THE LANCET
Diabetes & Endocrinology

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Baum A, Scarpa J, Bruzelius E, Tamler R, Basu S, Faghmous J. Targeting weight loss interventions to reduce cardiovascular complications of type 2 diabetes: a machine learning-based post-hoc analysis of heterogeneous treatment effects in the Look AHEAD trial. *Lancet Diabetes Endocrinol* 2017; published online July 12. http://dx.doi.org/10.1016/S2213-8587(17)30176-6.

Appendix

I. Details on underlying functioning of Causal Forest method

This paper applies recent advances in machine learning for causal inference to conduct a post-hoc analysis of a randomized controlled trial (RCT). The Action for Health in Diabetes ("Look AHEAD") clinical trial we focus on was halted early on the basis of a futility analysis.[1] However, we hypothesized that the null average treatment effect may mask clinically- and policy-relevant heterogeneity.

Causally interpreting post-hoc analyses of RCTs is challenging because investigators may test a large number hypotheses, but only report those with significant treatment effects. On the other hand, the small set of pre-specified hypotheses registered ex-ante by investigators may leave clinically useful relationships between interventions, outcomes, and subgroups undiscovered. Recognizing the limitations of conventional approaches to subgroup analyses, and the fact that many clinical trials will be underpowered to detect meaningful treatment variation, a number of newer approaches to identifying HTEs have been proposed.[2] These include a class of more data-driven predictive risk modeling tools such as Classification and Regression Trees (CART) which are typically most appropriate for early exploratory analyses.

Overview

The post-hoc analysis method we employ, called causal forest, extends classical recursive partitioning methods (e.g. random forest) to identify causally relevant subgroups defined by interactions of many variables, a combinatorial task for which human intuition and expertise is poorly suited. The initial, and conceptually important, step is to randomly split the data into two independent halves, using the first partition for hypothesis generation/tree construction (training data) and preserving the remainder of the data for statistically valid inference (testing data). The method first identifies subgroups with similar treatment effects in the training data, then tests the most promising heterogeneous treatment effect (HTE) hypotheses on the testing data to mitigate multiple testing concerns.

Identification of subgroups using the training data

To identify subgroups, we constructed an ensemble of causal trees [3], a type of decision tree. Decision trees are especially well-suited for identifying subgroups because they produce a partition of the sample in which subgroups share similar predictions or classifications that is not limited by model specification assumptions (as compared to several other approaches, e.g. [4] and [5]).In each causal tree, half the sample is randomly selected and its covariate space is sequentially partitioned into subspaces. Each split minimizes variation in the average treatment effect $\Delta y = \overline{y}_{treated} - \overline{y}_{control}$ within each subspace. A key output of each tree is the "variable importance" of each covariate, which reflects the covariate's inclusion in and order (depth) within the tree. Because the structure of a single tree depends on the training data, different training data may yield vastly different trees. To account for the high variance in any given tree, an ensemble of trees (a "forest") is often used. In this study, we constructed a forest of 1,000 trees and calculated the mean variable importance measure for 84 baseline covariates across the forest.

To generate testable HTE hypotheses from the results of the forest, we developed a heuristic to select the subgroups (leaves) most representative of the treatment effect heterogeneity identified by the forest. The heuristic had the following three criteria: (1) we identified trees split on covariates with the highest variable importance measures whose most important variable

was the covariate with the greatest variable importance measure across the forest; (2) from these trees, we identified the tree whose samples had the greatest variance in individualized treatment effects calculated by the causal forest, a surrogate of heterogeneity; (3) within this tree, each leaf with at least 10% of the total cohort sample size was considered for downstream analysis. This identifies a single tree with subgroups ("leaves") that reflects the forest's average output. Lastly, we prioritize HTE hypothesis testing on subgroups defined by the most important variables of the forest. These HTE hypotheses are tested on the half of the data that was preserved (testing data) when building this most representative tree.

Estimating HTE using the testing data

We tested two HTE hypotheses on the testing data. We calculated hazard ratios and 95% confidence intervals using likelihood-ratio tests from a Cox proportional-hazards regression, with a model containing terms for study-group assignment, a subgroup dummy, and their interaction. We additionally conducted robustness checks of the statistical significance of our findings using a bootstrapping procedure (**Figure A2**).

II. Exploratory analysis of mechanisms underlying heterogeneity

We investigated potential mechanisms through which differential intervention response across subgroups may have operated. We analyzed whether there were differences between treated and control participants in intermediate outcomes (e.g. traditional CVD outcomes) and process indicators (e.g. for intervention compliance) within each subgroup identified in the main analysis as experiencing differential long-term benefit from the intervention.

Recall, Subgroup 1 (15% of the overall sample; baseline HbA1C < 6•8% and baseline SF-36 General Health < 48) experienced no long-term benefit in terms of CVD-related morbidity and mortality from the intervention, whereas the remaining participants (85% of the overall sample) did. Using the testing dataset, we separately plot for Subgroup 1 vs. all remaining participants the monthly trends in the difference across treated and control participants to intermediate outcomes (**Figure A3**). We note suggestive evidence of less benefit in Subgroup 1 vs. all remaining participants over the short-term for HbA1c and self-reported mental health and over the long-term for blood pressure, with no evidence of heterogeneity for weight and self-reported general health. **Table A2** compares mean intervention compliance indicators among treated participants in Subgroup 1 vs. all remaining treated participants. Participants in Subgroup 1 reported fewer total minutes of exercise and fewer mean minutes of exercise in the first six months of the intervention year ($p < .05$) and the last six month of the intervention year ($p < .01$), suggesting differential compliance with the exercise components of the intervention.

Based on these exploratory analyses, the differential long-term intervention response across participants in Subgroup 1 vs. all remaining participants may have been driven by greater intermediate improvement in HbA1c, self-reported mental health, and blood pressure among those not in Subgroup 1, as well as by poorer intervention compliance among those in Subgroup 1. This suggests that those participants with less to gain in terms of HbA1c improvement who also had greater behavioral barriers to compliance were less likely to experience a long-term benefit from the intervention. These are exploratory analyses and the findings should be interpreted cautiously. The existing literature indicates that changes to HbA1c do not predict cardiovascular outcomes, though emerging data from new drug studies suggests the potential for cardiovascular risk reduction from improved HbA1c.[6] Therefore, the

use of HbA1c in our risk prediction is not necessarily causal, but rather could potentially be serving as an effective proxy measure for other factors, such as adherence.

Table A1. 84 baseline predictors from four major categories.

Number (corresponds to x-axis of Fig A1)	**Variable Name**	**Variable Label**	**Values**	**Category**
1	DIABETES_FAMILY	Family history of diabetes	Yes; No; Missing	History
2	CVDHIS	History of CVD	Yes; No; Missing	History
3	diab_dur	Self-Reported Duration of Diabetes (yrs)	numeric	History
4	met_syn_tot	Number of Metabolic syndrome criteria met	numeric [1,5]	History
5	MHAMP	Ever had amputation of feet or legs	Yes; No; Missing	History
6	MHANG	PTCA	Yes; No; Missing	History
7	MHAPNEA	Ever have sleep apnea	Yes; No; Missing	History
8	MHARTH	Ever had arthritis	Yes; No; Missing	History
9	MHBURN	Burning in legs/feet	Yes; No; Missing	History
10	MHCABG	CABG	Yes; No; Missing	History
11	MHCEND	Carotid endarterectomy	Yes; No; Missing	History
12	MHCRAMPS	Muscle cramps in legs/feet	Yes; No; Missing	History
13	MHDRY	Skin on feet so dry it cracks open	Yes; No; Missing	History
14	MHEMPH	Ever had emphysema	Yes; No; Missing	History
15	MHHURT	Legs hurt when walk	Yes; No; Missing	History
16	MHKDIS	Kidney disease	Yes; No; Missing	History
17	MHLOUD	How loud is snoring	Only slightly louder than heavy breathing	History
18	MHLOWER	Angioplasty of Lower extremity artery	Yes; No; Missing	History
19	MHMI	Myocardial Infarction (clinical)	Yes; No; Missing	History
20	MHNEUR	Had diabetic neuropathy	Yes; No; Missing	History
21	MHNUMB	Legs/feet numb	Yes; No; Missing	History
22	MHOFTEN	How often do you stop breathing during sleep	Sometimes (up to 2 nights a week)	History
23	MHPRCK	Prickling feelings in legs/feet	Yes; No; Missing	History
24	MHRETINA	Ever been told that diabetes affected the back of your eye	Yes; No; Missing	History
25	MHSENS	Feet too sensitive to touch	Yes; No; Missing	History
26	MHSLEEPY	How often do you feel overly sleepy during day	Never or rarely (1 day/month or less)	History
27	MHSNFREQ	How often do you snore	Do not snore any more	History
28	MHSNORE	Ever snored	Yes; No; Don't Know; Missing	History
29	MHSORE	Ever had open foot sore	Yes; No; Missing	History
30	MHSTPBTH	Ever stop breathing during sleep	Yes; No; Don't Know; Missing	History
31	MHSTROKE	Stroke	Yes; No; Missing	History
32	MHTELL	Tell hot from cold water	Yes; No; Missing	History

33	MHTOUCH	Hurt when bed covers touch skin	Yes; No; Missing	History
34	MHWALK	Sense feet when you walk	Yes; No; Missing	History
35	MHWEAK	Feel weak all over	Yes; No; Missing	History
36	MHWORSE	Symptoms worse at night	Yes; No; Missing	History
37	glucosemgDL	Fasting glucose (mgdl)	numeric	Molecular Traits
38	hba1cpct	Hemoglobin A1c %	numeric	Molecular Traits
39	hdlchlmgDL	HDL cholesterol (mgdl)	numeric	Molecular Traits
40	ldlchlmgDL	LDL cholesterol (mgdl)	numeric	Molecular Traits
41	screatmgDL	Serum Creatinine (mg/dL)	numeric	Molecular Traits
42	TrigmgDL	Triglycerides (mgdl)	numeric	Molecular Traits
43	ualbmgDL	Urine albumin (mg/dL)	numeric	Molecular Traits
44	ucreatmgDL	Urine creatinine (mg/dL)	numeric	Molecular Traits
45	avgabi	Ankle Brachial Index (average)	numeric	Molecular Traits
46	baselinewgt_kg	Computed Mean Weight in Kg	numeric	Molecular Traits
47	bmi	Calculated BMI (SAS)	numeric	Molecular Traits
48	bssbp_mean	Computed Mean Systolic Blood Pressure	numeric	Molecular Traits
49	eshgt_mean	Computed Mean Height in Cm	numeric	Molecular Traits
50	prob1yr_adj	Adjusted Framingham Risk at One Year	numeric	Molecular Traits
51	psothmd	Other Diabetes Meds (at Screening)	Yes; No; Missing	Molecular Traits
52	RHOVCNT	Do you still have both ovaries, one, or none?	None; Only one; Both; None; Not sure; Missing	Molecular Traits
53	RHPSTOP	How did your periods stop?	Naturally; By Surgery; Other; Missing	Molecular Traits
54	FEMALE	Female	Yes; No	Sociodemographic
55	RACEVAR	Race/Ethnicity	African American/Black (not Hispanic)	Sociodemographic
56	age	Age from Screening A or Prescreen	numeric	Sociodemographic
57	SDESTAT	Unemployed or laid off	Yes; No; Missing	Sociodemographic
58	SDHOUSE	Keeping house or rasing children full-time	Yes; No; Missing	Sociodemographic
59	SDINC1	Money Earned in past 12 months	Under $10,000; $10,000-$19,999; $20,000-$29,999; $30,000-$39,999; $40,000-$49,999; $50,000-$59,999; $60,000-$69,999; $70,000-$79,999; >$80,000; Missing	Sociodemographic
60	SDINC2	Net worth	0-$500; $501-$1,000; $1,001-$5,000; $5,001-$10,000; $10,001-$25,000; $25,001-$50,000; $50,001-$100,000; $100,001-$250,000; $250,001-$500,000; $500,001-$1,000,000; $1,000,001 or more; Missing	Sociodemographic
61	SDLOOKWRK	Looking for work	Yes; No; Missing	Sociodemographic
62	SDMARSTAT	Marital Status	Divorced	Sociodemographic

63	SDSTUDENT	Full or part-time student	Yes; No; Missing	Sociodemographic
64	SDUSUAL	Source of care	Community health center	Sociodemographic
65	SDWORKFT	Working full time for pay	Yes; No; Missing	Sociodemographic
66	SDWORKPT	Working part-time for pay	Yes; No; Missing	Sociodemographic
67	SMOKING	Smoking status	Never; Past; Present; Missing	Behavior
68	total_alcohol	Total alcohol consumption (oz/wk) (sum of beer, wine, and liquor per week)	numeric	Behavior
69	BINGE_EAT	Binge eating	Yes; No; Missing	Behavior
70	ACR	Albumin-Creatinine Ratio	numeric	Behavior
71	genhlth	SF-36 General Health	numeric	Behavior
72	ht	SF-36 Reported Health Transition	numeric	Behavior
73	mcs	SF-36 Mental Component Summary	numeric	Behavior
74	mentalhlth	SF-36 Mental Health	numeric	Behavior
75	pain	SF-36 Bodily Pain	numeric	Behavior
76	pcs	SF-36 Physical Component Summary	numeric	Behavior
77	phyfunc	SF-36 Physical Functioning	numeric	Behavior
78	roleemot	SF-36 Role-Emotional	numeric	Behavior
79	rolephy	SF-36 Role-Physical	numeric	Behavior
80	socialfunc	SF-36 Social Functioning	numeric	Behavior
81	vitality	SF-36 Vitality	numeric	Behavior
82	EDEAT2HR	Eat really big amount in short time?	Yes; No; Missing	Behavior
83	EDEAT6MO	Eat really big amount of food?	Yes; No; Missing	Behavior
84	bd_t	Beck Total score (entered at clinic)	numeric	Behavior

Table A2. Comparison of intervention compliance indicators for "Subgroup 1" (baseline HbA1C < 6•8% and baseline SF-36 General Health < 48) and Remaining Participants

	Subgroup 1	Remaining Participants	Difference
Months 0 – 6			
Number of sessions attended	22.8	22.5	0.657
Percentage of expected visits attended	91.5	89.9	3.032*
Total minutes of exercise	3899.2	4465.9	-494.9**
Total meal replacements	238.6	248.6	-6.999
Mean minutes of exercise	149.6	171.5	-19.11**
Mean meal replacements	9.5	9.9	-0.297
Months 7 – 12			
Number of sessions attended	12.9	13.1	-0.0151
Total minutes of exercise in months	3405.1	4381.4	-924.0***
Total meal replacements in months	121.7	140.9	-19.25**
Mean minutes of exercise in months	152.1	193.7	-39.12***
Mean meal replacements in months	5.9	6.5	-0.611*
Number of sessions	35.5	35.3	0.702
Months 0 – 12			
Percentage of expected visits attended	84.5	84.1	1.671

* p<0.05, ** p<0.01, *** p<0.001

Table A3. Comparison of intervention compliance indicators for "Subgroup 2" (baseline HbA1C < 6•8% and baseline SF-36 General Health >= 48) and Remaining Participants

	Subgroup 2	Remaining Participants	Difference
Months 0 – 6			
Number of sessions attended	22.6	23.2	-0.5
Percentage of expected visits attended	90.8	93.2	-2.2
Total minutes of exercise	4075.0	3731.7	343.7
Total meal replacements	241.1	239.5	3.3
Mean minutes of exercise	156.6	141.9	14.3
Mean meal replacements	9.6	9.4	0.2
Months 7 – 12			
Number of sessions attended	12.9	13.3	-0.2
Total minutes of exercise in months	3658.9	3418.6	286.2
Total meal replacements in months	126.1	125.9	1.4
Mean minutes of exercise in months	163.7	147.6	17.9
Mean meal replacements in months	6.1	5.9	0.2
Number of sessions	35.3	36.4	-0.8
Months 0 – 12			
Percentage of expected visits attended	84.0	86.6	-1.9

* p<0.05, ** p<0.01, *** p<0.001

Table A4. Baseline characteristics of Subgroup 1 and Remaining Participants across treated and control groups in the testing data

	Subgroup 1		Remaining Participants	
	Treated	**Control**	**Treated**	**Control**
Age	60.04	60.73	58.46	58.51
	(6.63)	(6.83)	(6.71)	(6.59)
Female	1.58	1.57	1.60	1.60
	(0.50)	(0.50)	(0.49)	(0.49)
Race	3.25	3.34	3.19	3.21
	(1.19)	(1.11)	(1.19)	(1.17)
CVD History	1.11	1.12	1.15	1.14
	(0.31)	(0.33)	(0.35)	(0.35)
Smoking	2.51	2.53	2.55	2.55
	(0.55)	(0.58)	(0.59)	(0.57)
Diabetes duration	5.22	5.32	7.36	7.38
	(6.05)	(6.25)	(7.10)	(6.36)
Weight (kg)	98.80	96.91	101.60	102.30
	(19.59)	(17.31)	(19.85)	(19.57)
BMI	34.86	34.56	36.16	36.39
	(6.08)	(5.15)	(6.01)	(6.02)
HbA1C	6.23	6.22	7.54	7.61
	(0.40)	(0.39)	(1.12)	(1.16)
Systolic BP	128.30	128.00	128.50	130.30
	(17.35)	(16.96)	(17.07)	(17.17)
HDL	45.17	43.99	43.06	43.33
	(12.17)	(11.63)	(12.20)	(11.72)

LDL	111.50	112.90	113.20	111.90
	(31.00)	(28.47)	(32.50)	(31.91)
Triglyceride (mg/dl)	165.30	163.60	190.80	183.20
	(103.80)	(100.30)	(121.20)	(127.00)
Observations	279	253	938	981

Table A5. Baseline characteristics of Subgroup 2 and Remaining Participants across treated and control groups in the testing data

	Subgroup 2		Remaining Participants	
	Treated	**Control**	**Treated**	**Control**
Age	58.24	57.92	58.92	59.15
	(6.50)	(6.40)	(6.76)	(6.74)
Female	1.53	1.64	1.60	1.59
	(0.50)	(0.48)	(0.49)	(0.49)
Race	3.40	3.30	3.17	3.23
	(1.07)	(1.13)	(1.20)	(1.17)
CVD History	1.15	1.09	1.14	1.15
	(0.36)	(0.28)	(0.34)	(0.35)
Smoking	2.52	2.52	2.54	2.55
	(0.54)	(0.58)	(0.59)	(0.58)
Diabetes duration	5.76	5.36	7.05	7.22
	(7.74)	(5.71)	(6.78)	(6.46)
Weight (kg)	106.00	104.70	100.10	100.60
	(19.65)	(19.81)	(19.73)	(19.09)
BMI	37.20	37.35	35.64	35.78
	(5.90)	(6.39)	(6.04)	(5.78)
HbA1C	6.22	6.29	7.41	7.50
	(0.39)	(0.35)	(1.14)	(1.19)
Systolic BP	127.10	129.00	128.70	130.00
	(16.64)	(17.47)	(17.21)	(17.09)
HDL	42.71	45.59	43.68	43.11
	(11.89)	(12.18)	(12.27)	(11.58)

LDL	108.40	111.30	113.50	112.30
	(33.26)	(30.34)	(31.93)	(31.38)
Triglyceride (mg/dl)	187.70	158.10	184.50	182.70
	(122.70)	(83.25)	(117.10)	(127.30)
Observations	171	179	1,046	1,055

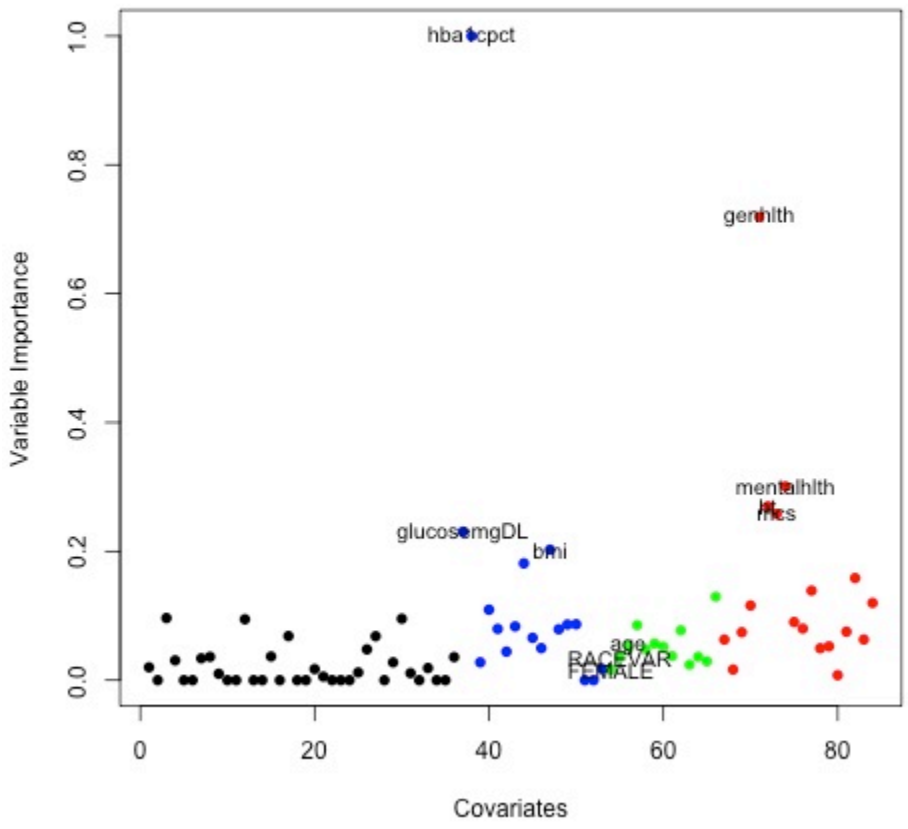

Figure A1: Mean Normalized Variable Importance across the Causal Forest
This figure shows the mean normalized variable importance for all covariates in the data. For ease of interpretation, we labeled covariates that were clear outliers, including HbA1C, self-reported general health (as reported on the SF-36), and others, as well as covariates commonly included in pre-analysis plans, including age, race, and gender. Medical history is labeled in black, laboratory values in blue, sociodemographic variables in green, and behavioral health variables in red.

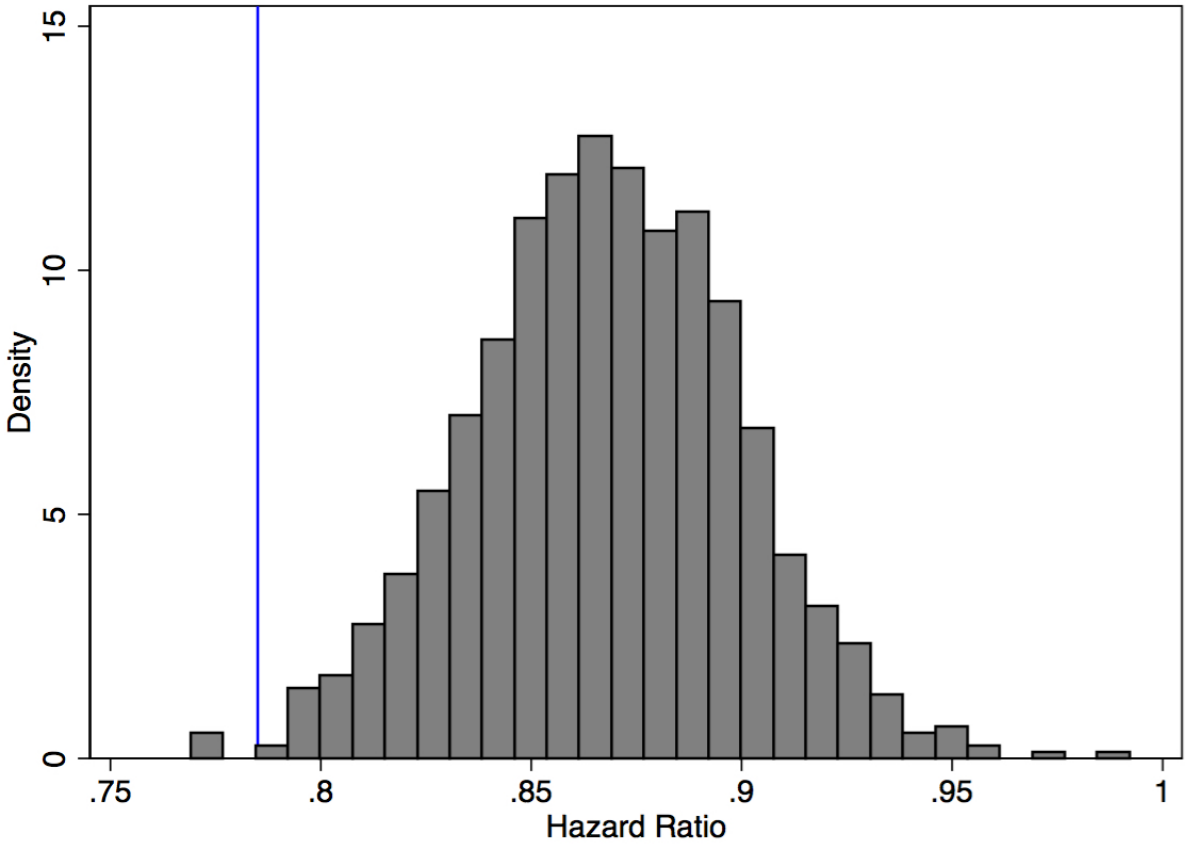

Figure A2. Results of bootstrapping procedure
Hazard ratios using likelihood-ratio tests from a Cox proportional-hazards regression across 1,000 replications of random subsamples of 84% of the testing data, with the blue line representing the hazard ratio result from a Cox proportional-hazards regression using the non-permuted data for the subgroup of participants not included in Subgroup 1 (the participants not in Subgroup 1 are those with both baseline HbA1C < 6•8% and baseline general health >= 48 or those with baseline HbA1C >= 6•8%).

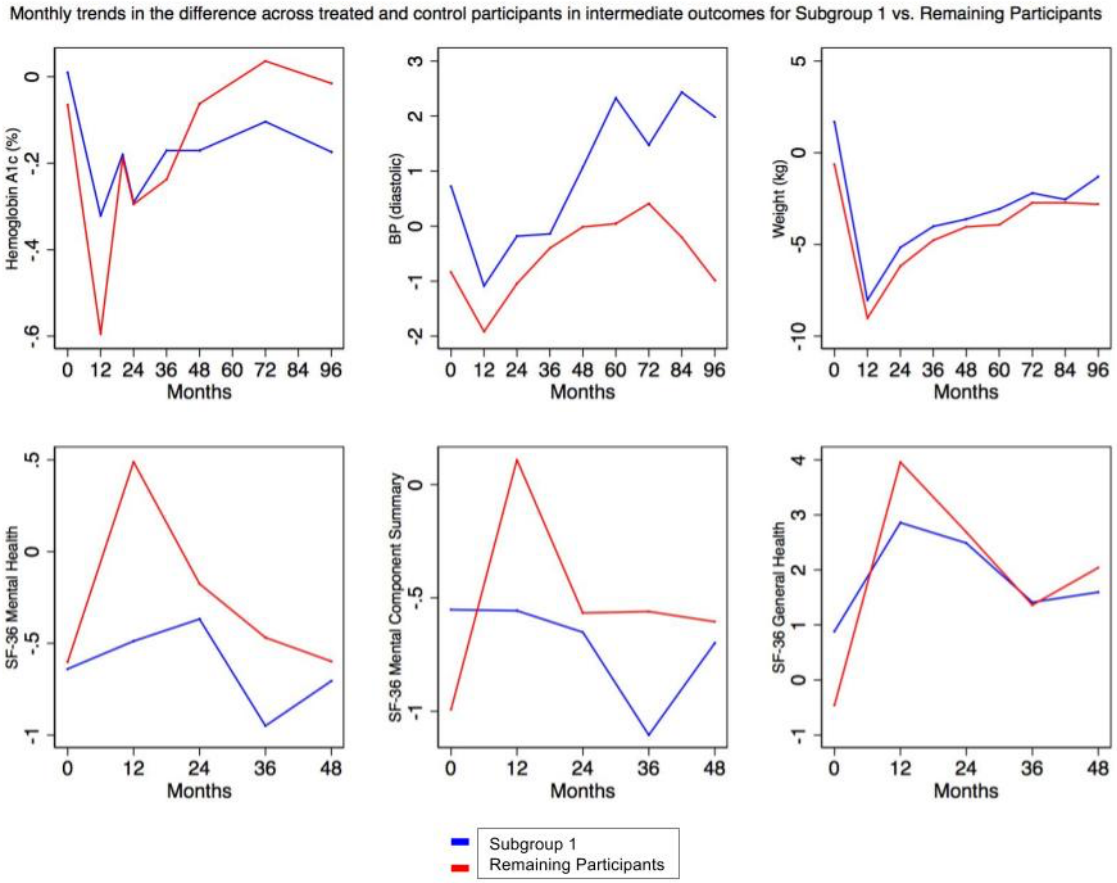

Figure A3. **Exploratory analysis of mechanisms underlying heterogeneity in Subgroup 1** Using the testing dataset, we separately plot for "Subgroup 1" (baseline HbA1C < 6•8% and baseline SF-36 General Health < 48) vs. all the remaining participants the monthly trends in the difference across treated and control participants for HbA1c, blood pressure, weight, self-reported mental health, and self-reported general health. Recall, Subgroup 1 (15% of the overall sample) experienced no long-term benefit in terms of CVD-related morbidity and mortality from the intervention, whereas the remaining participants (85% of the overall sample) did.

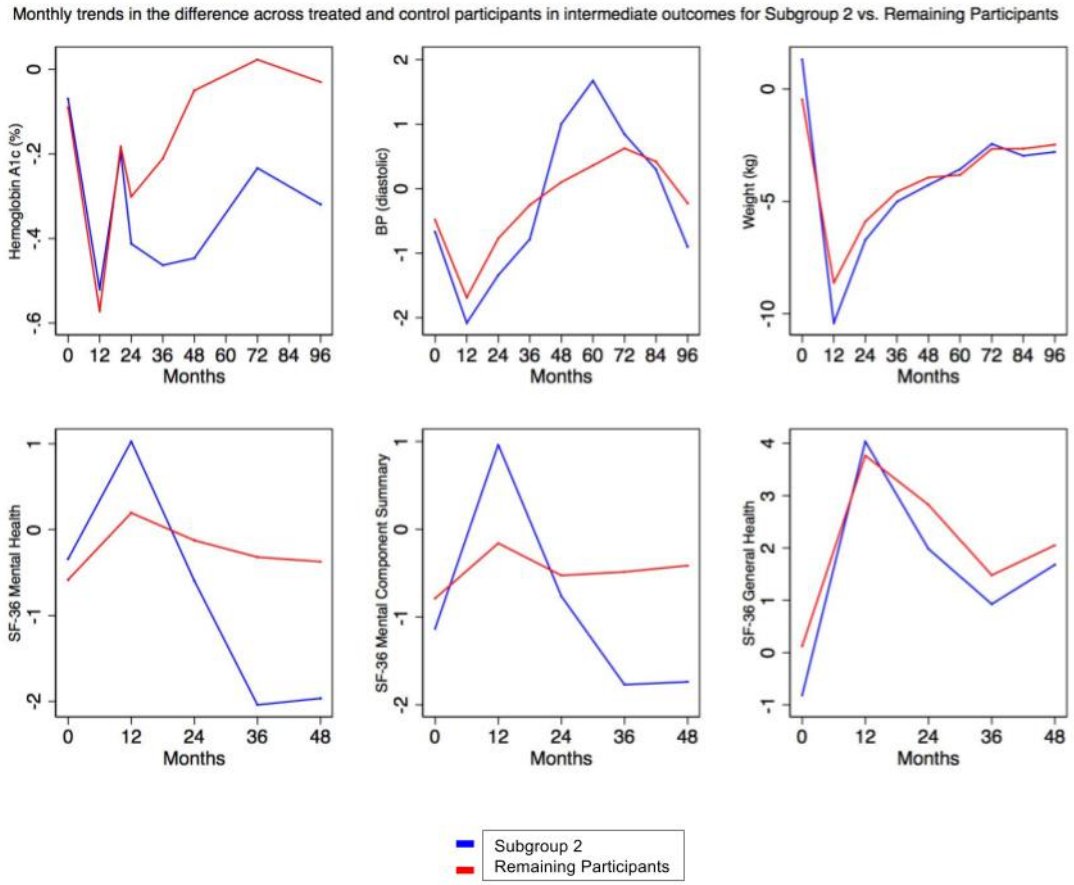

Figure A4. **Exploratory analysis of mechanisms underlying heterogeneity in Subgroup 2** Using the testing dataset, we separately plot for "Subgroup 2" (baseline HbA1C < 6•8% and baseline SF-36 General Health >= 48) vs. all the remaining participants the monthly trends in the difference across treated and control participants for HbA1c, blood pressure, weight, self-reported mental health, and self-reported general health. Recall, Subgroup 2 experienced greater long-term benefit in terms of CVD-related morbidity and mortality from the intervention than remaining participants.

References:

1. Look ARG, Wing RR, Bolin P, et al. Cardiovascular effects of intensive lifestyle intervention in type 2 diabetes. *N Engl J Med* 2013; **369**(2): 145-54.
2. Willke RJ, Zheng Z, Subedi P, Althin R, Mullins CD. From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC Med Res Methodol* 2012; **12**: 185.
3. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci U S A* 2016; **113**(27): 7353-60.
4. Breiman L, Friedman J, Stone CJ, RA O. Classification and regression trees. Boca Raton, FL: CRC press; 1984.
5. Breiman L. Random forests. *Machine Learning* 2001; **45**(1): 5-32.
6. Zinman B, Wanner C, Lachin JM, et al. Empagliflozin, Cardiovascular Outcomes, and Mortality in Type 2 Diabetes. *New England Journal of Medicine* 2015; **373**(22): 2117-28.