Targeting Healthcare Nudges*

Toshiaki Iizuka† — The University of Tokyo
Junya Kawamura‡ — The University of Tokyo
Hitoshi Shigeoka§ — The University of Tokyo, Simon Fraser University, IZA and NBER

First version: March, 2025

Abstract

Individuals self-select into healthcare treatment, but those who choose treatment may not be the ones who benefit the most. We examine this issue in the context of diabetes care in Japan, where those who pass the arbitrary-set diabetic threshold receive health warnings. We find that exceeding the threshold increases healthcare utilization and improves health outcomes. However, individuals who comply most with health warnings are observably healthier, while those who benefit the most are sicker, indicating negative selection on gains. Policy simulations suggest that targeting based on observable characteristics can substantially improve cost effectiveness, emphasizing the need to carefully account for heterogeneous treatment effects.

Keywords: Targeting, Health signals, Selection on gains, Regression discontinuity design, Conditional average treatment effect, Diabetes
JEL codes: I18, I13, I11, J13

* The authors thank Hunt Allcot, David Chan, Kuan-Ming Chen, Karen Eggleston, Koichiro Ito, Hiroyuki Kasahara, Toru Kitagawa, Jonathan Kolstad, Shiko Maruyama, Fabrizio Mazzonna, Magne Mogstad, David Molitor, Shosei Sakaguchi, Lin-Tung Tsai, and participants at ASHEcon 2024, ASSA 2025, and Kyoto Summer Workshop on Applied Economics 2024. We thank Japan Health Insurance Association for providing data. Shigeoka acknowledges financial support from JSPS KAKENHI (23H00828, 22H00057, 22H00847, 22H05009), and JST ERATO (JPMJER2301). Shun Takagi provided excellent research assistance. All remaining errors are our own.

† Faculty of Economics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN. Email: iizuka@e.u-tokyo.ac.jp

‡ Faculty of Economics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN. Email: iizuka@e.u-tokyo.ac.jp

§ Graduate School of Public Policy, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, Simon Fraser University, IZA, and NBER. Email: hitoshi.shigeoka@gmail.com

1. Introduction

Governments, health insurers, and healthcare providers frequently assess individuals' health status and encourage them to visit a physician if any health issues are detected. One example is cancer screening, which urges individuals to undergo further diagnostic tests if abnormalities are found in the initial screening. An important issue we believe is often overlooked is that the decision to respond to these nudges is typically voluntary and rests with the individual with limited medical knowledge. Consequently, those who "comply" with health signals may not necessarily be the ones who "benefit" the most from visiting a physician.

Targeting in health care involves a two-dimensional challenge: balancing costs and benefits. While nudges are low-cost, they can increase spending by prompting doctor visits, which may sometimes lead to unnecessary tests and treatments (Baicker et al. 2015). If those who respond to nudges are not the ones who benefit most, better targeting could improve efficiency by focusing on individuals with a higher net benefit from follow-ups. This contrasts with most other contexts, which focus only on benefits, such as nudges encouraging financial aid renewals (Athey et al., forthcoming).

We examine these issues in the context of targeting nudges following health screenings, where individuals receive a report card after an annual mandatory checkup and are encouraged to schedule a follow-up appointment with a physician if a biomarker exceeds a threshold for diabetes mellitus (DM) patients. Specifically, an individual with a fasting blood sugar level of 126 mg/dl or higher, a standard threshold for diagnosing diabetes, is considered to have diabetes and, consequently, receives the nudges.

We conduct our analysis in three steps. First, we apply a regression discontinuity design (RDD) to a diabetes diagnosis threshold to estimate the average treatment effect of the nudge at the threshold. Individuals just below and above the threshold have similar health statuses, but only those above the threshold receive a nudge to visit a physician for a consultation. Thus, by comparing individuals just below and above the threshold, we can identify the impact of the nudge. We estimate the impact of the nudge on i) healthcare utilization and spending during one year after an index checkup, ii) lifestyle habits, and iii) health outcomes at the checkup one year later.

While this approach identifies the average treatment effect of the nudges at the threshold, it does not address our primary interest in heterogeneity in who complies with nudges and who benefits from physician visits. Nor does it reveal discrepancies between responders and beneficiaries. To address this, in the second step of our analysis, we estimate the conditional

average treatment effect (CATE) using a machine learning method. CATE estimates clarify heterogeneous impacts and highlight key individual characteristics associated with the heterogeneous treatment effects. We use the causal forest model (Wager and Athey 2018), which incorporates the RDD structure when estimating heterogeneous treatment effects. Unlike conventional methods that divide samples by predefined arbitrary criteria (such as individuals over age 70 and those diagnosed with hypertension), this approach systematically identifies observable traits driving heterogeneous treatment effects.

The discrepancy between selection heterogeneity (who responds) and treatment heterogeneity (who benefits), if it exists, indicates the potential for targeting based on observable factors (Kitagawa and Tetenov 2018; Athey and Wager, 2021). We consider two scenarios: (1) a full-information case where the insurer nudges only those with expected positive treatment effects, aiming to reduce visits from individuals whose outcomes would not improve; and (2) a more feasible alternative that targets individuals based on a few key observable traits linked to improved health outcomes, as estimating the CATE may sometimes be impractical for all insurers. The first scenario may represent the optimal case regarding cost-effectiveness, aligning with the idea that targeting should be based on expected treatment effects rather than risk factors (e.g., Athey et al. forthcoming), while the second offers a practical targeting approach. We also implement a data-driven search for optimal policy rules based on individual estimates, referred to as a policy tree (Athey and Wager 2021).

Finally, while CATE analysis provides valuable targeting insights, it does not assess cost-effectiveness beyond compliers. For instance, it does not indicate whether health outcomes improve more or less for non-responders to nudges (i.e., never-takers). To address this, our third analysis applies the marginal treatment effect (MTE) framework (Björklund & Moffitt 1987; Heckman & Vytlacil 1999, 2005, 2007), linking treatment effect heterogeneity to both observed and unobserved variances in the propensity for compliance with the nudge. We estimate the impact of physician visits on health outcomes using an instrumental variable approach, where FBS above 126 serves as the binary instrument (Brinch et al. 2017). Based on these findings, we simulate a policy to improve cost-effectiveness by calculating the policy-relevant treatment effect (PRTE), which measures changes in average treatment effect on the treated (ATT) under an alternative policy. Specifically, we target patients with more severe conditions identified through CATE analysis and strengthen nudges only for them.

Our setting is Japan, where individuals are legally required to undergo annual mandatory health checkups. We obtain data from the Japan Health Insurance Association (JHIA), the largest health insurer in Japan, which covers 40 million members, approximately one-third of Japan's

total population. The large data size is crucial because we need many observations close to the threshold to estimate heterogeneous treatment effects to nudges accurately. The data consists of individual-level row data from 2016 to 2022, including itemized administrative claims data (utilization and spending), annual health checkup data (biomarker measures), and annual survey data collected at checkups (lifestyle habits such as smoking and drinking alcohol).

We have five main findings. First, estimating the average treatment effect at the diabetic-type threshold using RDD, we find that nudges induce individuals to visit a physician and increase healthcare spending. For example, nudges increase the probability of having at least one DM-related outpatient visit within a year of an index checkup by an average of 8.8 ppt (from 32.5 to 41.3 %), an increase of 27.2 % from the mean just below the cutoff. Similarly, nudges increase DM-related outpatient spending by an average of 35.3 USD (15.5 %). We also find evidence that health outcomes improve with nudges. For example, on average, the following year's FBS decreases by 0.71 mg/dl (0.57%), and LDL cholesterol (LDL-C) decreases by 1.59 mg/dl (1.20%) with nudges. These results indicate that nudges induce individuals to visit a physician and improve health outcomes *on average*. We note, however, that the improvement in health outcomes appears relatively small. One possible reason for the small health effect may be that those who comply with nudges may not benefit most from physician visits, possibly because they are healthy individuals who do not require medical treatments.

Second, our CATE results support this conjecture: those who "comply" most with nudges do not "benefit" most, indicating *reverse* selection on gains. Healthier individuals with better lifestyles are more likely to visit a physician in response to nudges. Specifically, compliers tend to have better health indicators (e.g., systolic blood pressure, triglycerides, BMI) and healthier habits, such as avoiding smoking, drinking and walking regularly. In contrast, those who experience greater health improvements tend to have poorer initial health and lifestyle habits. Higher-benefit individuals often have worse LDL/HDL cholesterol levels and unhealthy behaviors, such as late-night eating, weight gain, and reluctance to change habits—traits that differ from those who respond most to nudges.

Third, we find that targeting nudges to those who benefit more from physician visits yields similar improvements in health outcomes while reducing nudge-induced healthcare spending. For the first best-case scenario, the net benefit per person increased by 42% as the targeted nudges DM-related expenditure outpatient decreased, keeping health outcomes improvements unchanged. The net benefit per person for the second scenario increased by 13%. These results suggest an advantage of adopting an individual- or subgroup-specific targeting strategy over a "one-size-fits-all" approach.

Fourth, our MTE estimates provide evidence of a similar reverse selection on gains found for observed characteristics among compliers. The MTE curve decreases with the unobserved resistance to the nudge, suggesting that individuals most likely to comply with the nudge appear to benefit the least from seeing a physician. Thus, selection into treatment is negative based on both observables among the compliers and even beyond compliers that incorporate unobservable heterogeneity in this setting, suggesting that there may be room for further targeting beyond compliers.

Finally, we simulate the policy where we target nudges to very sick patients (SBP $\geq$ 160 or DBP $\geq$ 100 or LDL-C $\geq$ 180) who are identified as likely to benefit from the policy tree in CATE analysis. Specifically, we consider the scenario in which we increase only their propensity score by 0.2. However, the PRTE is only slightly higher than the average treatment effect and far lower than the average treatment effect for those untreated. This suggests that targeting based on observable traits beyond compliers does little to encourage hesitant patients to seek care, despite the potential benefits. A more effective alternative may be sending health information directly to primary care physicians instead of patients.

Related literature and our contributions.— Our paper contributes to several strands of the literature. First, this paper contributes to the literature on targeting the treatment assignment based on observables (Kitagawa and Tetenov 2018; Athey and Wager 2021). In the case of health care, Bonander and Svensson (2021), for example, use a simulated dataset and provide the framework to learn targeting policy from data using CATE analysis. We contribute to the literature by documenting the discrepancy between selection and treatment heterogeneity in real-world data and exploring welfare-maximizing targeting policies for healthcare nudges.

Second, our study relates to the recent literature investigating selection in take-ups (of health screening). The most closely related to ours are Einav et al. (2020) and Kowalski (2023), both of which study selection in mammography screening. Mammography is often recommended for women over a certain age threshold (e.g., 40 years), but the decision to undergo such a test is left to individual choice. They demonstrate selection into mammography screening and highlight the importance of considering heterogeneous responses when devising a screening program (Einav et al. 2020). Our results also suggest that reliance on self-selection into treatment may not be cost-effective, which echoes Jones (2019), who also found that program participants in the workplace wellness program had lower medical expenditures and healthier behaviors than nonparticipants. A key distinction between their studies and ours is that we employ a machine-learning approach to examine heterogeneous responses and systematically consider targeting policies to maximize welfare.

Third, we contribute to a growing body of literature investigating the causal impact of health signals, like report cards, on healthcare utilization and health outcomes (e.g., Kim et al. 2019; Fukuma et al. 2020; Iizuka et al. 2021). Like ours, these studies employ RDD to identify the causal effects of health signals exploiting arbitrary biomarker cutoffs. Our contribution is that, unlike previous studies that estimate average treatment effects at the threshold, we study heterogeneous treatment effects of a healthcare nudge by estimating CATE, which allows us to study potential mismatch between those who comply with nudges and those who benefit from nudges. Moreover, we advance the literature by examining how and whether targeting nudges can improve welfare.

Fourth, we contribute to the growing literature on applying machine learning to healthcare, particularly in treatment effect heterogeneity. For example, Goto et al. (2024) estimate CATE using data from the Oregon Health Insurance Experiment and find that health insurance benefits are heterogeneous, especially aiding those with low pre-insurance healthcare spending. Our study highlights that compliance with healthcare nudges is also heterogeneous, with healthier individuals and those with better habits more likely to comply. Moreover, we examine both who benefits most from treatment and who adheres to it, emphasizing that better alignment could improve efficiency. A recent working paper by Chen and Tsai (2024) estimates heterogeneous treatment effects in hyperlipidemia using Taiwanese health checkup data. Key differences in our work include 1) analyzing the mismatch between selection and treatment heterogeneity, 2) exploring targeting strategies, and 3) extending beyond local average treatment effects.

The remainder of the paper is organized as follows. In Section 2, we briefly discuss the institutional context of our study (e.g. mandatory health checkups in Japan, the key threshold value for DM diagnosis, and the contents of health signals). Section 3 describes the data, and Section 4 presents our empirical strategy. Section 5 reports our graphical and econometric results and additional analyses. Section 7 concludes.

2. Background

2.1. Annual mandatory checkup in Japan

Workplace annual health checkups are mandatory for both employers and employees in Japan. Under the Industrial Safety and Health Act of 1972, employers must provide annual health examinations at their expense, supervised by the Labor Standards Inspection Office. Failure to comply may result in fines of up to 500,000 JPY (about US$5,000).[1] While employees

[1] The exchange rate of JPY 100/$ is used throughout the paper.

are also required to participate in these checkups yearly, there are no legal penalties for individuals who do not comply. However, employers can take disciplinary action for non-compliance.

The health checkup aims to facilitate the early detection of potential health conditions and provide patients with health reports and guidance. By communicating the results, the checkup encourages individuals to seek medical care when necessary or adopt healthier behaviors. Ideally, patients will respond by initiating appropriate treatment and making lifestyle changes, both of which can help prevent more severe and costly complications in the future.

The checkup typically includes physical examinations (such as height and weight), lab tests to measure various biomarkers (including blood sugar, blood pressure, and cholesterol), and a survey of health behaviors (such as drinking and smoking).

2.2. Report card and follow-up nudge

After a checkup, individuals receive a report within one to two months. This report outlines the results of their health assessment from lab tests, along with the normal reference ranges of each biomarker. Figure 1-A shows an example. If a value exceeds a clinical threshold, the report includes a warning (e.g., 'H' for high or a grade from 'A' to 'E,' with 'E' indicating worse conditions) and advises consulting a physician (Iizuka et al. 2021).

Japan's largest insurer, the Japan Health Insurance Association ("Kyokai Kenpo"), covering about 40 million people, sends follow-up documents urging selected individuals to see a doctor. Selection is based on two criteria: not currently receiving diabetes treatment and meeting the Japan Diabetes Society (JDS) biomarker threshold. Specifically, for the former criteria, those taking hypertension or diabetes medication (self-reported) at the checkup are excluded, as they likely already see doctors. For the latter criteria, individuals with FBS being greater or equal to 126 mg/dl are classified as diabetic and targeted for the nudge.

Figure 1-B shows an example of such a reminder, which encourages individuals to see a doctor, along with some health warnings. While the wording varies slightly over the years, the core message remains the same. This 2018 example states, "If you do nothing, you risk developing arteriosclerosis, increase heart and cerebrovascular disease risk, and lose a healthy life."

Figure 2 outlines the checkup process. After the initial checkup at year t, all individuals receive their reports within one to two months. Those not in DM treatment and with FBS ≥126 mg/dl ("diabetic type") receive a reminder with a health warning in month 6. We then observe

their biomarkers at the year t+1 checkup for returning participants and healthcare utilization between checkups. Since health outcomes are observed only once at the annual checkup, we cannot isolate the effects of report cards and nudges, so we refer to them collectively as "nudges" unless stated otherwise.

Crucially, the decision to pursue medical care after noticing health signals rests solely with the individual since employers and health insurers cannot require doctor visits. Additionally, there is no gatekeeping by general practitioners (GPs), and hence individuals are not tied to a specific primary care physician. Consequently, primary care physicians do not monitor biomarkers or urge patients to come to their practice. Indeed, the adoption of electronic medical record systems has been slow in Japan, and physicians have only recently gained access to the results of annual health checkups. Therefore, it is up to the patient to decide whether to consult a healthcare provider upon receiving health warning signals. Treatment decisions are left to the physician's judgment, although the JDS offers some guidelines. Universal health insurance covers visit costs on a fee-for-service basis, with patients responsible for 30% of the expenses out-of-pocket.[2]

2.3. Treatment of Diabetes Mellitus

Diabetes mellitus (DM) is a costly, incurable chronic disease with rising incidence and prevalence. In Japan, it was the third largest disease category in 2019, with prevalence rates of 19.7% in men and 10.8% in women, increasing with age. Initially asymptomatic, DM eventually leads to serious complications affecting the eyes, heart, kidneys, nerves, and feet. However, it is preventable through early intervention targeting risk factors like smoking, poor diet, sedentary lifestyle, and obesity. Importantly, DM and prediabetes can be identified by elevated FBS levels, which is considered the gold standard for diagnosing diabetes, making clinical thresholds like 126 mg/dl common.

Treatment for DM or near-DM patients typically starts with lifestyle modifications, including healthy eating, exercise, weight management, and smoking cessation. If glycemic control is not achieved through these measures, first-line medications such as metformin and oral hypoglycemic agents are recommended (Haneda et al. 2018, p.664). Research suggests this comprehensive approach reduces the risk of diabetes-related complications (UK Prospective

[2] Japan has a highly regulated universal health insurance system overseen by the government. Every citizen must enroll in either an employment-based or a residential-based insurance program (see, for example, Ikegami and Campbell 1995; Kondo and Shigeoka 2013). All individuals receive the same benefits, irrespective of the insurance type, covering outpatient services, inpatient services, dental care, and prescription medications.

Diabetes Study Group 1998; Miller and Dunstan 2004).

3. Data

3.1. Data description

Our data is provided by the Japan Health Insurance Association (JHIA) through a joint research agreement with the University of Tokyo. JHIA typically covers employees of small and medium-sized enterprises (SMEs) and their families, as well as part-time and temporary workers. It does not include the elderly over 75 covered by a separate national health insurance scheme.

The dataset covers about 40 million people—roughly one-third of Japan's 120 million population. Our data spans April 2016 to March 2022, with approximately 48 million checkup-year records. It comprises three components: medical claims, health examination measurements, and health survey responses, all linked by patient ID. Medical claims data, collected monthly, allow for the tracking of individuals' healthcare utilization. Health outcomes are derived from health checkups, which are typically conducted annually, along with information on the year and month of the checkup. The health survey, administered during checkups, gathers information on self-reported health behaviors and disease history. Additionally, demographic data from the enrollment registry include gender, birth date, insurance premium category (a wage proxy), and, if applicable, death dates.

3.2. Outcome variables

We examine whether health signals affect three sets of outcome variables: (i) healthcare utilization, (ii) health behaviors, and (iii) health outcomes. Appendix D provides definitions for all the variables utilized in this study.

Healthcare utilization—. We analyze four variables: a) a binary indicator for at least one DM visit (extensive margin), b) DM visit frequency, c) DM-related outpatient medical spending (intensive margin), and d) DM-related drug spending, which separate costs incurred for DM-related medications.[3] These variables are derived by aggregating 12 months of claims data post-checkup. A visit is classified as DM-related if diabetes (E10-E14 in ICD-10) is listed as a condition.[4]

Health behaviors—. We create four dummy variables for (a) drinking daily, (b) smoking,

[3] We truncate spending records with values above 99.9 percentiles of spending-related outcomes to avoid the influence of outliers.

[4] In Japan, physicians must document all medical conditions (including suspected conditions) linked to a particular visit, and this ICD10 code is available in our dataset.

(c) not walking regularly, and (d) skipping breakfast regularly. These variables are derived from health surveys conducted during health checkups.

Health outcomes—. We use four biomarkers or physical health measures most closely related to DM: (a) FBS—the obvious targeted outcome, (b) low-density lipoprotein cholesterol (LDL-C), (c) systolic blood pressure (SBP), and (d) body mass index (BMI). These physical measures are taken from the annual checkup in the next year.

3.3. Sample selection

As mentioned, individuals taking hypertension or diabetes medications at the index checkup are excluded from the nudge. Figure 3 plots the probability of taking either medication. The probability of taking medications is monotonically increasing in the value of FBS as expected, but importantly, it is continuous at the FBS threshold of 126, as expected. The RD estimate from a local-linear regression using a triangular kernel, where a dummy for already taking medication is an outcome, is -0.00355 (SE= 0.00405) with a p-value of 0.407.

Since we are mostly interested in the (combined) effect of health signals, we focus on those not taking either of them (i.e., the sample *above* the line in Figure 3) as our main sample throughout the paper. Since the probability of taking either at 126 mg/dl is 54.1%, we roughly focus on half of the sample near 126 mg/dl. We use those already taking the medications (i.e., the sample *below* the line in Figure 3) as a "placebo sample" occasionally.

Also, we focus on those not being treated for DM at the time of the health checkup. We include a checkup in our analysis if it meets the following conditions: i) the individual was not diagnosed with DM before the checkup[5] ; ii) we have data for the individual at least 1 month before the checkup; and iii) we have data for the individual at least 12 months after the checkup.

Table A1 provides summary statistics of the main sample for the variables used in the analysis. Figure 4 plots the distribution of FBS values of the main sample. FBS values are available only as integers, so we use a bin size of one for the figures throughout the paper. Thanks to the large sample size, we have more than 1.31 million observations within the 10 mg/dl interval around FBS of 126 mg/dl threshold, from 116 to 135 mg/dl. This is roughly 30 times larger than the 44,000 observations in the same length of interval reported by Iizuka et al. (2021).

[5] Since those already taking diabetes medications are already excluded, this sample criteria does little to the further sample selection (1.48%).

4. Empirical strategy

4.1. Regression discontinuity design (RDD)

We exploit that the exact checkup threshold of FBS = 126 mg/dl is not scientifically optimized and that individuals with values just below and above the threshold are similar in their health status in regression discontinuity design (RDD). However, only those whose values cross the threshold will receive a health signal, and this discontinuity identifies the effect of health signals on the outcomes of our interest. The threshold is arbitrary as it reflects historical decisions on exact values about which many experts disagree. This arbitrariness is particularly evident in our context because the thresholds established in Japan were mainly selected to align with Western standards, even though optimal values may vary for Asian populations.

The RDD addresses the potential endogeneity between health signals and the outcomes of interest. For instance, if one simply regresses the amount of medical utilization on checkup values alone, the estimated effect of checkup values is likely biased due to omitted variables, such as an individual's unobserved health status, which may correlate with the checkup value. The RD approach mitigates endogeneity by comparing individuals who are similar in every respect except that their checkup values fall just above and below an arbitrary threshold.

We estimate the following local linear regression:

$$Y_{it+1}\ (or\ \Delta Y_{t+1}) = \alpha_0 + \alpha_1 1(FBS_{it} \geq 126) + \alpha_2(FBS_{it} - 126) + \alpha_3 1(FBS_{it} \geq 126)(FBS_{it} - 126) + \alpha_4 X_{it} + \delta_a + A_y + \varepsilon_{it}, \quad [1]$$

where Y_{it+1} represents healthcare utilization in $t+1$ and $\Delta Y_{t+1}(= Y_{it+1} - Y_{it})$ represents the change in either health behaviors or health otucomes between next year and this year for individual i.[6] $1(FBS_{it} \geq 126)$ takes the value 1 if individual i's FBS in year t is greater than or equal to 126 mg/dl, and zero otherwise. The next two terms are the linear line of FBS with respect to the threshold, allowing their effects to differ below and above the threshold. X_{it} is a vector of individual covariates, namely age, age squared, and gender. As discussed later in this section, the covariates are balanced across the threshold, our results quantitatively change little with or without the covariates. A_y is index checkup year fixed effect (from 2016 to 2022). ε_{it} is the error term. α_1 is our coefficient of interest.

We estimate local-linear regressions with a triangular kernel and use the optimal bandwidth that minimizes MSE (Calonico et al. 2019). The standard errors are estimated with the robust

[6] The change as an outcome essentially means including individual fixed effects. As shown before, since Y_{it} is smooth around the threshold, we obtain similar results if we use Y_{it+1} as an outcome instead. We prefer the difference as an outcome since we obtain slight efficiency by roughly 10-20%.

bias-corrected method (Calonico et al. 2014).

Density test—. A common issue in RDD is the manipulation of the running variable. In this context, FBS—our running variable—fluctuates over time, making precise control difficult. Moreover, doctors and checkup staff typically lack immediate access to blood sugar data, making it improbable for them to alter the measurements reliably. Our study focuses on individuals without a prior DM diagnosis, who are often unaware of their blood sugar levels and have no incentive to change them. Figure 4 shows a smooth distribution of FBS values with no visible discontinuity at the threshold. A local polynomial density test (Cattaneo et al. 2018) finds no evidence of manipulation or sorting at the threshold (p-value = 0.904).

Balanced checks—. Another check to validate the RD assumption is ensuring covariates remain continuous at the threshold. To test this, Figure A1 plots average values of selected variables—male, age, wage, and LDL-C (as a health outcome)—at the index time in year t for each FBS value as placebo outcomes. None exhibit noticeable discontinuities. Table A2 further confirms balance across all 10 examined variables, including dependent variables and covariates.

Attrition—. Our data allows analysis of medical utilization and health outcomes only for individuals who remain with the same employer and take their annual checkup the following year. To address potential selection bias, Figure A2 plots the probability of attending the next year's checkup. Attendance exceeds 90%, with no significant discontinuity at the thresholds, suggesting attrition is unrelated to receiving health signals. The RD estimate from a local-linear regression with a triangular kernel, using a dummy for year $t+1$ attendance as the outcome, is -0.00355 (SE=0.00405, p=0.407).

4.2. Causal forest (CF)

Although the RDD provides important insights into how nudges affect outcomes of interest, it only tells the *average* effects of passing the DM threshold, and important heterogeneity in treatment effects may be masked. In particular, we are interested in whether those who "complies" the most with health signals and visit doctors are the ones who "benefit" the most. If not, there may be a scope for the targeting.

To explore heterogeneous responses, we estimate the conditional average treatment effect (CATE) using machine learning. We estimate the generalized random forest (GRF) (Wager and Athey 2018 Athey et al. 2019), an extension of the random forest. The GRF method allows researchers to estimate heterogeneous treatment effects by incorporating identification restrictions corresponding to, for example, an RDD and an instrumental variable (IV) estimation. Also, instead

of arbitrarily splitting the sample by covariates as conventional studies perform in their heterogeneity analysis, the causal forest algorithm splits the sample repeatedly across multiple dimensions to maximize treatment effect heterogeneity between subsamples (leaves).

The heterogeneous treatment effect τ_i is given by the following (we omit t to for simplicity):

$$\left(\hat{\tau}, \hat{\theta}\right) \in \arg\min_{\tau,\theta} \quad \left\| \sum_i \alpha_i(x)\, \Psi_i(Y_i.) \right\|_2 \quad (1)$$

where $\alpha_i(x)$ is a "similarity weight," $\Psi_i(Y_i,)$ is a score function, and $\hat{\theta}$ is a nuisance parameter. $\alpha_i(x)$ takes a larger weight if individual i belongs more frequently to the same leaf as x while the algorithm grows trees. x includes information on past health outcomes, wage, gender, and age at year t.

Suppose the algorithm builds b trees where b = 1, 2, …, B. Then $\alpha_i(x)$ is defined as follows:

$$\alpha_{bi}(x) = \frac{1[X_i \in L_b(x)]}{|L_b(x)|}, \alpha_i(x) = \frac{1}{B} \sum_{b=1}^{B} \alpha_{bi}(x)$$

where $L_b(x)$ is the set of individuals who belong to the same leaf as x in tree b.

As noted, the GRF model can incorporate the structure of RDD when estimating the heterogeneous treatment effects. In particular, Athey et al. (2019) propose to estimate the following local linear model, called the linear model (LM) forest:[7]

$$Y_i = c(x) + \tau(x)Z + \beta(x)R$$

where Y_i is the outcome of interest (e.g., ΔY_{t+1}), Z is the treatment ($1(FBS_{it} \geq 126)$), and R is the running variable.

Then, predictions from multiple splits (trees) are averaged to obtain the CATE.[8] Once we obtain the CATE, we refine the raw predictions using the doubly robust score (Athey and Wager 2021) to mitigate regularization bias and overfitting (Chernozhukov et al. 2018).

We restrict the sample within 5 mg/dl of 126 thresholds to adjust to the RDD setting. The minimum node size of 5,000 is used. We estimate CATEs of receiving health signals separately for the healthcare utilization ($\hat{\tau}^{\Delta C}$) and health outcomes ($\hat{\tau}^{\Delta B}$). ΔC is the incremental cost from receiving nudges, and ΔB is the incremental benefit from receiving nudges, written by $\gamma \Delta H$, where γ is the value of statistical life (or willingness to pay) and ΔH is the health outcome improvement. We set γ to be \$50,000 per year, following the literature, with an annual discount rate of 3%.

[7] The LM Forest is part of the GRF package.

[8] Causal forest has been widely used in economics recently as a means to examine the heterogenous treatment effects. In healthcare, see, for example, Chen and Tsai (2024).

4.3. Targeting through policy learning

We consider several targeting scenarios. We start with two scenarios: (1) a full-information case, where the insurer nudges only those expected to see health improvements, aiming to reduce visits from healthier individuals whose outcomes change little; and (2) a more feasible alternative that excludes individuals lacking *any* characteristics linked to improved health outcomes, as estimating the CATE may sometimes be impractical. The first scenario may represent the optimal case in terms of cost-effectiveness, which aligns with the notion that targeting should be based on net benefits rather than risk factors, while the second provides a practical targeting approach.

Lastly, instead of identifying the subgroups to be targeted manually by exploring the results from causal forests, we implement a data-driven search for optimal policy rules based on individual estimates in order to maximize the net monetary benefit in the population, called policy trees (Athey and Wager 2021).

$$\hat{\pi} = argmax\left\{\frac{1}{n}\sum_{i=1}^{n}(2\pi x_i - 1)(\hat{\tau}^{\Delta B} - \hat{\tau}^{\Delta C})|\pi \in \Pi\right\}$$

The policy tree package for R (Zhou et al. 2020) finds the optimal tree of a pre-specified depth, which determines the number of subgroups allowed. Specifically, the policy trees provide estimates of the optimal cut points along the individual observable characteristics variables, given the individual-level estimates of net benefit from the causal forest. Then, the average welfare gain per individual member under the suggested targeted policy, as opposed to giving the nudge to all individuals above the threshold (e.g., treatment as usual), can be estimated.

4.4. Marginal treatment effect (MTE)

While the analysis of CATE offers valuable information for targeting based on observables, one limitation of CATE is that it does not provide insights beyond "compliers." For instance, CATE does not indicate whether the treatment effect is larger or smaller for those who do not respond to nudges (i.e., never-takers). To address this, we implement the marginal treatment effect (MTE) framework (Björklund and Moffitt 1987; Heckman and Vytlacil 1999, 2005, 2007). MTE enables us to connect the heterogeneity in the treatment effect to both observed and unobserved heterogeneity in the propensity for compliance with the nudge.

In our setting, a binary indicator that takes on a value of one if FBS is above 126 serves as the instrumental variable. Since our instrument is binary, we adopt the method proposed by

Brinch et al. (2017) to estimate MTE for a wide range of propensity to comply with the nudge. [9] A separability assumption is required for a binary instrument to estimate non-linear MTE. Under this framework, D is any DM-related visit within a year, Z is $1(FBS_{it} \geq 126)$ and Y is $\Delta\, FBS$.

Then, we conduct the policy simulation to evaluate whether we can improve the current status quo. To achieve this, we calculate a policy-relevant treatment effect (PRTE), which represents a weighted difference between the average treatment effect on the treated (ATT) under the alternative policy and the ATT under the baseline policy (Heckman and Vytlacil 2001, 2005; Carneiro et al. 2011; Cornelissen et al. 2016, 2018).

5. Results

5.1. Average effects from RDD

This section presents the baseline results of estimating the RD specification. We examine whether health signals affect three sets of outcome variables: (i) healthcare utilization, (ii) health behaviors, and (iii) health outcomes. Due to space constraints, we provide graphical evidence for two out of four outcome variables in each category here, while the supplementary graphs and results are available in Appendix Section A.

Healthcare utilization—. We begin with graphical evidence of medical care utilization. Figures 5-A and 5-B plot the probability of at least one DM-related doctor visit and outpatient spending within a year of the index checkup. Both measures show a clear discontinuous increase at the threshold. Table 1 summarizes RD estimates from equation [1]. Table 1-A shows an increase in physician visit probability of 8.8 percentage points (ppt), from 32.5% to 41.3%, representing a 27.2% increase (= 8.8/32.5). Similarly, DM-related outpatient spending increased by $35.3 at the cutoff, a 15.5% increase from the mean of $228.5. These findings suggest that health warnings significantly boost utilization.

So far, we have aggregated 12 months of post-checkup claims data, as health outcomes are observed annually. However, *monthly* data allows us to assess whether the nudge at 6 months has any additional effect on the utilization on top of the report card. Figure 6-A focuses on the main sample receiving the nudge six months after the index checkup. The outcome is a cumulative DM visit dummy, indicating at least one DM visit by month m. The left figure plots this for m = -1, 0, 1, 5, 6, 7, and 8, where m = -1 is before the checkup, and m = 0 is the checkup month. The right figure presents RD estimates from equation [1] for each month from -1 to 12.

The estimates for m = -1 and m = 0 are near zero, as individuals have yet to receive a

[9] Other applications of this binary IV to estimate MTE include Ito et al. (2019).

report card. At month 1, the probability of a DM visit rises by 2.5 ppt, increasing steadily before leveling off around month 4. A sudden 2.1 ppt jump occurs at $m = 6$ when the reminder with warning is sent, reaching 8.8 ppt in one year (Table 1-A).

Since other events may coincide with month 6, Figure 6-B examines a placebo sample that does not receive the nudge. A similar, though smaller, increase appears at $m = 1$, indicating that these individuals also respond to the report card and visit a physician. However, the impact is smaller, probably because these individuals regularly see a doctor, as shown by the higher mean, and the report card provides little additional information. Interestingly, unlike the main sample, they show no increase at $m = 6$, suggesting the observed jump in Figure 6-A reflects the reminder's additional effect.

Health behaviors—. Figure 5-B shows the mean change in daily drinking and smoking between years $t+1$ and t, with variables redefined so that higher values represent healthier behaviors. No noticeable decrease occurs at the cutoff for either outcome. Table 1-B indicates that only the reduction in "not walking regularly" is statistically significant at the 5% level. However, the magnitude is small, with a slight decline of 1.0 ppt from a 62.0% mean just below the cutoff. Therefore, we conclude that behavioral changes in response to health signals are minimal.

Health outcomes—. Figure 5-C shows the mean change in FBS and LDL-C between years $t+1$ and t, with higher values indicating *worse* health outcomes. Exceeding the threshold is associated with a discontinuous improvement in all examined biomarkers. Our reduced-form estimates in Table 1-C show that exceeding the threshold leads to an improvement of 0.92 mg/dl in FBS (0.57%), 1.59 mg/dl in LDL-C (1.20%), 0.537 mmHg in SBP (0.40%), and 0.097 in BMI (0.30%). Though these effects seem small at first glance, they are much larger for compliers, who visited a physician because they received the health signal, given the first-stage magnitude of about nine percentage points. Using the threshold as an excluded instrument[10], Table A3 finds the local average treatment effect (LATE) on FBS and LCL-C, for example, is 10.2 mg/dl (8.2% reduction), and 18.4 mg/dl (13.9% reduction).

However, these health improvements remain modest compared to the 27.2% rise in DM-related visits. One possible reason for the small health effect is that those who comply with nudges may be healthier individuals who gain little from medical treatment. To explore this, we report the conditional average treatment effect (CATE) to assess heterogeneous impacts and

[10] Regarding the exclusion restriction of the instrument, the channels other than medical treatment, such as behavioral changes, seem limited, as shown in Table 2-B.

identify key individual characteristics associated with such heterogeneity in the next subsection.

5.2. Heterogeneous treatment effects using CATE

This section presents the results of utilizing causal forest and the corresponding CATEs to explore the heterogeneous treatment effects. The supplementary results can be found in Appendix Section B.

CATE on D (physician visit)—. The top row of Figure 7-A shows the distribution of estimated CATEs, $\tau(X)$, of receiving health signals on making at least one DM-related visit within a year of the index checkup. The mode is around eight percentage points (ppt), similar to the mean treatment effect (8.7 ppt) from the RDD. The figure highlights that nudge effects on utilization are mostly positive and heterogeneous, with a 95% confidence interval ranging between 1.5 and 15.3 ppt. The middle row of Figure 7-A categorizes CATEs into four groups by applying Chernozhukov et al. (2019) that yield consistent estimates of average treatment effects for subsets of the sample (group-level treatment effects or GATE). We find significant heterogeneity, as indicated by the estimates' confidence intervals do not overlap. Figure B1 also presents the rank-weighted average treatment effects (RATE) as an additional method to evaluate and confirm the existence of heterogeneity (Yadlowsky et al. 2024).

Next, we examine which individuals are more likely to comply with nudges. The bottom row of Figure 7-A, the heat map, compares characteristics of high- (Q4) and low-compliance (Q1) groups. We note that the darker color represents better health outcomes and better lifestyle habits, except for age, gender, and wages. The high-compliance group (Q4) tends to be healthier, with lower LDL-C, SBP, and BMI, and exhibits better lifestyle habits, such as not smoking, not drinking daily, and not skipping breakfast. These findings suggest that healthier individuals with better habits are more likely to respond to nudges.[11]

CATE on Y (health outcomes)—. We next examine the heterogeneous effects of nudges on health outcomes in Figure 7-B, using the same approach but focusing on health outcomes one year after the index checkup. The top row shows the distribution of treatment effects on the (negative) change in FBS, redefined so positive values indicate improvement. The mode is around 1.02, similar to the RDD estimate (approximately 0.93), with a 95% confidence interval ranging between 0.05 and 9.03, highlighting heterogeneity. The middle row groups individuals

[11] We also report CATEs on other health utilization measures, specifically the frequency of DM-related outpatient visits and DM-related outpatient spending in Figure B2 for brevity. We find similar patterns, indicating that the effects of nudges on health utilization are heterogeneous, with larger effects observed among healthier individuals with better habits.

by treatment effect size and tests whether their group average effects differ from zero. The impact of nudges on (-)ΔFBS is heterogeneous, with a significant difference in the effect between groups 1 and 4. The bottom row compares the characteristics of individuals with high versus low CATEs for health outcomes. As before, individuals are grouped by the benefit of nudges on ΔFBS. The high-benefit group generally has worse health and poorer lifestyle habits.

Relationship between CATE on D and Y—. Comparing the heat maps reported at the bottom of Figures 7-A and 7-B reveals that high-compliance individuals (Q4 in Figure 7-A) tend to be healthier, while those benefiting most from doctor visits (Q4 in Figure 7-B) tend to be sicker. This suggests that those who "comply" most with health signals tend to "benefit" the least, indicating reverse selection on gains based on observables. To explore this further, we combine the two CATE estimates. Figure 8 presents a binned scatter plot showing the relationship between CATE for compliance (x-axis) and CATE for health improvements (y-axis) at the individual level. The negative correlation suggests that individuals more likely to comply with nudges experience smaller health improvements.[12]

5.3. Targeting

Our results, thus far, indicate important heterogeneity in who responds to nudges and who benefits from nudges. More importantly, there is a discrepancy between the two responses; those who comply with nudges are not those who benefit most from nudge-induced physician visits. Nudges in healthcare are generally costly as they increase physician visits and healthcare expenditures. They could even harm individuals in some cases, such as cancer screening if nudge-induced visits result in unnecessary tests and treatments. Therefore, there may be room for efficiency improvement by adequately targeting nudges to individuals whose benefits from a physician visit are higher.

Before we consider targeting, we must quantify the monetary values of both the benefits and costs. Thus, for the benefits, we examine the reduction in mortality rates over the six years since the index checkup, as this is the maximum duration we can observe in the data. Unlike the improvements in biomarkers like FBS, changes in mortality can translate into monetary benefits through the use of the value of statistical life (VSL). Figure B4 displays the CATEs on 6-year cumulative mortality. We observe reduced cumulative mortality among individuals with the

[12] We also report CATEs on other health outcomes, including LDL-C and SBP. For brevity, we present these results in Figure B3. Most results exhibit a pattern similar to the FBS findings reported earlier; the effects of nudges on health outcomes are heterogeneous and influence only a small fraction of individuals who visited physicians. Importantly, akin to the FBS case, we also observe discrepancies between the types of individuals who are more likely to comply with nudges and those who are more likely to benefit from them.

highest treatment effects (Q4). Figure A3 and the corresponding Table A4 confirm that the RD estimates for 6-year cumulative mortality among individuals in Q4 are indeed positive (i.e., improvement) and statistically significant (p-value <0.05).[13]

We then convert this into monetary benefits in Figure 9-A using the VSL at $50,000 per year, and an annual discount rate of 3%. On the cost side, we consider cumulative DM-related outpatient spending over the same six-year period with the same discount rates. Figure 9-B confirms substantial treatment heterogeneity in the CATEs on cost as well. Figure 9-C plots benefits from Figure 9-A on the x-axis and costs from Figure 9-B on the y-axis at each individual level, with the slope of the solid line representing a 45-degree angle. Thus, individuals below the solid line indicate that the *net* benefit of providing nudges is positive. Finally, Figure 9-D displays the distribution of such net benefit, defined as the difference between benefits and costs. The net benefit is positive for more than 70% of the sample.

With this individual-level net benefit information, we consider several targeting scenarios. The results are summarized in Table 2. Here, we consider the cost of implementing a target policy per se other than changes in healthcare utilization to be zero, as the cost of sending a nudge (e.g., the cost of the envelope and printing) is negligible.

We first consider the following targeting policy: nudge the high-benefit group only, defined as those whose predicted benefit from nudges outweighs the cost of nudges. This concept aligns with the idea that medical treatments should be assigned according to predicted potential benefits rather than risk factors to improve efficiency.

Another possible targeting rule is to select individuals with positive predicted benefits based on observable characteristics and to send health signals solely to those individuals. We find that targeting nudges to those who benefit more from physician visits yields similar improvements in health outcomes while reducing nudge-induced healthcare spending. Finally, we implement a data-driven search for optimal policy rules based on individual estimates called a policy tree. Figure 10 shows depth-2 policies based on the estimated net benefit.

In the first best-case scenario ("Non-parametric" in Table 2), the net benefit per person rose by 42% compared to the status quo of providing nudges to all individuals with an FBS of 126 or higher. This strategy effectively reduces DM-related outpatient expenses while keeping improvement in health outcomes. For the second scenario ("Policy tree" in Table 2), the net benefit per person increased by 13%. These findings indicate a preference for adopting an

[13] Figure B5 also shows a negative relationship between CATE for compliance and CATE for mortality improvement (y-axis), suggesting that individuals who are more likely to comply with nudges experience smaller mortality improvements.

individual- or subgroup-specific targeting strategy over a "one-size-fits-all" approach.

5.4. Supplementary analysis on MTE

So far, our discussion is limited to compliers' compliance/selection heterogeneity and treatment heterogeneity based on observable characteristics. We go beyond LATE and incorporate unobserved heterogeneity by estimating the marginal treatment effect (MTE). The supplementary results on MTE can be found in Appendix Section C. Figure C1 shows that under the separability assumption, there is a substantial overlap of propensity scores between treated and untreated. We trim the observations in the support outside of 0.1-99.9 percentiles.

Figure 11-A plots the MTE and 95% confidence intervals as a function of unobserved resistance to treatment (u_d) when the outcome is the change in FBS. The dotted line corresponds to the average treatment effect (ATE). The figure shows that MTE on ΔFBS decreases with unobserved resistance to treatment. This result is consistent with the result of CATE for compliers based on observables. Thus, selection into treatment is negative based on both observables and unobservables in this setting, suggesting that there may be room for targeting. Figure 11-B decomposes the MTE in Figure 11-A to the treated, $E(Y_1|\, U_d = u_d)$ and the untreated, $E(Y_0|\, U_d = u_d)$. The figure shows that the negative slope of MTE predominately stems from the negative trend of those treated.

Policy simulation—. Finally, we conduct the policy simulation to evaluate whether we can improve the current status quo. To achieve this, we calculate a policy-relevant treatment effect (PRTE), which represents a weighted difference between the average treatment effect on the treated (ATT) under the alternative policy and the ATT under the baseline policy.

We consider two counterfactuals. Counterfactual 1 is the scenario where we enhance the nudge to all individuals by adding 0.2 of the propensity scores (PS).[14] Counterfactual 2 targets the nudge to only the very sick. Specifically, counterfactual 2 is the scenario where we add 0.2 of the propensity scores (PS) only to the very sick (SBP ≥ 160 mmHg or DBP ≥ 100 mmHg or LDL ≥ 180). We are agnostic about how, but we assume that we can provide stronger nudges to prompt them to visit a physician for DM treatment.

Figure 12 plots the MTE and PRTE for the following two scenarios, and Table 3 provides PRTE from both counterfactuals. Comparing PRTE from two scenarios suggests that targeting in counterfactual 2 slightly does better than the case of counterfactual 1. However, PRTE of counterfactual 2 is significantly lower than the average treatment effect on those untreated

[14] We set it to 1 if PS becomes larger than one.

(ATOU). Thus, targeting based on observable characteristics does not effectively encourage patients who are hesitant to see a doctor but could greatly benefit from doing so, to make that visit.

Our findings indicate that when individuals self-select into treatment, targeting based on observable traits may have limited effects on improving health outcomes beyond compliers, at least in this context. An alternative approach, such as sending health information to primary care physicians directly, instead of patients, may be more effective.

6. Conclusion

The decision to seek healthcare services is typically at the discretion of the individual. A crucial yet underexplored issue in this context is that individuals who self-select into treatment may not necessarily be those who would derive the greatest benefit from it. If a disparity exists between these groups, efficiency in healthcare provision could be improved by, for example, prioritizing treatment for those most likely to benefit.

This study examines the marginal returns to healthcare within the context of diabetes management, utilizing data from annual health checkups in Japan. The analysis leverages that individuals with fasting blood subar levels just below and just above a clinical threshold are comparable in health status. However, only those exceeding the threshold receive a formal warning regarding the risk of diabetes mellitus and are advised to seek medical consultation.

This study shows that providing diabetic patients with health signals leads to increased doctor visits and improved health outcomes. However, individuals who respond most to health warnings are observably healthier, while those who benefit the most are sicker, indicating negative selection on gains.

We find that targeting nudges to those who would benefit more from physician visits results in similar improvements in health outcomes while reducing nudge-induced healthcare spending. However, targeting by observables beyond compliers does not effectively encourage sick patients, who are reluctant to visit the doctor but could benefit greatly from doing so, to seek medical care. Therefore, reliance on self-selection in treatment may be limited beyond compliers. A different approach might be more effective, such as sending health information to primary care providers instead of patients.

References

Athey, Susan, Niall Keleher, and Jann Spiess. 2025. "Machine learning who to nudge: Causal vs predictive targeting in a field experiment on student financial aid renewal." *Journal of Econometrics*, 105945.

Athey, Susan, and Stefan Wager. 2021. "Policy learning with observational data." *Econometrica*, 89(1): 133–161.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized random forests." *Annals of Statistics*, 47(2): 1148–1178.

Baicker, Katherine, Sendhil Mullainathan, Joshua Schwartzstein. 2015. "Behavioral hazard in health insurance." *The Quarterly Journal of Economics*, 130(4): 1623–1667.

Björklund, Anders, and Robert Moffitt. 1987. "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models." *Review of Economics and Statistics*, 69(1): 42–49.

Bonander, Carl, and Mikael Svensson. 2021. "Using causal forests to assess heterogeneity in cost-effectiveness analysis." *Health Economics*, 30(8): 1818–1832.

Brinch, Christian N., Magne Mogstad, and Matthew Wiswall. 2017. "Beyond LATE with a Discrete Instrument." *Journal of Political Economy*, 125(4): 985–1039.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust non-parametric confidence intervals for regression-discontinuity designs." *Econometrica*, 82(6): 2295–2326.

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. 2019. "Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs." *Econometric Journal*, 23(2): 192–210.

Carneiro, Pedro, and Sokbae Lee. 2009. "Estimating Distributions of Potential Outcomes Using Local Instrumental Variables with an Application to Changes in College Enrollment and Wage Inequality." *Journal of Econometrics*, 149(2): 191–208.

Cattaneo, Matias D, Michael Jansson, and Xinwei Ma. 2018 "Manipulation Testing Based on Density Discontinuity" *The Stata Journal*, 18(1): 234–261.

Chen, Kuan-Ming , and Lin-Tung Tsai. 2024. "Selecting the Patients Who Benefit the Most: Evidence from Marginal Patients in Health Checks." *Unpublished manuscript.*

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *Econometric Journal*, 21(1): C1–C68.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2019. "Generic machine learning inference on heterogenous treatment effects in randomized experiments." arXiv:1712.04802.

Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schönberg. 2016. "From LATE to MTE: Alternative methods for the evaluation of policy interventions." *Labour Economics*, 41: 47–60.

Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schönberg. 2018. "Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance." *Journal of Political Economy*, 126(6): 2356–2409.

Einav, Liran, Amy Finkelstein, Tamar Oostrom, Abigail Ostriker, and Heidi Williams. 2020. "Screening and Selection: The Case of Mammograms." *American Economic Review*, 110(12): 3836–3870.

Fukuma, Shingo, Toshiaki Iizuka, Tatsuyoshi Ikenoue, and Yusuke Tsugawa. 2020. "Association of the National Health Guidance Intervention for Obesity and Cardiovascular Risks With Health Outcomes Among Japanese Men." *JAMA Internal Medecine*, 180(12): 1630–1637.

Handel, Benjamin, and Jonathan Kolstad. 2017. "Wearable Technologies and Health Behaviors: New Data and New Methods to Understand Population Health." *American Economic Review Papers & Proceedings*, 107(5): 481–485.

Haneda, Masakazu, Mitsuhiko Noda, Hideki Origasa, Hiroshi Noto, Daisuke Yabe,

Yukihiro Fujita, Atsushi Goto, Tatsuya Kondo, and Eiichi Araki. 2018. "Japanese clinical practice guideline for diabetes 2016." *Journal of Diabetes Investigation*, 9(3): 657–697.

Heckman, James J., and Edward Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *PNAS*, 96(8): 4730–4734.

Heckman, James J., and Edward Vytlacil. 2001. "Local Instrumental Variables." In Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics; Essays in Honor of Takeshi Amemiya, edited by Cheng Hsiao, K. Morikune, and James L. Powell, 1–46. New York: Cambridge Univ. Press.

Heckman, James J., and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, 73(3): 669–738.

Heckman, James J., and Edward Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part II." In Handbook of Econometrics, vol. 6, edited by James J. Heckman and Edward E. Leamer, 4875–5143. Amsterdam: Elsevier.

Iizuka, Toshiaki, Katsuhiko Nishiyama, Brian Chen, and Karen Eggleston. 2021. "False alarm? Estimating the marginal value of health signals." *Journal of Public Economics*, 195, 104368.

Ikegami, Naoki, and John C. Campbell. 1995. "Medical care in Japan." *New England Journal of Medicine*, 333: 1295–1299.

Ito, Koichiro, Takanori Ida, and Makoto Tanaka. 2023. "Selection on Welfare Gains: Experimental Evidence from Electricity Plan Choice." *American Economic Review*, 113(11): 2937–2973.

Jones, Damon, David Molitor, and Julian Reif. 2019. "What do Workplace Wellness Programs do? Evidence from the Illinois Workplace Wellness Study." *The Quarterly Journal of Economics*, 134(4): 1747–1791.

Kim, Hyuncheol Bryant, Suejin A, Lee, and Wilfredo Lim. 2019. "Knowing is not half the battle: Impacts of information from the National Health Screening Program in Korea." *Journal of Health Economics*, 65: 1–14.

Kitagawa, Toru, and Aleksey Tetenov. 2018. "Who should be treated? Empirical welfare maximization methods for treatment choice," *Econometrica*, 86: 591–616.

Kondo, Ayako, and Hitoshi Shigeoka. 2013. "Effects of universal health insurance on health care utilization and supply-side responses: Evidence from Japan." *Journal of Public Economics*, 99: 1–23.

Kowalski, E. Amanda. 2023. "Behaviour within a Clinical Trial and Implications for Mammography Guidelines." *The Review of Economic Studies*, 90(1): 432–462.

Miller, Y.D., and Dunstan, D.W. 2004. "The effectiveness of physical activity interventions for the treatment of overweight and obesity and type 2 diabetes". *Journal of Science and Medicine in Sport*, 7(1): 52–59.

UK Prospective Diabetes Study (UKPDS) Group. 1998. "Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33)." *The Lancet*, 352(9131): 837–853.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association*, 113(523): 1228–1242.

Yadlowsky, Steve, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. 2024. "Evaluating treatment prioritization rules via rank-weighted average treatment effects." *Journal of the American Statistical Association*, 1–14.

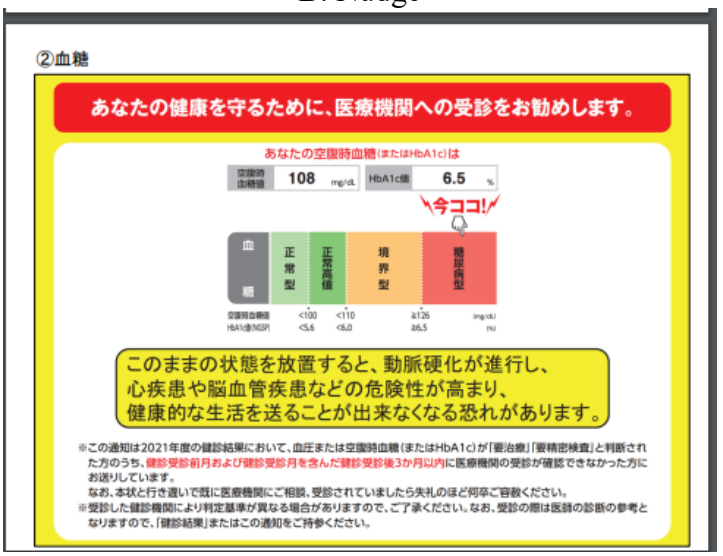
Zhou, Zhengyuan, Susan Athey, Stefan Wager, Ayush Kanodia, and Erik Sverdrup. 2020. "Multi-action treatment effect estimation and exact tree search (beta): policytree. R package version 0.1.0." Retrieved from https://grf-labs.github.io/policytree/ (on October 19, 2024)

Figure 1—An example of a report card and nudge

A. Report card

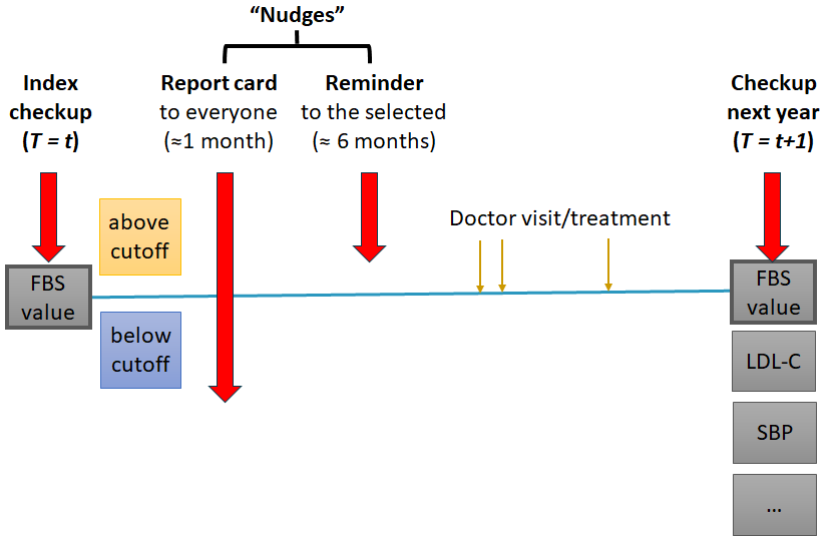


B. Nudge



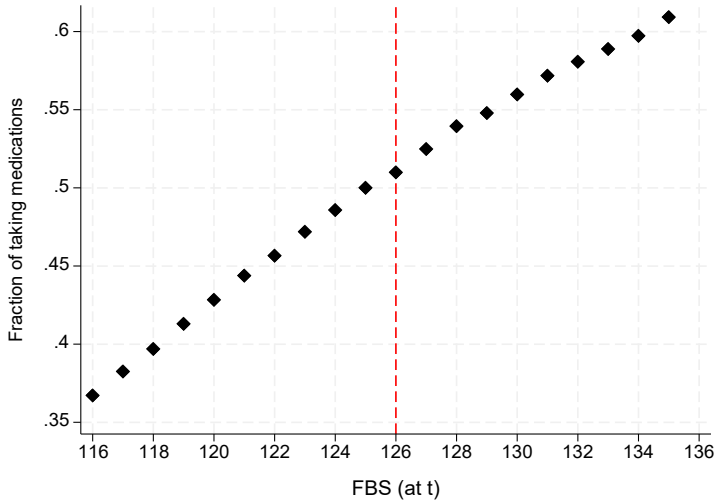
Notes: Panel A provides a typical checkup report that employees receive. If any value surpasses a clinical threshold, the report usually includes a warning (like 'H" for high) and recommends consulting a physician for further guidance. Panel B offers an example of a nudge sent by Japan Health Insurance Association (JHIA). The document includes a statement encouraging individuals to see a doctor, along with some warnings. While the wording of the warnings may slightly change over the years, the content remains similar. This particular example from 2018 states, "If you leave things as they are, you risk developing arteriosclerosis, increasing the risk of heart disease and cerebrovascular disease, and becoming unable to lead a healthy life." The original language is in Japanese.

Figure 2—Flowchart



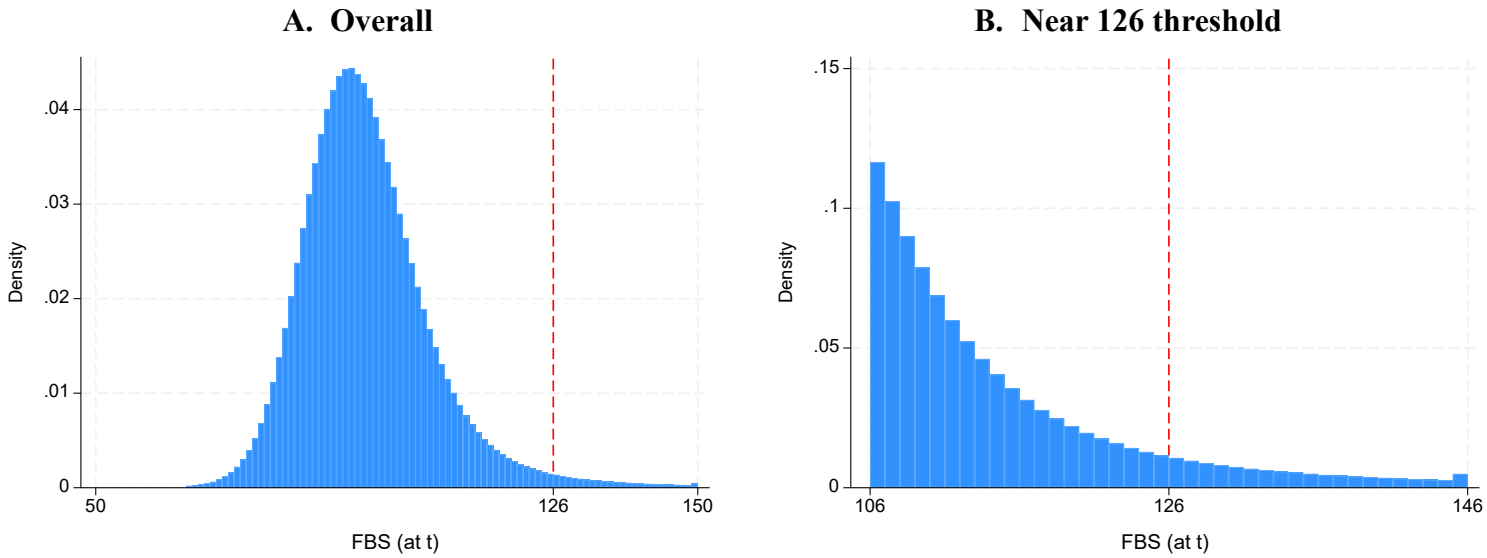
Notes: The figure illustrates the flowchart of the checkup process. After the initial checkup at time t, all individuals receive their report cards within one to two months. Subsequently, those not undergoing DM treatment and those with an FBS above 126 mg/dl receive a reminder with a health warning in month 6. We then observe their biomarkers during the next checkup at time *t+1* for those who participate again. Additionally, we monitor healthcare utilization between the initial checkup and the next one.

Figure 3—Probability of already taking DM medication at the index checkup



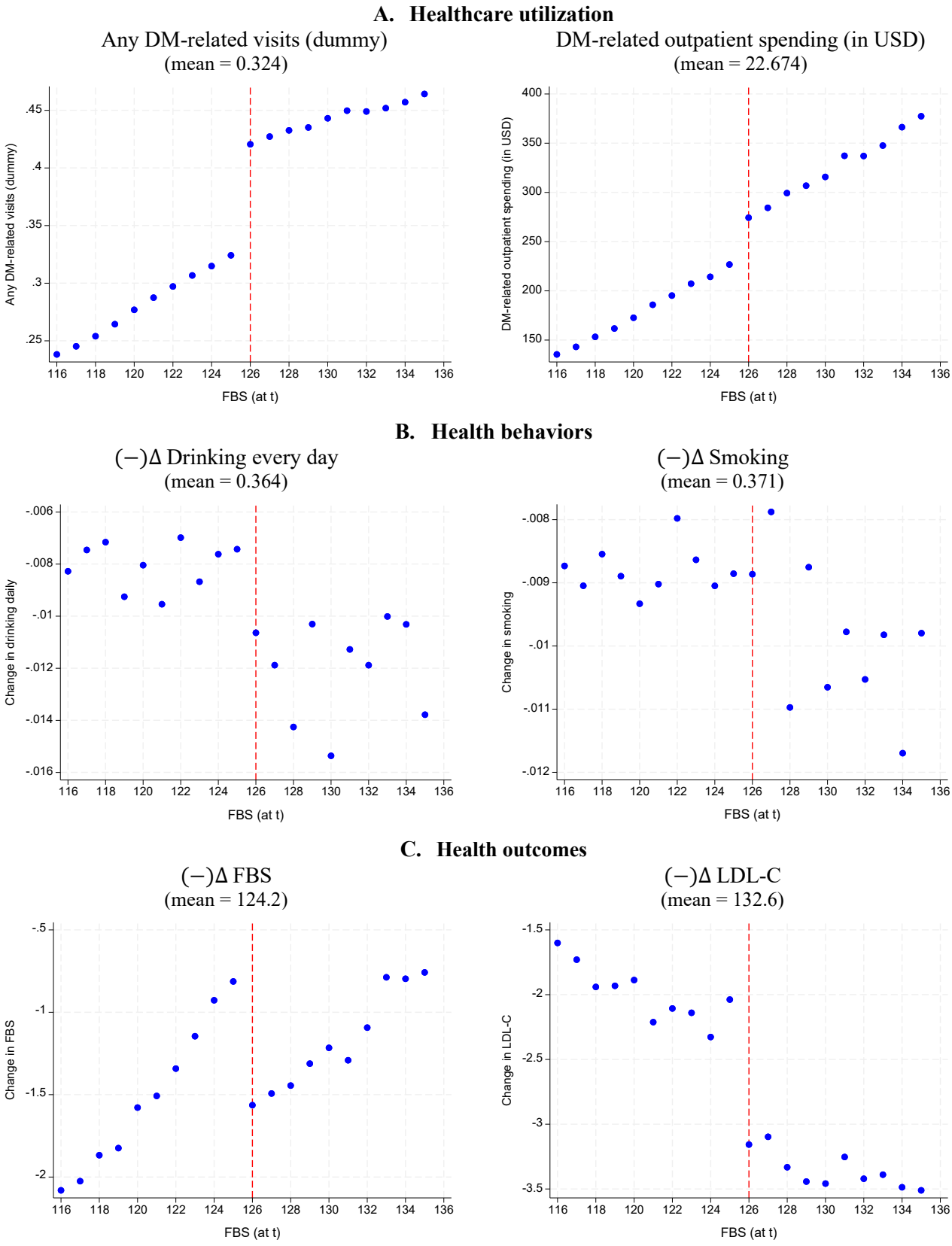
Notes: The figure plots the mean probability of taking either hypertension or diabetes medication at 1 point bins of the FBS value within 10 mg/dl around the FBS = 126 mg/dl threshold. The sample above the line is our "main sample," who received the nudge 6 months after the index checkup in year t. The sample below the line is our "placebo sample," which did not receive the nudge.

Figure 4—Distribution of FBS values



Notes: The data is restricted to the main sample. The histogram shows the density of FBS values at yare t at 1 point bins of the FBS value for the overall sample (panel A) and within 20 mg/dl around the FBS = 126 mg/dl threshold (panel B).

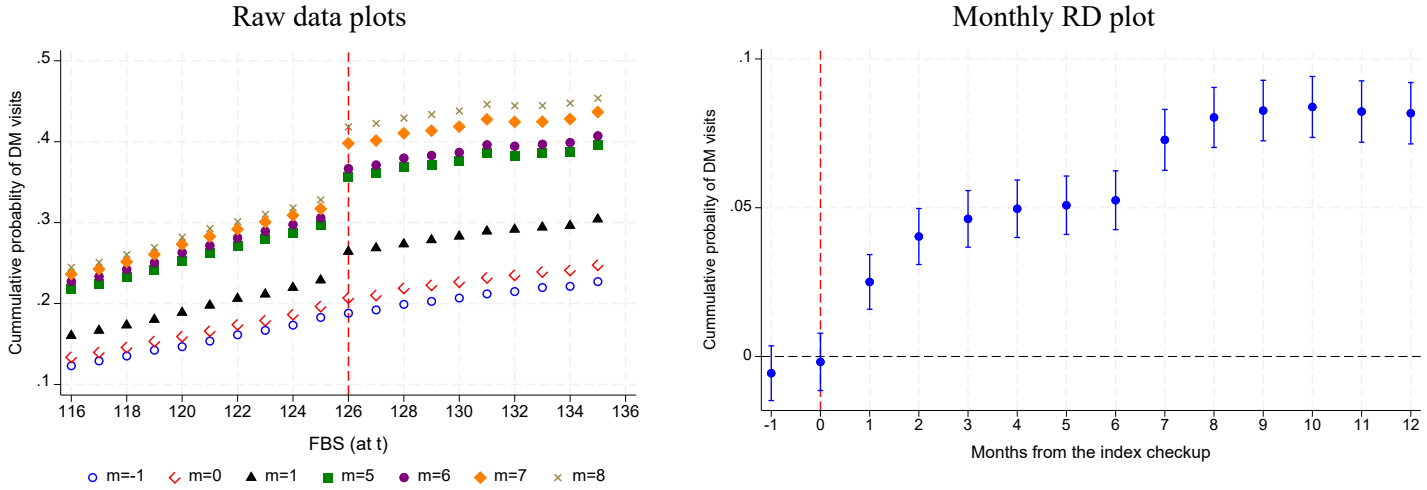
Figure 5—RD plots



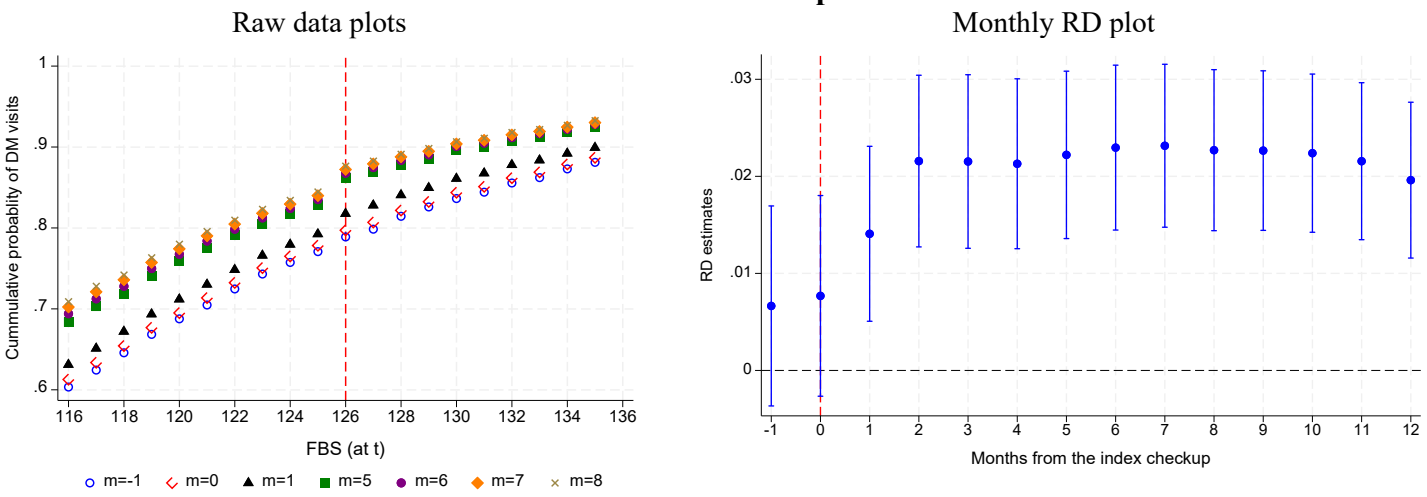
Notes: The data is limited to the main sample. The scatter plot shows the mean of each measure of healthcare utilization (panel A), health behaviors (panel B), and health outcomes (panel C) at 1-point bins of the FBS value within 10 mg/dl around the FBS = 126 mg/dl threshold at the time of index checkup in year *t*. See Appendix D for the definition of variables.

Figure 6—Monthly medical care utilization

A. Main sample

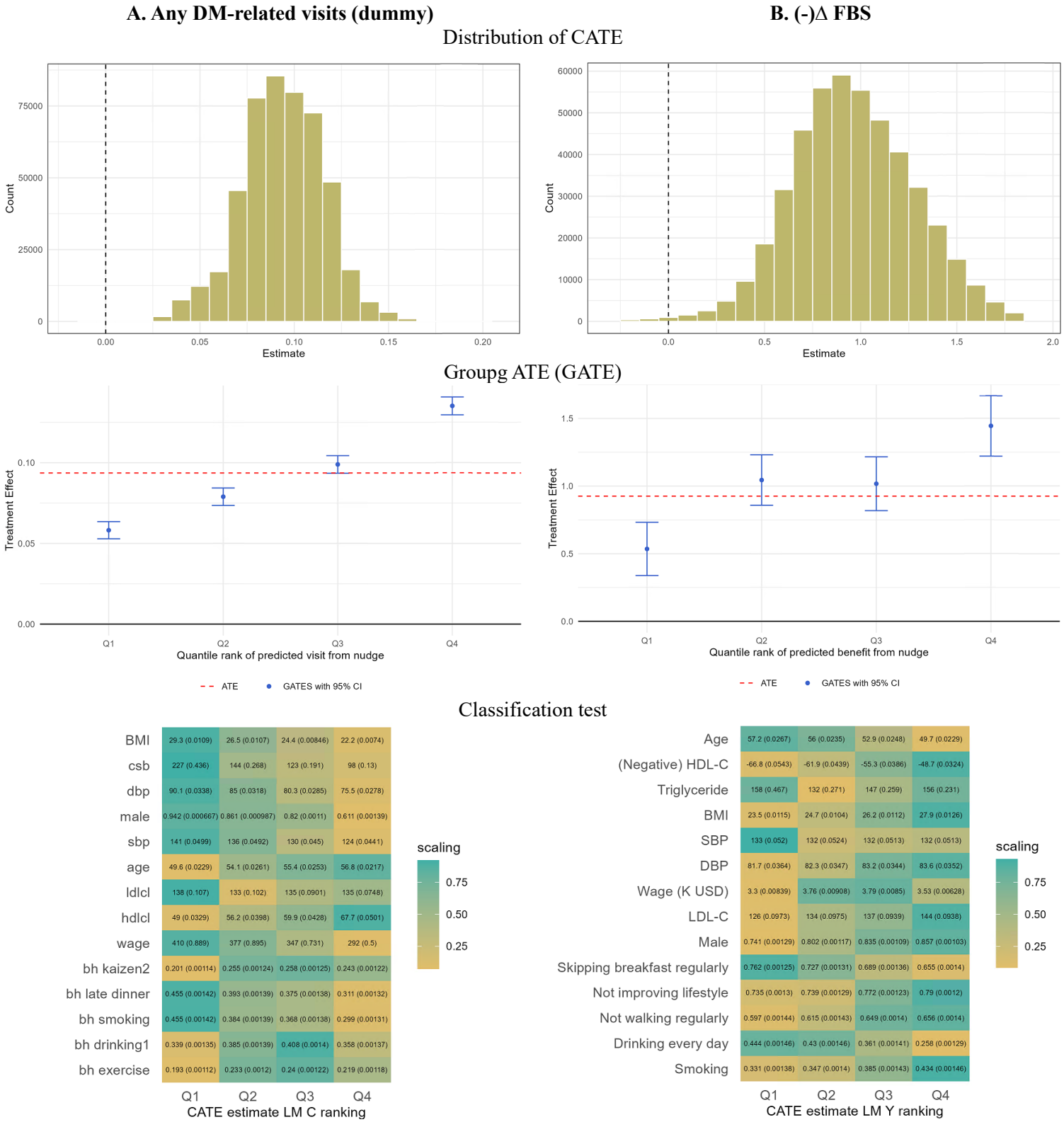


B. Placebo sample



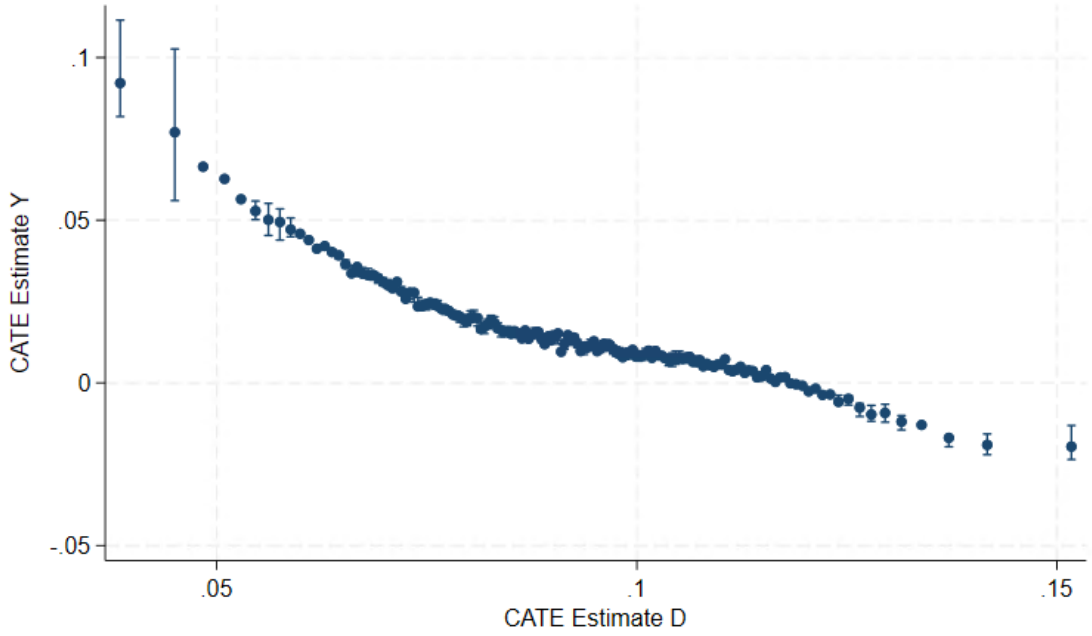
Notes: The data in panel A come from the main sample that receives the nudge six months after the index checkup, while the data in panel B are from the placebo sample that does not receive the corresponding nudge, as they are already taking either hypertension or diabetes medication, as shown in Figure 3. The outcome is a cumulative DM visit dummy, which equals one if an individual has at least one DM visit until month m from the index checkup and zero otherwise. For panels A and B, the left figure plots the outcome at cumulative months m = -1, 0, 1, 5, 6, 7, and 8, where m = -1 indicates a month before the index checkup and m = 0 is the month when the individual had the checkup. The figures on the right plot the RD estimates from equation [1], where the outcome is the cumulative DM visit dummy for each month m from -1 to 12. The nudge is provided to individuals in the main sample (panel A) at m = 6.

Figure 7—The CATEs of utilization and health outcome



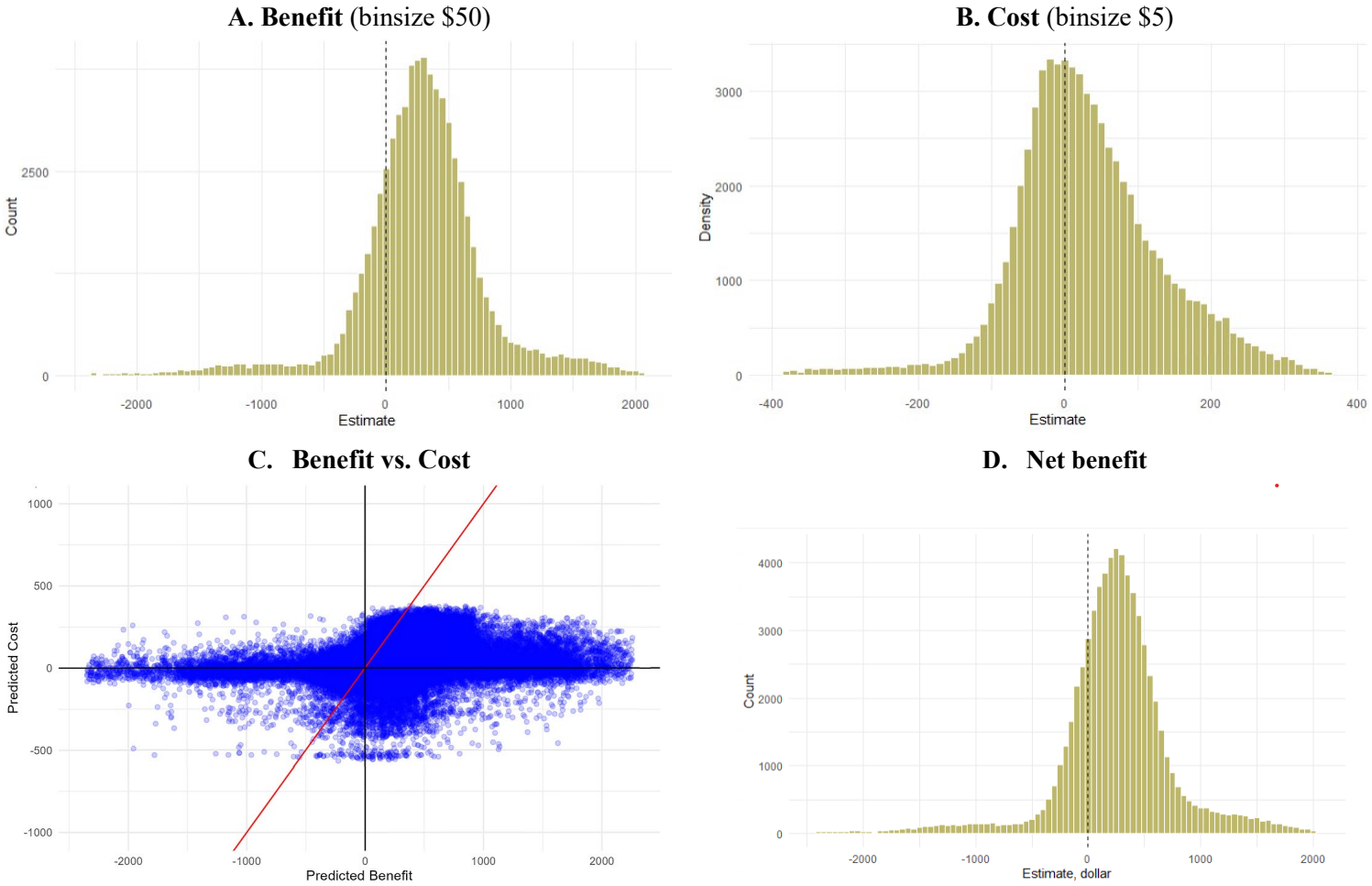
Notes: The data is limited to the main sample. The outcome in panel A is a binary indicator that equals one if an individual has at least one DM visit within a year following an index checkup and zero otherwise, and the outcome in panel B is the change of fasting blood sugar (Δ FBS) between year $t+1$ and year t. The top row plots the histogram of estimated $\tau(X)$ from equation [2], which is refined using the doubly robust score to mitigate regularization bias and overfitting. The middle row groups estimated CATEs in the upper row into four based on the estimated sizes and plots of the group-level treatment effects (GATE). The bottom row shows the classification test for causal forests to assess whether there is heterogeneity in estimated treatment effects by each observable. The LM Forest, which is part of the GRF (Generalized Random Forests) package in R, is used.

Figure 8—Relationship between CATEs for D and Y



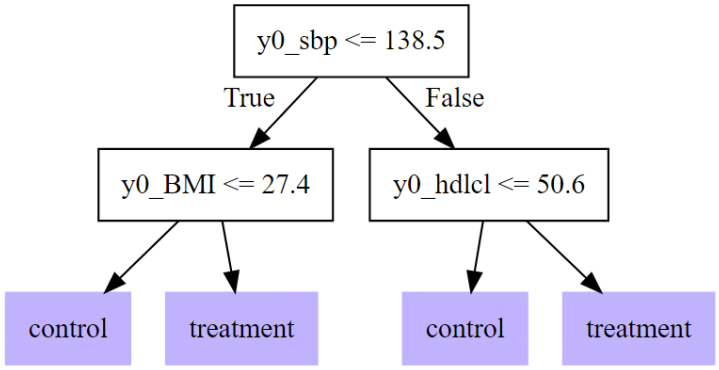
Notes: The data is limited to the main sample. The figure reports a binned scatter plot of the relationship between the estimated CATE for compliance (*x*-axis) and the CATE for health outcome improvements (*y*-axis) at the individual level.

Figure 9—Distribution of CATE on benefit, cost, and net benefit



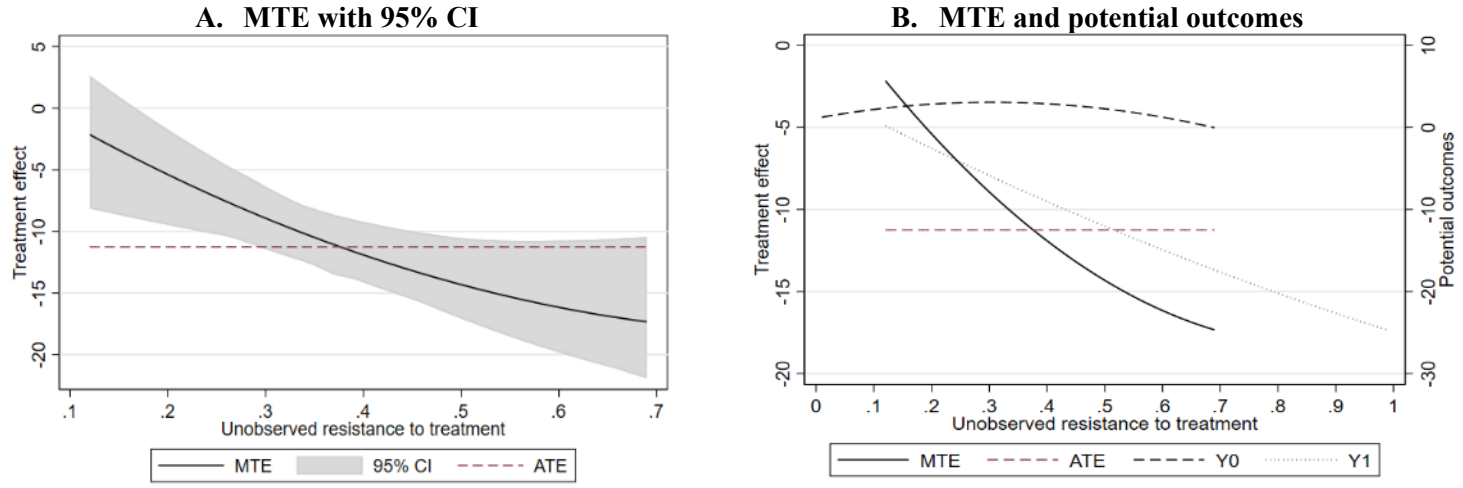
Notes: The data is limited to the main sample. The figures the histogram of estimated $\tau(X)$ from equation [2], which is refined using the doubly robust score to mitigate regularization bias and overfitting. For panel A, the benefit is defined by the 6-year mortality reduction times the value of statistical life per year, which is set at $50,000 per year. For panel B, the cost is defined by the cumulative DM-related outpatient spending during the 6 years from the initial index checkup. Both benefits and costs are adjusted by an annual discount rate of 3 %. Panel C plots the benefit from panel A on the *x*-axis, and the cost from panel B on the *y*-axis, where the slope of the solid line is $50,000, the value of statistical life used in this study. Finally, panel D is the net benefit, which is defined by the difference between the benefit and cost at each individual level.

Figure 10—Policy tree



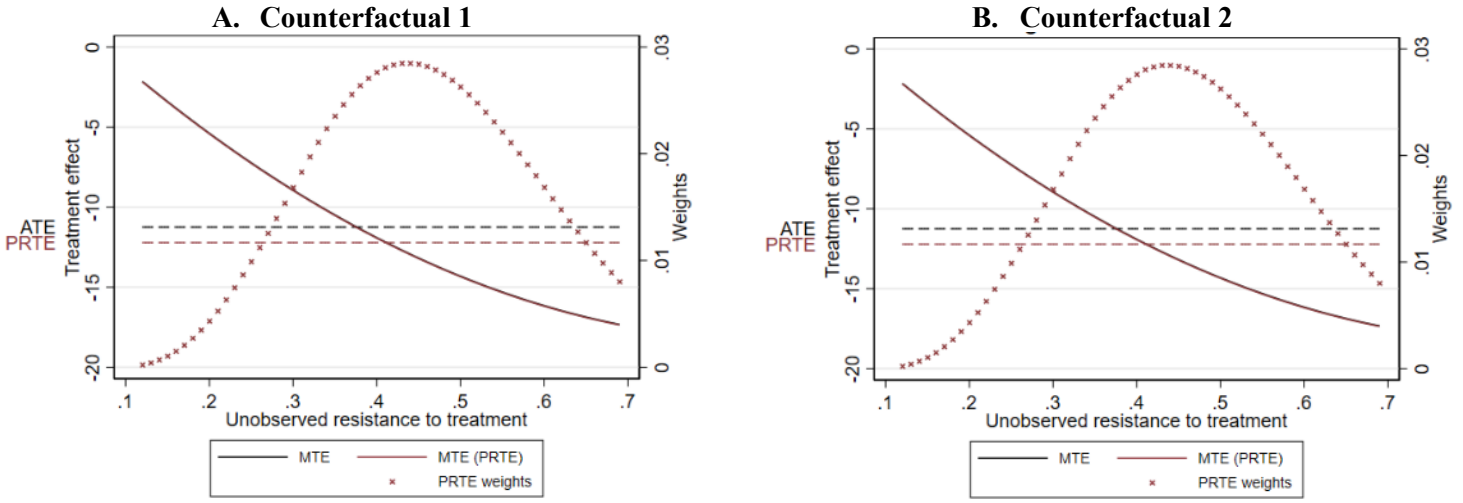
Notes: The data is limited to the main sample. Depth-2 policies based on the estimated net benefit using the policytree package for R.

Figure 11—MTE on ΔFBS



Notes: The data is limited to the main sample. The outcome is the change in fasting blood sugar (Δ FBS) between year $t+1$ and year t. Panel A plots the marginal treatment effect (MTE) and 95% confidence intervals as a function of unobserved resistance to treatment (u_d) The dotted line corresponds to the average treatment effect (ATE). Panel B decomposes the MTE in panel A to the treated, $E(Y_1|\ U_d = u_d)$ and the untreated, $E(Y_0|\ U_d = u_d)$.

Figure 12—Policy simulations



Notes: The data is limited to the main sample. The outcomes are the change between year $t+1$ and year t for fasting blood sugar (Δ FBS). The figures plot the marginal treatment effect (MTE), and a policy-relevant treatment effect (PRTE), which represents a weighted difference between the average treatment effect on the treated (ATT) under the alternative policy and the ATT under the baseline policy. Counterfactual 1 in panel A is the scenario where we add 0.2 of the propensity scores (PS) to all individuals (and set it to 1 if PS becomes ≥ 1. Counterfactual 2 in panel B is the scenario where we add 0.2 of PS only to the very sick only (SBP ≥ 160 or DBP ≥ 100 or LDL-C ≥ 180).

Table 1—RD Estimates

	(1) Coefficient	(2) SE	(3) Mean at 125	(4) Bandwidth	(5) Obs.
Panel A: Healthcare utilization					
Any DM-related visits (dummy)	0.088***	(0.005)	0.324	3.302	366,086
Number of DM visits	0.354***	(0.048)	1.877	3.314	366,086
DM outpatient spending	36.434***	(6.526)	226.739	3.330	365,757
DM drug spending	0.574***	(0.189)	2.673	3.347	366,038
Panel B: Health behaviors					
(−)Δ Drinking every day	0.0029	(0.003)	0.364	3.366	346,347
(−)Δ Smoking	0.0005	(0.002)	0.371	3.204	376,682
(−)Δ Not walking regularly	0.0103*	(0.0053)	0.620	3.073	344,257
(−)Δ Skipping breakfast regularly	-0.0025	(0.0045)	0.289	2.331	240,919
Panel C: Health outcomes					
(−)Δ FBS	0.934***	(0.197)	124.199	3.516	354,878
(−)Δ SBP	0.532***	(0.172)	132.687	2.950	256,551
(−)Δ BMI	0.082***	(0.012)	25.540	3.101	364,477
(−)Δ LDL-C	1.643***	(0.319)	132.702	2.243	255,592

Notes: The data is limited to the main sample. This table shows the estimates from local-linear regression in equation [1], using a triangular kernel and covariates. The optimal bandwidth is derived by Calonico et al. (2019), and the robust bias-corrected standard error by Calonico et al. (2014) is reported in parenthesis. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.10$.

Table 2—Targeting

	Net benefit per person ($)	Ratio to status quo
Status quo: FBS≥ 126	68.54	1.00
Policy tree	77.33	1.13
Non-parametric: All 1(net benefit))>0	97.33	1.42

Notes: The data is limited to the main sample. This table shows the per-person net benefit (AIPW) in different scenarios. See Figure 11 on policy trees.

Table 3—Policy simulations

	Effects	SE
ATE	-12.29	(1.96)
ATT	-4.92	(2.94)
ATU	-16.51	(2.68)
LATE	-11.06	(1.81)
PRTE (Counterfactual 1)	-13.39	(1.82)
PRTE (Counterfactual 2)	-14.07	(2.03)

Notes: The table summarizes several treatment effects (ATE, ATT, ATU, and LATE) and policy-relevant treatment effect (PRTE), which represents a weighted difference between the average treatment effect on the treated (ATT) under the alternative policy and the ATT under the baseline policy. Counterfactual 1 (see Figure 13-A) is the scenario where we add 0.2 of the propensity scores (PS) to all individuals (and set it to 1 if PS becomes ≥ 1. Counterfactual 2 (see Figure 13-B) is the scenario where we add 0.2 of the propensity scores (PS) only to the very sick (SBP ≥ 160 or DBP ≥ 100 or LDL ≥ 180).

Online Appendix

(Not for Publication)

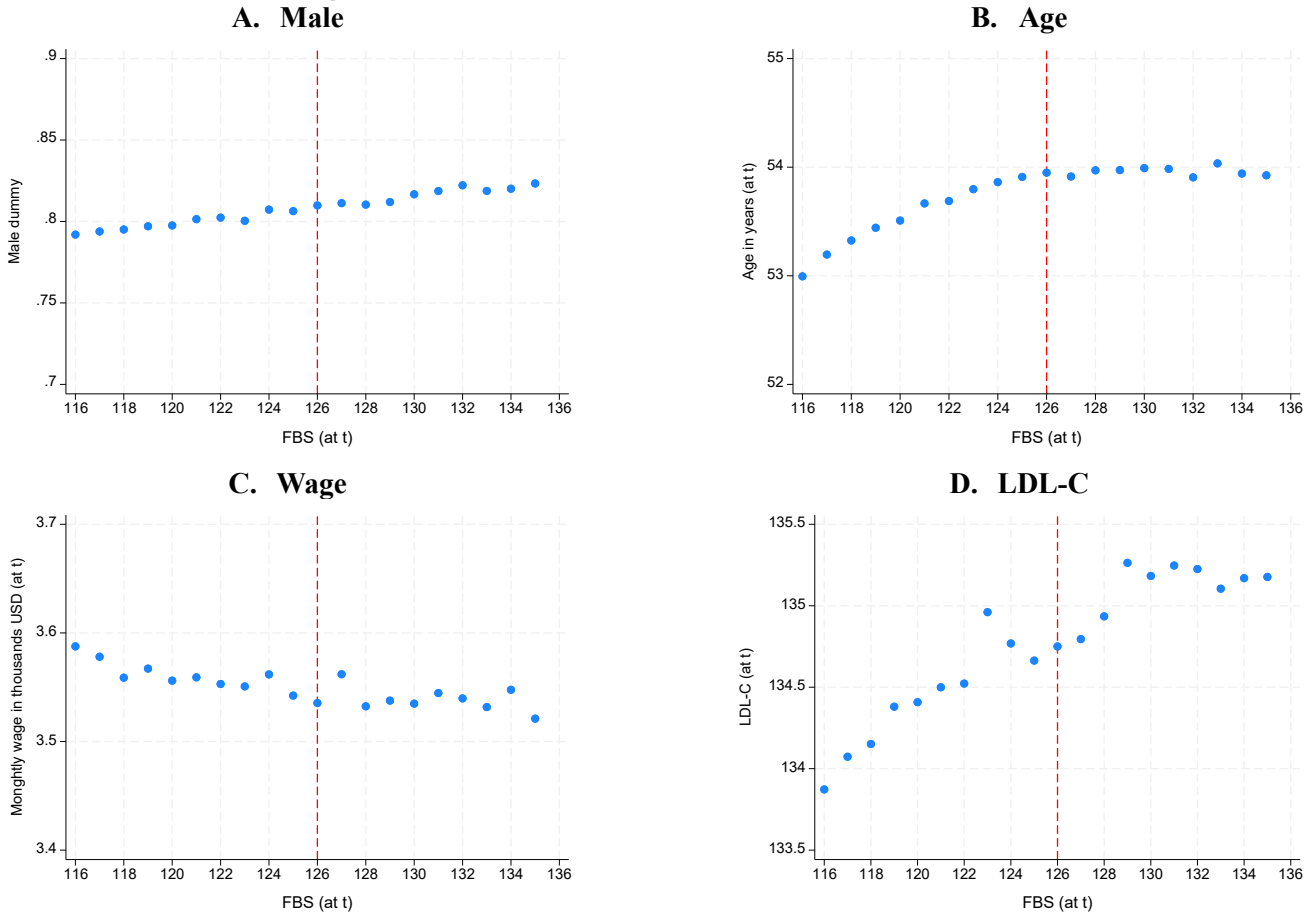
Targeting Healthcare Nudges

By Toshiaki Iizuka, Junya Kawamura, and Hitoshi Shigeoka

Section A RDD
Section B CATE
Section C MTE
Section D Data Appendix

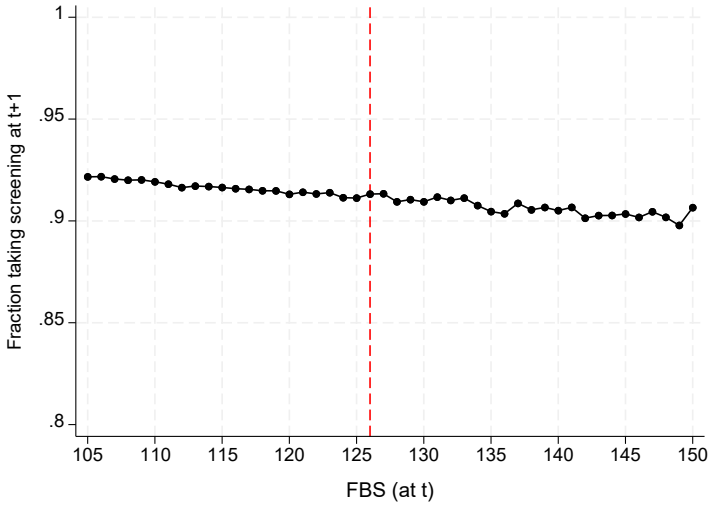
Appendix A: RDD

Figure A1—Balanced checks for selected variables



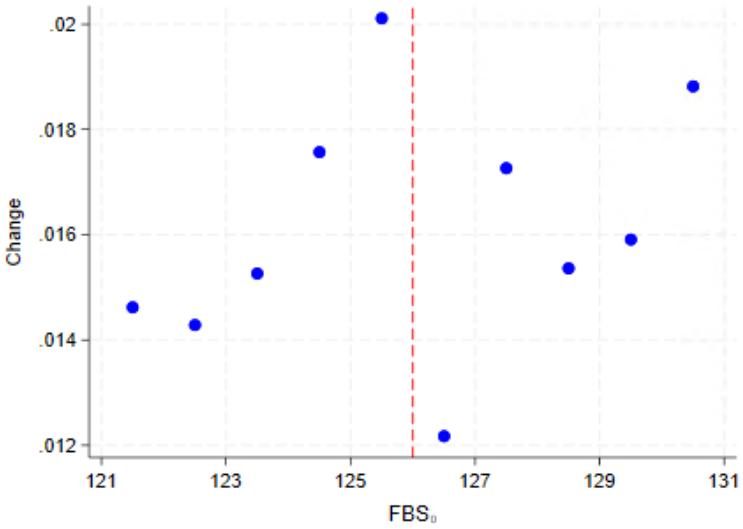
Notes: The data is limited to the main sample. The figures plot the mean of the variables at 1 point bins of the FBS value within 10 mg/dl around the FBS = 126 mg/dl threshold at the time of index checkup in year *t*. The outcomes are the male dummy (panel A), age (panel B), wage (panel C), and low-density lipoprotein cholesterol (panel D), measured at the index check-up in year *t*.

Figure A2—Attrition



Notes: The figure plots the mean of the probability of attending the next year's checkup (year *t+1*) at 1 point bins of the FBS value at the time of index checkup in year *t*. The dotted line corresponds to the FBS = 126 mg/dl threshold.

Figure A3—RD plots of 6-year mortality



Notes: The data is limited to the main sample and the individuals that belong to Q1 of GATE in Figure 9-B. The outcome is the cumulative mortality rate for 6 years from the index checkup.

Table A1—Summary statistics

	Full sample			Within FBS 116-135					
				Main sample			Placebo sample		
Variable:	N	Mean	SD	N	Mean	SD	N	Mean	SD
A. Running variable (at t)									
FBS (mg/dl)	48,017,588	98.46	19.17	1,312,491	122.37	5.34	1,310,744	124.21	5.71
B. Covariates									
Male (dummy)	48,017,588	0.64	0.48	1,312,491	0.80	0.40	1,310,744	0.81	0.39
Age	48,017,588	50.40	9.34	1,312,491	53.55	8.90	1,310,744	58.52	8.21
Monthly wage (in K USD)	47,744,932	3.33	2.35	1,305,851	3.58	2.74	1,302,876	3.51	2.94
C. Healthcare utilization (from t to t+1)									
Any DM-related visits (dummy)	48,017,588	0.20	0.40	1,312,491	0.32	0.47	1,310,744	0.82	0.39
Number of DM-related visits	48,017,588	1.28	4.48	1,312,491	1.74	4.15	1,310,744	8.29	9.38
DM-related outpatient spending (in USD)	47,978,760	201.39	724.01	1,311,369	210.12	586.75	1,305,502	1,429.59	1,535.98
DM-related drug spending (in USD)	47,972,208	56.97	324.72	1,312,382	23.93	169.48	1,299,910	556.98	863.42
D. Health behavior (at t+1)									
Drinking every day (dummy)	44,395,768	0.30	0.46	1,214,226	0.38	0.48	1,218,285	0.37	0.48
Smoking (dummy)	48,017,588	0.32	0.46	1,312,491	0.38	0.48	1,310,744	0.30	0.46
Not walking regularly (dummy)	44,117,568	0.62	0.49	1,207,622	0.63	0.48	1,212,542	0.66	0.47
Skipping breakfast regularly (dummy)	43,819,616	0.25	0.43	1,199,586	0.30	0.46	1,204,026	0.17	0.38
E. Health outcomes (at t+1)									
FBS (mg/dl)	46,398,308	98.92	18.97	1,273,870	120.83	18.16	1,266,300	126.61	19.89
SBP (mmHg)	47,915,440	123.89	17.10	1,307,787	132.09	17.65	1,308,076	132.31	16.32
BMI	47,916,612	23.64	3.93	1,308,401	25.36	4.20	1,305,743	26.29	4.26
LDL-C (mg/dl)	47,697,656	124.25	31.38	1,302,978	132.27	33.22	1,302,382	115.18	29.30

Notes: This table presents summary statistics for the full sample and the sample that falls within 10 mg/dl of the FBS = 126 mg/dl threshold. Monthly wage is measured in thousand USD, and DM-related outpatient spending and DM-related drug spending are measured in USD (100 JPY/USD). See Appendix D for the definition of variables.

Table A2—Balanced checks

Outcomes (at *t*)	(1) Coefficient	(2) SE	(3) Mean at 125	(4) Bandwidth	(5) Obs.
Male	0.006	(0.005)	0.806	2.894	258,303
Age	-0.008	(0.101)	53.884	3.118	366,086
Wage	0.013	(0.031)	3.562	3.174	364,254
Drinking every day	0.002	(0.006)	0.372	2.983	238,976
Smoking	-0.003	(0.005)	0.381	3.407	366,086
Not walking regularly	0.006	(0.007)	0.628	2.650	237,628
Skipping breakfast regularly	-0.008	(0.006)	0.299	2.926	236,070
LDL-C	0.338	(0.406)	134.714	2.885	256,486
SBP	0.189	(0.202)	132.871	3.113	364,593
BMI	-0.082	(0.058)	25.660	2.497	257,339

Notes: The data is limited to the main sample. The outcomes are measured at the index check-up in year *t*. This table shows the estimates from local-linear regression in equation [1], using a triangular kernel and year FE as covariates. The optimal bandwidth is derived by Calonico et al. (2019), and the robust bias-corrected standard error by Calonico et al. (2014) is reported in parenthesis. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.10$.

Table A3—IV Estimates (TOT)

	(1) Coefficient	(2) SE	(3) Mean at 125	(4) Bandwidth	(5) Obs.
Panel A: Endogenous variable = Any DM-related visit (dummy)					
(−)Δ FBS	10.177***	(2.490)	124.199	3.159	354,878
(−)Δ LDL-C	18.433***	(3.622)	132.702	2.324	255,592
(−)Δ SBP	5.086***	(1.669)	132.687	3.234	363,586
(−)Δ BMI	0.952***	(0.150)	25.540	3.014	364,477
Panel B: Endogenous variable = DM-related outpatient spending					
(−)Δ FBS	0.026***	(0.009)	124.199	3.006	354,564
(−)Δ LDL-C	0.046***	(0.012)	132.702	2.567	255,371
(−)Δ SBP	0.016***	(0.006)	132.687	2.745	256,327
(−)Δ BMI	0.002***	(0.001)	25.540	3.131	364,149

Notes: The data is limited to the main sample. This table shows the estimates from the two-stage linear regressions using a fuzzy regression discontinuity framework, using a triangular kernel and covariates. The estimation uses exceeding the threshold as an instrumental variable for two endogenous variables: the number of DM-related visits (panel A) and DM-related outpatient medical spending (panel B). Only the coefficients for the endogenous explanatory variables are reported. The optimal bandwidth is derived by Calonico et al. (2019), and the robust bias-corrected standard error by Calonico et al. (2014) is reported in parenthesis. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.10$.

Table A4—RDD on 6-year cumulative mortality

	(1) Coefficient	(2) SE
Bandwidth = ±3	-0.01182**	(0.00549)
Bandwidth = ±4	-0.01116**	(0.00472)
Bandwidth = ±5	-0.00987**	(0.00422)

Notes: The data is limited to the main sample and the individuals that belong to Q1 of GATE in Figure 9-B. This table shows the estimates from local-linear regression in equation [1], using a triangular kernel and covariates. The outcome is the cumulative mortality rate for 6 years from the index checkup. The bandwidth varies from 3-5 mg/dl, and the conventional standard error by Calonico et al. (2014) is reported in parenthesis. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.10$.

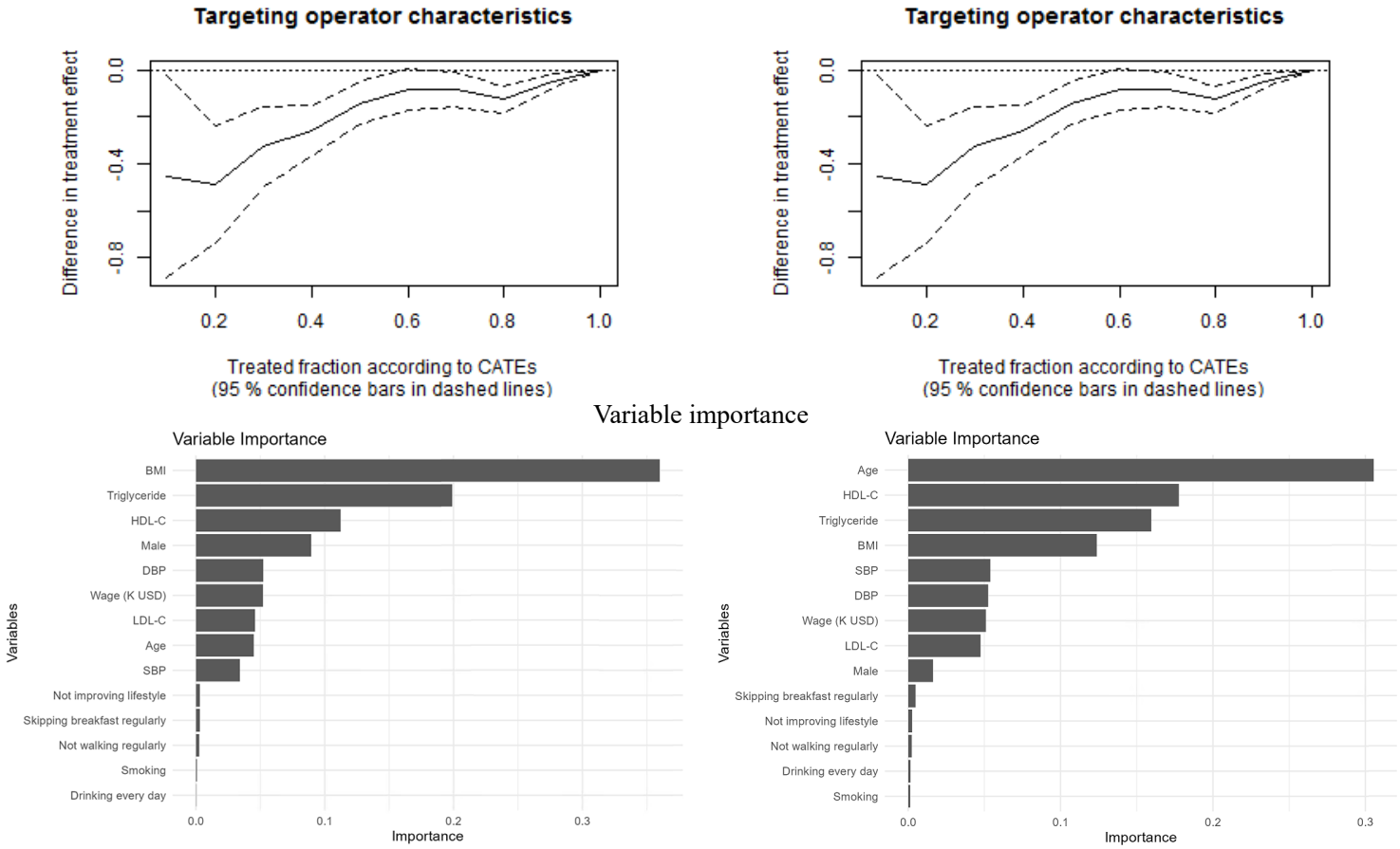
Appendix B: CATE

Figure B1—Other metrics from CATEs

(Any DM-related visit and ΔFBS)

A. Any DM-related visits (dummy) **B. (-)Δ FBS**

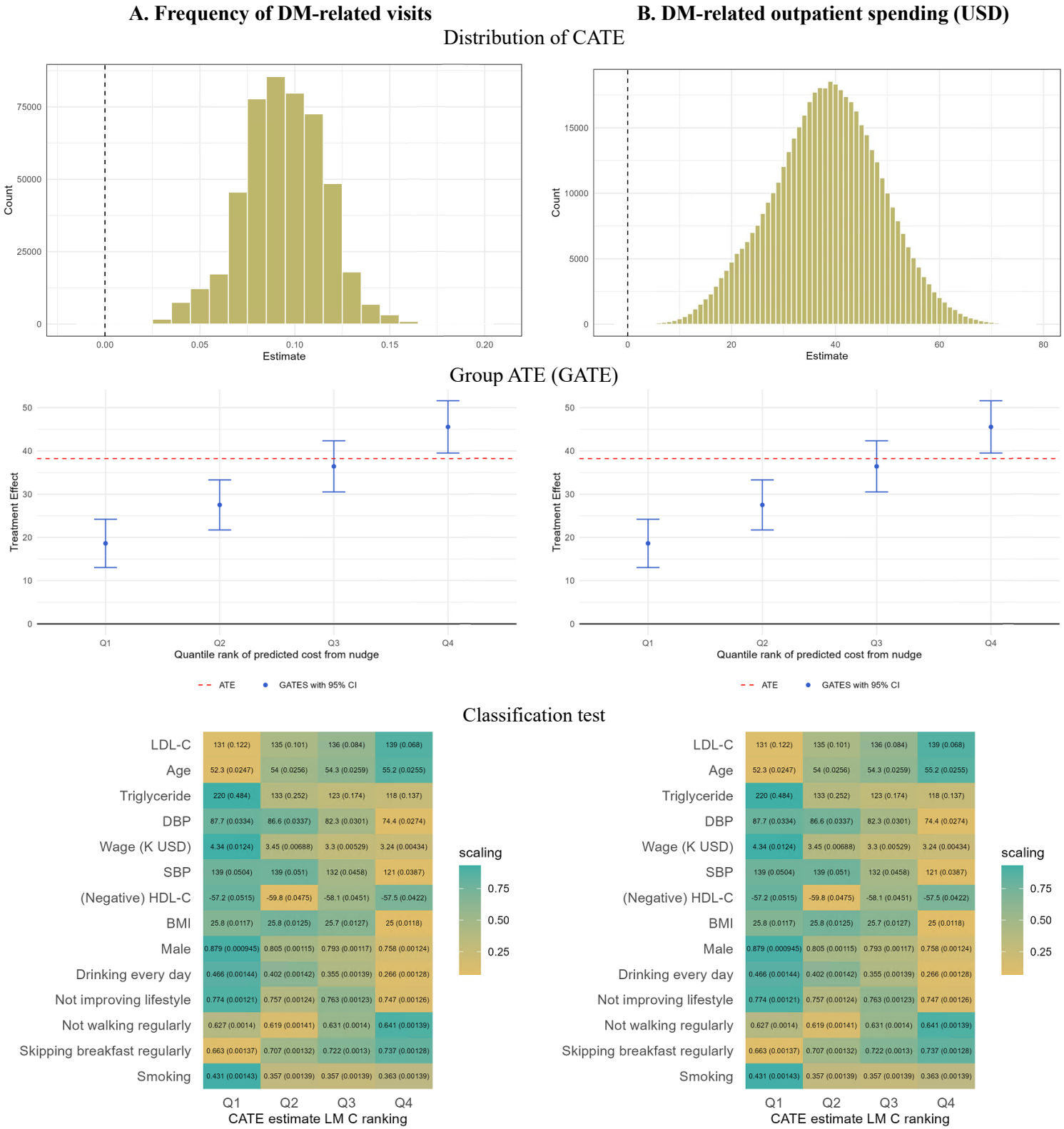
Rank-weighted ATE (RATE)



Variable importance

Notes: The data is limited to the main sample. The outcome in panel A is a binary indicator that equals one if an individual has at least one DM visit within a year following an index checkup and zero otherwise, and the outcome in panel B is the change of fasting blood sugar (Δ FBS) between year $t+1$ and year t. The top row reports the rank-weighted average treatment effects (RATE) proposed by Yadlowsky et al. (2024), and the bottom row reports the variable importance. The LM Forest, which is part of the GRF (Generalized Random Forests) package in R, is used.

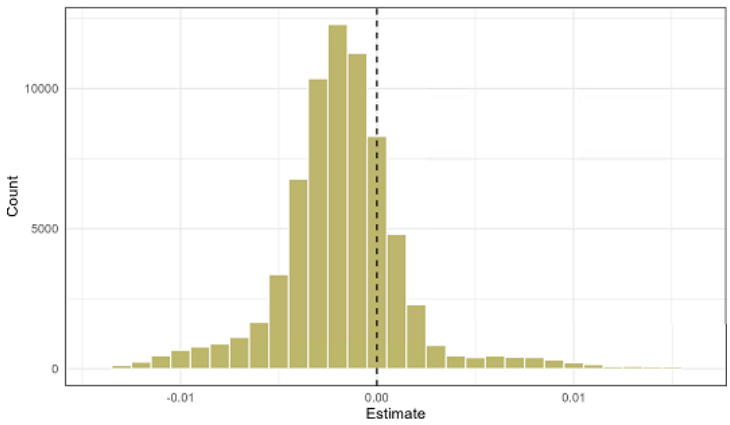
Figure B2—CATE for other utilization measures



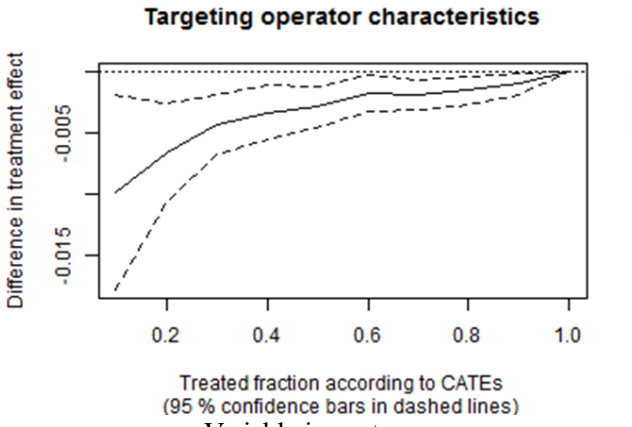
Notes: The data is limited to the main sample. The outcomes in panels A and B are the frequency of DM-related outpatient visits and DM-related outpatient spending aggregated for 12 months from the month of the checkup. The top row plots the histogram of estimated $\tau(X)$ from equation [2], which is refined using the doubly robust score to mitigate regularization bias and overfitting. The middle row groups estimated CATEs in the upper row into four based on the estimated sizes and plots of the group-level treatment effects (GATE). The bottom row shows the classification test for causal forests to assess whether there is heterogeneity in estimated treatment effects by each observable. The LM Forest, which is part of the GRF (Generalized Random Forests) package in R, is used.

Figure B3—The CATEs of 6-year cumulative mortality

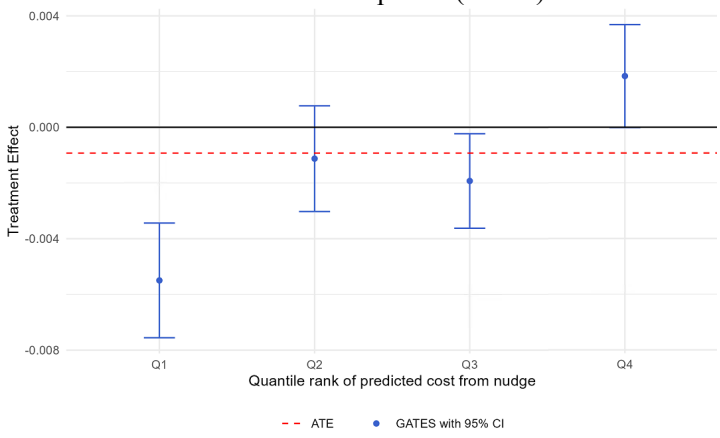
Distribution of CATE



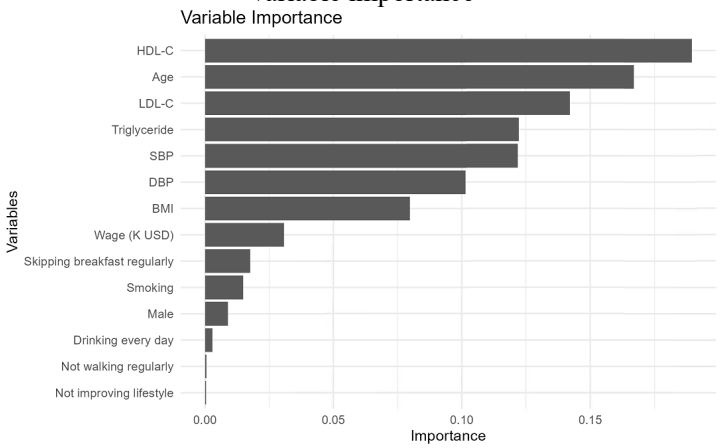
Rank-weighted ATE (RATE)



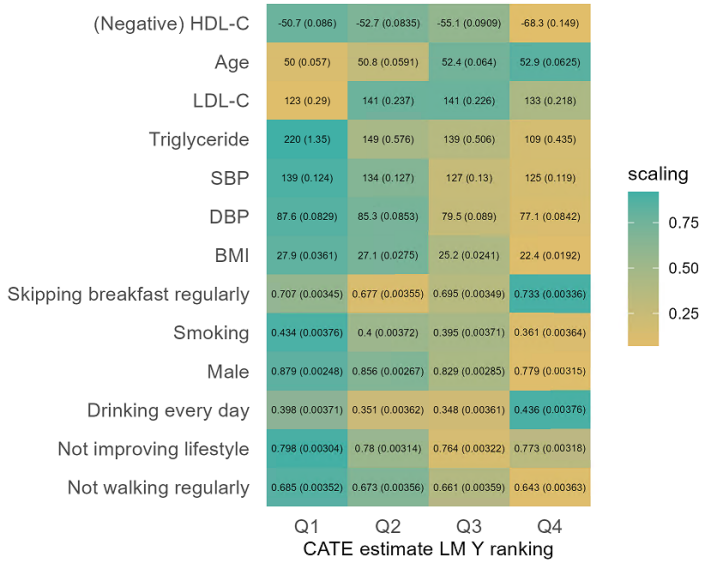
Group ATE (GATE)



Variable importance

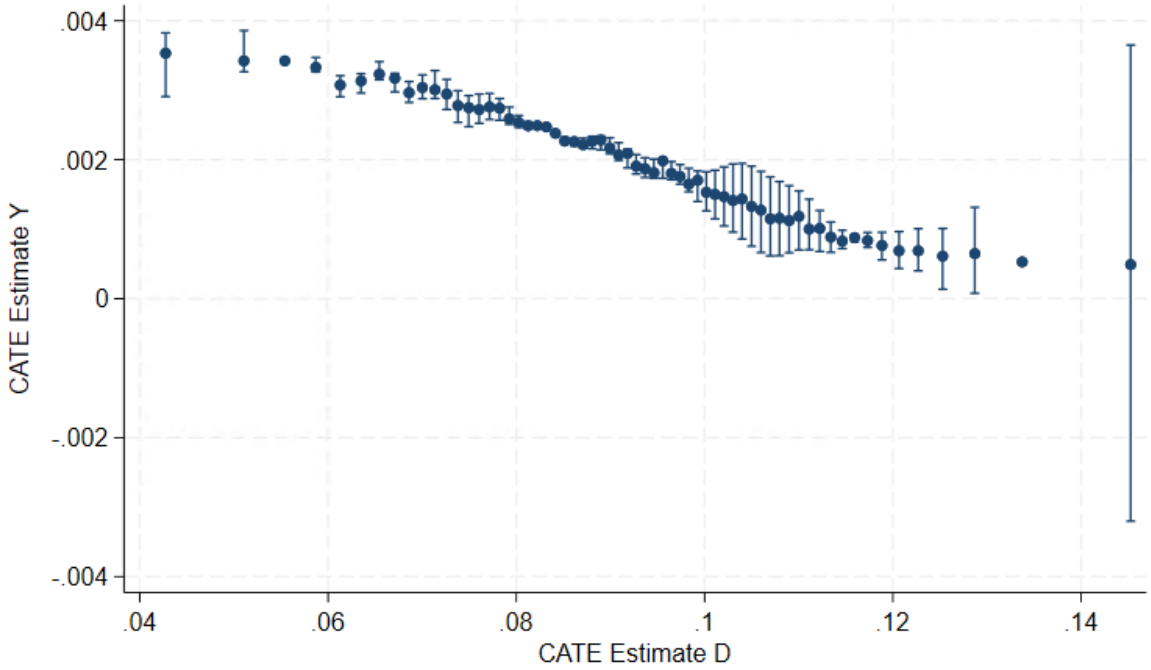


Classification test



Notes: The data is limited to the main sample. The outcome is the cumulative mortality for 6 years from the index checkup. The top row plots the histogram of estimated $\tau(X)$ from equation [2], which is refined using the doubly robust score to mitigate regularization bias and overfitting. The middle row groups estimated CATEs in the upper row into four based on the estimated sizes and plots of the group-level treatment effects (GATE). The bottom row shows the classification test for causal forests to assess whether there is heterogeneity in estimated treatment effects by each observable. The LM Forest, which is part of the GRF (Generalized Random Forests) package in R, is used. The minimum node size is 6,000, and the rep is 8,000.

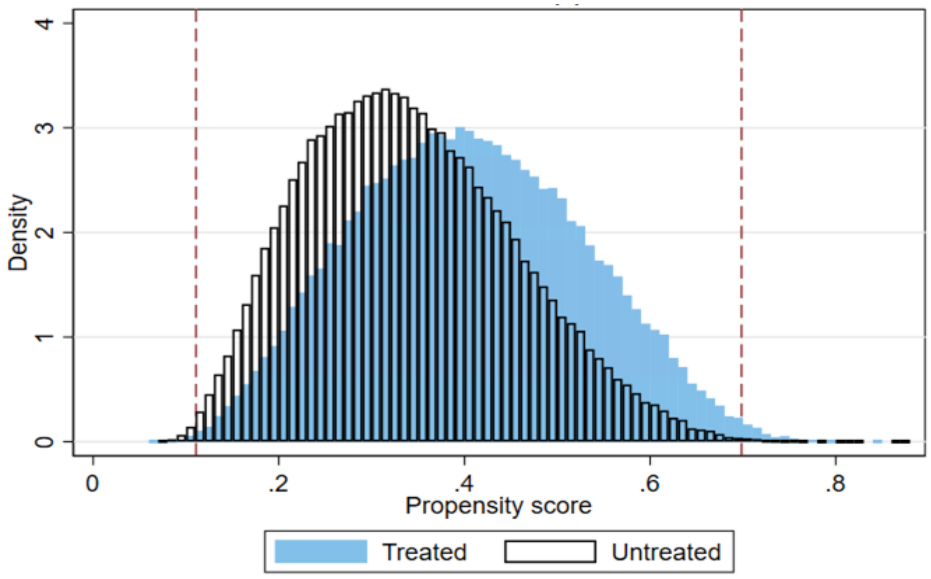
Figure B4—Relationship between CATEs for D and Y (6-year mortality rate)



Notes: The data is limited to the main sample. The figure reports a binned scatter plot of the relationship between the estimated CATE for compliance (*x*-axis) and the CATE for health outcome improvements (*y*-axis) at the individual level.

Appendix C: MTE

Figure C1—Overlap of propensity scores



Notes: The data is limited to the main sample. The figure shows the overlap of propensity scores between treated and untreated. The two vertical dotted lines correspond to the 0.1 and 99.9 percentiles.

Appendix D: Variable definition

Variables	Definition
Any DM-related visit	▪Indicator variable for any DM-related visit within one year from the index checkup ▪ DM is defined as ICD10 codes of E10~E14 as one of the conditions
Number of DM-related visits	▪Consultation days for DM within 1 year of the checkup
DM-related outpatient spending	▪Total spending on outpatient claims that include "DM" as one of the conditions
DM-related drug spending	▪Total spending on pharmacy claims that include a "DM drug" as one of the drugs.
Drinking everyday	▪ "Drinking every day" = 1 if he/she drinks an alcoholic beverage every day.
Smoking	▪"Smoking" = 1 if he/she has a smoking habit.
Not walking regularly	▪"Not walking regularly" = 1 if he/she does not walk or perform equivalent physical activity for at least one hour per day.
Skipping breakfast regularly	▪"Skipping breakfast regularly" = 1 if he/she skips breakfast 3 days or more per week.
Does not want to improve health	▪"Does not try to improve health" = 1 if he/she does not want to improve lifestyle habits, such as exercise and diet.
FBS / LDL-C / BMI / SBP	▪ The values of FBS / LDL-C / BMI / SBP TFigure A3. measured at the checkup a year after the index checkup.