Annals of Epidemiology xxx (xxxx) xxx

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Generalizability of heterogeneous treatment effects based on causal forests applied to two randomized clinical trials of intensive glycemic control

Sridharan Raghavan, MD, PhD[a,b,c,*], Kevin Josey, PhD[a,d], Gideon Bahn, PhD[e], Domenic Reda, PhD[e], Sanjay Basu, MD, PhD[f], Seth A. Berkowitz, MD, MPH[g,h], Nicholas Emanuele, MD[e,1], Peter Reaven, MD[i], Debashis Ghosh, PhD[d]

[a] Department of Veterans Affairs Eastern Colorado Healthcare System, Aurora, CO
[b] Division of Hospital Medicine, University of Colorado School of Medicine, Aurora, CO
[c] Colorado Cardiovascular Outcomes Research Consortium, Aurora, CO
[d] Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO
[e] Department of Veterans Affairs Hines VA Hospital, Hines, IL
[f] Center for Primary Care, Harvard Medical School, Boston, MA
[g] Division of General Medicine and Clinical Epidemiology, University of North Carolina School of Medicine, Chapel Hill, NC
[h] Cecil G. Sheps Center for Health Services Research, University of North Carolina at Chapel Hill, Chapel Hill, NC
[i] Department of Veterans Affairs Phoenix VA Medical Center, Phoenix, AZ

ARTICLE INFO

Article history:
Received 17 February 2021
Revised 4 June 2021
Accepted 9 July 2021
Available online xxx

Keywords:
Generalizability, Glycemic control, Causal forests, Heterogeneous treatment effects.

Abbreviations: ACCORD, Action to Control Cardiovascular Risk in Diabetes Study
BMI, Body mass index
eGFR, Estimated glomerular filtration rate
HbA1c, Hemoglobin A1c
HGI, Hemoglobin glycation index
HTE, Heterogeneous treatment effects
VADT, Veterans Affairs Diabetes Trial

ABSTRACT

Purpose Machine learning is an attractive tool for identifying heterogeneous treatment effects (HTE) of interventions but generalizability of machine learning derived HTE remains unclear. We examined generalizability of HTE detected using causal forests in two similarly designed randomized trials in type II diabetes patients.

Methods We evaluated published HTE of intensive versus standard glycemic control on all-cause mortality from the Action to Control Cardiovascular Risk in Diabetes study (ACCORD) in a second trial, the Veterans Affairs Diabetes Trial (VADT). We then applied causal forests to VADT, ACCORD, and pooled data from both studies and compared variable importance and subgroup effects across samples.

Results HTE in ACCORD did not replicate in similar subgroups in VADT, but variable importance was correlated between VADT and ACCORD (Kendall's tau-b 0.75). Applying causal forests to pooled individual-level data yielded seven subgroups with similar HTE across both studies, ranging from risk difference of all-cause mortality of -3.9% (95% CI -7.0, -0.8) to 4.7% (95% CI 1.8, 7.5).

Conclusions Machine learning detection of HTE subgroups from randomized trials may not generalize across study samples even when variable importance is correlated. Pooling individual-level data may overcome differences in study populations and/or differences in interventions that limit HTE generalizability.

Published by Elsevier Inc.

Introduction

Machine learning methods are powerful tools for predicting outcomes with ever-increasing accuracy. More recently, there has also been interest in using machine learning to identify heterogeneous treatment effects (HTE) of clinical interventions, particularly in settings in which univariate subgroup analyses are unrevealing [1–7]. However, the replicability of subgroup effects detected using

Declaration of interests: The authors declare the following financial interests and/or personal relationships which may be considered as potential competing interests: Sridharan Raghavan reports financial support was provided by American Heart Association. Debashis Ghosh reports financial support was provided by National Science Foundation. Medications and financial support for the VADT Study were provided by Sanofi, GlaxoSmithKline, Novo Nordisk, Roche, Kos Pharmaceuticals, Merck, and Amylin. This manuscript describes a secondary analysis of data from the VADT study but did not receive any financial or material support from the above companies.

* Corresponding author. Rocky Mountain Regional VA Medical Center, 1700 North Wheeling Street, Medicine Service (111), Aurora, CO 80045. Tel.: 415-254-3563
E-mail address: Sridharan.raghavan@cuanschutz.edu (S. Raghavan).

[1] Deceased.

https://doi.org/10.1016/j.annepidem.2021.07.003
1047-2797/Published by Elsevier Inc.

machine learning methods across study samples is not often tested as multiple studies investigating similar intervention-outcome effects are rarely available. Moreover, the absence of replicability could arise from differences in study design or study population characteristics even when the underlying machine learning method is valid.

One context in which multiple similarly designed intervention studies have been performed is the evaluation of the effect of glycemic control intensity on mortality and diabetes-related complications in type II diabetes patients. Three randomized trials compared clinical outcomes in type II diabetes patients treated to an intensive versus standard glycemic control target [8–10]. None of the three studies found an association of glycemic control intensity with the primary outcome of cardiovascular disease events, and one, the Action to Control Cardiovascular Risk in Diabetes (ACCORD) Study, found an association of intensive glycemic control with higher all-cause mortality [8–10]. To evaluate the mortality findings further, using the causal forests machine learning approach, three variables were identified that defined subgroups in the ACCORD study with different risk for mortality in response to intensive glycemic control [11,12]. Whether the HTE of intensive glycemic control in ACCORD generalize remains unexplored.

We evaluated the generalizability of HTE across two similarly designed studies in distinct diabetes patient populations. We used individual-level data from the ACCORD study and the Veterans Affairs Diabetes Trial (VADT), two of the three aforementioned randomized trials of intensive versus standard glycemic control, to examine HTE identified in ACCORD in the VADT, to compare variable prioritization by causal forests between the two trials, and to determine if HTE could be identified that generalize across both studies.

Material and methods

Study samples

Individual-level data from two randomized clinical trials were included in this study. The ACCORD and VADT studies have been described in detail previously [8,9]. We opted to analyze these trials as both studies were publicly funded and US-based with individual-level data publicly available (ACCORD) or accessible to US Department of Veterans Affairs investigators through a data use agreement (VADT). Both studies included adults with advanced type II diabetes and a hemoglobin A1c (HbA1c) $\geq$7.5%. The VADT study enrolled participants from December 2000 –May 2003, and follow-up continued through May 2008. Median follow-up time in the VADT study was 5.6 years. The ACCORD study intentionally enrolled participants at high cardiovascular risk from January–June 2001 and from February 2003–October 2005, and follow-up in the ACCORD study continued until June 2009. An interim evaluation in December 2007 suggested harm from intensive glycemic control, prompting discontinuation of that treatment arm; as a result, ACCORD study participants had a median on-protocol follow-up time of 3.7 years and a median total follow-up time of 4.9 years. Both studies randomized participants to receive intensive or standard glycemic control but the definitions of those targets differed. In the VADT study, different doses of oral diabetes medications and different thresholds for adding insulin were used between treatment arms to achieve a goal of at least 1.5% lower HbA1c in participants randomized to intensive control compared to standard control. In the ACCORD study, participants received open-label diabetes treatment to a target of HbA1c less than 6% for intensive glycemic control versus 7%–7.9% for standard glycemic control. In this secondary analysis, we included data from all 1791 VADT study participants and 10,251 ACCORD study participants. The Colorado Multiple Institutional Review Board and local VA Research and Development Committee provided human subjects oversight and approval of the study.

Outcome

The primary outcome was all-cause mortality as ascertained in each of the original studies. Major adverse cardiovascular events, defined as fatal or non–fatal myocardial infarction or stroke, was a secondary outcome.

Predictors

We included baseline variables that were common to the two studies, including patient demographics, comorbidities, diabetes and cardiovascular disease medications, and laboratory values (*Table 1*). Estimated glomerular filtration rate (eGFR) was calculated using the Modification of Diet in Renal Disease Study equation [*13*]. Hemoglobin glycation index (HGI) was estimated as the residual between measured HbA1c and HbA1c predicted by regressing on fasting glucose [14]. As the basis for this analysis was a prior HTE examination in the ACCORD study [11], we used the regression equation of HbA1c on fasting glucose from ACCORD study participants to estimate HGI. Body mass index (BMI) was calculated as the weight in kilograms divided by the height in meters squared.

Statistical approach

To evaluate previously published HTE subgroups from the ACCORD study [11], VADT study participants were separated into one of four subgroups identified by a representative tree found in the previous article. We then estimated the crude average treatment effect of all-cause mortality within each subgroup. To address differences in participant characteristics between the VADT and ACCORD study samples and thus improve external validation of the ACCORD HTE study [11], we weighted participants in the VADT based on their likelihood of being sampled for the ACCORD study and re-estimated the crude average treatment effect of all-cause mortality within each of the four subgroups in the weighted VADT study sample. This causes the VADT participants to appear more like the ACCORD participants based on the sample moments of the marginal covariate distributions. The weights themselves are reflective of the inverse odds of sampling and are estimated directly using a method of moments estimator [15].

Next, we fit causal forests on both the VADT and ACCORD samples with the goal of identifying variables that contribute the most to HTE. We used causal forests [12,16,17] for this analysis rather than alternative machine learning algorithms to facilitate direct comparability with previously published analysis of the ACCORD study [11]. Briefly, the causal forests method builds a series of decision trees by randomly splitting the study population and ranking a random subset of predictor variables based on their modification of the treatment effect (defined as the risk difference in all-cause mortality between intensive and standard glycemic treatment in our case) in the subsample. Each tree is then tuned in the second study subsample to identify cut-points for each variable that maximizes between-subgroup treatment effects in the terminal subgroups at the bottom of the tree. Variable importance can then be assessed across all trees in the forest based on their relative position in a tree which corresponds to their influence on HTE. When applying causal forests to the ACCORD and VADT study data, both forests contained 5000 trees and a minimum node size of approximately 5% of the total sample size. Again, these parameters were selected for comparability with the prior HTE analysis

Table 1
Study population characteristics

	ACCORD		VADT		ACCORD + VADT	
	Standard control	Intensive control	Standard control	Intensive control	Standard control	Intensive control
	N = 5123	N = 5128	N = 899	N = 892	N = 6022	N = 6020
Age, mean (SD)	62.8 (6.7)	62.8 (6.6)	60.3 (8.6)	60.5 (8.8)	62.4 (7.0)	62.4 (7.0)
Sex, n female (%)	1969 (38.4)	1983 (38.7)	26 (2.9)	26 (2.9)	1995 (33.1)	2009 (33.4)
Race, n (%)						
Black	956 (18.7)	997 (19.4)	147 (16.4)	152 (17.0)	1103 (18.3)	1149 (19.1)
Hispanic	379 (7.4)	358 (7.0)	136 (15.1)	155 (17.4)	515 (8.6)	513 (8.5)
HbA1c (% [mmol/mol]), mean (SD)	8.3 (1.1) [67 [9]])	8.3 (1.1) [67 [9]]	9.4 (1.6) [79 [13]])	9.4 (1.5) [79 [13]]	8.5 (1.2) [69 [10]]	8.5 (1.2) [69 [10]]
Glucose (mg/dL), mean (SD)	175.7 (56.4)	174.7 (55.9)	205.9 (69.0)	203.5 (67.8)	180.2 (59.5)	179.0 (58.7)
Hemoglobin glycation index (unitless), mean (SD)	-0.07 (0.9)	-0.08 (1.0)	0.8 (1.4)	0.8 (1.4)	0.06 (1.1)	0.05 (1.1)
Total cholesterol (mg/dL), mean (SD)	183.3 (41.6)	183.3 (42.1)	184.7 (52.7)	181.6 (40.4)	183.5 (43.5)	183.1 (41.8)
Triglycerides (mg/dL), mean (SD)	189.4 (148.6)	190.9 (148.2)	222.8 (351.8)	200.8 (161.8)	194.4 (193.5)	192.4 (150.3)
LDL cholesterol (mg/dL), mean (SD)	104.9 (33.8)	104.9 (34.0)	108.2 (34.0)	107.0 (30.9)	105.4 (33.9)	105.2 (33.6)
HDL cholesterol (mg/dL), mean (SD)	41.9 (11.5)	41.8 (11.8)	35.8 (10.4)	36.2 (9.9)	41.0 (11.5)	41.0 (11.7)
Creatinine (mg/dL), mean (SD)	0.9 (0.2)	0.9 (0.2)	1.0 (0.2)	1.0 (0.2)	0.9 (0.2)	0.9 (0.2)
eGFR (mL/min/1.73m^2), mean (SD)	91.3 (28.4)	90.8 (25.8)	87.5 (22.6)	87.3 (24.2)	90.7 (27.7)	90.3 (25.6)
ALT (mg/dL), mean (SD)	27.7 (14.9)	27.5 (17.4)	31.9 (17.4)	30.8 (15.2)	28.3 (15.3)	28.0 (17.1)
SBP (mm Hg), mean (SD)	136.5 (17.2)	136.2 (17.0)	131.8 (16.8)	131.4 (16.6)	135.8 (17.2)	135.5 (17.1)
DBP (mm Hg), mean (SD)	75.0 (10.7)	74.8 (10.7)	76.1 (10.2)	76.0 (10.4)	75.2 (10.6)	75.0 (10.6)
BMI (kg/m^2), mean (SD)	32.2 (5.4)	32.2 (5.4)	31.2 (4.4)	31.3 (4.4)	32.1 (5.3)	32.1 (5.3)
Diabetes duration (y), mean (SD)	10.9 (7.6)	10.7 (7.6)	11.5 (7.2)	11.5 (7.8)	11.0 (7.6)	10.9 (7.6)
Insulin use, n (%)	1832 (35.8)	1750 (34.1)	467 (51.9)	466 (52.2)	2299 (38.2)	2216 (36.8)
Sulfonylurea use, n (%)	2707 (52.9)	2767 (54.0)	561 (62.4)	529 (59.3)	3268 (54.3)	3296 (54.8)
Metformin use, n (%)	3285 (64.1)	3269 (63.7)	632 (70.3)	605 (67.8)	3917 (65.1)	3874 (64.4)
Glinide use, n (%)	131 (2.6)	126 (2.5)	4 (0.4)	5 (0.6)	135 (2.2)	131 (2.2)
Acarbose use, n (%)	45 (0.9)	50 (1.0)	16 (1.8)	20 (2.2)	61 (1.0)	70 (1.2)
Thiazolidinedione use, n (%)	1125 (22.0)	1133 (22.1)	171 (19.0)	166 (18.6)	1296 (21.5)	1299 (21.6)
History of amputation, n (%)	106 (2.1)	111 (2.2)	27 (3.0)	28 (3.1)	133 (2.2)	139 (2.3)
History of eye surgery, n (%)	1169 (22.9)	1119 (21.9)	150 (18.3)	152 (18.9)	1319 (22.3)	1271 (21.5)
Current smoker, n (%)	607 (11.8)	640 (12.5)	145 (16.2)	154 (17.3)	752 (12.5)	794 (13.2)
History of MI, n (%)	803 (15.7)	787 (15.3)	170 (19.0)	166 (18.6)	973 (16.2)	953 (15.8)
History of stroke, n (%)	325 (6.3)	305 (5.9)	41 (4.6)	56 (6.3)	366 (6.1)	361 (6.0)
History of CHF, n (%)	245 (4.8)	249 (4.9)	48 (5.3)	61 (6.8)	293 (4.9)	310 (5.2)
History of angina, n (%)	560 (10.9)	608 (11.9)	166 (18.5)	167 (18.7)	726 (12.1)	775 (12.9)
Prior coronary revascularization, n (%)	556 (10.9)	615 (12.0)	183 (20.4)	182 (20.4)	739 (12.3)	797 (13.2)

Abbreviations: ACCORD = action to control cardiovascular risk in diabetes study; VADT = veterans affairs diabetes trial; HbA1c = hemoglobin A1c; DBP = diastolic blood pressure; SBP = systolic blood pressure; eGFR = estimated glomerular filtration rate; BMI = body mass index; ALT = alanine amino transferase; HDL = high-density lipoprotein; LDL = low-density lipoprotein; MI = myocardial infarction; CHF = congestive heart failure

of the ACCORD study [11], and the 5% minimum node size was additionally selected to ensure that any detected HTE would potentially impact treatment decisions for a substantial number of diabetes patients. Each tree is fit using an honest splitting and estimation approach described briefly above [12,16,17] from random samples representing half of the stratified samples, respectively. Furthermore, to avoid overfitting, each tree only considers half of the covariates for splitting. These covariates are randomly selected from the set of covariates identified in Table 1. To evaluate which variables contribute the most to HTE, we employed a variable importance statistic included in the *grf* package in R which generates a weighted average of importance for each variable defined as

$$Imp(X_j) = \frac{\sum_{k=1}^{K} A_{jk}/(k^2 B_k)}{\sum_{k=1}^{K} 1/k^2}.$$

Here K is the maximum depth over all the causal trees, A_{jk} is the total number of splits of the variable X_j at depth k, where j is an index of covariates included in the algorithm and B_k is the total number of splits at depth k over every tree in the forest [16,17]. With the outputted variable importance statistics, we computed Kendall's tau-b between the two causal forests to evaluate the level of concordance of variables contributing to HTE.

Table 2
Replication of heterogeneous treatment effects of all-cause mortality from the ACCORD study in the VADT study

	ACCORD (Basu et al [11])			VADT unweighted			VADT weighted to ACCORD		
Subgroup	N(%)	Risk difference	95% CI	N	Risk difference	95% CI	N	Risk difference	95% CI
1	877(8.6)	-2.3%	-4.5, -0.2	140(7.8)	2.2%	-6.0, 10.4	NA	-6.7%	-19.7, 6.3
2	1717(16.7)	0.7%	-1.6, 3.1	192(10.7)	3.4%	-7.3, 14.2	NA	-0.8%	-10.0, 8.5
3	4678(45.6)	0.9%	-0.4, 2.1	517(28.9)	2.4%	-2.9, 7.7	NA	-0.2%	-2.8, 2.5
4	2529(24.7)	3.7%	1.5 6.0	940(52.5)	-0.8%	-4.8, 3.1	NA	1.3%	-5.1, 7.7

Subgroup 1: Hemoglobin glycation index (HGI) $<$ 0.44, Body mass index (BMI) $<$ 30kg/m^2, Age less than 61 years
Subgroup 2: HGI $<$ 0.44, BMI $<$ 30kg/m^2, Age $\geq$ 61 years
Subgroup 3: HGI $<$ 0.44, BMI $\geq$ 30kg/m^2
Subgroup 4: HGI $\geq$ 0.44
Risk difference: negative values indicate lower mortality in intensive glycemic control arm; positive values indicate higher mortality in intensive glycemic control arm.

We then pooled the ACCORD and VADT samples and fit causal forests on the combined sample using the same model parameters used for each study separately. We used the five most important variables to construct a single representative causal tree using honest cross-validation and once again requiring at least 5% of the total sample in every terminal node [18].

All analyses were conducted in R (version 3.5.3, R Foundation for Statistical Computing, Vienna, Austria). Statistical code is available upon request.

Results

Data from 10,251 ACCORD study participants and 1791 VADT study participants were included in this analysis. The ACCORD study included a larger proportion of women and a smaller proportion of individuals of Hispanic ancestry (*Table 1*). Compared to VADT study participants, ACCORD study participants at baseline were older on average, had lower HbA1c, were less likely using insulin, and less frequently had history of myocardial infarction, congestive heart failure, angina, and prior coronary revascularization (*Table 1*). The average treatment effects – risk differences – of intensive glycemic control on all-cause mortality were 1.2% (95% Confidence Interval [CI] 0.2, 2.3) in the ACCORD study and 0.9% (95% CI -2.1, 3.9) in the VADT study, with positive values indicating increased mortality in the intensive treatment arm.

HTE analysis using causal forests applied to the ACCORD study [11] found that HGI, BMI, and age could be used to divide the sample into four subgroups in which the intervention effect on all-cause mortality ranged from a risk difference of -2.3% (95% CI: -4.5, -0.2), indicating benefit from intensive glycemic control, in individuals with low glycemic variability (indicated by low HGI), BMI below the obese range, and of younger age (<61 years) to –3.7% (95% CI: 1.5, 6.0) in individuals with high glycemic variability (indicated by high HGI) (*Table 2*). When the VADT study sample was divided into four subgroups using the variables and cut-points derived in ACCORD, the same pattern of association of intensive glycemic control with all-cause mortality was not observed (risk difference of mortality of 2.2% [95% CI: -6.0, 10.4] and -0.8% [95% CI: -4.8, 3.1] in the two most extreme subgroups; *Table 2*). After reweighting the VADT sample to balance the likelihood of sampling between VADT and ACCORD, the trend in the risk differences of mortality between intensive and standard glycemic control across subgroups was similar to that observed in the ACCORD study, but with confidence intervals that consistently included the null (risk difference of mortality of -6.7% [95% CI: -19.7, 6.3] and 1.3% [95% CI: -5.1, 7.7] in the two most extreme subgroups; *Table 2*).

Next, we applied causal forests to ACCORD and VADT study data separately and after pooling data from both studies to compare variable importance for defining HTE of glycemic control on all-cause mortality. As in prior work, causal forests applied to the ACCORD study prioritized HGI, age, and BMI most highly for HTE of intensive glycemic control on all-cause mortality [11]. Of the ten most highly prioritized variables when applying causal forests to the pooled study data, the majority were also among the most highly prioritized variables when performing the analysis in each contributing study (*Table 3*). When an indicator variable for study (ACCORD or VADT) was included in the causal forest analysis of the pooled data, the study indicator variable had an importance score of 0.002 or 25th out of 46 variables. Comparing variable importance in ACCORD and VADT, the Kendall's tau-b correlation coefficient was 0.717 (*Fig. 1*).

Table 3
Variable importance scores and ranks for top ten variables influencing heterogeneous treatment effects of intensive glycemic control on all-cause mortality after pooling ACCORD and VADT study data

Variable	ACCORD + VADT		ACCORD		VADT	
	Score	Rank	Score	Rank	Score	Rank
Age	0.106	1	0.087	2	0.079	2
HGI	0.066	2	0.094	1	0.038	9
DBP	0.066	3	0.063	5	0.044	8
eGFR	0.0.06	4	0.058	6	0.069	3
BMI	0.06	5	0.068	4	0.018	15
HbA1c	0.052	6	0.0.68	3	0.021	14
Triglycerides	0.045	7	0.04	13	0.052	5
LDL cholesterol	0.044	8	0.044	10	0.088	1
Total cholesterol	0.044	9	0.04	12	0.055	4
Glucose	0.044	10	0.049	8	0.047	7

Abbreviations: ACCORD = action to control cardiovascular risk in diabetes study; VADT = veterans affairs diabetes trial; HGI = hemoglobin glycation index; DBP = diastolic blood pressure; eGFR = estimated glomerular filtration rate; BMI = body mass index; HbA1c = hemoglobin A1c; LDL = low-density lipoprotein

Using the five most highly prioritized variables from the causal forest analysis of the pooled ACCORD and VADT study data (age, HGI, diastolic blood pressure, eGFR, and BMI) to generate a representative causal tree yielded a summary tree that utilized only four of the five candidate variables (diastolic blood pressure, BMI, eGFR, and age) to split the pooled sample into seven subgroups (*Fig. 2*). Subgroup two, those with relatively normal diastolic blood pressure (≥65 mm Hg), at most class one obesity, younger age (<61 years), with low-normal or early-stage chronic kidney disease - comprising 11.5% of the pooled sample, demonstrated lower all-cause mortality from intensive glycemic control (risk difference of -3.0% [95% CI: -5.2, -0.8]; *Fig. 2*). Subgroup two also demonstrated consistent direction of effect of intensive glycemic control on all-cause mortality in the ACCORD (risk difference -2.9% [95% CI: -5.3, -0.5]) and VADT (risk difference -3.1% [95% CI: -8.4, 2.1]) studies but with 95% confidence intervals that included the null in VADT likely due to a smaller relative sample size. There was directional consistency of decreased major adverse cardiovascular events, the trial primary outcome, from intensive glycemic control in Subgroup two across the pooled sample, the ACCORD study, and the VADT study, though the 95% confidence intervals did not exclude the null (risk difference of -3.2% [95% CI: -6.3, 0.1] in pooled

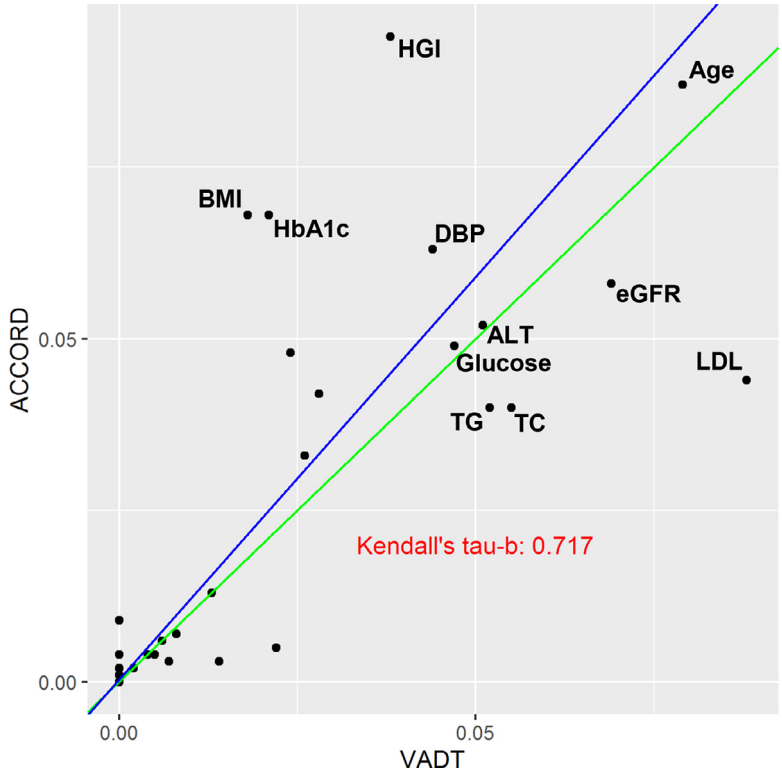


Fig. 1. Correlation of variable importance rank from causal forest in the VADT and ACCORD studies. Variable importance score in VADT study on x-axis and in ACCORD study on y-axis. Perfect correlation represented by green line; actual correlation represented by blue line. Key variables indicated (HbA1c = hemoglobin A1c; HGI = hemoglobin glycation index; DBP = diastolic blood pressure; eGFR = estimated glomerular filtration rate; BMI = body mass index; TG = triglycerides) (Color version of the figure is available online.)

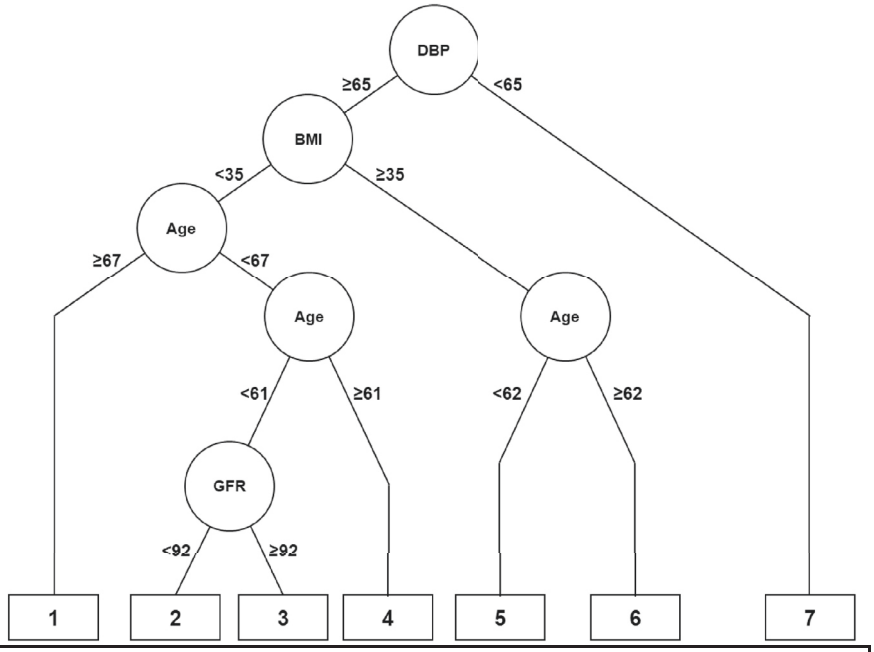
sample, -2.8% [95% CI: -6.2, 0.5] in ACCORD, and -4.2% [95% CI: -11.6, 3.2] in VADT; *Fig. 2*). A second subgroup comprising 14.8% of the pooled sample (Subgroup one, *Fig. 2)* defined by relatively normal diastolic blood pressure, at most class one obesity, and older age (≥67 years) also had lower all-cause mortality associated with intensive glycemic control but the 95% confidence intervals did not exclude the null in the pooled sample (risk difference -2.4% [95% CI -5.3, 0.6]), in the ACCORD study (risk difference -2.6% [95% -5.7, 0.5]), or in the VADT study (risk difference -1.7% [95% CI -10.0, 6.7]).

Three subgroups (subgroups four,six, and seven) demonstrated higher all-cause mortality from intensive glycemic control in the pooled sample. Subgroup four was defined by diastolic blood pressure ≥65 mm Hg, BMI below 35 kg/m^2, and a narrow age range of 61 –67 years, and the effect of intensive glycemic control on all-cause mortality differed between the ACCORD (risk difference 2.7% [95% CI: 0.5, 5.0]) and VADT (risk difference -1.1% [95% CI: -10.1, 8.0]) studies (*Fig. 2*). Among individuals with diastolic blood pressure ≥65 mm Hg, Subgroup six was defined by individuals with class two or greater obesity and older age (≥62 years). In contrast to Subgroup four, Subgroup six individuals exhibited consistent effects across the ACCORD (risk difference 4.5% [95% CI: 0.8, 8.1]) and VADT (risk difference 9.5% [95% CI: -5.8, 24.8]) studies, though with the confidence interval including null in VADT possibly owing to smaller sample size (*Fig. 2*). Subgroup seven, defined by low diastolic blood pressure (<65 mm Hg), had consistently increased mortality associated with intensive glycemic control in the ACCORD and VADT samples (risk difference of 4.7% [95% CI: 1.8, 7.5] in the pooled ACCORD+VADT sample, 3.8% [95% CI: 0.9, 6.7] in ACCORD, and 10.9% [95% CI: 1.2, 20.7] in VADT; *Fig. 2*).

Conclusions

A summary causal tree defining HTE subgroups for all-cause mortality associated with intensive glycemic control from one randomized trial did not generalize to a second randomized trial with a similar design. Weighting participants in the second trial based on resemblance to participants in the first improved replicability of HTE subgroup effects, suggesting differences in study samples at least partially contribute to differential subgroup effects. When performing identical analyses on both study samples, variable importance for defining HTE using causal forests was similar in each of the two studies, and a summary causal tree with several consistent HTE subgroups across both studies could be generated by applying causal forests to pooled individual-level data from the two trials. Taken together, the results suggest that applying causal forests to define HTE in a single study may yield results that cannot be directly applied to a second study or second population. This limitation may be due to spurious associations that would not be replicable under any conditions, to varying imbalance of prognostic factors in subgroups from different studies inducing confounding in the subgroup analyses [19,20], to sensitivity of the algorithm to user-defined parameters of the causal forests method, to sensitivity of the algorithm to differences in study populations, or to sensitivity of the algorithm to differences in study and/or intervention design even when seemingly similar. To overcome the latter two potential limitations, our results support the value of replication in multiple studies and/or pooling individual-level data when possible.

Much attention has recently focused on the challenges of reproducibility and replicability of machine learning applications in health care [21–23]. By reproducibility, we mean arriving at the



Mortality	**ACCORD + VADT**	N=1776	N=1385	N=1876	N=1774	N=2081	N=1202	N=1948
	Risk difference	-2.4%	-3.0%	0.7%	2.4%	1.1%	4.8%	4.7%
	(95% CI)	(-5.3, 0.6)	(-5.2, -0.8)	(-1.0, 2.4)	(0.2, 4.6)	(-0.8, 2.9)	(1.2, 8.5)	(1.8, 7.5)
	ACCORD	N=1479	N=1085	N=1516	N=1591	N=1800	N=1083	N=1697
	Risk difference	-2.6%	-2.9%	0.4%	2.7%	1.8%	4.5%	3.8%
	(95% CI)	(-5.7, 0.5)	(-5.3, -0.5)	(-1.4, 2.3)	(0.5, 5.0)	(-0.1, 3.8)	(0.8, 8.1)	(0.9, 6.7)
	VADT	N=297	N=300	N=360	N=183	N=281	N=119	N=251
	Risk difference	-1.7%	-3.1%	1.8%	-1.1%	-4.4%	9.5%	10.9%
	(95% CI)	(-10.0, 6.7)	(-8.4, 2.1)	(-2.0, 5.7)	(-10.1, 8.0)	(-10.8, 2.1)	(-5.8, 24.8)	(1.2, 20.7)
CVD	**ACCORD + VADT**							
	Risk difference	-2.0%	-3.2%	-1.3%	1.8%	-1.3%	0.0	-0.1%
	(95% CI)	(-5.3, 1.2)	(-6.3, 0.1)	(-3.5, 0.9)	(-0.9, 4.5)	(-3.7, 1.0)	(-3.7, 3.7)	(-3.3, 3.1)
	ACCORD							
	Risk difference	-2.8%	-2.8%	-0.8%	2.1%	-1.6%	-0.3%	0.1%
	(95% CI)	(-6.3, 0.7)	(-6.2, 0.5)	(-3.2, 1.7)	(-0.6, 4.9)	(-4.1, 0.9)	(-4.1, 3.4)	(-3.3, 3.5)
	VADT							
	Risk difference	1.7%	-4.2%	-3.6%	-1.6%	-0.3%	4.0%	-1.1%
	(95% CI)	(-6.9, 10.2)	(-11.6, 3.2)	(-8.6, 1.4)	(-12.4, 9.1)	(-8.0, 7.5)	(-10.3, 18.2)	(-10.6, 8.4)

Fig. 2. Summary causal tree of heterogeneous treatment effects of intensive glycemic control on all-cause mortality. Splitting variables and cut-points for each split are shown, resulting in seven terminal subgroups. Size, risk difference of all-cause mortality, and risk difference of major adverse cardiovascular events (CVD) from intensive versus standard glycemic control in each subgroup are shown for pooled data, the ACCORD study, and the VADT study. Negative risk differences indicate better outcomes in intensive glycemic control arm. Units for splitting variables are mm Hg for DBP (diastolic blood pressure), kg/m^2 for BMI (body-mass index), and years for age.

same result on multiple occasions using identical data and analyses. By replicability, we mean arriving at the same result in separate experimental studies with similar analyses. In this study, we were able to reproduce the causal forests variable importance results previously published for the ACCORD study [11], but were unable to replicate the HTE in the VADT study. The variable importance derived from causal forests was reasonably correlated between the ACCORD study and VADT study, suggesting at least partial replicability of the casual forests HTE algorithm in disparate samples. Furthermore, an indicator variable for the parent study (ACCORD or VADT) was not important for HTE detection when applying causal forests to pooled data from both studies, suggesting differences between the studies did not preclude identification of consensus HTE. Furthermore, the representative tree based on the 5 most highly ranked variables from causal forests analysis of the pooled data found that intensive glycemic control was associated with lower mortality in relatively leaner (BMI <35 kg/m^2) and younger (age <61 years) individuals – two features identified in the prior analysis of the ACCORD study [11]. That pooling individual-level data yielded consistent HTE subgroups across studies may suggest that the causal forests algorithm is sensitive to subtle between-study differences that were smoothed-out in the pooled data. Our results support the need for replication of HTE detection if multiple similar studies are available, with cautious interpretation or application to clinical care until results are confirmed.

While the results presented in the summary causal tree using pooled data from the ACCORD and VADT studies are potentially clinically relevant, they should be interpreted with caution. Neither the ACCORD study nor the VADT study, or the similarly designed ADVANCE study, found a benefit of intensive glycemic control for the primary outcome of cardiovascular events, and the ACCORD study found higher mortality associated with intensive control [8–10]. The initial split in the summary causal tree with individuals with diastolic blood pressure less than 65 mm Hg exhibiting higher mortality with intensive glycemic control fits with prior

studies of blood pressure and blood pressure variability and cardiovascular outcomes in the VADT study [24,25]. However, the summary causal tree presented here is otherwise difficult to interpret from the perspective of relating variable cut-points to physiology and clinical outcomes and would not be intuitive to institute into clinical practice in its current form. The analyses presented here highlight the generalizability challenges inherent to using machine learning for identifying HTE, and additional evaluation is needed to assess the performance of the HTE subgroups in the summary causal tree in diabetes patients drawn from the VA health system and in the general population.

There are several limitations to acknowledge. First, interpretation of the results should remain limited to the populations represented in the ACCORD and VADT studies. That is, though pooling data from the two studies broadens the general population representation in our analysis, the resulting pooled sample is still derived from a select randomized trial sample and does not necessarily better represent the diabetes patient population at risk. Second, we found that the causal forests results were sensitive to the minimum node size parameter, particularly if smaller nodes (terminal subgroups) were allowed (see Supplemental Material for comparison of variable importance across a range of minimum node sizes). We used a minimum node size of 5% of the total study population to align this study with the prior HTE analysis of the ACCORD study [11] and to retain a reasonably sized potential patient population who might be affected by any HTE. Third, while we demonstrate limitations to generalizability of causal forests for identifying HTE between two study samples, we do not propose specific strategies to overcome the limitations aside from pooling data from multiple studies, a solution that is often impractical. Whether methods for generalizability and/or transportability of trial data to a target population [26–30] can be tailored to machine learning algorithms for HTE detection will be explored in future work. Fourth, the VADT study was less than 20% the size of the ACCORD study, so some of the between-sample variation in significance of subgroup effects may be attributable to the differences in sample size. Finally, we evaluated only one machine learning algorithm for HTE detection – causal forests – in our analysis, based on previously published work. Therefore, we cannot necessarily generalize the findings with causal forests to other machine learning algorithms that can be used to identify HTE. While active research in refining machine learning algorithms has yielded improvements in HTE detection [31], it is unclear if better within-sample HTE identification would translate to better between-sample generalizability.

In conclusion, using data from two randomized trials of intensive glycemic control in type II diabetes patients, we found limited replicability or generalizability of HTE on all-cause mortality identified using the causal forests machine learning approach despite similar variable prioritization in each study. We speculate that differences in the study population characteristics or specifics of the intervention can undermine generalizability of HTE even for similarly designed randomized trials and urge caution in interpreting and/or applying HTE results in the absence of replication. While the limitations of causal forests could be overcome by pooling individual-level data from multiple studies, this solution is not always feasible. These findings motivate development of additional methods for meta-analyzing machine learning applications for HTE detection and for generalizing machine learning results from trials to target populations.

Acknowledgments

Funding

This work was supported by the US Department of Veterans Affairs (Award IK2-CX001907 to SR, Clinical Sciences Research & Development Service's Cooperative Studies Program #465-F to GB, and Cooperative Studies Program #2008 to NE); the American Heart Association (Award 17MCPRP33670728 to SR); the National Science Foundation (Award DMS 1914937 to DG); the National Cancer Institute of the US National Institutes of Health (Award R01 CA129102 to DG); and the National Institute of Diabetes and Digestive and Kidney Diseases of the US National Institutes of Health (Award K23DK109200 to SAB). The VADT Study was supported by the Veterans Affairs Cooperative Studies Program, Department of Veterans Affairs Office of Research and Development; the American Diabetes Association; and the National Eye Institute. Medications and financial support for the VADT Study were provided by Sanofi, GlaxoSmithKline, Novo Nordisk, Roche, Kos Pharmaceuticals, Merck, and Amylin.

Data availability

VADT study data were made available through a Data Use Agreement with the VA Cooperative Studies Program (clinical trials registration number NCT00032787). ACCORD study data is publicly available through the US National Institutes of Health, National Heart, Lung, and Blood Institute's Biologic Specimen and Data Repository Information Coordinating Center (https://biolincc.nhlbi.nih.gov/studies/accord/).

Study group members

All of the contributing investigators to the Veterans Affairs Diabetes Trial are acknowledged in a document for publication as an online-only supplement.

Author contributions

All authors contributed to critical review of the manuscript and approve of its submission. SR contributed to study design and analyses. KJ contributed to study design and analyses. GB contributed to data acquisition and availability and analysis plan development. SB contributed code for analyses and to study design. SAB contributed to study design. DR, NE, and PR all contributed to data acquisition and availability. DG contributed to study design and oversaw all analyses. SR, KJ, and DG drafted the initial manuscript with input from all co-authors. NE passed away during the preparation of this manuscript but approved of a draft version with only cosmetic differences from this submitted final manuscript.

Disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the US Department of Veterans Affairs, or the United States Government.

Declaration of Competing Interest

The authors have no conflicts of interest relevant to the work in this study. Medications and financial support were provided by Sanofi, GlaxoSmithKline, Novo Nordisk, Roche, Kos Pharmaceuticals, Merck, and Amylin. No other potential conflicts of interest relevant to this article were reported. These companies had no role in the design of the study, in the accrual or analysis of the data, or in the preparation or approval of the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.annepidem.2021.07.003.

References

[1] Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319(13):1317–18.
[2] Cikes M, Sanchez-Martinez S, Claggett B, Duchateau N, Piella G, Butakoff C, et al. Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. Eur J Heart Fail 2019;21(1):74–85.
[3] Duan T, Rajpurkar P, Laird D, Ng AY, Basu S. Clinical value of predicting individual treatment effects for intensive blood pressure therapy. Circ Cardiovasc Qual Outcomes 2019;12(3):e005010.
[4] Powers S, Qian J, Jung K, Schuler A, Shah NH, Hastie T, et al. Some methods for heterogeneous treatment effect estimation in high dimensions. Stat Med 2018;37(11):1767–87.
[5] Wendling T, Jung K, Callahan A, Schuler A, Shah NH, Gallego B. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. Stat Med 2018;37(23):3309–24.
[6] Fusar-Poli P, Hijazi Z, Stahl D, Steyerberg EW. The science of prognosis in psychiatry: a review. JAMA Psychiatry 2018;75(12):1289–97.
[7] Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2017;38(23):1805–14.
[8] HC Gerstein, Miller ME, et al., Action to Control Cardiovascular Risk in Diabetes Study Group Effects of intensive glucose lowering in type 2 diabetes. N Engl J Med 2008;358(24):2545–59.
[9] Duckworth W, Abraira C, Moritz T, Reda D, Emanuele N, Reaven PD, et al. Glucose control and vascular complications in veterans with type 2 diabetes. N Engl J Med 2009;360(2):129–39.
[10] Patel A, MacMahon S, et al., ADVANCE Collaborative Group Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. N Engl J Med 2008;358(24):2560–72.
[11] Basu S, Raghavan S, Wexler DJ, Berkowitz SA. Characteristics associated with decreased or increased mortality risk from glycemic therapy among patients with type 2 diabetes and high cardiovascular risk: machine learning analysis of the ACCORD trial. Diabetes Care 2018;41(3):604–12.
[12] Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proc Natl Acad Sci 2016;113(27):7353–60.
[13] Levey AS, Coresh J, Greene T, Stevens LA, Zhang YL, Hendriksen S, et al. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. Ann Intern Med 2006;145(4):247–54.
[14] Kim MK, Jeong JS, Kwon HS, Baek KH, Song KH. Concordance the hemoglobin glycation index with glycation gap using glycated albumin in patients with type 2 diabetes. J Diabetes Complications 2017;31(7):1127–31.
[15] Signorovitch JE, Wu EQ, Yu AP, Gerrits CM, Kantor E, Bao Y, et al. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. Pharmacoeconomics 2010;28(10):935–45.
[16] Athey S, Wager S. Estimating Treatment Effects with Causal Forests: An Application. https://arxiv.org/abs/1902.07409.
[17] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 2018;113(523):1228–42.
[18] Banerjee M, Ding Y, Noone AM. Identifying representative trees from ensembles. Stat Med 2012;31(15):1601–16.
[19] Wallach JD, Sullivan PG, Trepanowski JF, Sainani KL, Steyerberg EW, Ioannidis JP. Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. JAMA Intern Med 2017;177(4):554–60.
[20] Rigdon J, Baiocchi M, Basu S. Preventing false fiscovery of heterogeneous treatment effect subgroups in randomized trials. Trials 2018;19(382). doi:10.1186/s13063-018-2774-5.
[21] Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. JAMA 2020.
[22] Vollmer S, Mateen BA, Bohner G, Kiraly FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020;368:l6927.
[23] Li J, Liu L, Le DT, Liu J. Accurate data-driven prediction does not mean high reproducibility. Nat Mach Intell 2020;2:13–15.
[24] Anderson RJ, Bahn GD, Moritz TE, Kaufman D, Abraira C, Duckworth W, et al. Blood pressure and cardiovascular disease risk in the veterans affairs diabetes trial. Diabetes Care 2011;34(1):34–8.
[25] Nuyujukian DS, Koska J, Bahn G, Reaven PD, Zhou JJVADT Investigators. Blood Pressure Variability and Risk of Heart Failure in ACCORD and the VADT. Diabetes Care 2020;43(7):1471–8.
[26] Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. Am J Epidemiol 2010;172(1):107–15.
[27] Dahabreh IJ, Robertson SE, Tchetgen EJ, Stuart EA, Hernan MA. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. Biometrics 2019;75(2):685–94.
[28] Hong JL, Webster-Clark M, Jonsson Funk M, Sturmer T, Dempster SE, Cole SR, et al. Comparison of methods to generalize randomized clinical trial results without individual-level data for the target population. Am J Epidemiol 2019;188(2):426–37.
[29] Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. Epidemiology 2017;28(4):553–61.
[30] Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. Am J Epidemiol 2017;186(8):1010–14.
[31] Kunzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci 2019;116(10):4156–65.