Two Inefficiencies of Self-Selection: Evidence from Health Care*

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: August 2025

Abstract

Self-selection into social programs can lead to socially excessive or insufficient participation. We propose a framework to detect and address these inefficiencies, applying it to diabetes care, where individuals above a biomarker threshold receive nudges to seek care. Crossing the threshold increases healthcare utilization and improves health outcomes. However, those who opt into care—both compliers and beyond-compliers—are generally healthier and benefit less, indicating reverse selection on gains. Targeting based on observable characteristics reduces excessive participation and improves welfare, while outreach to individuals reluctant to seek care despite high potential benefits mitigates insufficient participation and may further raise welfare.

Keywords: Excessive participation; Insufficient participation; Selection on gains; Targeting; Nudges; Regression discontinuity design; Conditional average treatment effect; Diabetes
JEL codes: I18, I13, I11, J13

* The authors thank Naoki Aizawa, Hunt Allcot, David Chan, Kuan-Ming Chen, Karen Eggleston, Henry Fong, Joshua Gottlieb, Koichiro Ito, Hiroyuki Kasahara, Toru Kitagawa, Jonathan Kolstad, Shiko Maruyama, Fabrizio Mazzonna, Magne Mogstad, David Molitor, Shosei Sakaguchi, Lin-Tung Tsai, and participants at American-European Health Economics Study Group IX Edition 2025, ASHEcon 2024, ASSA 2025, HESG Summer 2025, Kyoto Summer Workshop on Applied Economics 2024, 2025 and LSE. We thank Japan Health Insurance Association for providing data and financial support. Kawamura acknowledges financial support from JSPS KAKENHI (24KJ0641), and JST SPRING (JPMJSP2108). Shigeoka acknowledges financial support from JSPS KAKENHI (23H00828, 22H00057, 22H00847, 22H05009, 22K21341), and JST ERATO (JPMJER2301). Shun Takagi provided excellent research assistance. The University of Tokyo's Institutional Review Board (IRB) approved this research (protocol number 21-434). 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

‡ Graduate School of Public Policy, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN. Email: kawamura-junya575@g.ecc.u-tokyo.ac.jp

§ Graduate School of Public Policy, The 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

When individuals self-select into social programs, two inherent inefficiencies may arise: socially *excessive* participation by low-benefit individuals and socially *insufficient* participation by high-need individuals. For example, excessive participation can occur in unemployment insurance if generous benefits reduce job search effort or extend unemployment spells (Schmieder and Von Wachter 2016). Conversely, insufficient participation may arise in welfare programs, such as the Supplemental Nutrition Assistance Program (SNAP) in the United States, due to stigma or complex application procedures (Currie 2006).

This paper proposes a framework that detects and addresses inefficiencies of both types and applies it to health care. Self-selection is a pervasive concern in health care, where patients—often with limited information about their health status and available medical options—decide whether to seek care, particularly for their initial visit. Our specific example is diabetes mellitus (DM) care. Healthcare utilization for DM could be socially excessive due to moral hazard when health insurance covers the cost of care (Arrow 1963). Conversely, DM care could also be socially insufficient when individuals underestimate the benefits of prevention (Baicker et al. 2015).

First, we examine who self-selects into DM care and whether these individuals are those who benefit the most from it. In our setting, individuals receive a nudge encouraging them to seek DM care if a biomarker exceeds the clinical threshold for DM. Those just below and above the threshold are otherwise similar in health status, but only the latter group receives the nudge. Comparing these groups helps identify who opts into care and whether they are the ones who benefit the most. To analyze heterogeneity in both selection and benefits, we employ the generalized random forest (GRF) (Athey et al. 2019), a machine learning method. Unlike conventional methods that divide samples using arbitrary predefined criteria (such as age over 70 or a hypertension diagnosis), this approach systematically detects observable characteristics driving heterogeneity.

The discrepancy between who selects into DM care and who benefits from it, if it exists, indicates the scope for targeting based on observable factors. Therefore, in our second step, we investigate whether policymakers can improve welfare by targeting nudges toward high-benefit individuals. To address socially excessive participation, we refine single-index eligibility rules to better target individuals whose expected benefits from care exceed the associated costs.

We broadly consider two types of targeting based on observable characteristics: *benefit-based* targeting, where we restrict nudges to individuals based on net benefits, and *risk-based*

targeting, where we do so based on conventional risk factors. The first approach aligns with the principle that targeting should be guided by expected benefits rather than by risk factors (e.g., Inoue et al. 2023), while the second offers a more practical approach often used in policy design. Among the benefit-based targeting, the best-case scenario targets only those whose expected net benefit from the nudge is positive. We also consider more feasible benefit-based approaches, including targeting based on a simple linear regression to predict individuals with positive net benefits (Kitagawa and Tetenov, 2018) and a data-driven search for optimal policy rules, referred to as a "policy tree" (Athey and Wager, 2021).

For socially insufficient participation, we examine whether welfare can be improved by reaching out to individuals who currently do not receive DM care. For this analysis, we apply the marginal treatment effect (MTE) framework (Björklund and Moffitt 1987; Heckman and Vytlacil 1999, 2005, 2007), which allows one to analyze the impact of receiving care beyond the group of "compliers." While our early analysis offers insights for targeting and helps address socially excessive participation by narrowing the segment that receives DM care, the MTE framework allows us to examine socially insufficient participation, specifically, whether welfare could increase by encouraging uptake among those who are otherwise reluctant to receive care. This approach informs policymakers whether additional measures, such as stronger nudges or more coercive interventions, could be welfare-improving. We estimate the MTE using an instrumental variable approach, with an FBS $\geq$ 126 mg/dL serving as the binary instrument (Brinch et al. 2017).

Our setting is Japan, where individuals are legally required to undergo annual health screenings. We obtain data from the Japan Health Insurance Association (JHIA), the country's largest health insurer, which covers approximately 40 million members, about one-third of Japan's population. The data consist of individual-level information from 2016 to 2023, including itemized administrative claims (utilization and spending), annual health screening results (biomarker measures), and annual survey information collected at screenings (lifestyle habits such as smoking and alcohol consumption).

Crucially, our ability to estimate the RDD, the heterogeneous effects of the nudge, and the MTEs relies on four key features of the empirical setting; (i) the exceptionally large sample size provides sufficient density near the discontinuity threshold, enabling precise estimation within the RDD framework; (ii) the exogenous variation created by an arbitrary cutoff ensures credible causal identification; (iii) the data contain an unusually rich set of predetermined covariates, including detailed histories of biomarkers and health behaviors; (iv) most importantly, there is substantial heterogeneity in receieving care, systematically related to such observables. This

heterogeneity permits flexible estimation of heterogeneous treatment effects and supports the extension of the propensity score required to identify MTEs across a broader population segment. It is rare for all of these conditions to be met simultaneously in a single setting.

We have four main findings. First, in our preliminary analysis, we apply a regression discontinuity design (RDD) to a diabetes diagnosis threshold—a fasting blood sugar (FBS) level of 126 mg/dl—to estimate the average effects of the nudge at the threshold. We find that nudges significantly increase DM care and healthcare spending. For instance, the probability of a DM-related outpatient visit within a year rises by 8.8 ppt (from 32.5% to 41.3%), a 27.2% increase, and DM-related outpatient spending increases by $35.3 (15.5%). Nudges also modestly improve health outcomes: the following year's FBS falls by 0.71 mg/dl (0.57%) and LDL cholesterol (LDL-C) by 1.59 mg/dl (1.20%) on average. All of these effects are precisely estimated thanks to the large dataset. Overall, nudges promote DM care and yield health improvements *on average*. While this analysis identifies the average effect of the nudges at the threshold, it does not address our primary interest in heterogeneity in who selects into DM care and who benefits from it. Nor does it reveal discrepancies between care receivers and beneficiaries.

Second, our analysis of who opts in and out of DM care, estimated using GRF, reveals that individuals who are most likely to self-select into DM care tend to benefit the least from receiving it, indicating *reverse* selection on gains—that is, selection on slopes in the sense of Einav et al. (2013). Interestingly, individuals who receive care tend to have better health—for example, lower triglycerides, BMI, diastolic and systolic blood pressures—and healthier lifestyles, such as not smoking or drinking and walking regularly. In contrast, those who experience greater health improvements from DM care tend to have poorer initial health, including worse LDL/HDL cholesterol levels, and exhibit less healthy behaviors, such as smoking, not walking regularly, and showing reluctance to change lifestyle habits—traits that distinguish them from those who self-select into DM care.

Third, we find that targeting nudges to individuals whose predicted benefits exceed the costs substantially improves welfare and mitigates socially excessive participation. In the first best-case scenario—where only those with an estimated positive *net* benefit receive a nudge—the net benefit per person increases by 180% compared to the status quo. Targeting based on a simpler linear prediction model or policy trees also yields welfare gains, though to a lesser extent. In contrast, targeting nudges based on conventional risk factors, such as elevated blood sugar or high blood pressure, does not necessarily improve welfare. This may happen because these risk factors and their thresholds do not necessarily identify individuals who would generate the greatest net benefit from follow-up care. Overall, these findings suggest that adopting a more

individualized or subgroup-specific targeting approach offers clear advantages over a "one-size-fits-all" approach when seeking to maximize the welfare impact of health interventions.

Fourth, our MTE analysis provides evidence of reverse selection on gains even beyond the group of compliers, indicating the possibility of socially insufficient participation. Specifically, the MTE curve for health improvement is *upward* sloping with respect to unobserved treatment resistance, implying that individuals who are least likely to receive medical care benefit the most from doing so. This result suggests that further encouraging these individuals to seek care—through stronger nudges or more intensive outreach—could be potentially welfare-improving, provided that the additional benefits outweigh the costs. This finding is also consistent with the result for compliers, which likewise reveals reverse selection on gains based on observable characteristics.

This paper demonstrates that substantial self-selection and associated inefficiencies exist in voluntary social programs, whereby some low-benefit individuals opt in (excessive participation) while some high-benefit individuals opt out (insufficient participation). Using DM care as an example, we develop a framework to assess both types of inefficiencies and address them. Specifically, we refine eligibility rules to reduce inefficiency arising from excessive participation and explore outreach strategies to reach high-benefit individuals who currently forgo care, thereby mitigating inefficiency from insufficient participation.

Our paper contributes to several strands of literature. Broadly speaking, our discussion of excessive participation is related to the literature on selection heterogeneity in program take-up and treatment effect heterogeneity, as well as the use of machine learning for policy targeting. Conversely, our analysis of insufficient participation is linked to the literature on MTE.

Specifically, first, it adds to the growing body of research on selection heterogeneity in program take-up (e.g., Finkelstein and Notowidigdo (2019) on SNAP; Deshpande and Li (2019) on Disability Insurance; Jones et al. (2019) on workplace wellness programs; Einav et al. (2020) and Kowalski (2023) on mammography screening). These studies provide evidence of heterogeneity in selection into programs and underscore the importance of accounting for heterogeneous selections when designing screening policies. Leveraging a large sample size and access to rich individual characteristics that affect selection and ultimate health outcomes, we can examine not only heterogeneity in selection but also heterogeneity in treatment effects. This dual analysis enables us to assess potential misalignment between selection and benefits and to evaluate targeting strategies aimed at improving overall welfare.

Second, this paper contributes to the literature on treatment effect heterogeneity and the use of machine learning for targeting policies (e.g., Knaus et al. (2022) on job search programs;

Johnson et al. (2023) on inspection; Burlat (2024) on training programs). Most of these cases involve perfect compliance, eliminating the need to account for selection heterogeneity. In contrast, our setting involves voluntary participation, introducing noncompliance. We explore the potential for targeting when treatment assignment (i.e., who receives a nudge) can be controlled, but treatment uptake (i.e., who visits a hospital and receives care) cannot. Moreover, we extend the analysis of treatment effect heterogeneity to encompass both benefits (improvements in health outcomes) and costs (increased healthcare utilization), a dimension often overlooked in existing studies. In many prior applications, such as training programs (Burlat 2024), the cost of treatment assignment is negligible or relatively homogeneous, allowing for a focus solely on benefits. Our context highlights the importance of jointly considering heterogeneous benefits and costs when designing targeted interventions.

Third, this paper relates to work that applies the MTE framework to extrapolate beyond LATEs (e.g., Heckman and Vytlacil 2005; Cornelissen et al. 2016; Kline and Walters 2016; Brinch et al. 2017). Health care is an interesting context for this approach, not only because patients commonly self-select into care, but also because the sign of selection on gains is ambiguous a priori. On the one hand, individuals who anticipate high benefits from care may be more likely to visit physicians, improving their health outcomes (positive selection on gains). On the other hand, individuals facing high unobserved costs of seeking care, such as the high cost of caring for their family, may be less likely to visit but could benefit more if they did (negative selection on gains). The direction of selection on gains has distinct welfare implications, and the MTE approach provides this important information. In our setting, we find negative selection on gains, indicating that stronger interventions encouraging hesitant individuals to seek care could improve welfare.

Finally, we contribute to the literature examining the causal impact of health signals—such as 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 causal effects of health signals by exploiting arbitrary biomarker thresholds.[1] In contrast to approaches that rely on a limited set of predefined criteria to divide the sample, we systematically identify observable characteristics driving heterogeneity in both selection and treatment effects using a machine learning framework. Moreover, we propose a framework to address inefficiencies arising from excessive and insufficient participation due to selection.

[1] Some studies examine supply-side behaviors using these biomarker thresholds instead of demand-side behaviors (Almond et al. 2010; Alalouf et al. 2024).

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 screenings in Japan, the key threshold value for DM diagnosis, and the contents of nudges). Section 3 describes the data, and Section 4 presents our empirical strategy. Section 5 reports our baseline graphical and econometric results, and Section 6 concludes.

2. Background

2.1. Annual mandatory screening in Japan

Workplace annual health screenings 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).[2] While employees are also required to participate in these screenings yearly, there are no legal penalties for individuals who do not comply. However, employers can take disciplinary action for non-compliance. Over 95% of employees in large corporations meet their annual checkup requirements. The rate is expected to be lower for smaller corporations, including ours, although an accurate estimate is unavailable.

The health screening aims to facilitate the early detection of potential health conditions and provide patients with health reports and guidance. By communicating the results, the screening 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 screening 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). The screening service, lab tests, and reporting are typically outsourced to healthcare providers.

2.2. Report card and follow-up nudge

Figure 1 outlines the health screening process. After the initial screening at year t, all 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 2-A shows an example of a report card. If a value exceeds a clinical threshold, the report

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

includes a warning (e.g., "*" in this example, or 'H' for high or a grade from 'A' to 'E,' with 'E' indicating worse conditions) and advises consulting a physician (Iizuka et al. 2021). In this paper, we focus on DM, and the Japan Diabetes Society determines that individuals with FBS greater than or equal to 126 mg/dl are diabetic type. Thus, those above the threshold are likely to receive a warning, although there might be variation depending on the provider.

In month 6, Japan's largest insurer, the Japan Health Insurance Association ("Kyokai Kenpo"), which covers approximately 40 million people, sends a follow-up document, or a reminder, with a warning urging selected individuals to consult a physician. Selection is based on two criteria: exceeding the diabetic-type threshold (FBS ≥126) and not currently receiving diabetes treatment. For the latter criterion, those who (i) were taking hypertension or diabetes medication (self-reported) at the initial year's screening and (ii) visited a physician for diabetes within one month before and three months after the initial screening, as observed in the claims data, are excluded. Figure 2-B shows an example of such a reminder. While the wording varies slightly over the years, the core message remains the same. This 2018 example states, "If left untreated, this condition may lead to arteriosclerosis, raising the risk of heart and brain diseases and harming your health."

We then observe their biomarkers at the year t+1 screening for returning participants and healthcare utilization between screenings. As discussed, individuals are encouraged to see a physician if their FBS is greater than or equal to 126, either through a report card, a reminder, or both. However, since health outcomes are observed only once at the annual screening, we cannot isolate the effects of report cards and the reminder, so we refer to them collectively as "nudges" unless stated otherwise.

Crucially, the decision to pursue medical care after receiving nudges rests solely with the individual since employers and health insurers cannot require physician visits. Additionally, there is no gatekeeping by general practitioners (GPs) in Japan, and hence individuals are not tied to a specific primary care physician. Consequently, primary care physicians do not monitor biomarkers or encourage patients to visit 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 screenings. Therefore, it is up to the patient to decide whether to consult a healthcare provider upon receiving nudges. On the other hand, whether and how to treat the patient, once they visit physicians, is left to the physician's judgment. Universal health insurance covers visit costs on a fee-for-service basis, with patients responsible for 30% of the

expenses out-of-pocket for most cases.[3]

2.3. Treatment of Diabetes Mellitus

In this paper, we focus on nudges sent to patients with diabetes mellitus (DM).

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 can be identified by elevated FBS levels, with a clinical threshold of 126 mg/dL commonly used for diagnosis.

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 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 who are covered by a separate national health insurance scheme.

The dataset includes approximately 40 million individuals—roughly one-third of Japan's 120 million population. It covers from April 2016 to March 2023 and comprises around 48 million screening-year records, with the analysis restricted to employees. It comprises three components: medical claims, health screening results, and health survey responses, all linked by anonymized patient ID. Medical claims data, collected monthly, allow for the tracking of

[3] 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.

individuals' healthcare utilization. Health outcomes are derived from health screenings, which are typically conducted annually, along with information on the year and month of the screening. The health survey, administered during screenings, gathers information on self-reported health behaviors and disease history. Additionally, demographic data from the enrollment registry include gender, birth date, monthly wage category, and, if applicable, death dates.

3.2. Outcome variables

We examine 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 separates costs incurred for DM-related medications.[4] These variables are derived by aggregating 12 months of claims data post-screening. A visit is classified as DM-related if diabetes (E10-E14 in ICD-10) is listed as a diagnosis.[5] DM-related drug spending includes spending for antidiabetic agents (Therapeutic Category Code or TCC, 396), pancreatic hormone preparations (TCC 2492), and GLP-1 receptor agonists (part of TCC 2499).

Health behaviors—. We create four dummy variables for (a) drinking daily, (b) smoking, (c) not walking regularly, and (d) skipping breakfast regularly. These variables are derived from self-reported health surveys conducted during health screenings.

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 screening in the next year.

3.3. Analysis sample

The analysis sample is constructed using the following criteria: (i) individuals must have data available from both the index screening and the subsequent year's screening; (ii) individuals taking either hypertension or diabetes medications at the time of the index screening are

[4] We truncate spending records with values above 99.9 percentiles of spending-related outcomes to avoid the influence of outliers.
[5] In Japan, physicians are required to document all medical conditions (including suspected ones) related to a patient's visit using ICD-10 codes, which are available in our dataset.

excluded, as they are not eligible for reminders[6]; (iii) individuals previously diagnosed with diabetes mellitus (DM) before the index screening are excluded to focus on those not under treatment for DM at the time[7]; and (iv) individuals must have healthcare utilization data covering at least one month prior to and twelve months following the index screening.

Table A1 provides summary statistics of the analysis sample. Figure A2 plots the distribution of FBS values of the analysis 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 the 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).

4. Empirical strategy

4.1. Regression discontinuity design (RDD)

We start our analysis by estimating the average effect of the nudge on utilization, health behaviors, and health outcomes. We exploit that the exact screening 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 a regression discontinuity design (RDD). However, only those whose values cross the threshold will receive a health signal, and this discontinuity identifies the effect of nudges 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 (Iizuka et al. 2021).

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

We estimate the following local linear regression:

[6] 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.
[7] Since those already taking diabetes medications are already excluded, this sample criteria does little to the further sample selection (1.48%).

$$Y_{it+1} = \alpha_0 + \alpha_1 Z_{it} + \alpha_2 R_{it} + \alpha_3 Z_{it} R_{it} + \alpha_4 X_{it} + A_y + \varepsilon_{it}, \quad [1]$$

where Y_{it+1} represents healthcare utilization in $t+1$ for individual i (as for health behaviors and health outcomes, we replace Y_{it+1} with $\Delta Y_{t+1} (= Y_{it+1} - Y_{it})$ which represents the change between next year and this year).[8] $Z_{it} = 1(FBS_{it} \geq 126)$ is the "nudge" dummy which takes the value 1 if individual i's FBS in year t is greater than or equal to 126 mg/dl, and zero otherwise. $R_{it} = FBS_{it} - 126$ is the running variable centered around 126, and thus the next two terms $(R_{it}, Z_{it} R_{it})$ 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, and thus our results quantitatively change little with the covariates. A_y is the index screening year fixed effect (from 2016 to 2022). ε_{it} is the error term. α_1 is our coefficient of interest (intention-to-treat (ITT) or reduced form estimates).

We also estimate the same model as [1], where an outcome is replaced by D_{it+1} which takes one if there is any DM-related visit within a year after the index screening. Taking the ratio of the two coefficients provides us with the local average treatment effects (LATE) or the average treatment effects for compliers.

We estimate local-linear regressions with a triangular kernel and use the optimal bandwidth that minimizes the mean squared error or MSE (Calonico et al. 2019). The standard errors are estimated with the robust 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, screening participants, physicians, and staff typically lack immediate access to blood sugar data, making it improbable for them to alter the measurements reliably. Moreover, our study focuses on individuals without a prior DM diagnosis, who are often unaware of their blood sugar levels and have no incentive to manipulate them. As expected, Figure A2 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 that covariates remain continuous at the threshold. To test this, Figure A3 plots average values of

[8] 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 achieve a slight efficiency gain of roughly 10-20%.

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 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 screening the following year. To address potential selection bias, Figure A4 plots the probability of attending the next year's screening. Attendance exceeds 90%, with no significant discontinuity at the thresholds, suggesting attrition is unrelated to receiving nudges. The RD estimate from a local-linear regression with a triangular kernel, using a dummy for year $t+1$ screening attendance as the outcome, is –0.00355 (SE= 0.00405, p-value= 0.407), indicating that attrition is unrelated to whether one receives a nudge.

4.2. Heterogeneous Effects of Nudges

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 the effect of the nudge may be masked. In particular, we are interested in whether those who "respond" the most with nudges and receive medical care (heterogeneity in selections) are the ones who "benefit" the most from receiving care (heterogeneity in benefits or treatment effects). If not, there may be a scope for targeting.

Estimating Heterogeneity—. To explore heterogeneity in selections and benefits, 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), which allows researchers to estimate heterogeneous treatment effects based on high-dimensional nonlinear functions of observable characteristics by incorporating identification restrictions corresponding to the RDD. Also, instead of arbitrarily splitting the sample by covariates as conventional studies perform in their heterogeneity analysis, the algorithm splits the sample repeatedly across multiple dimensions to maximize treatment effect heterogeneity between subsamples (leaves).

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

$$\left(\hat{\tau}, \hat{\theta}\right) \in \arg\min_{\tau,\theta} \quad \left\| \sum_i \alpha_i(x)\, \Psi_{\tau,\theta}(Y_i, Z_i, R_i) \right\|_2, \quad [2]$$

where $\alpha_i(x)$ is a "similarity weight," $\Psi_{\tau,\theta}(Y_i, Z_i, R_i)$ is a score function, where Y_i, Z_i, R_i are the

same as before, 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, past health behaviors, 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 heterogeneous effects of nudges. In particular, we estimate the following local linear model, called the linear model (LM) forest:[9]

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

In this case, the score function, $\Psi_{\tau,\theta}(Y_i, Z_i, R_i)$, in equation [2] is defined as the following:

$$\Psi_{\tau,\theta}(Y_i, Z_i, R_i) = (Y_i - c(x) - \tau(x)Z_i - \beta(x)R_i)\begin{bmatrix}1\\Z_i\\R_i\end{bmatrix}$$

Then, predictions from multiple splits (trees) are averaged to obtain the CATE. When aggregating trees, we employ a sample-splitting strategy known as "honesty" to reduce the risk of over-fitting (Wager and Athey 2018). This approach ensures that the data used to determine the structure of each tree is separate from the data used to estimate treatment effects. Under appropriate regularity conditions, this honest estimation procedure guarantees that the resulting estimates are consistent and asymptotically normal, facilitating valid statistical inference. In addition, we also use 10-fold cross-fitting to estimate CATEs. For each fold k, we train the GRF model on the observations outside fold k, and generate CATE predictions for the observations in fold k. We restrict the sample within 5 mg/dl of 126 thresholds to adjust to the RDD setting. The minimum node size of 1,000 is used.

We estimate CATE separately for Y_{it+1} or D_{it+1} as the outcome, in the above equation. Following Athey et al. (2025a), we define two CATEs (again, we omit t or t+1 for brevity):

$$\tau_D(x) = E(D_i|X_i = x, Z_i = 1) - E(D_i|X_i = x, Z_i = 0)$$
$$\tau_Y(x) = E(Y_i|X_i = x, Z_i = 1) - E(Y_i|X_i = x, Z_i = 0)$$

where $\tau_D(x)$ is the conditional response (compliance) rate or we refer to CATE D which is the fraction of the subpopulation with $X_i = x$ who receive care if the nudge is offered, and $\tau_Y(x)$ is the conditional ITT, or we refer to CATE Y, which is the expected change in health outcome

[9] The LM Forest is part of the *grf* package.

from offering the nudges to individuals with $X_i = x$.[10] From Imbens and Angrist (1994), we can define the Conditional Local Average Treatment Effect (CLATE) as:

$$\tau_{CL}(x) = (Y_{i1} - Y_{i0} | D_{i1} - D_{i0} > 0, X_i = x)$$
$$= \tau_Y(x)/\tau_D(x)$$

which is the local average treatment effect (LATE) for the subpopulation with $X_i = x$, that is, the change in health outcomes from receiving the treatment for the subpopulation with $X_i = x$ who receive the treatment only if $Z_i = 1$ (i.e., compliers).

The distinction in $\tau_Y(x)$ and $\tau_{CL}(x)$ is crucial when we discuss the scope for targeting in the next subsection, since we assume that policymakers can *only* change Z_i, i.e., who to nudge and cannot directly change who receive care D_i. Thus, we mainly focus on $\tau_D(x)$ and $\tau_Y(x)$—the causal effect of treatment *assignment* (i.e., nudges) rather than $\tau_{CL}(x)$—the causal effect of *receiving* the treatment (i.e., receiving care) unless specified otherwise.

Assessing Heterogeneity—. One concern of CATE analysis is that predicted treatment effects may reflect noise or sampling variation, as opposed to true heterogeneity. To address such a concern, we take several measures to ensure that our individual treatment effects capture genuine heterogeneity. First, we implement the Sorted Group Average Treatment Effect (GATES) method developed in Chernozhukov et al. (2018). Specifically, in each fold of the cross-fitting procedure, we rank individuals into groups based on the quantiles of their predicted CATEs, assigning higher ranks to individuals with higher predicted treatment effects. We divide the sample into four groups. Within each group, we compute the Augmented Inverse Probability Weighting (AIPW) scores using the predicted CATEs and estimate the Group Average Treatment Effect (GATE) as the weighted average of these scores. If the model is well-calibrated, we expect a monotonic relationship in which higher-ranked groups exhibit larger average treatment effects. Second, we separately estimate RDD using observed outcomes, such as the improvement in FBS, for each of the four groups defined by the quartile of predicted CATE, and check if the patterns of the RDD estimates obtained from actual data align well with those from GATE, which is based on predicted treatment effects from CATE. Finally, we implement the rank-weighted average treatment effects (RATE) method proposed by Yadlowsky et al. (2024) as an additional check to evaluate and confirm the existence of heterogeneity.[11]

[10] Athey et al. (2025a) refer to CATE D as "compliance CATE" and CATE Y as "outcome CATE" respectively.

[11] Specifically, we begin by constructing the Targeting Operator Characteristic (TOC) curve. For each proportion q, the TOC curve plots the average outcome gain among the top q fraction of units ranked by their estimated CATEs relative to the overall ATE. The RATE is then defined as the area between the TOC curve and the horizontal line at the ATE. A positive and statistically significant RATE implies the presence of treatment effect

4.3. Targeting based on observables

If heterogeneity exists in both who opt into care and the benefits derived from receiving it—and, more importantly, if a missalignment between the two is observed—then refining the existing criteria for sending nudges may lead to improved targeting and greater overall welfare.

To consider targeting, we must quantify the monetary values of both the benefits and costs. While our primary measure of health improvement is FBS, it cannot be directly translated into monetary terms. Therefore, to assess benefits, we examine the reduction in mortality over the six years following the index screening—the maximum follow-up period available in our data. Over the six-year observation period, we recorded 2,724 deaths among 208,848 individuals within ±5 mg/dL of the 126 mg/dL threshold, yielding a cumulative six-year mortality rate of 1.30%. Unlike improvements in biomarkers such as FBS, changes in mortality can be monetized using the value of a statistical life (VSL). We adopt a VSL of $50,000 per life-year. On the cost side, we consider cumulative DM-related outpatient spending over the same six-year period, applying the same annual discount rate of 3% to both benefits and costs.

We broadly consider two types of targeting or restricting the eligibility for receiving the nudges: *benefit-based* targeting, where we restrict the individuals who receive the nudge based on net benefit from targeting, and *risk-based* targeting, where we restrict the individuals who receive the nudge based on conventional risk factors, in addition to high blood sugar (i.e., FBS ≥ 126).

Benefit-based targeting—. We consider several scenarios of targeting based on the net benefit. We start with the benchmark case where the insurer targets individuals whose predicted benefits from nudges exceed their costs based on the CATE estimates. This scenario may represent the optimal case regarding cost-effectiveness, which aligns with the notion that targeting should be based on benefits instead of risks. Second, in the spirit of Kitagawa and Tetenov (2018), we estimate a linear model of net benefits on observables and target those with positive predicted values. This approach offers a more practical alternative for policymakers, such as municipal governments, who may lack the capacity to estimate CATE and hence rule out the complex non-linear policy rules, but can apply coefficients to observable characteristics to guide targeting.

Lastly, rather than manually identifying target subgroups from GRF prediction results, we adopt a data-driven approach—called policy trees (Athey and Wager 2021)—to derive optimal

heterogeneity (e.g., Athey et al. 2025b).

targeting rules that maximize population-level net monetary benefit based on individual CATE estimates. The policy tree algorithm finds the optimal treatment assignment rule as a decision tree of a pre-specified depth (Zhou et al. 2023).[12] Specifically, the policy trees provide estimates of the optimal cut points along the individual observable characteristics, using the individual-level estimates of net benefit from the GRF. Then, the average welfare gains per individual member under the suggested targeted policy, as opposed to nudging all individuals above the threshold (e.g., treatment as usual), can be estimated.

Risk-based targeting—. As an alternative to benefit-based targeting, we consider targeting if one meets additional risk factors. Specifically, we assume that health insurers nudge individuals if their blood pressure and/or cholesterol are high in addition to high blood sugar (i.e., FBS ≥ 126). We use the thresholds suggested by the Ministry of Health, Labour, and Welfare to determine high blood pressure (SBP ≥140 mmHg or DBP ≥ 90) and high cholesterol (LDL-C ≥ 140 or triglycerides ≥ 300).[13]

4.4. Marginal treatment effect (MTE)

While the analysis of CATE provides valuable insights for targeting based on observable characteristics, it does not reveal whether the potential gains from receving medical care are greater or smaller among individuals beyond the group of compliers.

To address this, we implement the marginal treatment effect (MTE) framework (Björklund and Moffitt 1987; Heckman and Vytlacil 1999, 2005, 2007) that extends analysis beyond "compliers" by connecting the heterogeneity in the treatment effect to both observed and unobserved variation in the propensity to receive care. While exceeding the FBS threshold increases the likelihood of receiving care, as reported in Figure 3A, "compliers" consist of a relatively small share, and the majority of individuals who receive nudges do not comply with them. The MTE analysis enables us to examine whether reaching out to those reluctant individuals is worthwhile.

Setup—. Consider an individual who decides whether to receive care ($j = 1$) or not ($j = 0$). A binary variable D equals 1 if one receives care, and 0 otherwise.[14] The observed health outcome (Y) can be linked to potential outcomes (Y_0, Y_1) as

$$Y = DY_1 + (1 - D)Y_0$$

[12] We use the R-package *policytree* provided by Sverdrup et al. (2020).
[13] Source: https://www.mhlw.go.jp/bunya/kenkou/seikatsu/pdf/ikk-a05.pdf (Accessed April 28, 2025; in Japanese)
[14] We omit an individual i and time t or t+1 for brevity in this subsection.

where $Y_j = \mu_j(X) + U_j$ for $j \in \{0,1\}$, in which $\mu_j(\cdot)$ is an unspecified function of observables X, and U_j are unobservables with $E[U_j|X = x] = 0$ for $j \in \{0,1\}$.

The individual's net benefit of receiving care depends on observed variables (X, Z) and an unobserved continuous variable (U_D):

$$D = 1[D^* \geq 0], D^* = \mu_D(X, Z) - U_D$$

where $\mu_D(\cdot)$ is an unspecified function, and Z is the excluded instrument (i.e., $1(FBS_{it} \geq 126)$).

With the normalization of $U_D \sim Unif[0,1]$, where higher values close to 1 represents greater unobserved resistance to treatment (i.e., less likely to visit physicians and receive care for unobservable reasons), $\mu_D(X, Z)$ can be interpreted as the propensity score as:

$$P(X, Z) \equiv \Pr(D = 1|X, Z) = \mu_D(X, Z)$$

so that $D = 1$ if $P(X, Z) > U_D$.

With the assumption of conditional independence $(U_1, U_0, U_D) \perp Z|X$, MTE is defined as

$$\begin{aligned} MTE(x, p) &\equiv E(Y_1 - Y_0|X = x, U_D = p) \\ &= \mu_1(x) - \mu_0(x) + E(U_1 - U_0|X = x, U_D = p) \\ &= \mu(x) + k(x, p) \end{aligned}$$

where the first and second term represents heterogeneity in observable and unobservables.

Binary instrument—. In our setting, a binary indicator that takes on a value of one if FBS is equalt to or above 126 serves as the instrumental variable.[15] The challenge with a binary instrument is that $P(X, Z)$ only takes two values given X, which precludes the use of a conventional local instrumental variables (LIV) approach to estimate MTE (e.g., Heckman and Vytlacil 1999). However, Brinch et al. (2017) show that with a separability assumption[16], MTE is additively separable on X and U_D as

$$MTE(x, p) = \mu_1(x) - \mu_0(x) + E(U_1 - U_0|U_D = p),$$

so that we can estimate a flexible MTE model by separately estimating the following two equations (called a separate estimation approach)

$$\begin{aligned} E(Y|X = x, P(Z) = p, D = 1) &= \mu_1(x) + E(U_1|X = x, U_D = p) \\ E(Y|X = x, P(Z) = p, D = 0) &= \mu_0(x) + E(U_0|X = x, U_D = p) \end{aligned}$$

where MTE is simply the difference between these two.

The advantage of this estimation approach using a binary instrument is that for each value of X, we now have two values of $P(X, Z)$ for the treated (always-takers vs. combined group of always-takers and compliers) and two values of $P(Z)$ for the untreated (never-takers vs.

[15] Other applications of this binary IV to estimate MTE include Ito et al. (2019) and Khanna (2023).
[16] Technically, $E(Y_j|U_D, X = x) = \mu_j(x) + E(U_j|U_D)$ for $j \in \{0,1\}$.

combined group of never-takers and compliers). As shown in CATE analysis (Panel A of Figure 5), the effects of discrete instrument (Z) on treatment takeup (D) are heterogeneous among individuals with different values of observables X, indicating that variation in Z creates many values of $P(X,Z)$ in the range of $U_D[0,1]$. We can exploit this variation in $P(X,Z) = p$ to estimate the above equations with flexible polynomials in p or semiparametrically over a wide range of U_D that is covered by the common support of $P(X,Z|D = 1)$ and $P(X,Z|D = 0)$. For estimation, we follow the two-step approach of Brinch et al. (2017). See Brinch et al. (2017) for details.[17]

5. Results

5.1. Average effects from RDD

This section presents the baseline results of estimating the RD specification. We examine whether nudges, *on average*, 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 3-A and 3-B plot the probability of at least one DM-related physician visit and DM-related outpatient spending within a year of the index screening. 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 increases by $35.3 (15.5% from the mean of $228.5) and DM-related drug spending also increases by $5.7 (21.5% from the mean of $26.7) at the cutoff. Thanks to the large dataset, these effects are precisely estimated. These findings suggest that nudges significantly boost utilization.

So far, we have aggregated 12 months of post-screening claims data. The monthly data further allows us to assess whether the reminder with a warning at the 6th month has any additional effect on the utilization on top of the report card. The outcome is a cumulative DM visit dummy, indicating at least one DM visit by month m. Figure 4-A plots this for m = –1, 0, 1, 5, 6, 7, and 8, where m = –1 is before the screening, and m = 0 is the screening month. Figure 4-B presents the corresponding RD estimates from equation [1] for each month from –1 to 12.

[17] We use the Stata package *mtefe* provided by Andresen (2018).

The estimates in Figure 4-B at $m = -1$ and $m = 0$ are close to zero because individuals have not yet received a report card. In month 1, when one is supposed to receive the report cards, the probability of a DM visit increases by 2.5 ppt and rises steadily before leveling off around month 4. A sudden jump of 2.1 ppt (from 5.2% to 7.3%) occurs at $m = 6$, when the reminder with warning is sent, reflecting the additional effect of the reminder. It finally reaches 8.8 ppt increase in one year (as reported in Table 1-A), suggesting that the 3.6 ppt (=8.8–5.2) increase is due to the reminders. Thus, about 40% (=3.6/8.8) of the nudge effects come from the reminders, and the remaining 60% (=5.2/8.8) from the report card.

Health behaviors—. Figure 3-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 increase 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 increase of 1.0 ppt from a 62.0% mean just below the cutoff. Therefore, we conclude that behavioral changes in response to nudges are minimal.

Health outcomes—. Figure 3-C shows the mean change in FBS and LDL-C between years $t+1$ and t, again with higher values indicating *better* health outcomes (e.g., $(-)\Delta$FBS). 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 reflect the intention-to-treat (ITT) effects. The effect of visiting a physician and receiving care on health outcome is much larger for compliers—who visited physicians because they received the health signal—given the first-stage magnitude of about nine percentage points. Using the threshold as an excluded instrument for a dummy for any DM-related visits, Table 1-D reports the local average treatment effect (LATE) on the same health outcomes. The improvements in FBS and LCL-C, for example, are 10.2 mg/dl (8.2% reduction) and 18.4 mg/dl (13.9% reduction) for compliers, which are sizeable. In terms of the mechanism by which health outcomes improved, as we have shown above, the increase in healthcare utilization, rather than changes in health behaviors, appears to be the main contributor.

5.2. Heterogeneous effects of nudges

This section presents the results from CATEs to explore the heterogeneity in who selects

into receiving medical care and who benefits from doing so. The supplementary results can be found in Appendix Section B.

Heterogeneity in selections or CATE D—. The top row of Figure 5-A shows the distribution of estimated CATEs, $\tau_D(x)$, of receiving nudges on making at least one DM-related visit within a year of the index screening. The mode is around eight percentage points (ppt), similar to the mean treatment effect (8.8 ppt) from the RDD. The figure highlights that nudge effects on utilization are mostly positive and heterogeneous, with a 95% confidence interval between 1.5 and 15.3 ppt. The middle row of Figure 5-A categorizes CATEs into four groups by applying Chernozhukov et al. (2019) or GATE. We find significant heterogeneity, as indicated by a monotonic relationship where higher-ranked groups exhibit larger average treatment effects and non-overlap of the estimates' confidence intervals. Figure A5 and Table A3 (both in panels A) confirm that the patterns of the actual RDD estimates from each group are very similar to those of GATE. Finally, Figure B1 also presents the RATE (Yadlowsky et al. 2024) as an additional check to confirm the existence of heterogeneity. Consistent with the TOC curve, the RATE—equivalently, the area under the TOC—is positive and statistically significant at the 5 % level, providing additional evidence of heterogeneity in responses to nudges.

Next, we explore the characteristics of individuals who are more likely to respond to nudges (called the classification test). The bottom row of Figure 5-A, the heat map, compares characteristics of high-response (Q4) and low-response (Q1) groups. The lighter color represents better health outcomes and better lifestyle habits, except for age, gender, and wages. We find that the high-response group (Q4) tends to be healthier, with lower LDL-C, SBP, and BMI, and exhibits better lifestyle habits, such as not currently smoking, not drinking daily, and walking regularly. These findings suggest that healthier individuals with better health and better lifestyle habits—whose health return from receiving care is likely small—are more likely to respond to nudges (selection on *levels*).[18]

Heterogeneity in benefits or CATE Y—. We next examine the heterogeneous effects of nudges on health outcomes, $\tau_Y(x)$, in Figure 5-B, using the same approach but focusing on health outcomes one year after the index screening. The top row shows the distribution of health improvements regarding the (negative) change in FBS, redefined so that positive values indicate improvement. The mode is around 1.02, similar to the RD estimate (approximately 0.93), with a

[18] 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.

95% confidence interval ranging between 0.05 and 9.03, highlighting heterogeneity. The middle row groups individuals by effect size and tests whether their group average effects differ from zero. The impact of nudges on $(-)\Delta$FBS is heterogeneous, with a significant difference in the effect between groups 1 and 4.[19] 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 $(-)\Delta$FBS. In sharp contrast to receipt of medical care (D), the high-benefit group generally has worse health and poorer lifestyle habits (darker color).

Discrepancy between selections and benefits—. Comparing the heat maps reported at the bottom of Figures 5-A and 5-B reveals that high-response individuals (Q4 in Figure 5-A) tend to be healthier, while those benefiting most from physician visits (Q4 in Figure 5-B) tend to be sicker. This result suggests a discrepancy between who responds to nudges and who benefits from nudges. To see this relationship more clearly, Figure 6 combines the two CATE estimates and presents a binned scatter plot showing the relationship between CATE for selections D (x-axis) and the ratio of CATE Y and CATE D (y-axis), which can be interpreted as the CLATE (Athey et al. 2025a), capturing health improvements among compliers at the individual level.[20] The negative slope indicates that individuals who are more likely to receive care tend to experience smaller health improvements.[21] This figure confirms that those who respond the most to nudges tend to benefit the least from visiting physicians and receiving medical care, indicating *reverse* selection on gains based on observables (selection on *slopes*).

In sum, our results indicate important heterogeneity in who responds to nudges and visits physicians (selection heterogeneity) and who benefits from doing so (treatment effect heterogeneity). More importantly, there is a misalignment between the two effects; those who respond to nudges are not those who benefit most from nudge-induced physician visits—inefficiency from excessive participation.

5.3. Targeting

Nudges in healthcare are generally costly, as they tend to increase physician visits and

[19] See Figure A5 and Table A3 (both panels B) for the actual RDD estimates from each group and RATE as an additional check to confirm the existence of heterogeneity.

[20] Another approach to estimating CLATE directly, using IV forest (the part of the GRF package), currently cannot include covariates, including the slope of the running variable, which undermines the virtue of the RD setting.

[21] 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, Figure B4 also observes discrepancies between the types of individuals who are more likely to receive care and those who are more likely to benefit from them.

overall healthcare expenditures. In some cases—such as cancer screening—they may even cause harm if nudge-induced visits lead to unnecessary tests or treatments. These potential downsides highlight the room for efficiency improvement by adequately targeting nudges toward individuals for whom the net benefits of physician visits are higher.

Figure B5 displays the CATEs on 6-year cumulative mortality. We observe reduced cumulative mortality among individuals with the highest benefits from nudge (Q4).[22] We then convert this into monetary benefits in Figure 7-A, using the VSL of $50,000 per year and an annual discount rate of 3%. Figure 7-B confirms substantial heterogeneity in the impact of nudge on cost as well. Figure 7-C plots the benefits from Figure 7-A on the *x*-axis and costs from Figure 7-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 7-D displays the distribution of such net benefits, 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, stamp, and printing) is negligible. For reference, Row (1) reports the results of the status quo, where all individuals whose FBS is equal to or greater than 126 received the nudge. Below, we restrict the individuals who will receive the nudge based on net benefit (benefit-based targeting) in Rows (2)-(5) and based on conventional risk factors (risk-based targeting) in Rows (6)-(8).

Benefit-based targeting—. In the first best-case scenario, where we nudge the positive net benefit group only (Row (2): "Non-parametric"), the fraction nudged decreases to 57% and the net benefit per person rose by 180% compared to the status quo of providing nudges to all individuals with an FBS of 126 or higher. This strategy increases welfare because those whose costs from nudges outweigh the benefits—those above the 45-degree line in Figure 7-C—will not receive a nudge. For the second scenario (Row (3): "Linear prediction"), the fraction nudged decreases to 62%, while the net benefit increases by 76%. This result indicates that using a simpler linear prediction model for targeting still improves welfare relative to the status quo, but the gains are smaller compared to the non-parametric model, as expected.

[22] Figure B6 also shows a negative relationship between CATE for responses (*x*-axis) and CATE for mortality improvement (*y*-axis), suggesting that individuals who are more likely to receive care experience smaller mortality improvements.

In the third scenario, we implement the policy learning approach of depth one and two policies with 10-fold cross-fitting. In both depth one and depth two, the choice and order of covariates used to construct trees in k-fold approaches (k=10) are identical, but the values are slightly different. Therefore, Figure 8 shows the variables selected as optimal and the average values of the covariates in 10-fold samples. In Table 2, "Policy tree one" and "Policy tree two" show that the level one and two policies reduce the fraction nudged to 88% and 68%, respectively, relative to the status quo. Although these policies use only up to three variables to optimize welfare, the net benefit per person increases by 61% and 92%, respectively, relative to the status quo. Overall, these findings suggest that adopting an individual- or subgroup-specific targeting strategy has an advantage over a "one-size-fits-all" approach in maximizing the welfare impact of health interventions.

Risk-based targeting—. Row (6) reports results when we nudge individuals with high blood pressure (in addition to high blood sugar of $FBS_{it} \geq 126$). The fraction nudged decreases to 35%, but unlike the benefit-based targeting, the net benefit per person decreases by 6%. The results are similar for targeting based on high cholesterol (Row 7) and combining the two (Row 8). These results suggest that targeting nudges based on conventional risk factors may not necessarily increase welfare, as past biomarkers and health behaviors are noisy signals for predicting those who will benefit the most, and that selecting individuals who benefit the most from nudges is crucial.

5.4. Beyond compliers

This section reports results from the MTE analysis, which examines whether potential gains from visiting physicians are larger or smaller beyond compliers. This analysis is important because, if the former is true, more intensive policies that encourage reluctant individuals to visit physicians may improve welfare. We address these issues by estimating MTE. The supplementary results on MTE can be found in Appendix Section C.

Table C1 presents the estimation results of the selection equation, where the outcome D_{t+1} takes one if there is any DM-related visit within a year after the index screening. The table shows the marginal effects at the means of the covariates from logit estimation. As expected, the coefficient (0.094) of the "nudge" dummy or our instrument, $Z_{it} = 1(FBS_{it} \geq 126)$, is positive and statistically significant, and the magnitude is very similar to the RD estimates reported in Panel A of Table 1 (0.088). Moreover, the coefficients for most biomarkers are negative, which is consistent with the pattern of selection heterogeneity observed in our CATE analysis in

Section 5.2: individuals in better health and with healthier lifestyle habits are more likely to receive DM care.

Most importantly, the table reveals substantial heterogeneity in selection, enabling the expansion of propensity scores needed to identify MTEs in a broader population segment beyond compliers, as shown in Figure C1. Moreover, the figure illustrates a substantial overlap of propensity scores between the treated and untreated groups. We trim observations with estimated propensity scores below 0.18 or above 0.58, corresponding to the 1st percentile of the propensity score distribution among treated units and the 99th percentile among untreated units, respectively.

Table C2 shows the estimates of $\mu_1(x) - \mu_0(x)$—the observable part of MTE function—on the improvement in fasting blood sugar ($(-)\Delta$ FBS) between year $t+1$ and year t. Columns (1) and (2) present the cases of a quadratic and semi-parametrically estimated function of unobserved resistance to treatment, yielding very similar results. The coefficients on all biomarkers at the index checkup, including FBS, our running variable, LDL-C/$(-)$HDL-C, SBP/DBP, BMI, and triglycerides, are positive and statistically significant (except for DBP), suggesting that individuals with poorer health conditions at the initial checkup are more likely to benefit from receiving care. These findings are again consistent with the patterns of heterogeneity in benefits observed in our CATE analysis presented in Section 5.2.

We now move on to the unobservable part of the MTE function. Figure 9-A plots the MTE and 95% confidence intervals as a second-order polynomial function of unobserved resistance to physician visit (u_d) when the outcome is $(-)\Delta FBS$. The dotted line corresponds to the average treatment effect (ATE). The figure shows that MTE is *upward* sloping with unobserved resistance to treatment, indicating that the impact of receiving medical care on health outcomes is greater for individuals with high unobserved resistance to treatment.[23] This positive slope with respect to u_d suggests potentially substantial gains from an alternative policy if it can reach those less likely to seek care—inefficiency from insufficient participation.

Figure 9-B decomposes the MTE in Figure 9-A to the average treated outcome, $E(Y_1|\, U_d = u_d)$ and untreated outcome, $E(Y_0|\, U_d = u_d)$. The figure shows that the untreated outcome (Y_0) is relatively flat regardless of unobserved resistance to treatment. In contrast, the treated outcome (Y_1) is larger for those with higher unobserved resistance. These results further suggest that visiting physicians and subsequent medical care are indeed medically meaningful for individuals with high resistance to treatment, helping to improve their health outcomes.

[23] Figure C2 shows that semiparametric estimation yields similar results.

Another advantage of estimating MTEs is that we can obtain an estimate of economic interest, such as ATT, ATU, and LATE, by integrating relevant regions of the MTE curve. Table 3 reports the estimates, where the outcome is again $(-)\Delta FBS$. Importantly, ATU (the average effect of receiving medical care for those reluctant to do so) is greater than LATE (the average effect of receiving medical care for those who complied with a nudge), further suggesting that there may be welfare gains by reaching out to individuals reluctant to visit physicians.

These findings suggest that alternative approaches—such as sending stronger nudges or providing health information directly to primary care physicians rather than to patients—that shift the propensity scores, may improve welfare by persuading hesitant individuals to seek care, as long as the resulting benefits outweigh the associated costs.

6. Conclusion

Many, if not all, social programs rely on individuals to self-select into programs. The unintended consequences of self-selection into programs are twofold: some individuals whose social costs outweigh benefits may choose to participate, resulting in wasted resources, while others who could benefit from participating may decide not to participate, leading to missed opportunities. This paper develops a framework to address both types of inefficiency and applies it to health care.

Exploiting the quasi-experimental setup in diabetes screening—where individuals are nudged to seek care if a biomarker exceeds a threshold—we examine who self-selects into DM care and whether these individuals are those who benefit the most from receiving it. We then refine eligibility rules to reduce inefficiency from excessive participation and explore outreach strategies to address inefficiency resulting from insufficient participation among high-benefit individuals who currently forgo care.

Utilizing exceptionally large-scale data from annual health screenings in Japan, we find that the status quo policy of nudging individuals at the diabetic threshold increases physician visits and improves health outcomes on average. However, individuals who opt into care—both compliers and beyond-compliers—benefit the least from care, suggesting *reverse* selection on gains. Accordingly, targeting nudges based on observable characteristics to individuals with positive net benefits can reduce excessive participation and substantially enhance welfare.

Furthermore, we find that individuals who are currently reluctant to seek care under existing nudges have the greatest potential to benefit from doing so. This suggests that more intensive interventions—such as stronger nudges or delivering health information directly to

primary care physicians instead of patients—could enhance welfare by encouraging these hesitant individuals to seek medical care. However, this study does not assess the cost-effectiveness of such alternative strategies, and we leave this important question for future research.

Another limitation of this research is that we do not know why some people do not receive care even when the returns from doing so seem high. Some potential reasons may include transaction costs (they are too busy, it's too far away, they can't take their children with them), myopia (they heavily discount the long-term benefits of visiting physicians today), and information adoidance (they are afraid of bad news and don't want to admit that they are sick). Understanding these reasons or mechanisms is crucial for devising policies to bring them to physician offices, and this important issue is also left for future research.

References

Alalouf, Mattan, Sarah Miller, and Laura R. Wherry. 2024. "What Difference Does a Diagnosis Make?: Evidence from Marginal Patients." *American Journal of Health Economics*, 10(1): 97–131.

Almond, Douglas, Joseph J. Doyle, Amanda E. Kowalski, Heidi Williams. 2010. "Estimating Marginal Returns to Medical Care: Evidence from At-risk Newborns." *The Quarterly Journal of Economics*, 125(2): 591–634.

Andresen, Martin E. 2018. "Exploring marginal treatment effects: Flexible estimation using Stata." *The Stata Journal*, 18(1): 118–158.

Arrow, Kenneth J. 1963. "Uncertainty and the Welfare Economics of Medical Care." *American Economic Review*, 53: 941–973.

Athey, Susan, Kosuke Inoue, and Yusuke Tsugawa. 2025a. "Targeted Treatment Assignment Using Data from Randomized Experiments with Noncompliance." *AEA Papers and Proceedings*, 115: 209–214.

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

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

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

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.

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

Burlat, Héloïse. 2024. "Everybody's got to learn sometime? A causal machine learning evaluation of training programmes for jobseekers in France." *Labour Economics*, 89: 102573.

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. https://arxiv.org/abs/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.
Currie, Janet. 2006. "The Take-Up of Social Benefits." In *Public Policy and Income Distribution*, edited by Alan J. Auerbach, David Card, and John M. Quigley, 80–148. New York: Russell Sage Foundation.
Deshpande, Manasi, and Yue Li. 2019. "Who Is Screened Out? Application Costs and the Targeting of Disability Programs." *American Economic Journal: Economic Policy*, 11(4): 213–248.
Einav, Liran, Amy Finkelstein, Stephen Ryan, Paul Schrimpf, and Mark R. Cullen. 2013. "Selection on Moral Hazard in Health Insurance." *American Economic Review*, 103(1): 178–219.
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.
Finkelstein, Amy, and Matthew J Notowidigdo. 2019. "Take-Up and Targeting: Experimental Evidence from SNAP." *The Quarterly Journal of Economics*, 134(3): 1505–1556.
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.
Erik Sverdrup, Ayush Kanodia, Zhou, Zhengyuan, Susan Athey, and Stefan Wager. 2020. "*policytree*: Policy learning via doubly robust empirical welfare maximization over trees." *Journal of Open Source Software*, *5*(50): 2232.
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.
Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and estimation of local average treatment effects." *Econometrica*, 62 (2): 467–475.

Inoue, Kosuke, Susan Athey, and Yusuke Tsugawa. 2023. "Machine-learning-based high-benefit approach versus conventional high-risk approach in blood pressure management." *International Journal of Epidemiology*, 52(4): 1243–1256.

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.

Johnson, Matthew S., David I. Levine, and Michael W. Toffel. 2023. "Improving regulatory effectiveness through better targeting: Evidence from OSHA." *American Economic Journal: Applied Economics*, 15(4): 30–67.

Khanna, Gaurav. 2023. "Large-Scale Education Reform in General Equilibrium: Regression Discontinuity Evidence from India." *Journal of Political Economy*, 131(2): 549–591.

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.

Kline, Patrick, and Christopher R. Walters. 2016. "Evaluating Public Programs with Close Substitutes: The Case of Head Start." *The Quarterly Journal of Economics*, 131(4): 1795–1848.

Knaus, Michael C., Michael Lechner, and Anthony Strittmatter. 2022. "Heterogeneous employment effects of job search programs." *Journal of Human Resources*, 57(2): 597–636.

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.

Schmieder, Johannes F., and Till Von Wachter. 2016. "The Effects of Unemployment Insurance Benefits: New Evidence and Interpretation." *Annual Review of Economics*, 8: 547–581.

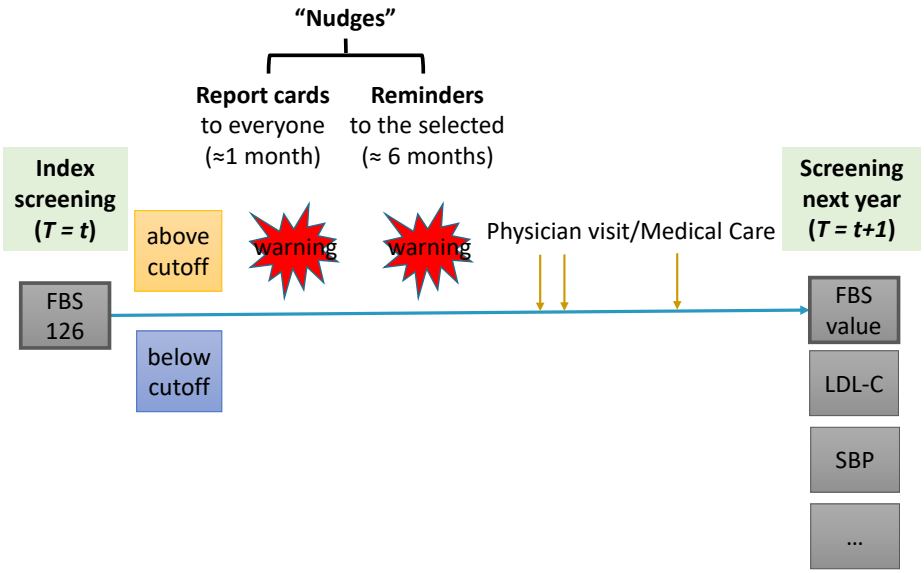
UK Prospective Diabetes Study (UKPDS) Group. 1998. "Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33)." *The Lancet*, 352(9131): 837–853.

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, and Stefan Wager. 2023. "Offline multi-action policy learning: Generalization and optimization." *Operations Research*, *71*(1): 148–183.

Figure 1—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 2—An example of a report card and nudge

A. Report card

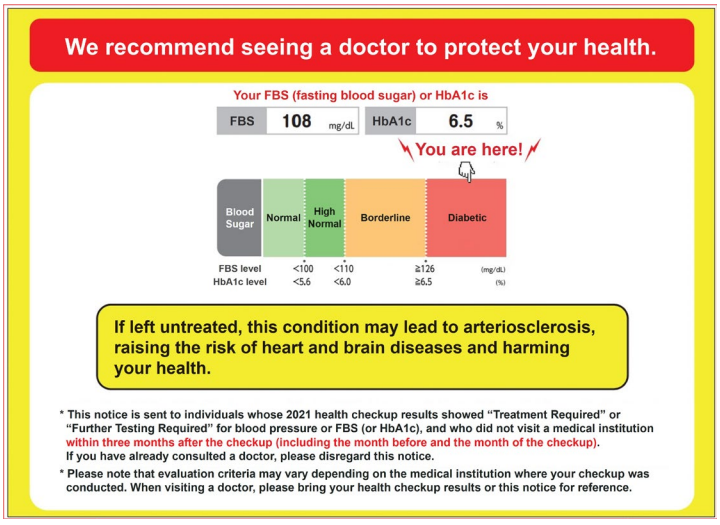
Health Screening Report

Name	**********	Date of Birth	19**/**/**	Gender	****	Age	**

	Test Items	Reference Range	This Time	Previous Time
Body Measurements	Height(cm)		182.2	182.1
	Weight(kg)		80.7	83.5
	BMI	18.5–24.9kg/㎡	24.3	25.1
	Waist circumference	Men: <85 cm, Women: <90 cm	* 89.6	* 90.2
Blood Pressure	SBP	<130 mmHg	120	128
	DBP	<85 mmHg	70	73
Lipid Profile	Total Cholesterol	140–199 mg/dL	190	196
	Triglycerides	<150 mg/dL	120	124
	HDL Cholesterol	≧40 mg/dL	55	51
	LDL Cholesterol	<120 mg/dL	110	117
Liver Function	GOT(AST)	≦30 U/L	25	28
	GPT(ALT)	≦30 U/L	28	26
	γ-GPT(γ-GT)	≦50 U/L	35	44
Glucose Metabolism	FBS	<100 mg/dL	* 100	98
	HbA1c	<5.6%	* 5.8	5.1
Upper GI X-ray	Esophagus		Normal	Normal
	Stomach		Normal	Normal
	Duodenum		Normal	Normal
...	...	...	...	...

Comments	Waist circumference and HbA1c are slightly elevated. Please review your diet and exercise habits, and work on weight management. Other results are generally good; please maintain your current condition.

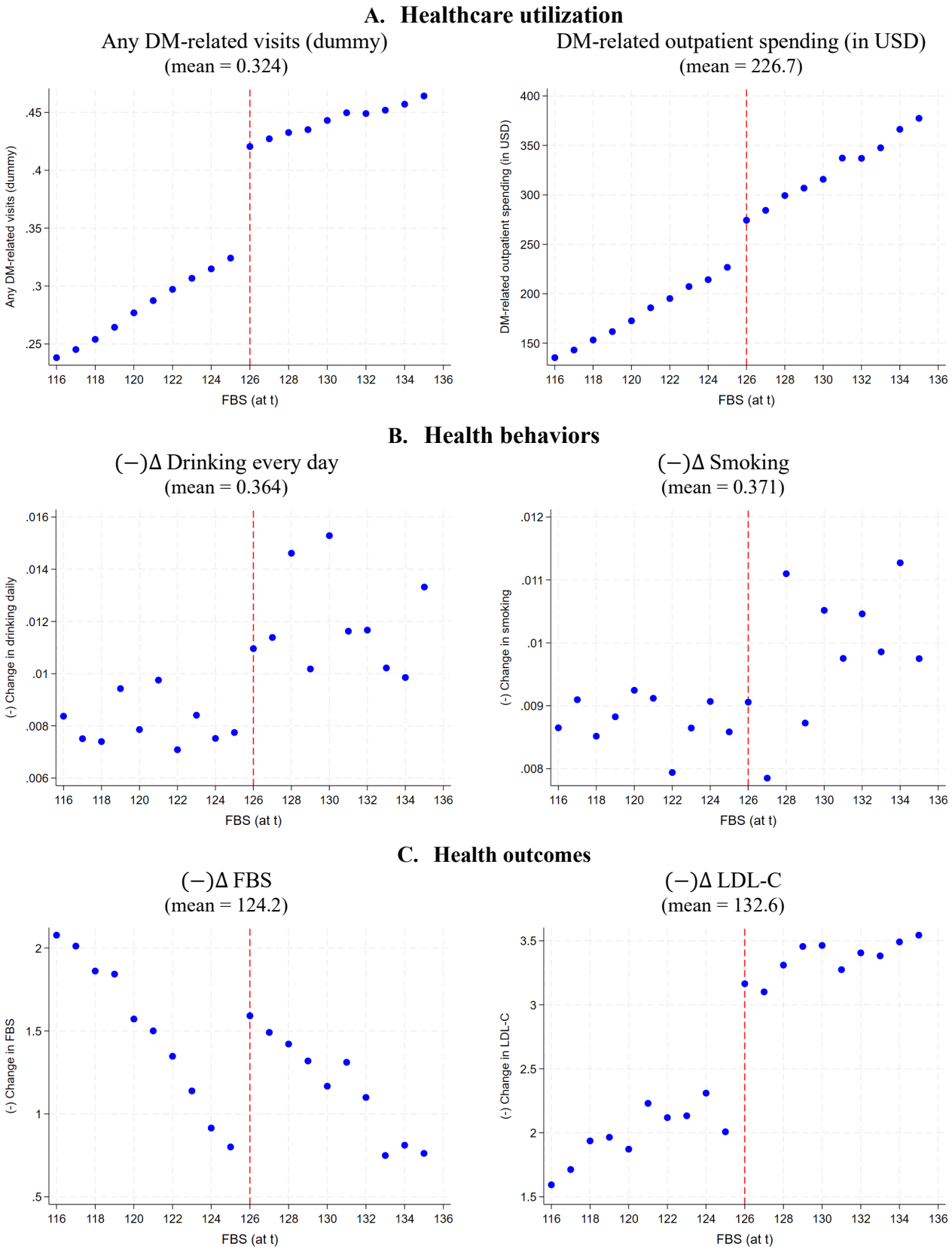
B. Reminder



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 "*" in this case or "H" for high) and recommends consulting a physician for further guidance. Panel B offers an example of a nudge sent by the Japan Health Insurance Association (JHIA). The document includes a statement encouraging individuals to see a doctor, along with strong warnings. While the wording of the warnings may slightly change over the years, the content remains similar. This particular example from 2018 states, "If left untreated, this condition may lead to arteriosclerosis, raising the risk of heart and brain diseases and harming your health."

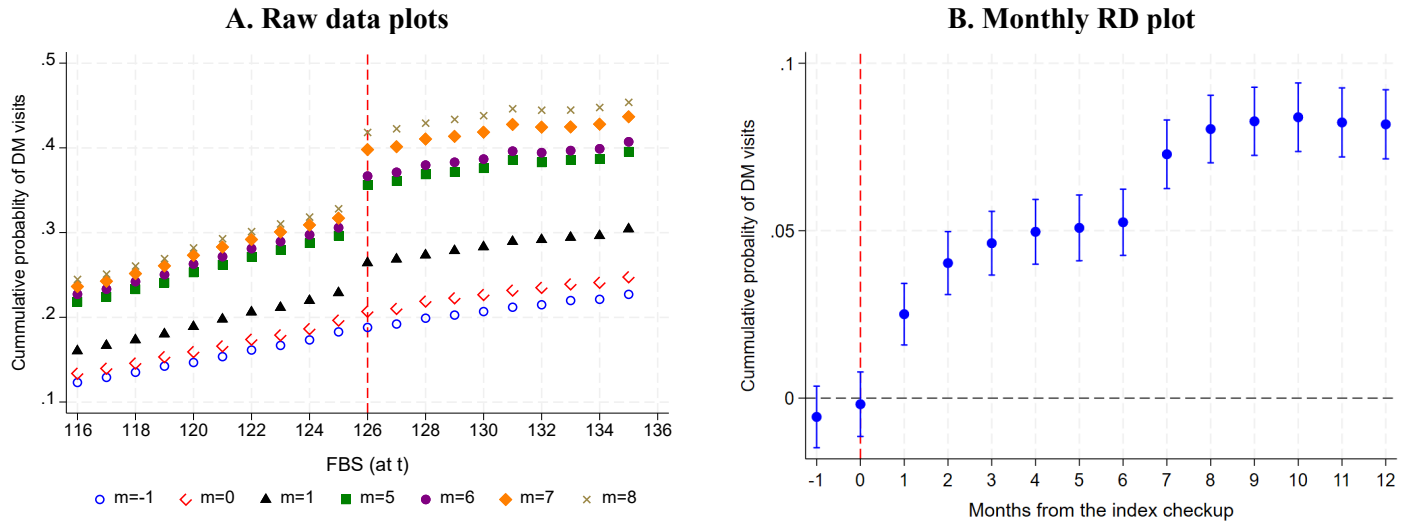
Source: https://www.kyoukaikenpo.or.jp/~/media/Files/honbu/kenshin/2022/jyusinnkannsyoutuuti_2021-2022.pdf (last accessed on May 9, 2025). The original language was in Japanese, which was translated by the authors. See Figure A1 for the original versions.

Figure 3—RD plots



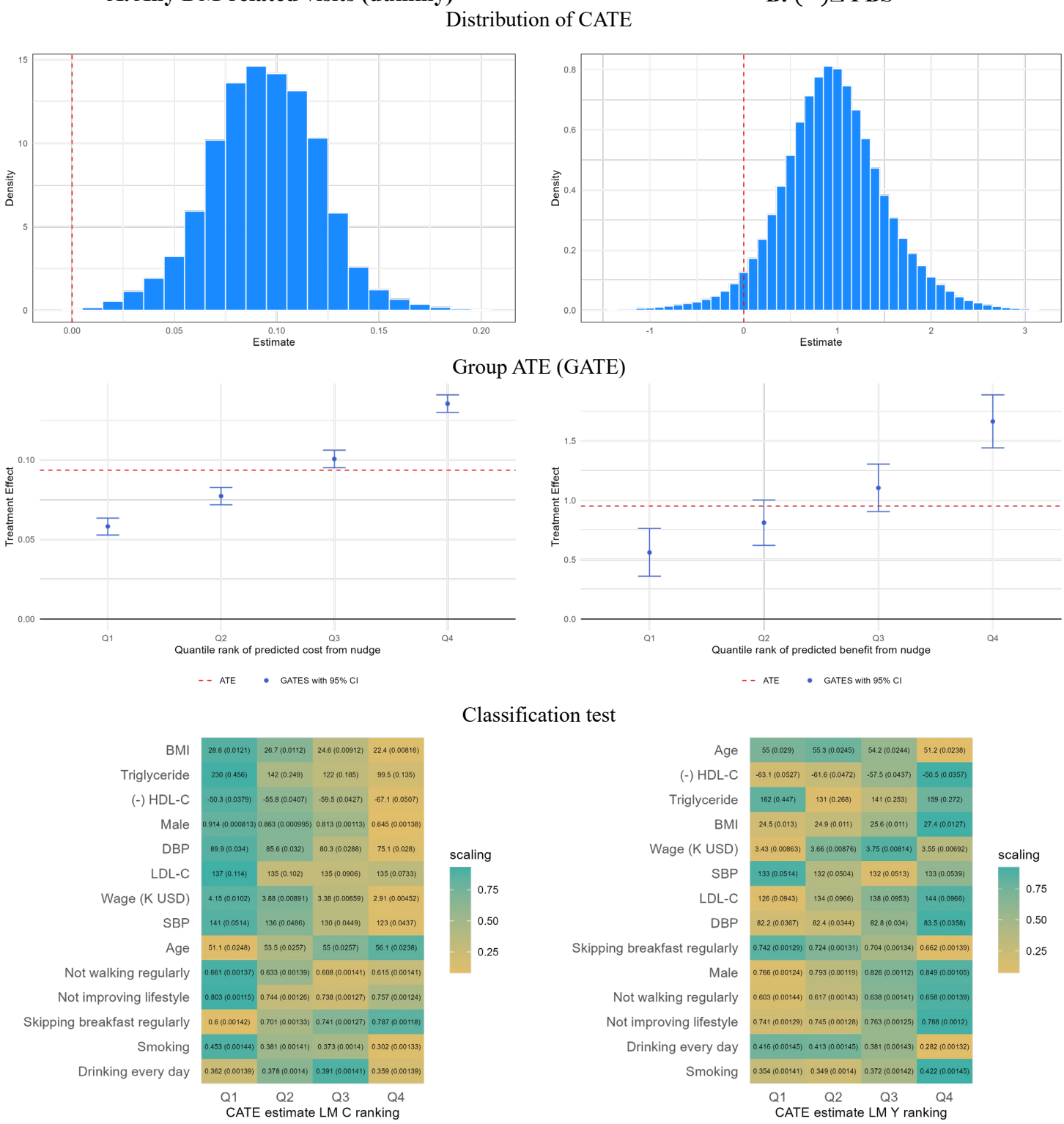
Notes: The scatter plot shows the mean of each measure of healthcare utilization (panel A), health behaviors (panel B), and health outcomes (panel C) at 1-point bins of the FBS value within 10 mg/dl around the FBS = 126 mg/dl threshold at the time of index checkup in year t. See Appendix D for the definition of variables. The outcomes in panels B and C are redefined so that higher values represent healthier behaviors and better health outcomes.

Figure 4—Monthly healthcare utilization



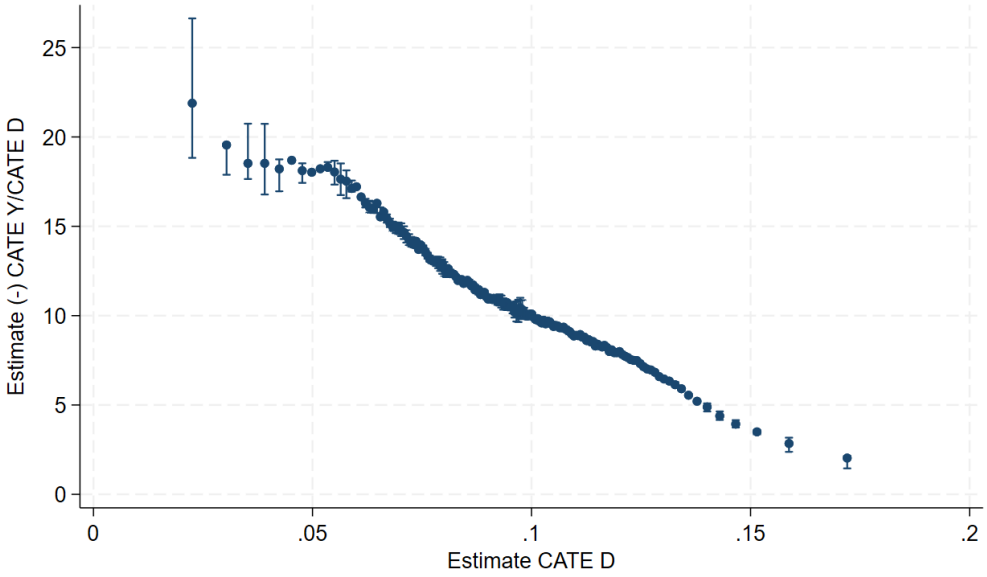
Notes: The outcome is a cumulative DM visit dummy, which equals one if an individual has at least one DM visit until month m from the index checkup and zero otherwise. Panel A plots the outcome at cumulative months m = -1, 0, 1, 5, 6, 7, and 8, where m = -1 indicates a month before the index checkup and m = 0 is the month when the individual had the checkup. Panel B plots the corresponding RD estimates from equation [1], where the outcome is the cumulative DM visit dummy for each month m from -1 to 12. The reminders are provided at m = 6.

Figure 5—The CATEs of utilization and health outcome



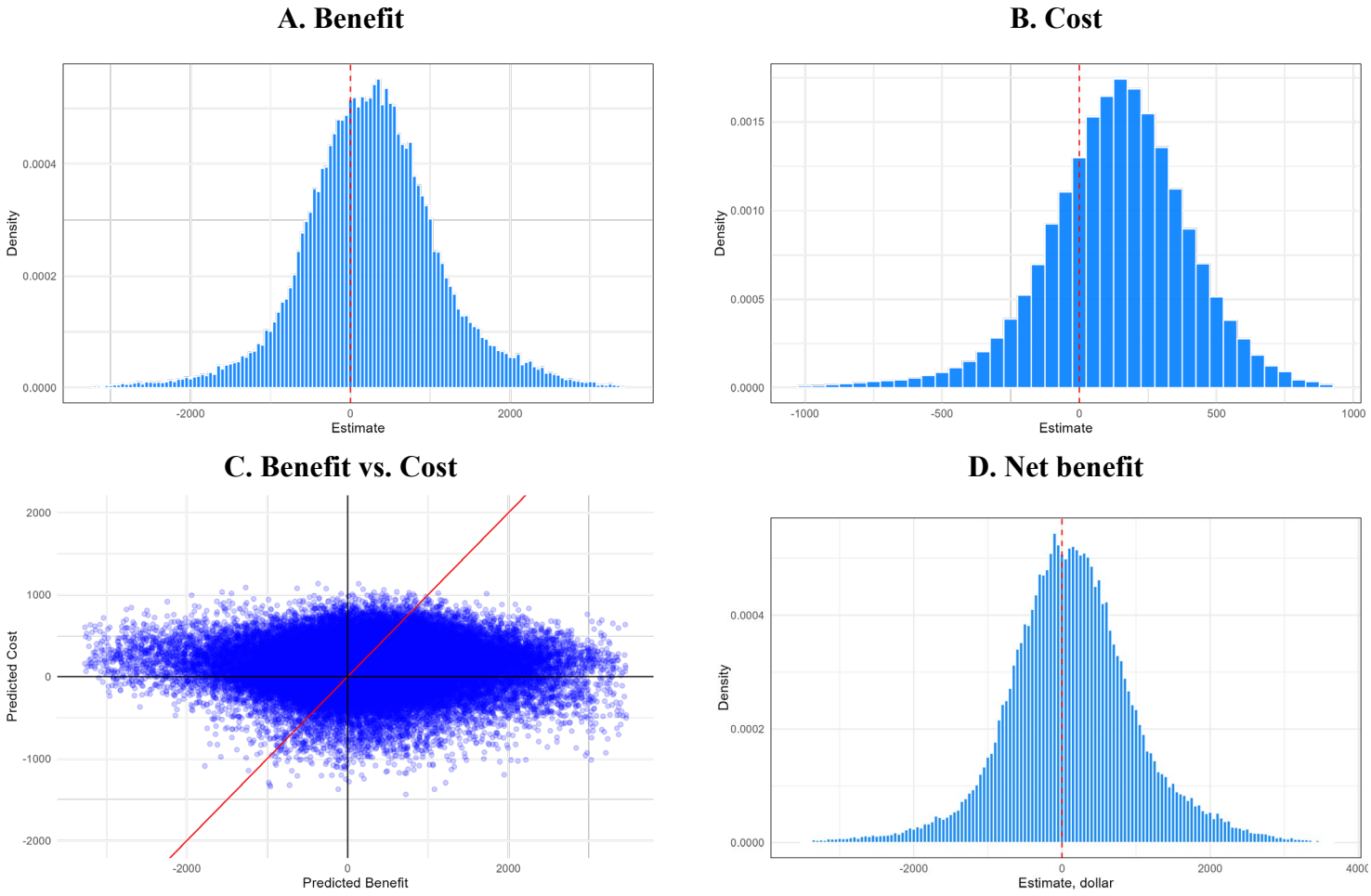
Notes: The outcome in panel A is a binary indicator that equals one if an individual has at least one DM visit within a year following an index checkup and zero otherwise, and the outcome in panel B is the change of fasting blood sugar ((−)Δ FBS) between year $t+1$ and year t, which is redefined so that higher values represent better health outcome. 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 order of the variables follows the order of the variable importance, as shown in Figure B1. The LM Forest, which is part of the GRF (Generalized Random Forests) package in R, is used.

Figure 6—Relationship between CATEs for selections and benefits



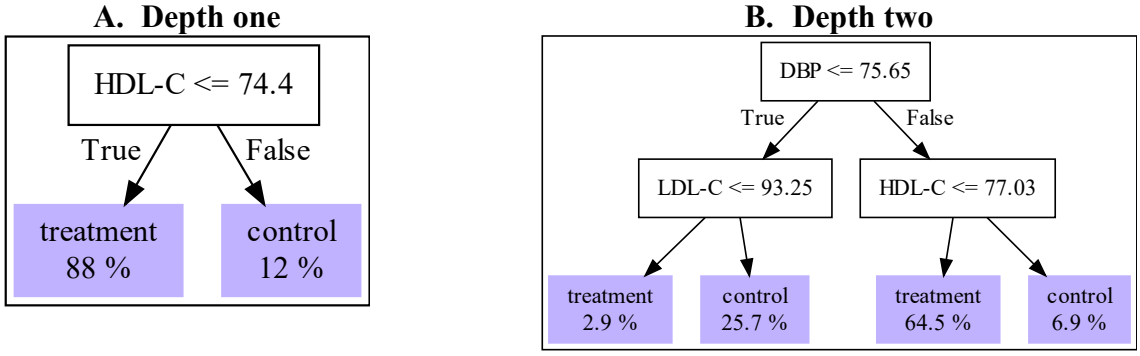
Notes: The figure reports a binned scatter plot of the relationship between the estimated CATE for selection D on the x-axis and the CATE for health benefits Y ($(-)\Delta$ FBS) divided by the CATE for D that measures the health return on compliers (treatment-on-treated) on the y-axis at the individual level.

Figure 7—Distribution of CATEs on benefit, cost, and net benefit



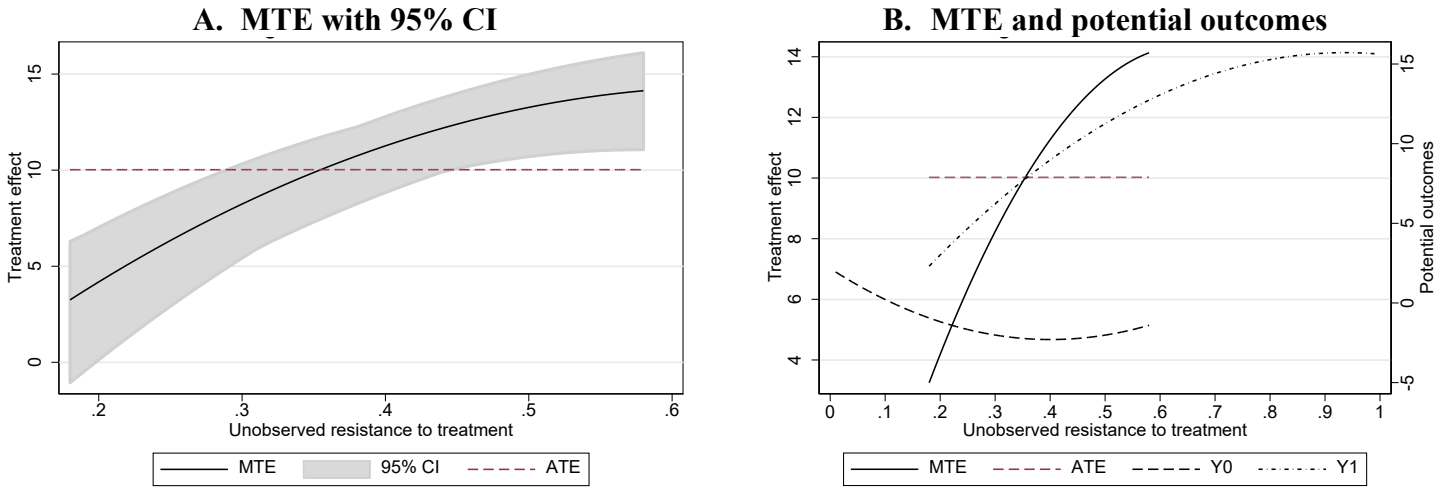
Notes: . The figures plots the histogram of estimated $\tau(X)$ from equation [2], which is refined using the doubly robust score to mitigate regularization bias and overfitting. For panel A, the benefit is defined by the 6-year mortality reduction times the value of statistical life per year, which is set at $50,000 per year. For panel B, the cost is defined by the cumulative 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 8—Policy tree



Notes: The figure shows the policy trees of depth one (panel A) and depth two (panel B) based on the estimated net benefit using the *policytree* package for R. In both depth one and depth two, the choice and order of covariates used to construct trees in k-fold approaches (k=10) are identical, but the values are slightly different. Therefore, we report the average values of the covariates in 10-fold samples.

Figure 9—MTE on $(-)\Delta$FBS



Notes: The outcome is the improvement in fasting blood sugar ($(-)\Delta$ FBS) between year $t+1$ and year t. Panel A plots the marginal treatment effect (MTE) and the 95% confidence intervals as a second-order polynomial function of unobserved resistance to treatment. The dotted line represents the average treatment effect (ATE). Panel B decomposes the MTE in panel A to the treated, $E(Y_1|\, U_d = u_d)$ which is labeled as Y_1, and the untreated, $E(Y_0|\, U_d = u_d)$ which is labeled as Y_0 on the right *y*-axis, and the two treatment effects (MTE and ATE) from panel A on the left *y*-axis.

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	5.737***	(1.887)	26.731	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 (ITT)					
(−)Δ FBS	0.934***	(0.197)	124.199	3.516	354,878
(−)Δ LDL-C	1.643***	(0.319)	132.702	2.243	255,592
(−)Δ SBP	0.532***	(0.172)	132.687	2.950	256,551
(−)Δ BMI	0.082***	(0.012)	25.540	3.101	364,477
Panel D: Health outcomes (LATE)					
(−)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

Notes: Panels A-C show the estimates from local-linear regression in equation [1], using a triangular kernel and covariates. Panel D 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 a dummy for any DM-related visits. 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 parentheses. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.10$.

Table 2—Targeting

		Fraction receiving nudges	Net benefit /person ($)	
(1)	Status quo	**100%**	109	**100%**
	Benefit-based targeting:			
(2)	Non-parametric: (Net benefit)>0	57%	305	280%
(3)	Linear prediction: $\widehat{Prob}$(1(net benefit))>0	62%	192	176%
(4)	Policy tree: depth1	88%	175	161%
(5)	Policy tree: depth2	68%	209	192%
	Risk-based targeting:			
(6)	Blood pressure (SBP ≥ 140 or DBP ≥ 90)	35%	102	94%
(7)	Fat (LDL-C≥140 or Triglycerides≥300)	48%	84	77%
(8)	Both blood pressure and fat	18%	72	66%

Notes: This table shows the fraction of individuals among FBS ≥ 126 who received the nudge and the corresponding net benefit per person in different scenarios. The net benefit is calculated using the inverse probability weighting (IPW) method, which adjusts each individual's CATE estimate by their propensity score. Row (1) is the status quo, where the nudge is sent to those FBS ≥ 126. Rows (2)-(5) are benefit-based targeting. Row (2) is a scenario where the insurer nudges only those individuals whose predicted benefit from the nudge exceeds the cost of the nudge, in addition to FBS ≥ 126. Row (3) is a scenario where we predict those with positive net benefits by estimating a simple linear regression of net benefits on observables and then target only those with positive expected net benefits, in addition to FBS ≥ 126. Rows (4) and (5) are scenarios based on the policy tree (Athey and Wager 2021) with the optimal tree of a pre-specified depth to be one and two, in addition to FBS ≥ 126, respectively. See Figure 8 for the observable characteristics used to define subgroups. Rows (6)-(8) are risk-based targeting. In addition to FBS ≥ 126, nudges are sent to individuals with high blood pressure (SBP ≥140 mmHg or DBP ≥ 90), high cholesterol (LDL-C ≥ 140 or triglycerides ≥ 300), and those with both conditions in rows (6), (7), and (8), respectively.

Table 3—Policy simulations

	Effects	SE
ATE (average treatment effect)	10.02	(1.01)
ATT (Average treatment effect on treated)	7.43	(1.18)
ATU (Average treatment effect on untreated)	11.91	(1.01)
LATE (Local average treatment effect)	9.88	(0.99)

Notes: The outcome is the improvement in blood sugar ($(-)\Delta$ FBS) between year $t+1$ and year t. The table summarizes several treatment effects (ATE, ATT, ATU, and LATE). The bootstrapped standard errors are in parentheses.

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—An example of a report card and nudge (in Japanese)

A. Report card



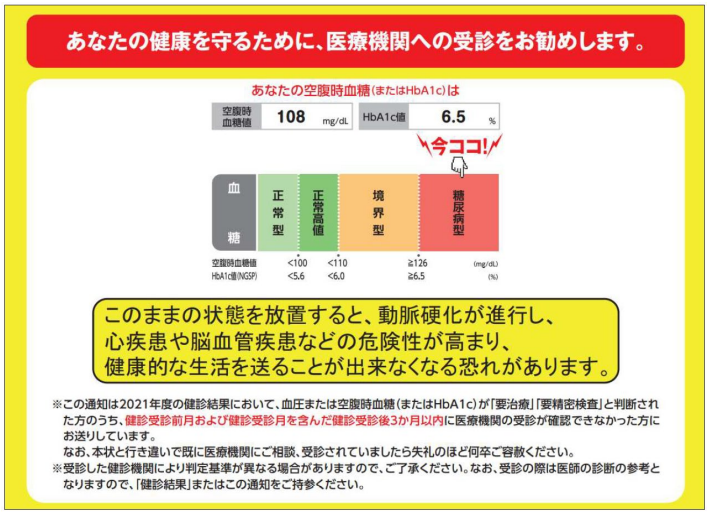
健診結果通知表

氏名	**********	生年月日	19**/**/**	性別	**	年齢	**

	項目	基準範囲	今回	前回
身体計測	身長(cm)		182.2	182.1
	体重(kg)		80.7	83.5
	BMI	18.5 ~ 24.9kg/㎡	24.3	25.1
	腹囲	男性 85cm 未満, 女性 90cm 未満	＊89.6	＊90.2
血圧	収縮期血圧	130mmHG 未満	120	128
	拡張期血圧	85mmHG 未満	70	73
脂質代謝	総コレステロール	140 ~ 199mg/dl	190	196
	中性脂肪	150mg/dl 未満	120	124
	HDL-コレステロール	40mg/dl 以上	55	51
	LDL-コレステロール	120mg/dl 未満	110	117
肝機能	GOT(AST)	30U/L 以下	25	28
	GPT(ALT)	30U/L 以下	28	26
	γ-GPT(γ-GT)	50U/L 以下	35	44
糖代謝	空腹時血糖	100mg/dl 未満	＊100	98
	HbA1c	5.6% 未満	＊5.8	5.1
胃部X線	食道		異常なし	異常なし
	胃		異常なし	異常なし
	十二指腸		異常なし	異常なし
…	…	…	…	…

コメント	腹囲とHbA1cにやや高値が見られます。食事・運動習慣を見直し、体重管理に努めましょう。その他の項目はおおむね良好ですので、現状を維持してください。

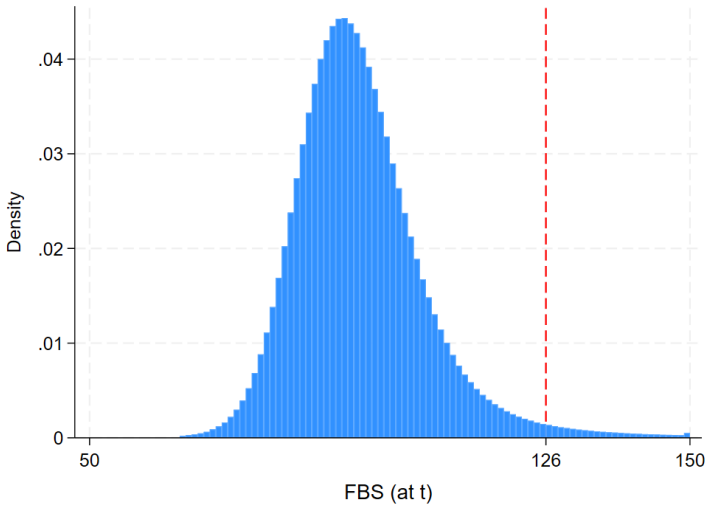
B. Reminder



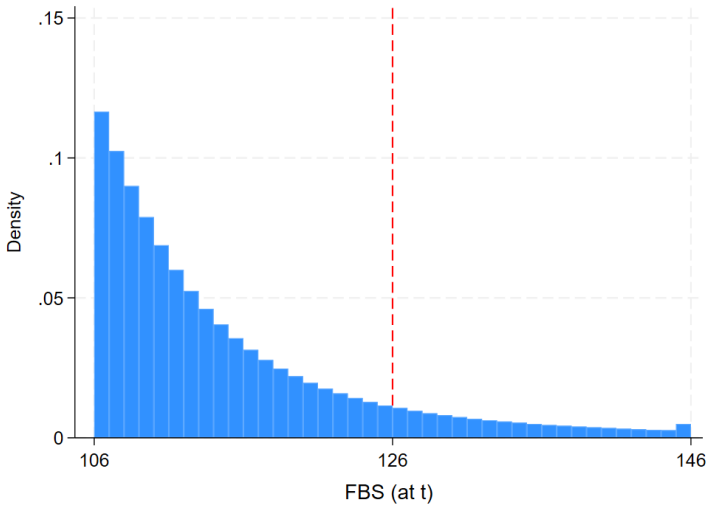
Notes: Original version in Japanese.

Figure A2—Distribution of FBS values

A. Overall

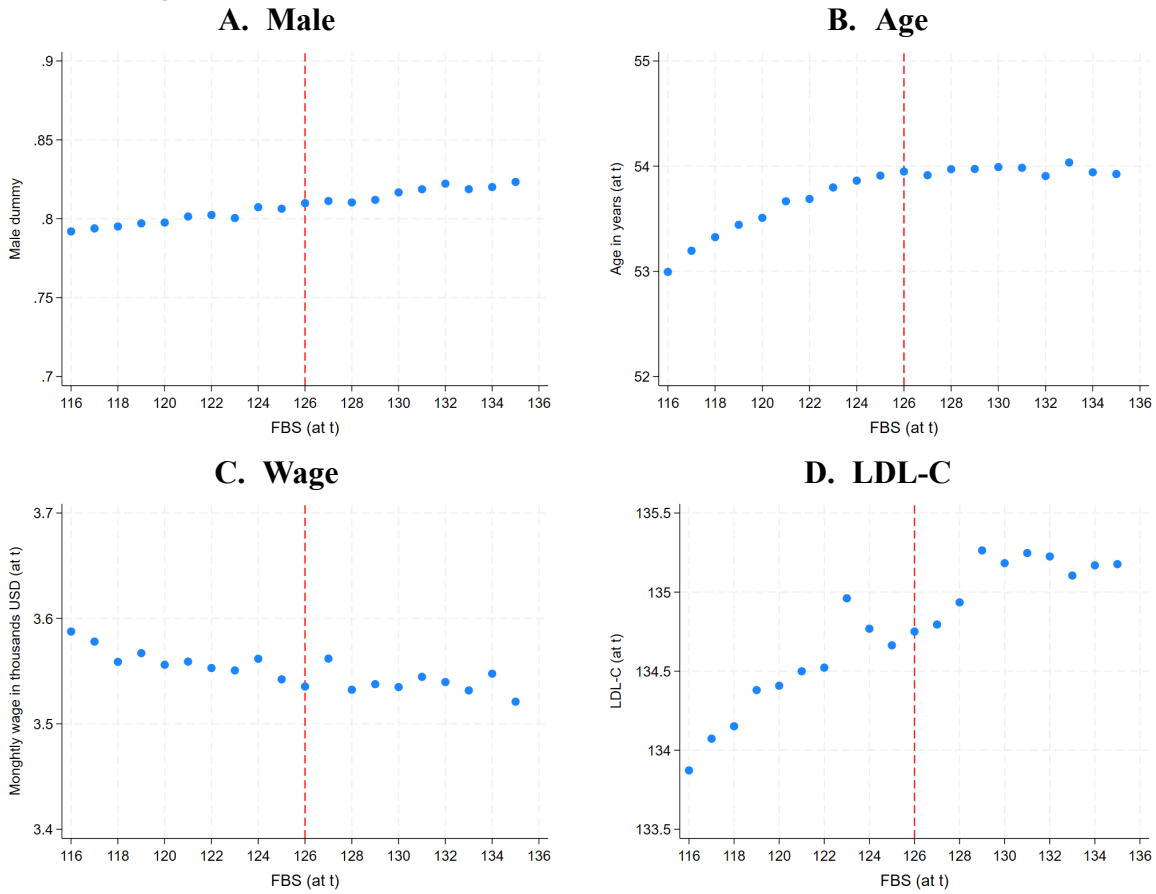


B. Near 126 threshold



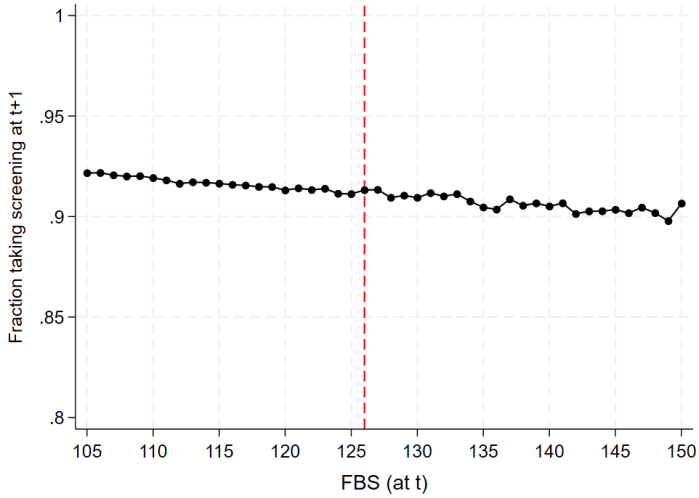
Notes: The histogram shows the density of FBS values at year t at 1-point bins of the FBS value for the overall sample (panel A) and within 20 mg/dl around the FBS = 126 mg/dl threshold (panel B).

Figure A3—Balanced checks for selected variables



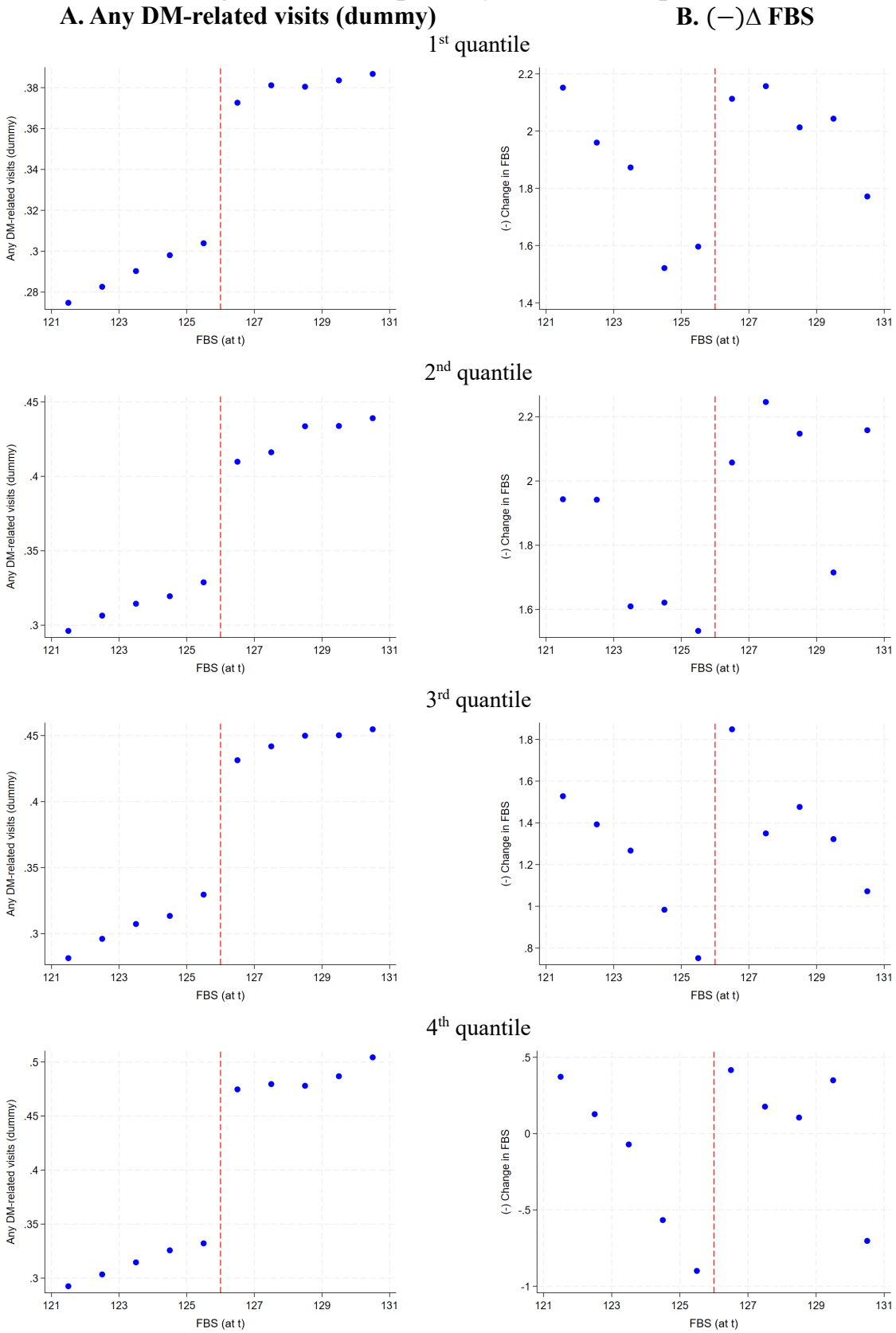
Notes: The figures plot the mean of the variables at 1-point bins of the FBS value within 10 mg/dl around the FBS = 126 mg/dl threshold at the time of index checkup in year t (dotted line). The outcomes are the male dummy (panel A), age (panel B), wage (panel C), and low-density lipoprotein cholesterol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

Figure A4—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 the index checkup in year t (dotted line).

Figure A5—RD plots by the CATEs quantiles



Notes: The scatter plot shows the mean of each measure of any DM-related visits (panel A), and (−)Δ FBS (panel B), separately for each quantile of CATEs distributions in Figure 5, 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* (dotted line). See Appendix D for the definition of variables.

Table A1—Summary statistics

	Full sample			Analysis sample (FBS 116-135)		
Variable:	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
B. Covariates						
Male (dummy)	48,017,588	0.64	0.48	1,312,491	0.80	0.40
Age	48,017,588	50.40	9.34	1,312,491	53.55	8.90
Monthly wage (in K USD)	47,744,932	3.33	2.35	1,305,851	3.58	2.74
C. Healthcare utilization (at t+1)						
Any DM-related visits (dummy)	48,017,588	0.20	0.40	1,312,491	0.32	0.47
Number of DM-related visits	48,017,588	1.28	4.48	1,312,491	1.74	4.15
DM-related outpatient spending (in USD)	47,978,760	201.39	724.01	1,311,369	210.12	586.75
DM-related drug spending (in USD)	47,972,208	56.97	324.72	1,312,382	23.93	169.48
D. Health behavior (at t+1)						
Drinking every day (dummy)	44,395,768	0.30	0.46	1,214,226	0.38	0.48
Smoking (dummy)	48,017,588	0.32	0.46	1,312,491	0.38	0.48
Not walking regularly (dummy)	44,117,568	0.62	0.49	1,207,622	0.63	0.48
Skipping breakfast regularly (dummy)	43,819,616	0.25	0.43	1,199,586	0.30	0.46
E. Health outcomes (at t+1)						
FBS (mg/dl)	46,398,308	98.92	18.97	1,273,870	120.83	18.16
LDL-C (mg/dl)	47,697,656	124.25	31.38	1,302,978	132.27	33.22
SBP (mmHg)	47,915,440	123.89	17.10	1,307,787	132.09	17.65
BMI	47,916,612	23.64	3.93	1,308,401	25.36	4.20

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 outcomes are measured at the index checkup 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 parentheses. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.10$.

Table A3—RD Estimates by CATEs quantiles

	(1) 1st quantile	(2) 2nd quantile	(3) 3rd quantile	(4) 4th quantile
Panel A: Any DM-related visits (dummy)				
Bandwidth = ±3	0.063***	0.072***	0.092***	0.137***
	(0.008)	(0.008)	(0.008)	(0.008)
Bandwidth = ±4	0.064***	0.072***	0.095***	0.136***
	(0.007)	(0.007)	(0.007)	(0.007)
Bandwidth = ±5	0.064***	0.073***	0.096***	0.134***
	(0.006)	(0.006)	(0.006)	(0.006)
Panel B: (−)Δ FBS				
Bandwidth = ±3	0.616*	0.513*	1.333***	1.607***
	(0.324)	(0.302)	(0.320)	(0.356)
Bandwidth = ±4	0.674**	0.650**	1.236***	1.518***
	(0.279)	(0.261)	(0.276)	(0.307)
Bandwidth = ±5	0.716***	0.699***	1.179***	1.509***
	(0.249)	(0.233)	(0.246)	(0.274)

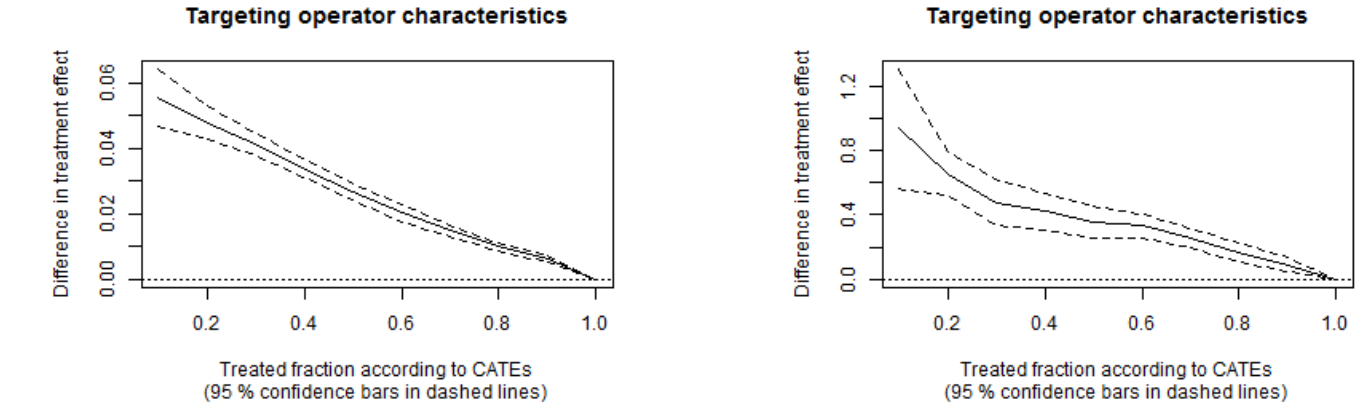
Notes: This table shows the estimates from local-linear regression in equation [1], using a triangular kernel and covariates, separately for each quantile of CATEs distributions in Figure 5. The outcomes are any DM-related visits (panel A), and (−)Δ FBS (panel B). The bandwidth is set at 5 mg/dl, and the conventional standard error by Calonico et al. (2014) is reported in parentheses. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.10$.

Appendix B: CATE

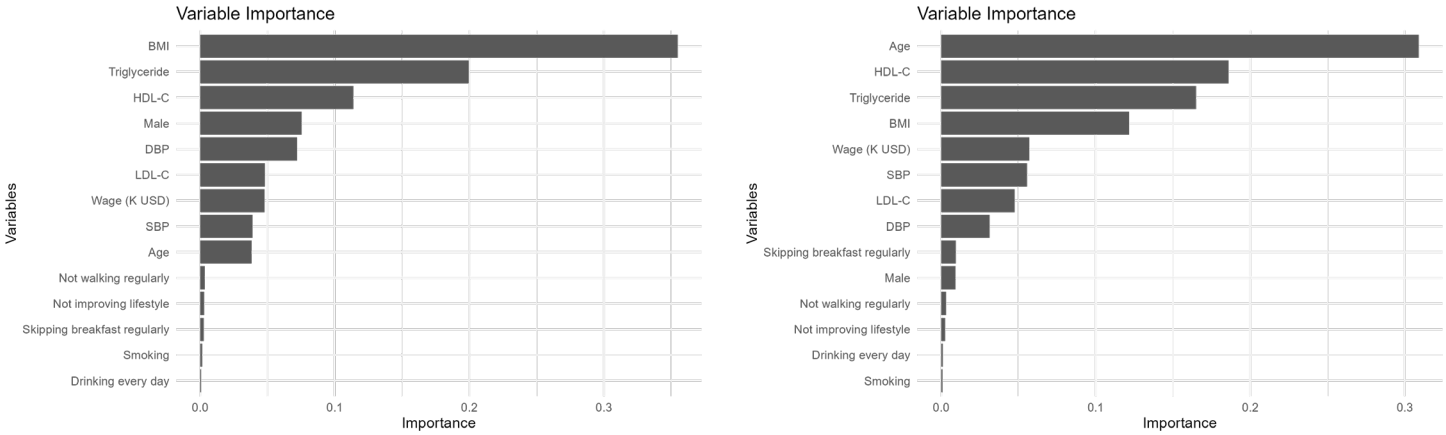
Figure B1—Other metrics from CATEs

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

Rank-weighted ATE (RATE)



Variable importance



Notes: 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 Targeting Operator Characteristic (TOC) curve is constructed by ranking all observations according to their predicted CATEs for the test set—from a model trained on the training set—and, for each treated fraction q, comparing the average treatment effect among the top q percentile to the overall ATE. Point estimates of the TOC are plotted as a solid black line as q varies, and 95 % confidence intervals are shown as dashed lines. RATE is the area under the TOC. The RATE estimates for both panels A and B are statistically significant at the 5% level. We use 10-fold cross-fitting for RATE estimation to split the data into training and test sets. The LM Forest, part of the *grf* (Generalized Random Forests) package in R, is used.

Figure B2—CATEs for other utilization measures



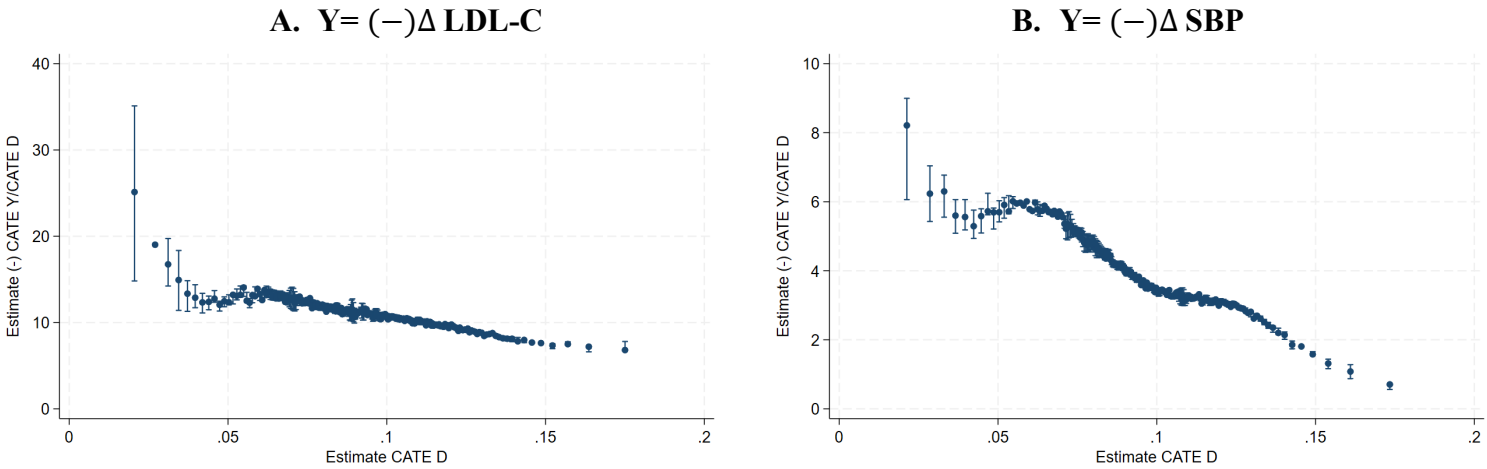
Notes: The outcomes in panels A and B are the frequency of DM-related outpatient visits and DM-related outpatient spending aggregated for 12 months from the month of the checkup. The top row plots the histogram of estimated $\tau(X)$ from equation [2], which is refined using the doubly robust score to 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—CATEs for other health outcomes



Notes: The outcomes in panels A and B are the change of low-density lipoprotein cholesterol ((−)Δ LDL-C) and systolic blood pressure ((−)Δ SBP) 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.

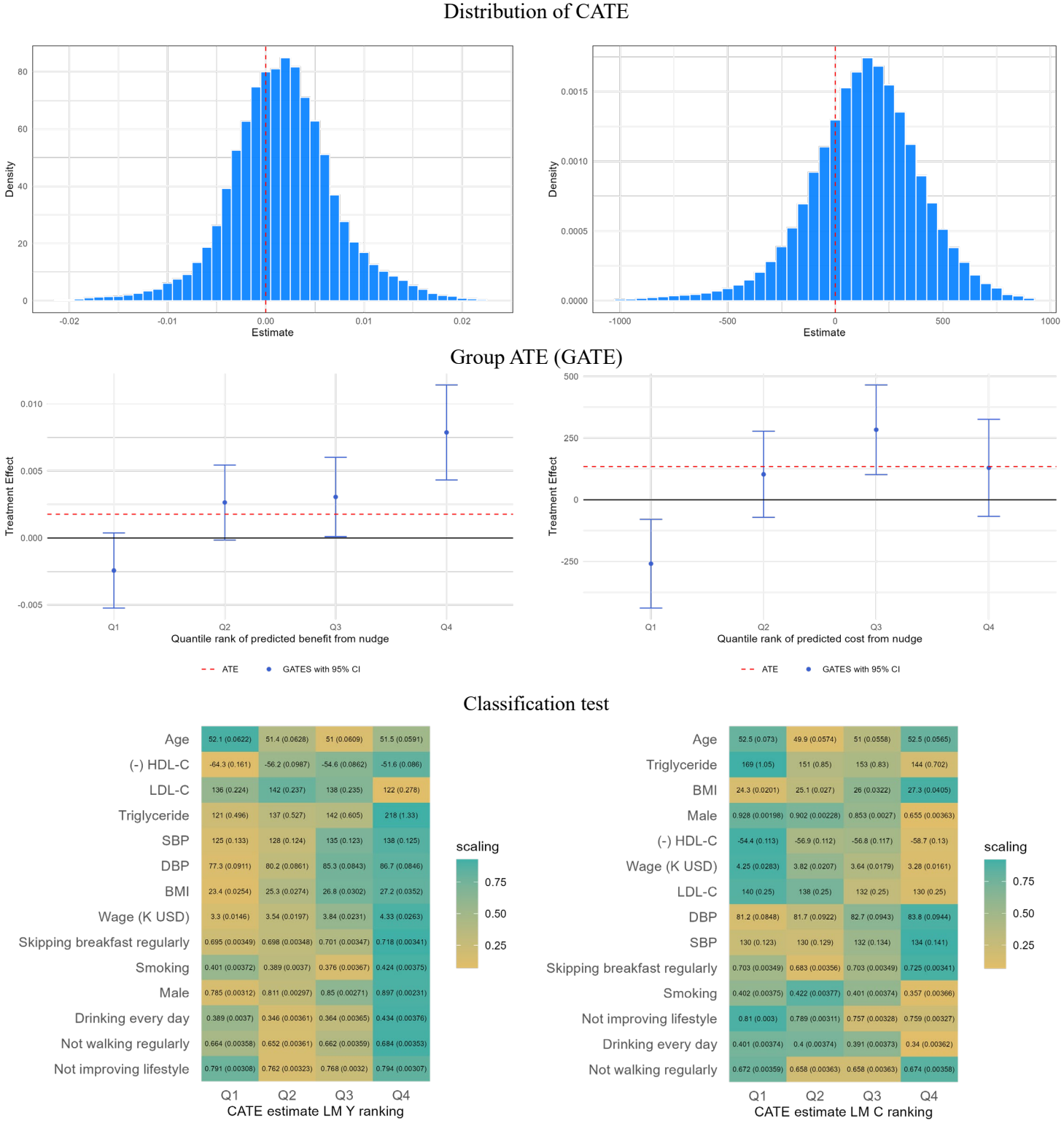
Figure B4—Relationship between CATEs for selections and benefits



Notes: The figure reports a binned scatter plot of the relationship between the estimated CATE for seletion D on the x-axis and the CATE for health benefits Y ($(-)\Delta$ LDL-C for panel A and $(-)\Delta$ SBP for panel B) divided by the CATE for D that measures the health return on compliers (treatment-on-treated) on the y-axis at the individual level.

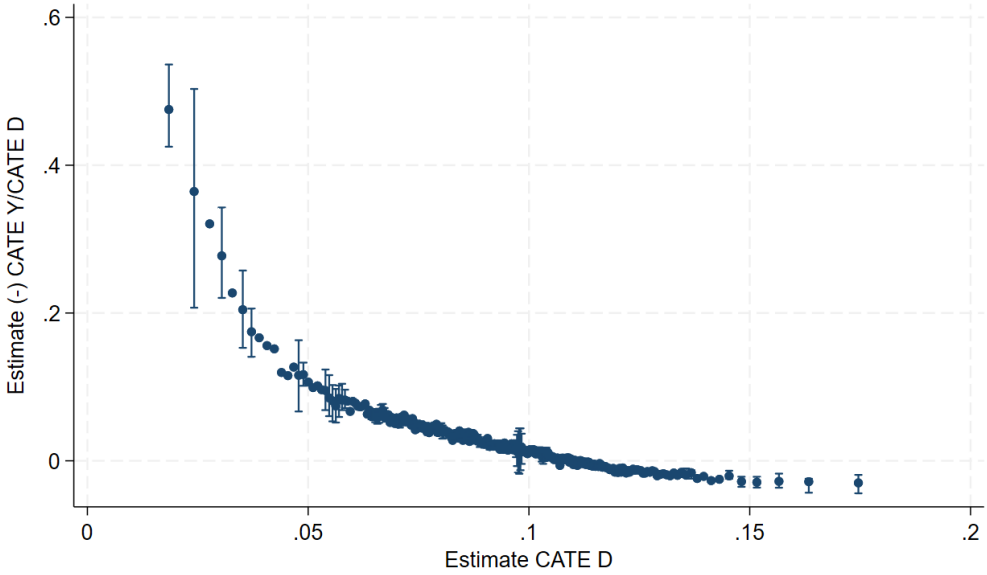
Figure B5—The CATEs for cumulative outcomes

A. (−)Δ 6-year cumulative mortality **B. 6-year cumulative outpatient spending**



Notes: The outcomes in panels A and B are the 6-year cumulative mortality from the index checkup and corresponding 6-year cumulative outpatient spending (in USD). 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 1,000, and the rep is 8,000.

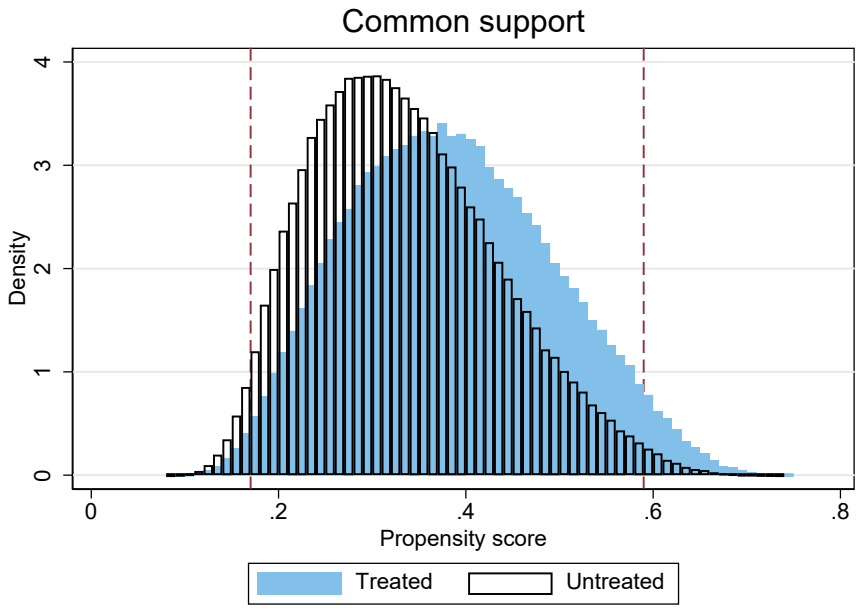
Figure B6—Relationship between CATEs for D (selection) and Y (6-year cumulative mortality rate)



Notes: The figure reports a binned scatter plot of the relationship between the estimated CATE for compliance D on the x-axis and the CATE for health outcome improvements Y ($(-)$6-year cumulative mortality rate) divided by the CATE for D that measures the health return on compliers (treatment-on-treated) on the y-axis at the individual level.

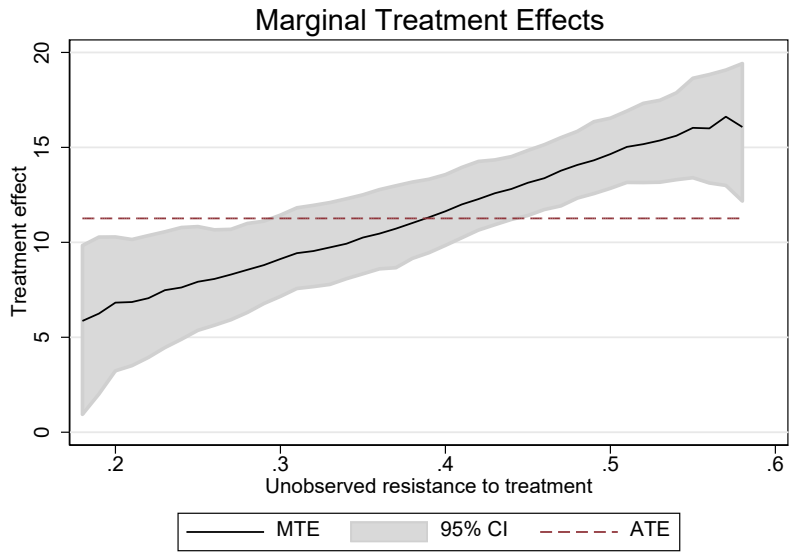
Appendix C: MTE

Figure C1—Overlap of propensity scores



Notes: The figure shows the overlap of propensity scores between treated and untreated. The two vertical dotted lines represent the 1 and 99 percentiles, which correspond to the propensity score of 0.18 and 0.58, respectively.

Figure C2—MTE on (−)ΔFBS (semi-parametric)



Notes: The outcome is the improvement in fasting blood sugar ((−)Δ FBS) between year $t+1$ and year t. Panel A plots the marginal treatment effect (MTE) and the 95% confidence intervals as a semi-parametrically estimated function of unobserved resistance to treatment. The dotted line represents the average treatment effect (ATE).

Table C1—Selection equation (logit)

Variables (at t)	(1)
$Z = 1(FBS \geq 126)$	0.0944***
	(0.0028)
$R = FBS - 126$	0.0089***
	(0.0005)
Age	0.0101***
	(0.0001)
LDL-C	-0.0013***
	(0.0000)
$(-)$HDL-C	0.0001***
	(0.0000)
SBP	-0.0015***
	(0.0001)
DBP	-0.0006***
	(0.0001)
BMI	0.0075***
	(0.0002)
Triglyceride	-0.0001***
	(0.0000)
log-likelihood	-282,660
Observations	458,380

Notes: This table shows marginal effects at the means of the covariates from logit estimation where the outcome is D_{t+1}, which takes one if there is any DM-related visit within a year after the index screening. We use the delta method to obtain standard errors and report them in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table C2—Observables in the MTE

	Quadratic	Semi-parametric
Variables	(1)	(2)
$R = FBS - 126$ (at t)	0.091**	0.092*
	(0.043)	(0.054)
Age (at t)	-0.112***	-0.111***
	(0.021)	(0.025)
LDL-C (at t)	0.036***	0.036***
	(0.003)	(0.004)
$(-)$HDL-C (at t)	0.034***	0.034***
	(0.004)	(0.003)
SBP (at t)	0.026***	0.025***
	(0.006)	(0.007)
DBP (at t)	0.006	0.006
	(0.008)	(0.007)
BMI (at t)	0.282***	0.282***
	(0.021)	(0.027)
Triglyceride (at t)	0.002***	0.002***
	(0.001)	(0.001)
Observations	458,380	458,380

Notes: This table shows the estimates of $\mu_1(x) - \mu_0(x)$—the observable part of MTE function—on the improvement in fasting blood sugar ($(-)\Delta$ FBS) between year $t+1$ and year t. The bootstrapped standard errors are in parentheses. Columns (1) and (2) are the cases of a quadratic function and a semi-parametrically estimated function of unobserved resistance to treatment.

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 were measured at the checkup a year after the index checkup.