To Treat (ITT) impact of access to and use of plans. However, not all individuals in the control group actually enrolled in a plan and a small fraction in the control group were able to access a plan. To get an estimate of the impact of plans themselves on outcomes, we rescale the estimates using treatment as an instrument for plan take-up.

Table 1 presents the main results. Panel A presents the means for each of our outcome variables of interest for both treatment and control populations. The third column then presents the difference, the ITT estimate for the impact of wearable planning tools. Finally, column 4 presents the IV estimate using being in the treatment group as an instrument for plan enrollment.

The difference in average daily steps between treatment and control of 51.85 steps per day is statistically significant and precisely estimated. The magnitude of the effect is small, both relative to the mean and in aggregate. This constitutes about a 1 percent change in total steps and is roughly the equivalent of walking an additional 130 feet per day. The associated IV estimate for the impact of planning is larger, reflecting the less than full take-up of the activity plan. However, the effect on behavior remains relatively small, 150 additional steps per day or approximately 2.5 percent of the mean steps per day.

The second row presents similar estimates for the cardiovascular exercise score. We find, again, statistically significant and economically small impacts.

Rows 3 and 4 present our estimates of the impact on sleep. We find a statistically significant difference in sleep hours for treatment and control of 0.02 and an IV impact of 0.06. These effects are precise and small (1.2 and 3.6 minutes, respectively). We find no significant difference between the groups in the measured quality of sleep.

The fifth row of panel A presents the differences in the rates with which enrollees actually wore their wristband. This is an outcome of interest in and of itself. Many technology companies focus on finding ways to increase user *engagement* with products. Providing enhanced experience or allowing an individual to improve behavior and customize a plan is one way this could be done. In addition, differences in rates of wear are important as they could induce measurement error in our outcome variable. The results of the fifth column in panel A of Table 1 show that being in the treatment group is associated with an increase in hours worn of 0.13 (0.8 percent of the control mean). The IV estimate for the effect is 0.35 (2 percent of the control mean).[1]

Panel B presents the same approach but limits the sample to only the period from January 15 to February 15. We do this for two reasons. First, we might expect an initial period of high engagement where we are particularly interested in effects. Second, differential attrition is less likely to affect results in the early periods of

[1] In order to address the relationship between wearing and measurement, we estimate a series of regressions including controls for hours worn for each individual in each day. The results are remarkably similar to the raw mean results.

Table 1—Mean Outcome for Treatment and Control with Different Sample Restrictions

	Treatment mean	Control mean	Diff	IV
Panel A. Base results				
Steps per day	6,040.32	5,988.47	51.85	154.8
Cardio score	1,069.00	1,056.38	12.62	37.55
Sleep time (hours)	6.30	6.28	0.02	0.0627
Sleep recovery quality	43.02	42.92	0.10	0.300
Hours worn per day	16.15	16.02	0.13	0.349
Panel B. First month of treatment				
Steps per day	5,390.95	5,322.77	68.18	230.3
Cardio score	953.44	928.34	25.10	84.19
Sleep time (hours)	6.35	6.32	0.03	0.114
Sleep recovery quality	43.92	43.83	0.08	0.301
Hours worn per day	16.42	16.28	0.13	0.386
Panel C. Low groups				
Steps per day	3,885.81	3,852.78	33.03	103.0
Cardio score	588.84	614.39	−25.55	−80.00
Sleep time (hours)	4.92	4.99	−0.07	−0.252
Sleep recovery quality	39.25	37.87	1.37	4.415
Hours worn per day	15.21	14.67	0.54	1.573

the study. The results are similar, though slightly larger in magnitude for both sleep and exercise.

Panel C presents results that limit the analysis to those with a baseline level of behavior that is low ($< 3{,}000$ steps per day or < 4.5 hours per night averaged between December 1 and January 15). For this group, we find somewhat different effects. If anything, rates of cardio score and sleep hours *decline* with plan enrollment. Treatment does seem to be a greater influence on hours worn in this group. Finally, we find some evidence that for those with low sleep at baseline, the quality of sleep does seem to improve for the treatment group (by 11 percent). These low exercise and low sleep populations are especially interesting groups to study in future work on health behavior changes, and may be groups whose behavior is especially difficult to alter.

Together, our results suggest small effects of access to planning tools associated with wearable technology we study on either exercise or sleep outcomes. Based on the scale of the study, we have precision to capture small differences between treatment and control. Both ITT and IV estimates based on actual plan enrollment for the treatment group suggest statistically significant but economically small changes in behavior after three months.

B. *Heterogeneous Treatment Effects*

The scale of the intervention as well as the rich data in our setting provide a unique opportunity to assess the heterogeneous impacts of wearable technologies and IT. In addition to being an attractive empirical setting to assess this question, response heterogeneity is an important economic question in this market. The value of wearable technologies and the associated data-driven tools to change behavior hinge on the ability to customize tools. We expect some groups to respond to the implemented plan if subgroup-specific experiences are a key driver of value in wellness and wearable health technologies. Furthermore, as a growing number of assessments by both academic researchers and industry groups find little value in population-based wellness initiatives, speculation has increased that these (often imprecise) zero effects might mask groups for whom outcomes are improving.

To assess these questions, we implement a machine learning approach developed by Athey and Imbens (2015) to identify heterogeneous causal effects. We refer the interested reader to their paper for details but note two key features that make it a particularly attractive methodology in our setting: we can overcome the

concern about *over fitting* and the end nodes of a regression tree (or random forest) have standard asymptotic properties for hypothesis testing without imposing a sparsity assumption.

We fit highly flexible models (tree-based and random forest) of the relationship between rich observables and treatment effects and simply ask whether there is any relationship. In both cases, as complexity is increased we find little improvement in out-of-sample (OOS) fit for treatment effects. Including demographic, survey, and baseline behavioral data, we find the OOS R^2 for the treatment effect remains below 1 percent. This demonstrates that, in our setting, the role of treatment effect heterogeneity as a function of observables is unlikely to play a substantial role in the impact of access to plans.

Our findings also suggest that efforts to further personalize or target planning tools, at least in the context we study, are unlikely to yield either improved outcomes or make the program more efficient (e.g., by spending resources only on a subset of the population). The application of this method in other RCT settings has the potential to be a fruitful way to test for a role of heterogeneity without parametric assumptions or ad hoc analysis.

IV. Conclusion

Our study has a number of important caveats. First, we cannot distinguish the impact of simply being able to see personal data on behavior, a plausible intrinsic motivator for improvement. Second, we study a specific tool in a specific setting. We cannot rule out different results for a different user experience or alternate form of planning or personalization. Finally, the approach to identifying heterogeneous treatment effects is a flexible function of observables. We cannot rule out unmeasured differences that drive important differences in user experience.

Despite these limitations, we believe wearable data has important potential for measurement of key economic behaviors and outcomes (e.g., sleep, stress). Furthermore, as firms and policymakers increasingly implement randomized control trials to evaluate programs, particularly in health and wellness, our approach demonstrates a feasible pathway for future evaluation to tractably study average and heterogeneous treatment effects.

REFERENCES

Athey, Susan, and Guido Imbens. 2015. **"**Recursive Partitioning for Heterogeneous Causal Effects." https://arxiv.org/pdf/1504.01132v3.pdf.

Baicker, Katerine, Sendhil Mullainathan, and Joshua Schwartzstein. 2015. "Behavioral Hazard in Health Insurance." *Quarterly Journal of Economics* 130 (4): 1623–67.

Jackevicius, Cynthia A., Muhammad Mamdani, and Jack V. Tu. 2002. "Adherence with Statin Therapy in Elderly Patients with and without Acute Coronary Syndromes." *Journal of the American Medical Association* 288 (4): 462–67.

Loewenstein, George, Troyen Brennan, and Kevin G. Volpp. 2007. "Asymmetric Paternalism to Improve Health Behaviors." *Journal of the American Medical Association* 298 (20): 2415–17.

Milkman, Katherine L, John Beshears, James J. Choi, David Laibson, and Brigitte C. Madrian. 2011. "Using Implementation Intentions Prompts to Enhance Influenza Vaccination Rates." *Proceedings of the National Academy of Sciences* 108 (26): 10415–20.

Organisation for Economic Co-operation and Development. 2016. "OECD Health Statistics." Organisation for Economic Co-operation and Development. www.oecd.org/health/health-data.htm (accessed January 3, 2017).