NBER WORKING PAPER SERIES

MEAN REVERSION IN RANDOMIZED CONTROLLED TRIALS: IMPLICATIONS FOR PROGRAM TARGETING AND HETEROGENEOUS TREATMENT EFFECTS

Marcella Alsan
John Cawley
Joseph J. Doyle Jr.
Nicholas Skelley

Working Paper 33369
http://www.nber.org/papers/w33369

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2025

We thank Martin Munoz for excellent research assistance, Joshua Angrist, Susan Athey, Amanda Kowalski, Adrienne Sabety, Jesse Shapiro, and Sophie Sun for helpful comments and suggestions. We gratefully acknowledge funding support from the MIT Sloan Health Systems Initiative and the Jameel Poverty Action Lab. The project was pre-registered: AEA: AEARCTR-0004098 and ClinicalTrials.gov: NCT03718832. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Mean Reversion in Randomized Controlled Trials: Implications for Program Targeting and Heterogeneous Treatment Effects

Marcella Alsan, John Cawley, Joseph J. Doyle Jr., and Nicholas Skelley

NBER Working Paper No. 33369
January 2025
JEL No. C9, D04, I12

ABSTRACT

Eligibility criteria for interventions can induce an Ashenfelter Dip, and subsequent mean-reversion may result in improvement over time even absent the intervention. We investigate these dynamics for a food-as-medicine program to treat diabetes, where eligibility required elevated hemoglobin A1c (HbA1c). Both treatment and control groups experienced significant improvements in HbA1c, resulting in an estimated null effect. When we predict improvement using baseline characteristics, we find that subjects unlikely to improve on their own appear to benefit from the program. Our findings have implications for program targeting and estimating heterogeneous treatment effects.

Marcella Alsan
Kennedy School of Government
Harvard University
79 John F. Kennedy St.
Rubenstein Bldg R403
Cambridge, MA 02138
and NBER
marcella_alsan@hks.harvard.edu

John Cawley
2312 MVR Hall
Jeb E. Brooks School of Public Policy
and Department of Economics
Cornell University
Ithaca, NY 14853
and NBER
JHC38@cornell.edu

Joseph J. Doyle Jr.
MIT Sloan School of Management
100 Main Street, E62-516
Cambridge, MA 02142
and NBER
jjdoyle@mit.edu

Nicholas Skelley
Cornell University
Department of Economics
Uris Hall, 401A
109 Tower Rd.
Ithaca, NY 14853
ns977@cornell.edu

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/4098
https://clinicaltrials.gov/study/NCT03718832?cond=NCT03718832
A Dataverse entry is available at
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LZUTU0

1 Introduction

It is common to target interventions to subjects who have an extreme value of the lagged outcome. For example, a reading program might be targeted to students who performed poorly on a literacy assessment, or a job training program might be targeted to those who are unemployed. In healthcare, a "hotspotting" program intended to reduce health care spending was targeted to "superutilizers" with frequent hospital visits (Finkelstein, 2020). In the present study, we examine a food-as-medicine program that sought to improve diabetes control by targeting those with elevated HbA1c, which is a measure of average blood sugar over the past 2-3 months that is used to define diabetes status.[1]

When targeting programs in this way, there can be mean reversion; individuals with an elevated value of the lagged outcome may experience improvement over time even in the absence of treatment. A null result in an RCT could mean that the intervention truly does not work for anyone (the sharp null hypothesis), but it is also consistent with the intervention being effective for a subset of subjects whose outcomes would not improve on their own; i.e. those who would be durably eligible for the program. This raises the question of whether it is possible to identify which subjects with extreme lagged values of the outcome are likely to mean-revert, and which are likely to respond to the treatment. If it is possible, it could provide more clarity in the interpretation of RCT results and allow more cost-effective targeting of the treatment.

2 Setting: a Food as Medicine RCT

As described in more detail elsewhere (Doyle et al., 2024), the program we study is a clinic-based food-as-medicine program. It is designed for food-insecure patients with type 2 diabetes. The program is intensive, with patients visiting a clinic each week to collect enough healthy ingredients and recipes for ten meals for their entire household. The program is also comprehensive, as the clinic is staffed by a dietitian to provide consultations, a nurse for primary care, and a community health worker for social support. The program has no pre-determined end date, and typical engagement lasts approximately one year, with an average cost of approximately $2000 per year.

The trial took place at a large healthcare system at two sites in the mid-Atlantic region of the U.S. Eligiblity criteria included an HbA1c ≥ 8.0, and the primary outcome is HbA1c at 6

[1]The American Diabetes Association has the following guidelines: HbA1c $< 5.7\%$ is normal, $5.7\% \geq$ HbA1c $< 6.5\%$ is prediabetes, and HbA1c $\geq 6.5\%$ is diabetes (American Diabetes Association Professional Practice Committee, 2024).

months. A waitlist design was used, motivated by equity considerations. Consenting subjects were randomized, stratified by site and HbA1c category of above or below 9.5, to the treatment group (who started the program immediately) or the control group (who started the program in 6 months). During the first six months, the control group received (1) usual care, (2) a letter describing the locations of local food banks, and (3) the option to join the program after six months. We leverage Electronic Health Records, claims, and surveys to track both treatment and control groups before and after randomization. Subjects were compensated $50 for completing surveys and labs at 6 months and 12 months after trial entry.

As detailed in Doyle et al. (2024), those randomized to the treatment group did in fact accept the treatment; nearly all received healthy food, and 70% remained active through the first six months. Relative to the control group, the treatment group received in the first six months significantly more dietitan consultations, diabetes education trainings, and primary care such as foot exams. Survey data indicate that the treatment group reports eating healthier than the control group.

Figure 1 shows that both groups have similar, elevated levels of HbA1c at baseline, averaging roughly 10.3. At six months, the treatment group has experienced a substantial reduction in HbA1c to 8.8. In a study without a control group, this might be perceived as evidence of program effectiveness. However, the control group experienced the same reduction, with the result that the program has no detectable effect on HbA1c. The estimate is reasonably precise, with the 95% confidence interval ruling out improvements greater than 0.3, while interventions typically seek an improvement in HbA1c of 1.0 (Davies, 2022).

3 Mean Reversion and Heterogeneous Treatment Effects

In a stationary distribution, the change in the outcome across successive draws will be negatively correlated with the base. In our current setting where eligibility draws from an extreme value of the population distribution, we may observe subsequent draws tending toward the mean due to statistical noise in the initial draw or behavior change prompted by the high reading.

We can examine mean reversion in historical data from the same healthcare provider in the same geographic areas by selecting patients who have an HbA1c above a given threshold at a point in time (a mock eligibility date). We selected these samples using the dates November 1, 2014 and November 1, 2015. Figure 2 shows HbA1c levels measured within 3-month windows relative to the mock eligibility date. These individuals experience the inverse of an "Ashenfelter's dip" (Ashenfelter, 1978) - a rise in their HbA1c before the eligibility

determination. They then soon experience a decline in their HbA1c despite the entire sample receiving no intervention beyond usual care. The decline is steeper as the elgibility threshold increases from 8 to 11, a pattern where the baseline is negatively correlated with the change and passes statistical tests of mean reversion. This evidence raises the question of whether one could identify those likely to improve on their own *ex ante*. If so, one could target the program to those who may need additional support in order to achieve an improvement.

4 Conceptual Framework

Consider a simple setting in which success is defined by binary outcome, Y, such as a patient with elevated HbA1c getting it "under control" (e.g. below 8). Further, consider a binary treatment, D, which is randomly assigned as in an RCT with no compliance issues. There are four types of people defined by their potential outcomes, analogous to the LATE framework (Imbens and Angrist, 1994; Joshua D. Angrist and Rubin, 1996):

$Y_1 = 1$ and $Y_0 = 1$	$\Delta Y = 0$	Always improvers
$Y_1 = 1$ and $Y_0 = 0$	$\Delta Y = 1$	Responders
$Y_1 = 0$ and $Y_0 = 0$	$\Delta Y = 0$	Never improvers
$Y_1 = 0$ and $Y_0 = 1$	$\Delta Y = -1$	Derailers

Always Improvers improve with or without the treatment, which includes those who mean revert; *Responders* improve only if they are treated; *Never Improvers* do not improve with or without the treatment; and *Derailers* would have improved in the absence of the treatment, but the treatment derails their improvement.

A monotonicity assumption that the program can only improve outcomes would preclude the existence of Derailers. Similar to the LATE framework, under that assumption we could estimate the share of Responders (i.e. the ATE) and mean characteristics of the Always Improvers. We can also estimate the mean characteristics of Responders, as long as their share is positive. When setting eligibility criteria for an RCT, or when evaluating heterogeneous treatment effects, it would be useful to predict those likely to be Always Improvers, as they cannot improve any further with the program.

5 Heterogeneous Treatment Effects in the Presence of Mean Reversion

5A. Baseline Outcome Trajectory

In the presence of mean reversion, a natural first approach to examine heterogeneous treatment effects is to examine differences across patients with different trajectories of the outcome prior to entering the trial. Using data from 12 months prior to trial entry, we calculate the average HbA1c for all tests taken more than three months prior to the test that determined program eligibility. We then calculate the change in HbA1c between that earlier average and the test that determined eligibility at baseline. When calculating terciles among the subsample where the pre-trial trajectory can be calculated, the middle (relatively stable) tercile ranges from -0.68 to +0.71, while the other terciles have patients who were on a steeper upward or downward trajectory.

For each tercile, Figure 4 reports ITT estimates from the pre-specified linear regression model of HbA1c at six months on the treatment indicator, strata controls and baseline HbA1c. We find that for those in the middle tercile - i.e. those who were relatively stable in their HbA1c prior to study enrollment - the program looks to have improved HbA1c by a substantial 1.4 points. In contrast to those with more volatile levels of HbA1c prior to entering the trial, those with a stable (while elevated) level of the outcome may benefit more from this program.

5B. Heterogeneous Treatment Effects using Predicted Mean Reversion

We estimated predicted improvement—i.e. change in HbA1c over time absent the treatment. Specifically, using only the control group we regressed change in HbA1c from baseline to 6 months on measures observed at baseline: demographics, indicators for taking the four most common diabetes medications, indicators for the top 50 diagnoses, and the patient's most recent biometrics for LDL cholesterol, triglycerides, weight and blood pressure using leave-out regressions (Abadie, Chingos and West, 2018). The characteristics that predicted improvement included baseline HbA1c—consistent with mean reversion—a diagnosis of heart disease or chronic renal failure in the prior year, and an indicator the subject was referred to the program by their primary care physician, a marker for active engagement via "usual care".

Figure 3 reports results for quartiles of predicted improvement in HbA1c. For those in the bottom two quartiles of predicted improvement (i.e. those less likely to mean revert during the study), the program looks to have been effective, with clinically meaningful reductions in HbA1c of 1 and 0.7 points, respectively. For those in the top two quartiles (i.e. those more

likely to mean revert), the program yields no further improvement in HbA1c beyond what patients achieve in the absence of the program.

Caution is warranted when interpreting these results, however, as they were not pre-specified, and the results are sensitive to the method of predicting the likelihood of improvement in the control group. Nevertheless, these findings provide a proof of concept when exploring RCT results for evidence of heterogeneous treatment effects along a particular dimension: those with different predicted improvements in the outcome of interest, especially in an environment characterized by mean reversion.

6 Conclusion

When program eligibility is based on an extreme measure of the primary outcome at baseline, the resulting sample selection may result in predictable dynamics in the subsequent outcome. In the context of mean reversion, this typically means that the subjects will improve over time, confounding pre-post comparisons. This highlights the need for a credible control group when estimating causal effects of an intervention. If the outcome is bounded, such as a binary outcome, mean reversion may make it particularly difficult to detect any program effectiveness.

If mean reversion is predictable, there are implications for targeting the program *ex ante* or exploring heterogeneous treatment effects *ex post*. Historical data can be used to estimate future changes in the "untreated outcomes" (Kowalski, 2022). Going forward, applying machine learning methods to this prediction problem could enable researchers and policymakers to design eligibility criteria that prioritize enrolling Responders and excluding Always Improvers—those who would not be durably eligible for the program. As with other algorithmic approaches, care would need to be taken to avoid inequities that could arise. Recognizing and addressing the dynamics of mean reversion can lead to more effective program design.

References

Abadie, Alberto, Matthew M. Chingos, and Martin R. West. 2018. "Endogenous Stratification in Randomized Experiments." *The Review of Economics and Statistics*, 100(4): 567–580.

American Diabetes Association Professional Practice Committee. 2024. "Diagnosis and Classification of Diabetes: Standards of Care in Diabetes-2024." *Diabetes Care*, 47(Suppl 1): S20–S42.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *The Review of Economics and Statistics*, 60(1): 47–57.

Davies, Melanie J. et al. 2022. "Management of Hyperglycemia in Type 2 Diabetes." *Diabetes Care*, 45(11): 2753–2786.

Doyle, Joseph, Marcella Alsan, Nicholas Skelley, Yutong Lu, and John Cawley. 2024. "Effect of an Intensive Food-as-Medicine Program on Health and Health Care Use: A Randomized Clinical Trial." *JAMA Internal Medicine*, 184(2): 154–163.

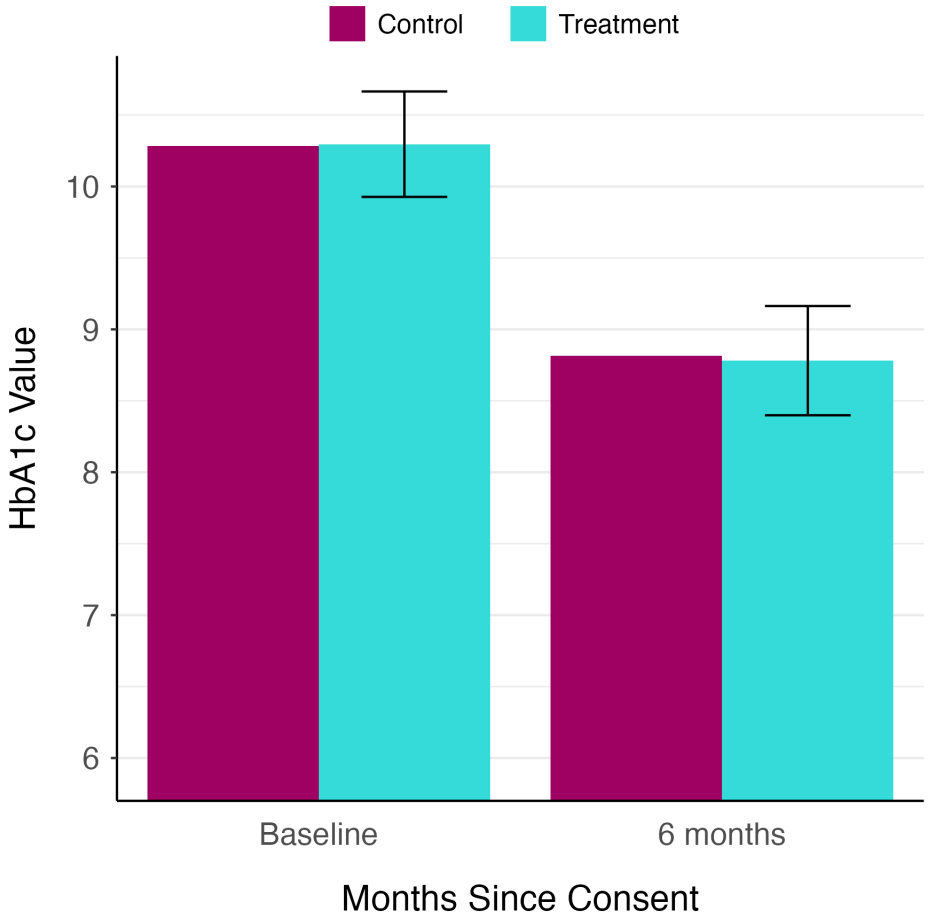
Finkelstein, A., Zhou A. Taubman S. Doyle J. 2020. "Health Care Hotspotting - A Randomized, Controlled Trial." *The New England journal of medicine*, 382(2): 152–162.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, 91(434): 444–455.

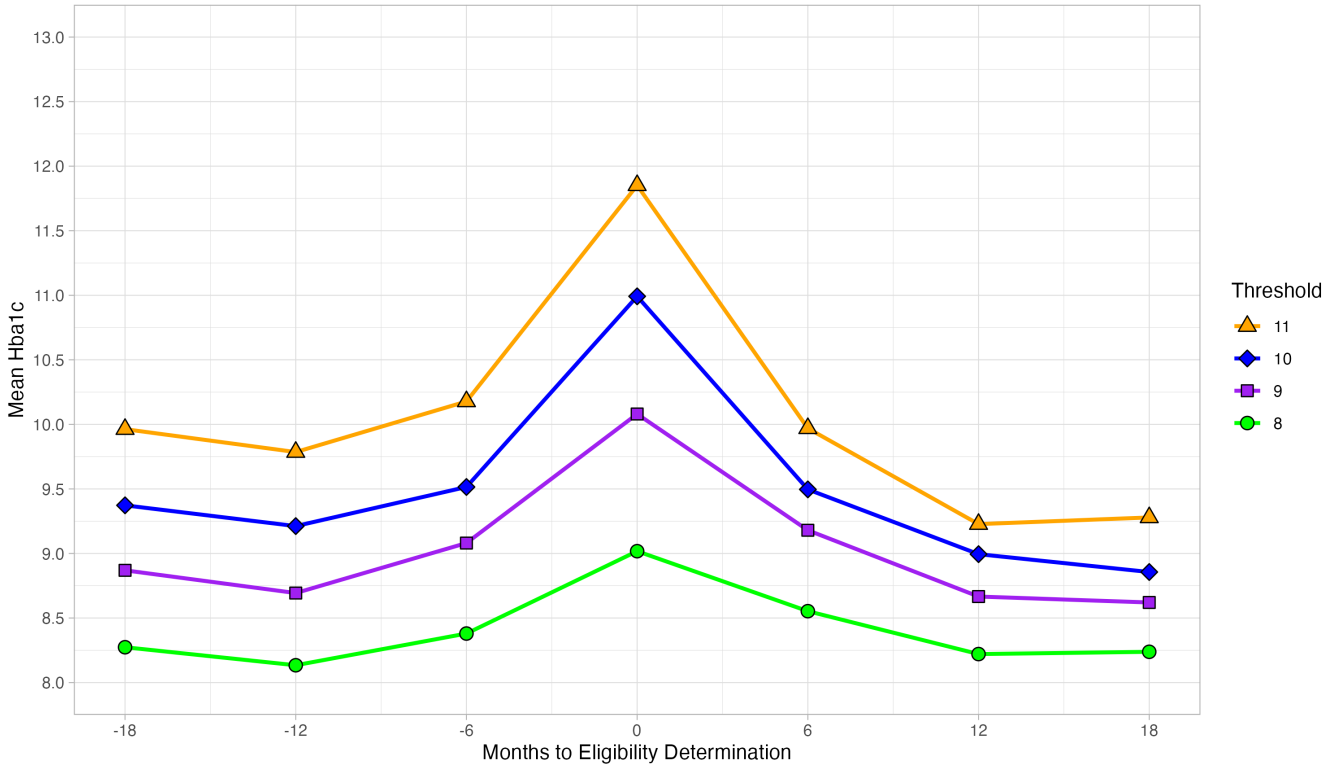
Kowalski, Amanda E. 2022. "Behaviour within a Clinical Trial and Implications for Mammography Guidelines." *The Review of Economic Studies*, 90(1): 432–462.

Figure 1: RCT Main Result: HbA1c Over Time



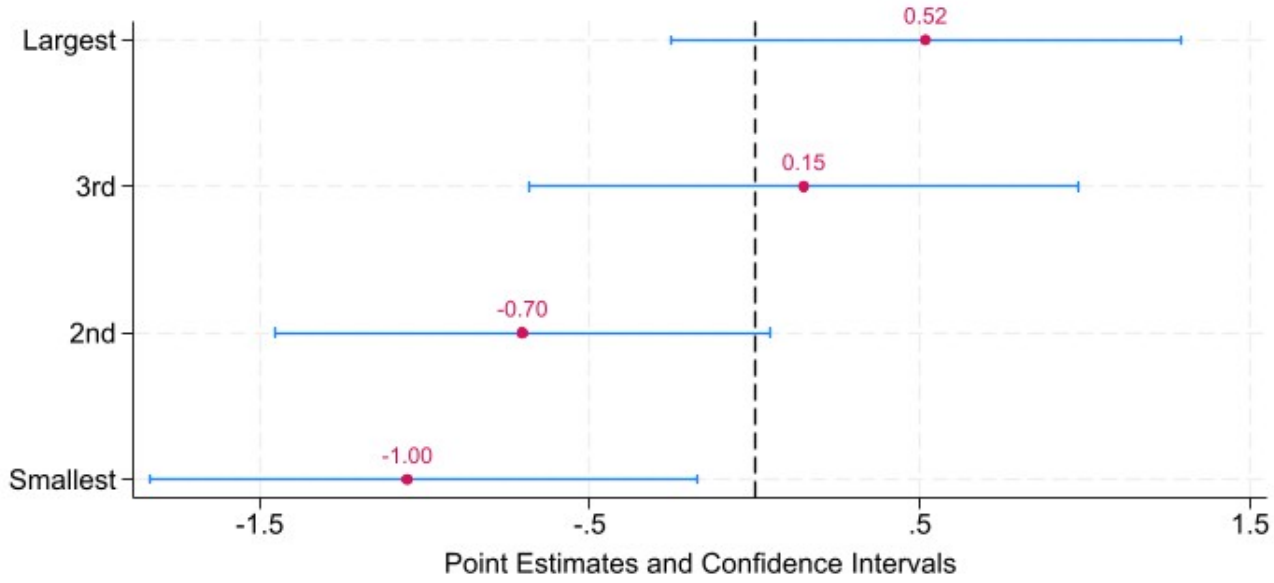
Note: This figure presents the primary outcome, HbA1c, at baseline, 6 months and 12 months separately for the treatment and control groups. N=349 at 6 months and N=325 at 12 months. HbA1c is a measure of average blood sugar over the past 2-3 months and is used to diagnose diabetes.

Figure 2: HbA1c Over Time Conditional on Crossing Thresholds



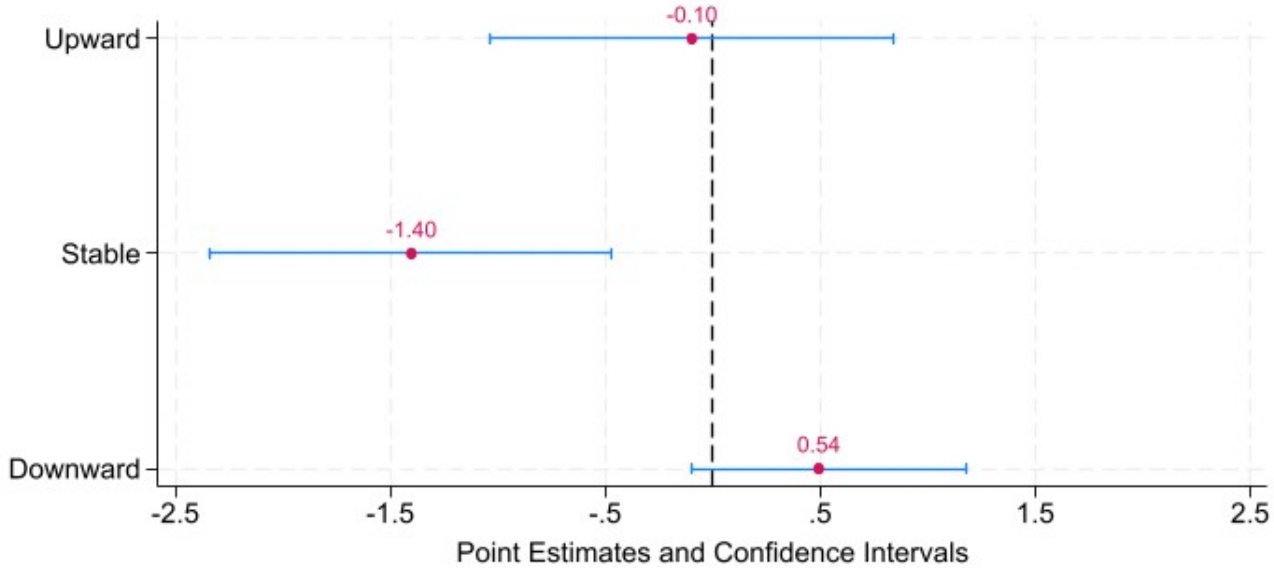
Note: This figure presents HbA1c in 6 month windows, radiating out from when an individual is observed with an HbA1c above a given threshold ranging from 8 to 11. For the four thresholds, N = 2993, 1701, 969, and 568 respectively. HbA1c is a measure of average blood sugar over the past 2-3 months and is used to diagnose diabetes.

Figure 3: ITT Estimates by Predicted HbA1c Improvement Quartile



Note: This figure presents the intent-to-treat (ITT) estimates by quartile of the predicted Change in HbA1c estimated using the control group as described in the text. N = 275 subjects with available data for the prediction model including baseline lab results for cholesterol, triglycerides, weight, and blood pressure. HbA1c is a measure of average blood sugar over the past 2-3 months and is used to diagnose diabetes.

Figure 4: ITT Estimates by Baseline HbA1c Trajectory



Note: This figure presents the intent-to-treat (ITT) estimates by tercile of the Change in HbA1c from an average HbA1c across all tests taken 3 months prior the baseline test to the baseline. N = 181 subjects whose prior trajectory can be calculated with available lab results.