Health Services and Outcomes Research Methodology
https://doi.org/10.1007/s10742-021-00243-x

Which patients benefit most from completing health risk assessments: comparing methods to identify heterogeneity of treatment effects

Maren K. Olsen, et al. ***[full author details at the end of the article]***

Received: 12 September 2019 / Revised: 22 January 2021 / Accepted: 28 January 2021

Abstract
Methods for identifying heterogeneity of treatment effects in randomized trials have seen recent advances, yet applying these methods to health services intervention trials has not been well investigated. Our objective was to compare two approaches—predictive risk modeling and model-based recursive partitioning—for identifying subgroups of trial participants with potentially differential response to an intervention involving health risk assessment completion alone (n = 192) versus health risk assessment completion plus telephone-delivered health coaching (n = 173). Notably, these approaches have been developed by investigators from distinct disciplines and reported in separate literatures and have generally not been compared in prior work. Furthermore, these methods approach subgroup identification differently and answer related but slightly different questions. The primary outcome for both approaches was prevention health program enrollment by six months. The predictive risk model was developed in two steps, where, first, a single risk score was derived from a logistic regression model with 12 a priori chosen covariates by the scientific investigator team (c-statistic = 0.63). Then, the treatment effect was calculated within quartiles of risk via interaction in a logistic regression model (c-statistic = 0.69; c-for-benefit = 0.43). The greatest treatment effect was in the second quartile, in which 54% (22 of 41) of intervention patients and 10% (5 of 50) of control patients reported prevention program enrollment. In contrast, with the data-driven approach of model-based recursive partitioning, all 28 baseline covariates were considered, with the algorithm selecting covariates and optimal split points. Final model results had a c-statistic of 0.69 and a c-for-benefit of 0.55 (optimism-corrected c-statistic = 0.62 and c-for-benefit = 0.53) and identified 4 subgroups, with the greatest treatment effect among patients with lower mean numeracy, education less than a bachelor's degree, and diabetes, in which 54% (15 of 28) of intervention patients reported prevention program enrollment versus 7% (3 of 41) of control patients. While there is increasing interest in discovering heterogeneity of treatment effects, our analyses highlight the important differences between these approaches, both from questions answered, model development, and results obtained. Specifying goals of treatment heterogeneity analyses, choosing the appropriate method to best address the goals, and external validation of results are important steps when applying these methods.
Clinicaltrials.gov identifier: NCT01828567

Keywords Veterans · Health risk assessment · Subgroup · Treatment effect heterogeneity

1 Introduction

Randomized trials provide the strongest evidence about intervention effectiveness, but there is growing recognition that the average treatment effect generated from a trial does not generalize to most patients eligible for the intervention (Rothwell 2005; Kent, Hayward 2007). A principled approach to identifying heterogeneity of treatment effects (HTE) is needed. Historically, HTE was assessed by identifying subgroups stratified by one variable (e.g., male vs. female) and with statistical tests for treatment by subgroup interactions, which are easy to implement and intuitive to understand (Lagakos 2006; Alosh et al. 2017). However, this approach often does not fully characterize the multivariable risk and/or benefit of treatment (Rothwell, Warlow 1999); additionally, there is a risk of false negatives due to lack of statistical power in small samples and a risk of false positives as the number of stratified analyses grows (Hayward et al. 2005). To avoid some of these pitfalls, multivariable predictive risk modeling approaches (Rothwell, Warlow 1999; Kent, Hayward 2007) as well as data-driven (Lipkovich et al. 2017) approaches have been developed in parallel, from distinct disciplines, to more systematically discover and describe HTEs. The predictive risk framework has been largely developed in the medical literature and focused on regression models developed on the basis of clinical intuition and prior evidence; the data-driven framework encompasses a wide variety of approaches, including model-based recursive partitioning, the simultaneous threshold interaction modeling algorithm, and the generalized unbiased interaction detection and estimation approach, is primarily in the statistical literature and has arisen from statistical classification methods (Dusseldorp et al. 2010b; Loh et al. 2015; Seibold et al. 2016). We describe an approach from each of these frameworks here.

In predictive risk modeling (PRM), patients are first grouped together within strata based on their risk from a pre-specified risk score (e.g., Framingham Risk Score [FRS] or an internally derived score constructed from a priori patient factors that have a plausible clinical relationship with the outcome) (Rothwell 2005). Treatment effects are then assessed within risk strata. This approach for describing treatment heterogeneity was first introduced in a landmark study showing that the clinical benefits accrued to patients randomized to receive carotid endarterectomy were entirely driven by 16% of the treatment group that was at highest risk for stroke (Rothwell, Warlow 1999).

Alternatively, data-driven methods identify subgroups with similar responses to treatment whose treatment effects vary from other subgroups. These methods generally consider all available baseline covariates to classify patients into discrete, intuitive subgroups (e.g., men age >57, men ≤ 57, women age >57, women ≤ 57) (Lipkovich et al. 2017). Many data-driven methods are derived using statistical classification methods (Dusseldorp et al. 2016) that are well suited to situations with many predictors with potentially complex interactions and little a priori knowledge concerning which subgroups may benefit most (Lipkovich et al. 2017; Strobl et al. 2009). The underlying search, optimization, and modeling algorithms vary by method and answer subtly different questions, often yielding varying results even when applied to the same dataset (Doove et al. 2014; Alemayehu et al. 2018). For example, the simultaneous threshold interaction modeling algorithm searches for the subgroups of patients that yield the largest differential treatment effect upon the outcome (Dusseldorp et al. 2010a); model-based recursive partitioning (MoB) searches for subgroups of treatment-covariate interactions that yield a better fitting model than the overall treatment effect model (Seibold et al. 2016; Sies et al. 2019). MoB was chosen as the data-driven approach for this paper because it can be applied to a dichotomous outcome, is

easily implemented in a well-documented R package, and allows discovery of both large and small treatment effects, which is similar to PRM, and was of overall interest to the investigator team (Doove et al. 2014).

This paper compares PRM and MoB to determine whether there were differential treatment effects in the primary outcome of a recent behavioral intervention (ACTIVATE) trial where patients were randomized to receive phone-based health coaching following completion of a health risk assessment (HRA) or to HRA completion alone (Oddone et al. 2018). In September 2014, the Department of Veterans Affairs (VA) implemented a comprehensive web-based HRA.(MyHealtheVet: HealtheLiving Assessment. 2018) HTE assessment of the ACTIVATE trial may help identify which Veterans would benefit most and least from phone-based coaching after completion of the HRA.

This analysis is novel in several respects. Few studies have applied HTE methods to identify subgroup effects within a behavioral intervention trial (Sussman et al. 2015; Baum et al. 2017); most HTE evaluations have examined surgical or medication (Kent et al. 2003) interventions in which there is potentially wide variability in response to treatment and a plausible risk of harm. Additionally, data-driven and PRM methods have been developed by investigators from distinct disciplines and reported in separate literatures. This is the first analysis to compare findings using these two distinctly different HTE approaches.

2 Methods

2.1 Trial design, participants, outcomes and covariates

The ACTIVATE trial randomized Veterans to receive a comprehensive HRA and telephone coaching intervention or an HRA alone (Oddone et al. 2018). Veterans were eligible if they were enrolled in primary care at the study sites and had at least one modifiable risk factor: body mass index (BMI) ≥ 30, current smoker, and/or less than 150 min of moderate/vigorous physical activity per week. Sociodemographic and clinical characteristics were measured prior to randomization.

All randomized Veterans (n = 417) completed the VA's web-based HRA at baseline (MyHealtheVet: HealtheLiving Assessment. 2018), which uses a proprietary risk modeling algorithm to provide patients with "health age" based on lifestyle choices, family risk, and biological values, as well as information about the degree to which lifestyle changes can lower their "health age". The intervention group received two telephone calls delivered within one month after baseline by a health coach. The control group received a printed copy of their HRA and were encouraged to discuss questions with their primary care team. Further details regarding randomization and baseline characteristics of the groups are included in Oddone et al. (2018). The primary outcome was self-reported enrollment in a structured prevention program within 6 months after randomization, and higher rates of enrollment (0.51 vs 0.29; $p < 0.0001$) were found in the HRA + coaching arm (Oddone et al. 2018). Additionally, prior analyses via logistic regression examined pre-specified subgroups of health literacy and numeracy and showed that HRA + coaching had a greater effect on the probability of enrollment in prevention programs for patients with low numeracy (intervention vs control difference of 0.31 95% CI: 0.18, 0.45) as compared to those with high numeracy (0.13, 95% CI: -0.01, 0.27) (Nouri et al. 2019).

In this secondary analysis using PRM and MoB, 52 of 417 participants were excluded due to missing data on the outcome and/or covariates used in model building resulting in

365 participants (173 randomized to intervention and 192 randomized to control). Baseline characteristics were similar between those included and excluded in the analysis (see Appendix Table 5). Additionally, results of the original trial were robust to missing data assumptions (Oddone et al. 2018).

2.2 Predictive risk modeling

PRM was implemented in two steps. In the first step, all selected risk predictors were distilled into a single predicted risk score, thus eliminating issues of "one-variable-at-a-time" subgroup analyses described above. In the second step, patients were stratified by quartiles of predicted risk and the treatment effect was estimated within each stratum. Note that our outcome—prevention program enrollment—is a positive rather than a negative outcome, so the common terminology of "predictive risk model" is a misnomer. To be consistent with the literature, we retained this term.

For the first step, externally validated risk models should be used to characterize potential risk and benefit of treatment; an internally derived score is developed only if a validated score is unavailable (Kent et al. 2010). There was no externally validated predictive model for enrollment in prevention programs, so we fit a logistic regression model to predict the enrollment outcome from the study's baseline covariates on the entire study cohort. Following the guidelines presented by Kent and colleagues (Kent et al. 2010), the study team identified predictors of adverse event risk in the absence of intervention, and health and non-health predictors of responsiveness to the behavioral intervention (Kravitz et al. 2004). Specifically, selection of health-related patient factors that might prompt a patient to enroll in a prevention program in this analysis was informed by a prior analysis of predictors of MOVE! participation (McVay et al. 2014). Non-health patient-based barriers to enrollment in a structured prevention program were expected to be related to lack of motivation/activation, time or income. A consensus of twelve covariates were chosen a priori and entered as main effects into the model, including the Patient Activation Measure (PAM) which assesses activation of an individuals' knowledge, skills, beliefs, and confidence for managing their health, (Hibbard et al. 2005, 2007) and the FRS, a validated and widely used assessment of cardiovascular risk (D'Agostino et al. 2008). Other health-related predictors were validated measures of general health (excellent/very good vs. good/fair/poor), alcohol use, depression from the Patient Health Questionnaire (Kroenke et al. 2001), self-reported pain in the past week, and the Medical Outcomes Study 6-item sleep quality measure (Hays, Stewart 1992). We also included a variable indicating the difference between each Veteran's "health age" as estimated from the HRA and their chronological age as Veterans with health age greater than chronological age may be more motivated to enroll in a prevention program. Non-health-related factors included measures of income, computer literacy, general literacy, and numeracy – assessed via a modified 3-item version of the subjective numeracy scale (McNaughton et al. 2015). Veterans with low levels of these factors might have greater difficulty completing the HRA accurately or understanding its results.

In the second step, the predicted probability of enrollment based on the regression model from the first step was used to group participants into quartiles (n = 91–92 Veterans per quartile). A logistic regression was then fit to estimate the treatment effect on enrollment in each quartile via PROC LOGISTIC in SAS (version 9.4, SAS Institute, Cary, NC). Model coefficients included dummy-coded treatment arm, dummy-coded quartile, and the treatment by quartile interaction. For the final model, both the conventional c-statistic for

risk and the novel concordance statistic for benefit were calculated (van Klaveren et al. 2018). The c-for-benefit assesses how well the model discriminates participants who benefit from the ACTIVATE intervention, i.e., have a positive treatment effect. Because the c-for-benefit requires equivalent sample sizes between the intervention and control groups, the predicted values from the first step model were used to create a 1:1 match between the control and intervention group participants for the calculation of that statistic. The research question addressed by PRM is whether differences in enrollment between the treatment and control arms differed between subgroups characterized by predicted risk quartiles, based on a logistic regression model of 12 covariates chosen a priori.

2.3 Data-driven approach: model-based recursive partitioning

MoB and other data-driven methods have been developed from statistical classification methods that lend themselves well to situations with many predictors with potentially complex interactions (Lipkovich et al. 2017; Strobl et al. 2009). The basic premise of MoB is that, rather than one overall treatment effect model, it may be possible to partition patients into subgroups based on the full set of available covariates, resulting in a better fitting model for each respective subgroup, often defined by multiple covariates (Zeileis et al. 2008). In doing so, treatment by covariate interactions are modeled. To assess whether or not a split on a covariate improves model fit, MoB first performs a fluctuation test for parameter instability across all values of the covariates to identify a split variable (i.e., the covariate with the lowest p-value). If there is instability, the algorithm selects the split-value of the of the identified covariate by minimizing an objective function. For example, a test statistic of instability may identify patient age as the first split variable with the objective function indicating an optimal split at age 57. Thus, the algorithm determines the optimal split variables and split points rather than selecting covariates and split points a priori. A regression model of the treatment effect upon the outcome is fit within each subgroup (e.g., patients >57; patients ≤ 57). The process is repeated within each of the resulting subgroups until the best model fit is achieved, implicitly conducting variable selection. MoB yields a regression-based tree with each leaf or terminal node representing a subgroup experiencing different effects of health coaching on prevention program enrollment rates. Therefore, the research question answered by MoB is which subgroups of patients experience differential treatment effects on the outcome of prevention program enrollment.

MoB was implemented via the mob function in the R package partykit 1.2–8 (R version 3.6.2) (Hothorn, Zeileis 2015). We specified a logistic regression model with the outcome of prevention program enrollment by 6 months. A total of 28 baseline covariates were included as potential partitioning variables. Results are reported graphically as a regression-based tree; details of function specification are included in the figure footnote of enrollment for patients randomized to treatment (HRA + coaching) compared to control (HRA alone) within each subgroup. Similar to the final PRM model, both the conventional c-statistic and the novel c-for-benefit was calculated for the final MoB solution. The same 1:1 matched sample from step 1 of the PRM was used for the MoB c-for-benefit calculations. Finally, we conducted an internal validation by applying the MoB steps to 100 bootstrap samples, with c-statistics and the c-for-benefit calculated for each sample. The resulting model from each was also then applied to the original sample and the corresponding c-statistics and c-for-benefits calculated. The average difference between the two c-statistics and two c-for-benefits (van Klaveren et al. 2018) provided estimates of optimism (i.e.,

correcting for original model performance being too optimistic) for both indices, respectively (Harrell Jr et al. 1996).

For both methods, we report the absolute benefit by quartile or terminal node (difference in enrollment rates between the treatment (HRA + coaching) and control (HRA alone) arms) and relative benefits via odds ratios. This study was approved by the Institutional Review Board of the Durham VA Health Care System.

3 Results

3.1 Patient characteristics

Study participants' (n = 365) mean age was 56.4, mean HRA-generated "health age" was 60.9, mean BMI was 33.9, mean 10-year cardiovascular risk score from the FRS was 22.4%, mean PAM score was 61.5, and mean numeracy was 4.6 (Table 1). The majority of participants (86.3%) were male.

3.2 Results from predictive risk modeling

The predictive risk model (first step) had discrimination (c-statistic) of 0.63 and 95% confidence intervals of all odds ratios including 1.0, showing that none of the 12 a priori selected covariates were strongly associated with the outcome (Table 2). Yet, enrollment rates still varied across risk strata quartiles. The treatment-quartile model (second step) had a c-statistic of 0.69, a c-for-benefit of 0.43, and the quartile-by-treatment interaction (3df) p-value was 0.06. Only 32.6% of Veterans in the first quartile enrolled in a prevention program compared to 53.8% of Veterans in the highest risk quartile (Table 3). Compared to control group patients, intervention group patients in the first quartile were twice as likely to enroll (43.5 vs. 21.7%; absolute benefit = 21.7%), five times more likely (53.7 vs. 10.0%; absolute benefit = 43.7%) in the second quartile, and almost two times more likely in the third quartile (52.2 vs. 33.3%; absolute benefit = 18.8%). The intervention effect was most modest in the highest risk quartile (60.0 vs. 49.0%; absolute benefit = 11.0%).

3.3 Results from MoB

MoB resulted in a tree with 4 subgroups based on splits of 3 variables—mean numeracy score, highest education level, and diabetes status (Fig. 1), with numeracy being the first splitting variable. The number of subjects in each ranged from 50 to 146 (Fig. 1). For patients with lower mean numeracy, the coaching intervention had a significantly greater effect on enrollment than the control intervention, but to varying degrees depending upon additional patient characteristics (Table 4). The greatest effect (absolute benefit = 46.3; 53.6 vs. 7.3%) was among patients with lower mean numeracy, less than a bachelor's degree, and diabetes (terminal node 2, Fig. 1 and Table 4).

The second greatest effect (absolute benefit = 39.4; 76.9 vs. 37.5%) was for patients with lower mean numeracy and at least a bachelor's degree (terminal node 3, Fig. 1 and Table 4). Finally, the largest subgroup identified (n = 146) were patients with lower numeracy, less than a bachelor's degree, and no diabetes who realized a more modest effect (absolute benefit = 24.4; 51.4 vs. 27.0%) (terminal node 1, Fig. 1 and Table 4). This subgroup had an effect size (OR = 2.9) similar to the average treatment effect found in the

Table 1 Descriptive statistics of ACTIVATE arms at baseline

	Overall N = 365	Intervention: HRA + Coaching N = 173	Control: HRA only N = 192
Age, mean (SD)	56.4 (11.7)	56.2 (12.0)	56.5 (11.5)
HRA Health age, mean (SD)[1]	60.9 (12.0)	60.9 (12.1)	60.9 (12.0)
Difference in age (Health age—Actual age), mean (SD)	4.5 (5.7)	4.7 (5.5)	4.3 (5.9)
PAM score, mean (SD)	61.5 (12.5)	62.5 (13.0)	60.6 (12.1)
Framingham 10-year cardiovascular risk score, mean (SD)	22.4 (16.4)	22.2 (16.3)	22.6 (16.5)
General health, n (%)			
Excellent	23 (6.3)	13 (7.5)	10 (5.2)
Very good	88 (24.1)	37 (21.4)	51 (26.6)
Good	144 (39.5)	70 (40.5)	74 (38.5)
Fair	85 (23.3)	39 (22.5)	46 (24.0)
Poor	25 (6.8)	14 (8.1)	11 (5.7)
Employed full/part-time, n (%)	126 (34.5)	58 (33.5)	68 (35.4)
Inadequate income, n (%)	99 (27.1)	48 (27.7)	51 (26.6)
Married/living as married, n (%)	191 (52.3)	81 (46.8)	110 (57.3)
Non-Hispanic white race, n (%)	172 (47.1)	78 (45.1)	94 (49.0)
Male gender, n (%)	315 (86.3)	146 (84.4)	169 (88.0)
Education, n (%)			
High school or less	66 (18.1)	33 (19.1)	33 (17.2)
Some college, Associate's degree, or trade school	211 (57.8)	95 (54.9)	116 (60.4)
Bachelor's degree or higher	88 (24.1)	45 (26.0)	43 (22.4)
Assistance required for reading, n (%)			
Never	251 (68.8)	122 (70.5)	129 (67.2)
Rarely	67 (18.4)	32 (18.5)	35 (18.2)
Sometimes/often/always	47 (12.9)	19 (11.0)	28 (14.6)
Mean numeracy score (range 1–6), mean (SD)[2]	4.6 (1.2)	4.6 (1.3)	4.6 (1.2)
Computer literacy—ability to use, n (%)			
Do not use computer	22 (6.0)	13 (7.5)	9 (4.7)
Basic	70 (19.2)	36 (20.8)	34 (17.7)
Moderate	133 (36.4)	62 (35.8)	71 (37.0)
Advanced	98 (26.8)	44 (25.4)	54 (28.1)
Expert	42 (11.5)	18 (10.4)	24 (12.5)
Body mass index, mean (SD)	33.9 (6.3)	33.8 (6.3)	34.1 (6.2)
Current smoker of cigarettes or other tobacco, n (%)	140 (38.4)	72 (41.6)	68 (35.4)
Minutes of physical activity in past week, median (IQR)	150.0 (380.0)	175.0 (370.0)	130.0 (380.0)
Total number of inclusion criteria,[3] n (%)			
1	157 (43.0)	76 (43.9)	81 (42.2)
2	167 (45.8)	79 (45.7)	88 (45.8)
3	41 (11.2)	18 (10.4)	23 (12.0)
MOS-6 Sleep Scale Score, mean (SD)	61.0 (21.7)	61.1 (21.5)	60.9 (21.8)
Pain in past week[4], mean (SD)	4.6 (2.7)	4.5 (2.7)	4.8 (2.7)
PHQ-8 Total Score, mean (SD)	7.2 (5.5)	7.0 (5.5)	7.4 (5.4)

Table 1 (continued)

	Overall N = 365	Intervention: HRA + Coaching N = 173	Control: HRA only N = 192
Alcohol consumption, n (%)			
Never	147 (40.3)	62 (35.8)	85 (44.3)
Monthly or less	80 (21.9)	41 (23.7)	39 (20.3)
2–4 times a month	53 (14.5)	22 (12.7)	31 (16.1)
2–3 times a week	52 (14.2)	26 (15.0)	26 (13.5)
4 or more times a week	33 (9.0)	22 (12.7)	11 (5.7)
Total cholesterol (md/dL), mean (SD)	178.5 (41.9)	178.3 (42.4)	178.8 (41.5)
Diabetes diagnosis, n (%)	103 (28.2)	45 (26.0)	58 (30.2)
Average systolic blood pressure (mm Hg), mean (SD)	129.9 (15.3)	129.9 (15.3)	129.9 (15.3)
High-density lipoprotein (mg/dL), mean (SD)	45.9 (14.2)	46.8 (14.8)	45.1 (13.6)
Take blood pressure medications, n (%)	228 (62.5)	109 (63.0)	119 (62.0)
OUTCOME			
Enrollment in prevention program at 6 months, n (%)	145 (39.7)	90 (52.0)	55 (28.6)

SD Standard deviation, *HRA* Health Risk Assessment via the Health*e*Living Assessment, *PAM* Patient Activation Measure, *IQR* Interquartile range, *MOS* Medical Outcomes Study, *PHQ* Patient Health Questionnaire

[1]HRA Health Age: The HRA uses a proprietary risk modeling algorithm to determine patients' "health age" based on lifestyle choices, family risk, and biological values, as well as information about the degree to which lifestyle changes can lower their "health age."

[2]Numeracy variables: Skill with fractions, skill with percentages, and usefulness of numerical information in making health decisions; each variable is on a 1–6 scale, with a value of 1 anchoring "not at all good" and 6 anchoring "extremely good"

[3]To be included in the study, Veterans had to have at least one of the following modifiable risk factors: body mass index (BMI) ≥ 30, current smoker, or < 150 min of moderate/vigorous physical activity per week

[4]Range is 0–10, with 0 representing no pain

[5]Mean of two systolic blood pressure measurements

main study (OR = 2.5). (Oddone et al. 2018). Conversely, the intervention effect was null (absolute benefit = − 5.1; 38.3 vs. 43.4%) among patients with the highest mean numeracy scores (terminal node 4, Fig. 1 and Table 4).

The original model had a c-statistic = 0.69 and a c-for-benefit = 0.55. The estimated optimism from the 100 bootstrap samples was 0.07 for the c-statistic and 0.02 for the c-for-benefit, resulting in an optimism-corrected c-statistic of 0.62 and an optimism-corrected c-for-benefit of 0.53.

4 Discussion

Treatment effect estimates from randomized trials do not generalize to individual patients. Thus, there is increasing interest in principled HTE assessment within trials, which would be more meaningful for patients, providers, and health systems. HTE can be used to identify patients who may be at higher risk for harm or who may experience especially strong treatment effects. In behavioral interventions like the ACTIVATE trial with minimal risk of harm (Oddone et al. 2018), HTE assessment can inform which subgroups to prioritize for

Table 2 Logistic regression (predictive risk model) of enrollment in a prevention program

	Full model odds ratio (95% Confidence interval)
Framingham risk score (FRS)	1.01 (0.99, 1.02)
Patient activation measure (PAM)	1.01 (0.99, 1.03)
General health: fair, poor, or good	1.62 (0.96, 2.74)
Computer literacy: advanced or expert	1.21 (0.73, 1.98)
General literacy: never needing assistance	1.11 (0.68, 1.81)
Numeracy	1.06 (0.88, 1.28)
Alcohol use: 2–4 times a month or more	0.82 (0.52, 1.30)
Inadequate income	0.77 (0.46, 1.30)
Depression (PHQ-8)	1.04 (0.98, 1.10)
Pain in past week	0.93 (0.85, 1.02)
Difference between health age and actual age	0.97 (0.93, 1.01)
MOS-6 Sleep Scale Score, 2nd and 3rd quintiles[1]	0.73 (0.38, 1.41)
MOS-6 Sleep Scale Score, 4th and 5th quintiles[1]	1.31 (0.58, 2.96)
c-statistic	0.627
Sample size	365

[1]Non-linearity in the logit was observed for MOS-6 Sleep Scale Score when included in the model as a continuous variable. Therefore, the measure was split into quintiles, and further reduced into three groups. Reference = 1st quintile

future dissemination efforts. The PRM approach to HTE assessment was developed by clinicians to understand whether average treatment effects were driven by a subgroup of participants who realized a disproportionate benefit from treatment (Rothwell, Warlow 1999), in which subgroups were defined by multiple clinical measures of risk. Data-driven methods were developed by statisticians who approached subgroup identification as a model selection problem and considered all baseline covariates for classifying patients into discrete and intuitive subgroups (Lipkovich et al. 2017).

To date, no studies have applied both approaches to the same trial and only recently have these methods been applied to behavioral intervention trials (Sussman et al. 2015; Baum et al. 2017). This is also one of the first studies to apply HTE methods to identify subgroup effects in a behavioral intervention trial (Sussman et al. 2015; Baum et al. 2017), as most prior HTE evaluations have examined subgroup effects from surgical or medication interventions. The purpose of this study was to identify whether there were subgroups of participants in the ACTIVATE trial who varied in their responsiveness to an HRA + coaching intervention. Implementing these two methods also had the potential to highlight some of the advantages and disadvantages of each method.

It is notable that numeracy was the initial branching variable from the MoB analysis because this result is consistent with a prior one-variable-at-a-time analysis examining the impact of the intervention among health literacy and numeracy subgroups (Nouri et al. 2019). That "traditional" subgroup analysis found that the coaching intervention had a greater effect on the probability of enrollment in prevention programs for patients with low mean numeracy than among Veterans with high mean numeracy, and, by design, only identified two subgroups, instead of four identified using MoB that allows all baseline covariates to be considered.

Table 3 Effect of HRA + coaching intervention on enrollment in a prevention program, by quartile of predicted risk

Strata	N	% (N) Enrolled overall	% (N) Enrolled intervention	% (N) Enrolled control	Absolute benefit	OR (95% CI)
Quartile 1 (lowest predicted enrollment)	92	32.6 (30 of 92)	43.5% (20 of 46)	21.7% (10 of 46)	21.7%	2.8 (1.1, 6.9)
Quartile 2	91	29.7 (27 of 91)	53.7% (22 of 41)	10.0% (5 of 50)	43.7%	10.4 (3.4, 31.6)
Quartile 3	91	42.9 (39 of 91)	52.2% (24 of 46)	33.3% (15 of 45)	18.8%	2.2 (0.9, 5.1)
Quartile 4 (highest predicted enrollment)	91	53.8 (49 of 91)	60.0% (24 of 40)	49.0% (25 of 51)	11.0%	1.6 (0.7, 3.6)

Absolute Benefit = enrollment rate of intervention—enrollment rate of control

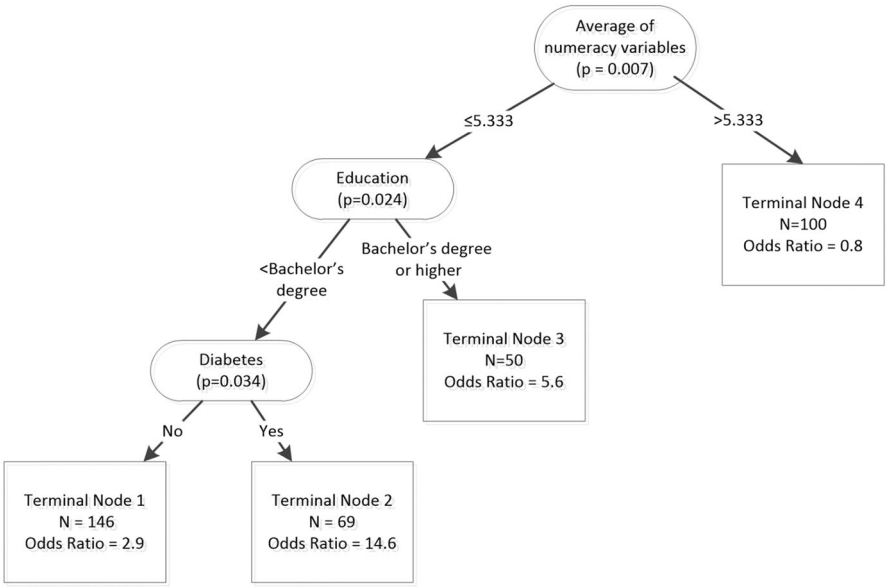

Fig. 1 Data-driven Method (MoB) Results Showing Subgroups with Differential Treatment Effects. We used the default value of statistical significance for the fluctuation tests (alpha = 0.05). Instead of specifying a Bonferroni correction (which would have altered the statistical significance to 0.05/28), we chose to post prune by Akaike's Information Criteria fit index and set the minimum node sample size as 40 (~ 10% of the overall sample size). Finally, we specified maxLM-type test as the fluctuation test for ordered factor variables. All other control parameters were kept at their default values.

There are several limitations that should be acknowledged in this HTE analysis. First, the PRM was internally developed and had limited discrimination, which may not generalize to HTE method comparisons in other trials. Note that an additional limitation was that diabetes was not chosen as an a priori predictor in the predictive risk model, yet it was identified as an important subgroup identifier in the data-driven analysis. Furthermore, the sample size was small (n = 365), so there were limits to the number of terminal nodes from the data-driven methods and number of strata from the predictive risk model that were possible. The limited sample size also precluded split-sample validation that would have been useful in refining the predictive risk model and confirming data-driven method results. Instead, as validation for the process, we estimated optimism of the MoB steps and corresponding c-statistic and c-for-benefit via bootstrap samples. Prior work has produced c-for-benefit values to be between 0.5 and 0.6 (van Klaveren et al. 2018), and while benchmarks for this novel index have not yet been developed, our c-for-benefit values are in that range for MoB but suggest poorer performance for the PRM approach. An important future step would be to assess the performance of these methods in an external sample. Finally, the validity of the confidence intervals from the regression models the subgroups found by MoB are unclear, as noted by Seibold et al. (Seibold et al. 2016). This further emphasizes the exploratory nature of the analyses and need for future validation.

The two HTE approaches differed in the clinical actionability of model results. MoB generated discrete and intuitive subgroups (e.g., patients with low numeracy, low education and diabetes) due to the binary splits imposed by recursive partitioning. The

Table 4 Effect of HRA + coaching intervention on enrollment in a prevention program, by terminal node from model-based recursive partitioning

Terminal node	N	% (N) Enrolled overall	% (N) Enrolled intervention	% (N) Enrolled control	Absolute benefit	OR (95% CI[1])
1: Low numeracy[2]; low education[3]; no diabetes	146	39.0% (57 of 146)	51.4% (37 of 72)	27.0% (20 of 74)	24.4%	2.9 (1.4, 5.7)
2: Low numeracy; low education; diabetes	69	26.1% (18 of 69)	53.6% (15 of 28)	7.3% (3 of 41)	46.3%	14.6 (3.6, 58.7)
3: Low numeracy; high education[3]	50	58.0% (29 of 50)	76.9% (20 of 26)	37.5% (9 of 24)	39.4%	5.6 (1.6, 19.0)
4: High numeracy[2]	100	41.0% (41 of 100)	38.3% (18 of 47)	43.4% (23 of 53)	-5.1%	0.8 (0.4, 1.8)

Absolute Benefit = enrollment rate of intervention—enrollment rate of control

[1]95% confidence intervals are calculated from logistic regression models generated post terminal node creation and should be interpreted with caution

[2]Low numeracy ≤ 5.333; High numeracy > 5.333

[3]Low education < bachelor's degree; high education = bachelor's degree or higher

subgroups identified via PRM, however, are a function of 12 baseline variables that were used to create a single risk score in the first step, and then stratified based on quartiles in the second step. This internally-derived model does not have widely accepted and well validated thresholds like the FRS, so subgroups did not have clinically discrete characteristics. With its limited discrimination ($c = 0.63$), it is possible that classification of individual patients within subgroups from this predictive risk model would change if the specification incorporated more predictive variables. Differences in the number of subgroups and treatment effect estimates across the two methods is unsurprising because these two methods address related but distinct questions. MoB can be used to discover subgroups showing differential treatment effects, while PRM assessed whether differences in enrollment between the treatment and control arms differed between subgroups characterized by quartiles of risk based on patient characteristics defined a priori. Future work is needed to understand the comparative value of PRM and data-driven approaches to understanding heterogeneity of treatment effects, which are becoming increasingly recognized as an important complement to the average treatment effect.

Yet, both methods imply a potential decision rule for prioritizing who may benefit the most from HRA + coaching. In a setting of constrained resources (i.e., how many patients can be managed by a health care coach), the data-driven method suggests priority should be given to patients with low numeracy, low education, and diabetes. The predictive risk model showed that those with lower overall risk of enrollment would have greater benefit from the intervention, so coaching resources should be directed to these patients. However, because risk was defined by an internally developed prediction model, identifying these patients would be more difficult to operationalize. Both approaches require validation in future work. Data-driven and predictive risk methods approach subgroup identification differently, answer related but slightly different questions, and differ in the heterogeneity observed in the effect of an intervention that combined a HRA with health coaching.

Appendix 1

See Table 5.

Table 5 Participant baseline characteristics stratified by missing data status

	Complete Case N = 365	Not Complete Case[1] N = 52
Age, mean (SD)	56.4 (11.7)	51.8 (14.8)
HRA Health age, mean (SD)[2,3]	60.9 (12.0)	56.6 (16.0)
Difference in age (Health age—Actual age), mean (SD)[2]	4.5 (5.7)	4.8 (5.8)
PAM score, mean (SD)	61.5 (12.5)	61.5 (12.4)
Framingham 10-year cardiovascular risk score, mean (SD)[2]	22.4 (16.4)	20.8 (20.1)
General health, n (%)[2]		
Excellent	23 (6.3)	0 (0)
Very good	88 (24.1)	10 (19.6)
Good	144 (39.5)	22 (43.1)
Fair	85 (23.3)	15 (29.4)
Poor	25 (6.8)	4 (7.8)
Employed full/part-time, n (%)[2]	126 (34.5)	27 (52.9)
Inadequate income, n (%)	99 (27.1)	12 (23.1)
Married/living as married, n (%)	191 (52.3)	22 (42.3)
Non-Hispanic white race, n (%)	172 (47.1)	28 (53.8)
Male gender, n (%)	315 (86.3)	41 (78.8)
Education, n (%)		
High school or less	66 (18.1)	9 (17.3)
Some college, Associate's degree, or trade school	211 (57.8)	29 (55.8)
Bachelor's degree or higher	88 (24.1)	14 (26.9)
Assistance required for reading, n (%)		
Never	251 (68.8)	41 (78.8)
Rarely	67 (18.4)	9 (17.3)
Sometimes/often/always	47 (12.9)	2 (3.8)
Mean numeracy score (range 1–6), mean (SD)[4]	4.6 (1.2)	4.4 (1.2)
Computer Literacy—ability to use, n (%)		
Do not use computer	22 (6.0)	2 (3.8)
Basic	70 (19.2)	11 (21.2)
Moderate	133 (36.4)	20 (38.5)
Advanced	98 (26.8)	12 (23.1)
Expert	42 (11.5)	7 (13.5)
Body mass index, mean (SD)	33.9 (6.3)	32.8 (7.2)
Current smoker of cigarettes or other tobacco, n (%)	140 (38.4)	23 (44.2)
Minutes of physical activity in past week, median (IQR)	150.0 (380.0)	112.5 (425.0)
Total number of inclusion criteria,[5] n (%)		
1	157 (43.0)	17 (32.7)
2	167 (45.8)	32 (61.5)
3	41 (11.2)	3 (5.8)
MOS-6 Sleep Scale Score, mean (SD)	61.0 (21.7)	61.1 (22.5)
Pain in past week[6], mean (SD)	4.6 (2.7)	4.1 (2.9)
PHQ-8 Total Score, mean (SD)	7.2 (5.5)	6.2 (5.3)

Table 5 (continued)

	Complete Case N = 365	Not Complete Case[1] N = 52
Alcohol consumption, n (%)		
Never	147 (40.3)	21 (40.4)
Monthly or less	80 (21.9)	14 (26.9)
2–4 times a month	53 (14.5)	6 (11.5)
2–3 times a week	52 (14.2)	7 (13.5)
4 or more times a week	33 (9.0)	4 (7.7)
Total cholesterol (md/dL), mean (SD)[2]	178.5 (41.9)	192.2 (40.2)
Diabetes diagnosis, n (%)	103 (28.2)	13 (25.0)
Average systolic blood pressure (mm Hg), mean (SD)[7]	129.9 (15.3)	130.0 (16.4)
High-density lipoprotein (mg/dL), mean (SD)[2]	45.9 (14.2)	45.2 (11.9)
Take blood pressure medications, n (%)	228 (62.5)	29 (55.8)

SD Standard deviation, *HRA* Health Risk Assessment via the HealtheLiving Assessment, *PAM* Patient Activation Measure, *IQR* Interquartile range, *MOS* Medical Outcomes Study, *PHQ* Patient Health Questionnaire

[1]40 Veterans missing the outcome (enrollment in prevention at 6 months) only, 4 Veterans missing both the outcome and at least one baseline characteristic, and 8 Veterans missing at least one baseline covariate only

[2]Missing values in not complete case (n = 52) data: HRA health age (n = 1), difference between age and HRA health age (n = 1), Framingham (n = 9), general health (n = 1), employed (n = 1) total cholesterol (n = 7), high-density lipoprotein (n = 7). Observations with missing data excluded from calculations

[3]HRA Health Age: The HRA uses a proprietary risk modeling algorithm to determine patients' "health age" based on lifestyle choices, family risk, and biological values, as well as information about the degree to which lifestyle changes can lower their "health age."

[4]Numeracy variables: skill with fractions, skill with percentages, and usefulness of numerical information in making health decisions; each variable is on a 1–6 scale, with a value of 1 anchoring "not at all good" and 6 anchoring "extremely good"

[5]To be included in the study, Veterans had to have at least one of the following modifiable risk factors: body mass index (BMI) ≥ 30, current smoker, or < 150 min of moderate/vigorous physical activity per week

[6]Range is 0–10, with 0 representing no pain

[7]Mean of two systolic blood pressure measurements

Appendix 2: Statistical code for MoB analyses

Part 1: R code

```
################################################################################
####
#Specify dichotomous variables as unordered factor, and ordinal variables as
#ordered factors
################################################################################

hte_Analyze <- hte_activatecc

hte_Analyze$EVER_ENROLLED_A <-factor(hte_Analyze$EVER_ENROLLED_A,ordered=FALSE)
hte_Analyze$armLID <-factor(hte_Analyze$armLID,ordered=FALSE)
hte_Analyze$GENHEALTH <- factor(hte_Analyze$GENHEALTH, ordered=TRUE)
hte_Analyze$EMPLOYMENT2 <- factor(hte_Analyze$EMPLOYMENT2, ordered=FALSE)
hte_Analyze$FINANCE2 <- factor(hte_Analyze$FINANCE2, ordered=FALSE)
hte_Analyze$MARITAL2 <- factor(hte_Analyze$MARITAL2, ordered=FALSE)
hte_Analyze$RACEW <- factor(hte_Analyze$RACEW, ordered=FALSE)
hte_Analyze$SEX <- factor(hte_Analyze$SEX, ordered=FALSE)
hte_Analyze$EDUC3 <- factor(hte_Analyze$EDUC3, ordered=TRUE)
hte_Analyze$LITERACY3<- factor(hte_Analyze$LITERACY3, ordered=TRUE)
hte_Analyze$COMPUTERLITERACY_R<- factor(hte_Analyze$COMPUTERLITERACY_R,
ordered=TRUE)
hte_Analyze$INCLUDE_SMOKE <- factor(hte_Analyze$INCLUDE_SMOKE, ordered=FALSE)
hte_Analyze$SUM_INCLUDE <- factor(hte_Analyze$SUM_INCLUDE, ordered=TRUE)
hte_Analyze$ALCOHOL <- factor(hte_Analyze$ALCOHOL, ordered=TRUE)
hte_Analyze$DMDX <- factor(hte_Analyze$DMDX, ordered=FALSE, levels = c(0,1),
labels = c("No", "Yes"))
hte_Analyze$Diabetes<- hte_Analyze$DMDX
hte_Analyze$BPMED <- factor(hte_Analyze$BPMED, ordered=FALSE)


#load libraries

library(vcd)
library(partykit)
library(strucchange)
library(Hmisc)
library(DescTools)

logit <- function(y, x, start = NULL, weights = NULL, offset = NULL, ...) {
  glm(y ~ 0 + x, family = binomial, start = start, ...)
}
#GLMTREE code for model

Ever_ENR_A_GLM <- glmtree(EVER_ENROLLED_A ~ armLID | AGE + HLA_HEALTH_AGE +
AGEDIFF + FRAMINGHAM_Score + GENHEALTH +
                          + EMPLOYMENT2 + FINANCE2 + MARITAL2 + RACEW  + SEX   +
                          + EDUC3 + LITERACY3 + NumMean+ COMPUTERLITERACY_R +
                          + BMI  + INCLUDE_SMOKE + PHYS_ACTIVITY + SUM_INCLUDE +
                          + MOS_Score +PAINPASTWEEK + PHQ_Score +
                          + ALCOHOL + CHOLESTEROL  + Diabetes + SBP  + HDL +
BPMED+
                          + PAM_Score  , data=hte_Analyze, family=binomial,alpha
= 0.05, bonferroni = FALSE, minsize = 40, prune = "AIC", verbose=TRUE,
                        catsplit = "binary", vcov = "opg", ordinal = "L2")
Ever_ENR_A_GLM
plot(Ever_ENR_A_GLM ,terminal_panel=NULL, main="Enrolled A: glmtree Min size 40
post pruning aic catsplit=binary vcov=opg ordinal=L2")


##c-statistic##
predicttest<-predict(Ever_ENR_A_GLM, newdata=hte_Analyze, type="response")
testnew <- cbind(hte_Analyze,predicttest)
mobsd <- SomersDelta(testnew$predicttest,testnew$EVER_ENROLLED_A)
mobcindex <- (mobsd/2) + 0.5
```

Part 2. SAS Code for c-for-benefit Predicted Risk Model (PRM)(PREDICTED_EVRENR is from step 1 of the PRM model)

```
ods graphics on;
proc psmatch data=PRMpreddata region=cs;
   class ARMLID;
   psmodel ARMLID(Treated='1')= PREDICTED_EVRENR;
   match method=optimal(k=1)
         distance=mah(lps var=(PREDICTED_EVRENR)) caliper=.
         weight=none;
   assess lps var=(PREDICTED_EVRENR);
   output out(obs=match)=OutEx7 matchid=_MatchID;
run;
proc sort data = outex7 out=outex7sort;
by _MatchID;
run;
proc sort data =outex7sort out= PRMPred_InterSort;
by _MatchID;
where ARMLID = 1;
run;
proc sort data= outex7sort out = PRMPred_ControlSort;
by _MatchID;
where ARMLID = 0;
run;

*merge together by _MatchID;
data PredTogetherForRand;
merge PRMPred_InterSort (in=a rename = (predicted=PredictedInter
EVER_ENROLLED_A = EVER_ENROLLED_A_INTER))
      PRMPred_ControlSort (in=a rename = (predicted=PredictedControl
EVER_ENROLLED_A = EVER_ENROLLED_A_Control));
by _MatchID;
       *averge the two predicted variables (inter and control);
       PredBenAverage = mean(PredictedInter, PredictedControl);

       *ObsBen is the difference between the enrollment_A of inter and
enrollment_A of control;
       ObsBen = EVER_ENROLLED_A_INTER- EVER_ENROLLED_A_Control;
run;
          proc print data=PredTogetherForRand;
          var _MATCHID  PredBenAverage PredictedInter PredictedControl
ObsBen EVER_ENROLLED_A_INTER EVER_ENROLLED_A_Control;
          run;

          proc freq data=PredTogetherForRand;
               tables ObsBen*PredBenAverage/measures nocol norow
nopercent;
               output out=somerD SMDCR;
          run;

data AllSomer;
set somerD;
Cbenefit = _SMDCR_/2 + 0.5;
run;

          proc freq data= AllSomer;
          tables Cbenefit;
          run;
```

Acknowledgements This work was supported by HSR&D funding (CRE 12-306, RCS 10-391) and by the Durham Center of Innovation to Accelerate Discovery and Practice Transformation (ADAPT), (CIN 13-410) at the Durham VA Health Care System. The views represented in this article represent those of the authors and not those of the VA or the United States Government. We are grateful for helpful comments on an earlier draft from Drs. Daniel Almirall, Sandeep Vijan, and David Kent.

References

Alemayehu, D., Chen, Y., Markatou, M.: A comparative study of subgroup identification methods for differential treatment effect: performance metrics and recommendations. Stat Methods Med. Res. **27**(12), 3658–3678 (2018)

Alosh, M., Huque, M.F., Bretz, F., D'Agostino, R.B., Sr.: Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. Stat. Med. **36**(8), 1334–1360 (2017)

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. The Lancet. Diabetes and Endocrinol. **5**(10), 808–815 (2017)

D'Agostino, R.B., Sr., Vasan, R.S., Pencina, M.J., Wolf, P.A., Cobain, M., Massaro, J.M., Kannel, W.B.: General cardiovascular risk profile for use in primary care: the framingham heart study. Circulation **117**(6), 743–753 (2008)

Doove, L.L., Dusseldorp, E., Van Deun, K., Van Mechelen, I.: A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. Adv. Data Anal. Classif. **8**(4), 403–425 (2014)

Dusseldorp, E., Conversano, C., Jan Van Os, B.: Combining an additive and tree-based regression model simultaneously: STIMA. J. Comput. Gr. Stat. **19**(3), 514–530 (2010)

Dusseldorp, E., Conversano, C., Van Os, B.J.: Combining an additive and tree-based regression model simultaneously: STIMA. J. Comput. Gr. Stat. **19**(3), 514–530 (2010)

Dusseldorp, E., Doove, L., Mechelen, I.: Quint: an R package for the identification of subgroups of clients who differ in which treatment alternative is best for them. Behav. Res. Methods. **48**(2), 650–663 (2016)

Harrell, F.E., Jr., Lee, K.L., Mark, D.B.: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat. Med. **15**(4), 361–387 (1996)

Hays, R.D., Stewart, A.L.: Sleep Measures. In: Stewart, A.L., Ware, J.E.J. (eds.) Measuring functioning and well-being; the medical outcomes study approach, pp. 235–259. Duke University Press, Durham (1992)

Hayward, R.A., Kent, D.M., Vijan, S., Hofer, T.P.: Reporting clinical trial results to inform providers, payers, and consumers. Health Aff. (Project Hope) **24**, 1571–1581 (2005)

Hibbard, J.H., Mahoney, E.R., Stock, R., Tusler, M.: Do increases in patient activation result in improved self-management behaviors? Health Serv. Res. **42**(4), 1443–1463 (2007)

Hibbard, J.H., Mahoney, E.R., Stockard, J., Tusler, M.: Development and testing of a short form of the patient activation measure. Health Serv. Res. **40**(6), 1918–1930 (2005)

Hothorn, T., Zeileis, A.: partykit: a modular toolkit for recursive partytioning in R. J. Mach. Learn. Res. **16**, 3905–3909 (2015)

Kent, D.M., Hayward, R.A.: Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA **298**(10), 1209–1212 (2007)

Kent, D.M., Rothwell, P.M., Ioannidis, J.P., Altman, D.G., Hayward, R.A.: Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials **11**, 85 (2010). https://doi.org/10.1186/1745-6215-11-85

Kent, D.M., Ruthazer, R., Selker, H.P.: Are some patients likely to benefit from recombinant tissue-type plasminogen activator for acute ischemic stroke even beyond 3 hours from symptom onset? Stroke **34**(2), 464–467 (2003)

Kravitz, R.L., Duan, N., Braslow, J.: Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q **82**(4), 661–687 (2004). https://doi.org/10.1111/j.0887-378X.2004.00327.x

Kroenke, K., Spitzer, R.L., Williams, J.B.: The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. **16**(9), 606–613 (2001)

Lagakos, S.W.: The challenge of subgroup analyses–reporting without distorting. N. Engl. J. Med. **354**(16), 1667–1669 (2006)

Lipkovich, I., Dmitrienko, A., Agostino, B.R.D.S.: Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat. Med. **36**(1), 136–196 (2017)
Loh, W.Y., He, X., Man, M.: A regression tree approach to identifying subgroups with differential treatment effects. Stat. Med. **34**(11), 1818–1833 (2015)
McNaughton, C.D., Cavanaugh, K.L., Kripalani, S., Rothman, R.L., Wallston, K.A.: Validation of a short, 3-item version of the subjective numeracy scale. Med. Decis. Making **35**(8), 932–936 (2015)
McVay, M.A., Yancy, W.S., Jr., Vijan, S., Van Scoyoc, L., Neelon, B., Voils, C.I., Maciejewski, M.L.: Obesity-related health status changes and weight-loss treatment utilization. Am J Prev Med **46**(5), 465–472 (2014)
MyHealtheVet: HealtheLiving Assessment. https://www.myhealth.va.gov/mhv-portal-web/web/myhealthevet/ss20170509-birds-eye-view-of-your-wellness-and-your-health-risks (2018). Accessed 28 November, 2019
Nouri, S.S., Damschroder, L.J., Olsen, M.K., Gierisch, J.M., Fagerlin, A., Sanders, L.L., McCant, F., Oddone, E.Z.: Health coaching has differential effects on veterans with limited health literacy and numeracy: a secondary analysis of ACTIVATE. J. Gen. Intern. Med. **34**(4), 552–558 (2019)
Oddone, E.Z., Gierisch, J.M., Sanders, L.L., Fagerlin, A., Sparks, J., McCant, F., May, C., Olsen, M.K., Damschroder, L.J.: A coaching by telephone intervention on engaging patients to address modifiable cardiovascular risk factors: a randomized controlled trial. J. Gen. Intern. Med. **33**(9), 1487–1494 (2018)
Rothwell, P.M.: Treating individuals Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet **365**(9454), 176–186 (2005). https://doi.org/10.1016/S0140-6736(05)17709-5
Rothwell, P.M., Warlow, C.P.: Prediction of benefit from carotid endarterectomy in individual patients: a risk-modelling study. European carotid surgery trialists' collaborative group. Lancet **353**(9170), 2105–2110 (1999)
Seibold, H., Zeileis, A., Hothorn, T.: Model-based recursive partitioning for subgroup analyses. Int. J. Biostat. **12**(1), 45–63 (2016)
Sies, A., Demyttenaere, K., Van Mechelen, I.: Studying treatment-effect heterogeneity in precision medicine through induced subgroups. J. Biopharm. Stat. **29**(3), 491–507 (2019)
Strobl, C., Malley, J., Tutz, G.: An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol. Methods **14**(4), 323–348 (2009). https://doi.org/10.1037/a0016973
Sussman, J.B., Kent, D.M., Nelson, J.P., Hayward, R.A.: Improving diabetes prevention with benefit based tailored treatment: risk based reanalysis of diabetes prevention program. BMJ **350**, h454 (2015). https://doi.org/10.1136/bmj.h454
van Klaveren, D., Steyerberg, E.W., Serruys, P.W., Kent, D.M.: The proposed 'concordance-statistic for benefit'provided a useful metric when modeling heterogeneous treatment effects. J. Clin. Epidemiol. **94**, 59–68 (2018)
Zeileis, A., Hothorn, T., Hornik, K.: Model-based recursive partitioning. J. Comput. Graph Stat. **17**, 492–514 (2008)

Authors and Affiliations

Maren K. Olsen[1,2] · Karen M. Stechuchak[1] · Eugene Z. Oddone[1,3] · Laura J. Damschroder[4,5] · Matthew L. Maciejewski[1,3,6]

✉ Maren K. Olsen
maren.olsen@duke.edu

[1] Durham Center of Innovation To Accelerate Discovery and Practice Transformation, Durham Veterans Affairs Health Care System, Durham VA Medical Center (152), 508 Fulton Street, Durham, NC 27705, USA

[2] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

[3] Division of General Internal Medicine, Department of Medicine, Duke University, Durham, NC, USA

[4] Ann Arbor VA HSR&D/Center for Clinical Management Research, P.O. Box 130170, Ann Arbor, MI, USA

[5] VA MIDAS QUERI Program, Ann Arbor, MI, USA

[6] Department of Population Health Sciences, Duke University, Durham, NC, USA