Journal of Clinical Epidemiology 176 (2024) 111538

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

Machine learning approaches to evaluate heterogeneous treatment effects in randomized controlled trials: a scoping review

Kosuke Inoue[a,b,*,1], Motohiko Adomi[c,1], Orestis Efthimiou[d,e], Toshiaki Komura[f], Kenji Omae[g,h], Akira Onishi[i], Yusuke Tsutsumi[j,k], Tomoko Fujii[l,m], Naoki Kondo[a], Toshi A. Furukawa[m]

[a]*Department of Social Epidemiology, Graduate School of Medicine, Kyoto University, Kyoto, Japan*
[b]*Hakubi Center, Kyoto University, Kyoto, Japan*
[c]*Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA*
[d]*Institute of Primary Health Care (BIHAM), University of Bern, Bern, Switzerland*
[e]*Institute of Social and Preventive Medicine (ISPM), University of Bern, Bern, Switzerland*
[f]*Department of Epidemiology, School of Public Health, Boston University, Boston, MA, USA*
[g]*Department of Innovative Research and Education for Clinicians and Trainees, Fukushima Medical University Hospital, Fukushima, Japan*
[h]*Center for Innovative Research for Communities and Clinical Excellence, Fukushima Medical University, Fukushima, Japan*
[i]*Department of Advanced Medicine for Rheumatic Diseases, Kyoto University Graduate School of Medicine, Kyoto, Japan*
[j]*Human Health Sciences, Kyoto University Graduate School of Medicine, Kyoto, Japan*
[k]*Department of Emergency Medicine, National Hospital Organization Mito Medical Center, Ibaraki, Japan*
[l]*Intensive Care Unit, Jikei University Hospital, Tokyo, Japan*
[m]*Departments of Health Promotion and Human Behavior and of Clinical Epidemiology, Kyoto University Graduate School of Medicine/School of Public Health, Kyoto, Japan*

Accepted 16 September 2024; Published online 19 September 2024

Abstract

Background and Objectives: Estimating heterogeneous treatment effects (HTEs) in randomized controlled trials (RCTs) has received substantial attention recently. This has led to the development of several statistical and machine learning (ML) algorithms to assess HTEs through identifying individualized treatment effects. However, a comprehensive review of these algorithms is lacking. We thus aimed to catalog and outline currently available statistical and ML methods for identifying HTEs via effect modeling using clinical RCT data and summarize how they have been applied in practice.

Study Design and Setting: We performed a scoping review using prespecified search terms in MEDLINE and Embase, aiming to identify studies that assessed HTEs using advanced statistical and ML methods in RCT data published from 2010 to 2022.

Results: Among a total of 32 studies identified in the review, 17 studies applied existing algorithms to RCT data, and 15 extended existing algorithms or proposed new algorithms. Applied algorithms included penalized regression, causal forest, Bayesian causal forest, and other metalearner frameworks. Of these methods, causal forest was the most frequently used (7 studies) followed by Bayesian causal forest (4 studies). Most applications were in cardiology (6 studies), followed by psychiatry (4 studies). We provide example R codes in simulated data to illustrate how to implement these algorithms.

Conclusion: This review identified and outlined various algorithms currently used to identify HTEs and individualized treatment effects in RCT data. Given the increasing availability of new algorithms, analysts should carefully select them after examining model performance and considering how the models will be used in practice.

Keywords: Heterogeneous treatment effect; Individualized treatment effect; Machine learning; Randomized controlled trial; Personalized medicine; Scoping review

Funding: Kosuke Inoue (KI) was supported by grant from the Japan Society for the Promotion of Science (22K17392 and 23KK0240), the Japan Agency for Medical Research and Development (AMED; JP22rea522107), the Japan Science and Technology (JST PRESTO; JPMJPR23R2), the Japan Health Insurance Association, and the Program for the Development of Next-generation Leading Scientists with Global Insight (L-INSIGHT) sponsored by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. Study sponsors were not involved in study design, data interpretation, writing, or the decision to submit the article for publication.

[1] equally contributed.

* Corresponding author. Department of Social Epidemiology, Graduate School of Medicine, Kyoto University, Floor 2, Science Frontier Laboratory, Yoshida-konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan.
E-mail address: inoue.kosuke.2j@kyoto-u.ac.jp (K. Inoue).

https://doi.org/10.1016/j.jclinepi.2024.111538

What is new?

Key findings

- In this scoping review, we identified 32 studies focusing on statistical and machine learning (ML) algorithms for assessing heterogeneous treatment effects (HTEs) in randomized controlled trial (RCT) data.
- By the end of 2022, this review categorized 17 studies as application papers and 15 studies as methodology papers.
- Although topics and algorithms varied across the studies, cardiology was the most popular field of application, and the causal forest was the most frequently applied model in healthcare literature.

What this adds to what was known?

- Despite the rapid development of statistical and ML methods to assess HTEs, evidence is limited regarding the commonality of each method's application in clinical research.
- This scoping review extends existing literature by detailing the practical application of various ML methods for HTE assessment in RCTs, offering guidance and example R codes in simulated data for implementation.
- We also described the strengths and limitations of each method, which will help researchers choose appropriate algorithms for investigating HTEs based on their research design and research purposes.

What is the implication and what should change now?

- When investigating HTEs in clinical epidemiology, researchers should carefully select algorithms based on the causal estimands of interest, the performance of the algorithms, and the practical application perspectives.

1. Introduction

Average treatment effect (ATE) is the primary focus of randomized controlled trials (RCTs), because establishing ATE is often required to obtain regulatory approval or change clinical guidelines through informing regulatory agencies and health-care practitioners about the expected treatment effects of interventions in the target population. When applying estimated ATEs to make treatment decisions, we implicitly assume that these estimated effects are applicable to all individuals in the population. This, however, may not be a plausible assumption to make in some cases, for example, when the magnitude and the direction of treatment effect vary substantially according to individual's baseline characteristics [1]. Even when RCTs report no evidence of a treatment effect on average, there may still be some individuals who benefit from treatment. For example, recent post hoc analyses have reported that a certain subpopulation may have a decreased risk of cardiovascular diseases through policy intervention [2], lifestyle intervention [3], intensive glucose control [4], and pharmacological therapy [5], while the original RCT reported a null ATE. Moreover, some participants could be harmed even when ATE indicates beneficial effects. If we only focus on ATE, such patients will miss the opportunity to receive benefit or avoid harm of the treatment. As the concept of personalized medicine has emerged over the years, the importance of assessing heterogeneous treatment effects (HTEs) has been widely recognized [6–8]. HE refers to the situation when the effect of treatment at individual levels, known as conditional average treatment effect (CATE) [9], is different across individuals or across patient subgroups. Estimating CATE allows us to prioritize individuals with high expected benefits from the intervention under the strong assumption that the results are unlikely to be false positive [10,11]. This is implemented in the "*high-benefit approach*" [12] and "*optimal treatment regimes*" [13] in the prior literature.

Over the last decades, a range of statistical and machine learning (ML) methods have been developed for assessing HTE and CATE [14], and have been implemented in open-source software packages such as R and Python [15,16]. Compared to modern methods, traditional approaches like 'one-variable-at-a-time' subgroup analysis have several limitations. The latter considers only one variable to create patient subgroups, and estimate treatment effects therein. Although straightforward and easy to use, the method has several limitations, such as an increased risk of false positives when considering many variables and the loss of statistical power of finding true effects [7]. Moreover, results from the analysis cannot be easily used to guide personalized choice of treatment; for example, if the analysis shows that the treatment is beneficial for males and older patients and harmful for females and younger patients, what about young males or older females? This issue could be addressed by dividing patients into additional groups based on more than one covariate, but such approach would intensify the multiple-testing problem and raise concerns about cherry-picking. Therefore, a principled approach is needed to determine which groups should be considered. To date, a wide range of statistical and ML methods have been proposed, which could be useful in modeling not only linear but nonlinear relationships and high-order interactions between covariates. However, the variety of methods may create confusion among epidemiologists regarding the optimal choice for practical applications. One caveat for the assessment of HTEs is that

the high predictive performance of the model does not necessarily correspond to accurate effect estimation. Although some reviews have summarized the characteristics of currently available methods for HTEs assessment [17−19], the evidence as to how common these methods are in epidemiologic research is limited.

In the Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement, two primary methods are described for assessing HTEs: risk modeling and effect modeling [8]. Risk modeling employs a multivariable approach to predict the outcome risk, followed by stratification of individuals based on the predicted risk. Effect modeling, alternatively, involves the development of models that incorporate interaction terms between treatment and baseline patient characteristics. While both approaches offer unique advantages and should not be exclusively favored, this review emphasizes effect modeling given the rapid advancement in data-driven methods for this approach. Specifically, we aimed to (1) summarize currently available statistical and ML methods for assessing HTEs via effect modeling that have been applied to RCT data and (2) provide a summary of how each algorithm works along with code for implementing it in R, exemplifying it with the use of simulated data. Our overall objective is to provide readers with guidance on how to apply methods to assess HTEs in large clinical RCTs.

2. Methods

This scoping review (ScR) was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Extension for ScRs [20] and the protocol was registered in Open Science Framework on April 2, 2023 [21]. The terminology we used is based on the PATH Statement [8].

2.1. Inclusion and exclusion criteria

Inclusion criteria are as follows: 1) no language restriction; 2) studies published from 2010 until 2022; 3) studies that developed, evaluated, or applied statistical or ML algorithms to predict HTE or CATE; 4) studies that applied an existing method to RCT datasets (ie, data involving random allocation of treatment strategies at individual levels); and 5) studies that conceptualized or modeled treatment strategies based on predicted CATE. In this review, we defined that CATE refers to the CATE, which is the treatment effect conditional on an individual's characteristics. More formally, within the counterfactual framework, CATE can be written as

$$\mathrm{E}[\mathrm{Y}_{t=1} - \mathrm{Y}_{t=0} \mid \mathrm{Z} = \mathrm{z}]$$

where Y_t denotes potential outcome Y under treatment $T = t$, and Z denotes a set of baseline characteristics.

Exclusion criteria are as follows: 1) conference abstract; 2) studies that only used summary data from RCTs; 3) studies that used datasets from cluster-randomized trials, cross-over trials, single-arm trials, or observational studies; 4) studies that only used simple regression-based methods (eg, linear or logistic regression without penalization) even with effect modeling approach such as the metalearner framework; 5) studies that conducted standard subgroup analysis (ie, stratified analysis by a single or some variables such as age, sex, and race); 6) studies that developed risk modeling and stratified patients solely based on prognostic models (ie, assess HTEs based on the predicted risk of outcome); and 7) studies that conceptualized treatment prioritization without specifying HTEs or CATE.

2.2. Search strategy

The search was conducted on two databases: MEDLINE and Embase via OVID. The search terms were determined through the meetings among all authors (Supplementary Tables 1 and 2). The search was performed on March 8, 2023 (MEDLINE), and on April 28, 2023 (Embase).

2.3. Selection of study and data extraction

After all duplicates were removed in the identified studies, six independent reviewers (KI, MA, KO, AO, YT, and TF) screened titles and abstracts. Full texts of the candidate studies were retrieved and underwent full-text screening. We also retrieved citations suggested by the authors of this review. The full-text screening was performed by the two reviewers (KI and MA). Disagreements between reviewers were resolved through consensus-driven discussion. After selecting studies, we categorized them into application studies (which aimed to assess HTEs by applying statistical and ML methods to RCT datasets) and methodological studies (which aimed to propose new methods and used RCT data as an example illustration).

The following basic information was extracted from the included studies after full-text screening and summarized in tables:

- Authors, year of publication (if the study was published online first, the corresponding year was regarded as the year of publication)
- Name, medical area, and sample size of RCT.
- Treatments randomized and outcome examined in RCT.
- Targeted measure of CATE (eg, risk difference, risk ratio, etc.)
- Analysis method used to assess HTEs (eg, decision tree, regularization, targeted learning, etc.)

For application papers, we additionally extracted the information on (i) whether ATE was significant or not in the original RCT and (ii) whether HTEs were identified or not. We also assessed whether each application paper

assessed the calibration and discrimination performance of the model on the treatment effect scale and took some approaches to avoid overfitting.

3. Results

After removing duplicates, a total of 3969 citations were identified in MEDLINE and Embase. After screening titles and abstracts, 3864 citations were excluded as they did not meet eligibility criteria. After seven citations hand-searched by the authors were added, the remaining 112 citations were reviewed in full text, 79 of which were excluded. The main reasons for exclusion were 1) did not use RCT datasets ($N = 29$); 2) did not predict CATE ($N = 13$); and 3) used an outcome prediction approach which was clarified during the full-text review ($N = 9$). Among the 33 citations included in this review, one citation was retracted in September 2023 [22], therefore, we reviewed the remaining 32 articles. The PRISMA flow diagram is shown in Figure 1.

3.1. Study characteristics

A total of 32 studies were included in the review. Of these, 17 (53%) focused on the application of existing algorithms to RCT datasets [3,23–38], and 15 (47%) focused on developing methods for HTE assessment [10,39–52]. Hereafter, we summarize the characteristics of studies separately by study type (ie, application paper or methodology paper).

3.1.1. Application papers

The characteristics of studies that used statistical and ML methods for an applied project ($N = 17$) are shown in Table 1. The most frequently used algorithms were causal forest ($N = 7$) and Bayesian Additive Regression Trees (BART) ($N = 4$). Additional algorithms included XGBoost, penalized regression, SuperLearner, support vector machines, and random forest in metalearner frameworks. The causal forest was applied multiple times to specific RCT datasets such as the Systolic Blood Pressure Intervention Trial [53] or the Action to Control Cardiovascular Risk in Diabetes study [54], and, as result, a total of six studies were in cardiovascular medicine. The remaining studies were in a variety of medical fields: geriatrics, intensive care, neurology, nutrition, psychiatry, respiratory medicine, and sociology. Regarding targeted measures of CATE, risk difference was specified in six studies, odds ratio in two studies, and risk ratio in one study. Other measures included hazard ratio ($N = 3$), difference in survival time ($N = 2$), and difference in the score of a continuous outcome ($N = 6$). The calibration performance of the model was assessed in only four studies, while discrimination performance was not formally assessed in any studies. Most studies (15 in total) employed cross-validation or similar approaches to avoid overfitting.

3.1.2. Methodology papers

The characteristics of studies that developed new methods for CATE ($N = 15$) are shown in Table 2. In most of the included methodological studies, the authors proposed the extension of a pre-existing method and compared the performance of multiple algorithms in simulations, using mean squared error of predicted CATE or population average outcome under the derived individualized treatment rule as performance measures. For example, Spanbauer et al proposed the extension of BART (mixedBART) to incorporate random effects and clustering of outcomes, and

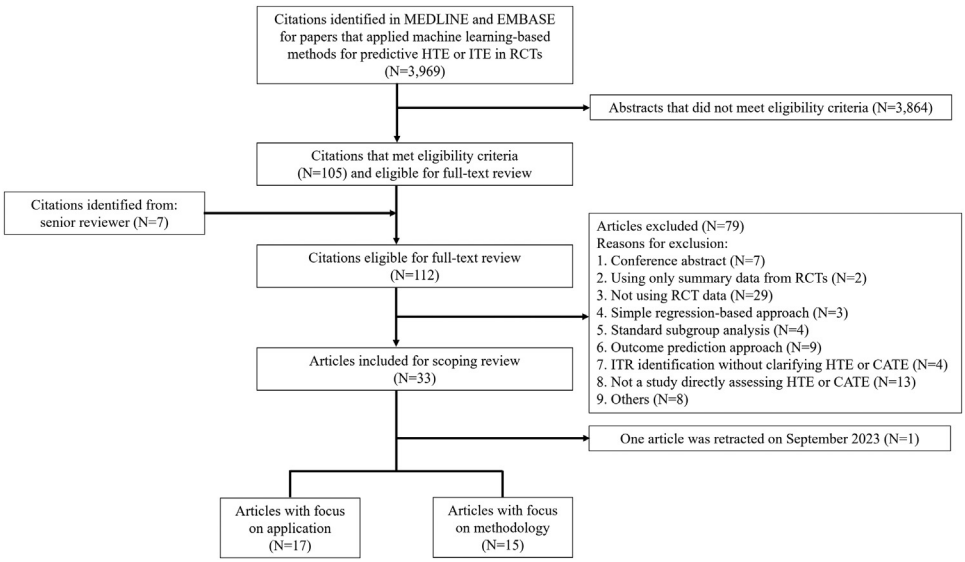

Figure 1. Study flow chart Footnote: HTE, heterogeneous treatment effect; CATE, conditional average treatment effect; RCT, randomized controlled trial; ITR, individualized treatment rule.

Table 1. Characteristics of studies that applied ML algorithms in RCTs

Author	Year	Field	Trial (sample size)	Treatment	Outcome	Causal estimand[a]	Method/Base-learner	Significant ATE reported in the original RCT?	HTE identified?
Edward	2022	Cardiovascular	1. ACCORD (N = 10,251) 2. VADT (N = 1,791)	Intensive glycemic control	MACE	RD	Causal forest	1. No 2. No	Yes
Falet	2022	Neurology	1. OPERA I (N = 821) 2. OPERA II (N = 835) 3. BRAVO (N = 1,331) 4. ORATORIO (N = 661) 5. OLYMPUS (N = 331) 6. ARPEGGIO (N = 318)	Anti-CD20 antibody	Disability progression	Survival time, HR	Deep learning	1. Yes 2. Yes 3. No 4. Yes 5. No 6. No	Yes
Kianmehr	2022	Cardiovascular	1. ACCORD (N = 10,251) 2. ACCORD-BP (N = 4,733)	1. Intensive glycemic control 2. Intensive BP control	Incident heart failure	RD, RR	Causal forest	1. No 2. No	1. Yes 2. Yes
Oikonomou	2022	Cardiovascular	1. SPRINT (N = 9,361) 2. ACCORD-BP (N = 4,733)	Intensive BP control	MACE	HR	XGBoost (along with the Gower method to define phenotypical neighborhood)	1. Yes 2. No	Yes
Sadique	2022	Intensive care	The 65 Trial (N = 2,463)	Permissive hypotension strategy	90-day mortality	RD	Causal forest	No	No
Hu	2021	Respiratory	The National Lung Screening Trial (N = 53,454)	Screening with low-dose CT vs CXR	1. Lung cancer mortality 2. Overall survival	The ratio of survival time	Accelerated failure time model with BART	Yes	Yes
Jiang	2021	Nutrition	IDEA trial (N = 343)	Exercise/diet	1. PCS 2. Weight loss 3. WOMAC scores 4. Compressive force 5. Plasma IL-6 level	RD	1. Penalized regression 2. Kernel ridge regression 3. Random forests 4. Reinforcement learning trees 5. List-based dynamic treatment regime 6. Residual weighted learning 7. BART	Yes	Yes
Kessler	2021	Psychiatry	SUN☺D (N = 1,549)	1. Sertraline only 2. Mirtazapine only 3. Sertraline + Mirtazapine	Depression remission at week 9	Difference in outcomes	SuperLearner	Yes	Yes
Raghavan	2021	Cardiovascular	1. ACCORD (N = 10,251) 2. VADT (N = 1,791)	Intensive glycemic control	All-cause mortality	RD	Causal forest	1. No 2. No	Yes
Sinha	2021	Intensive care	1. ALVEOLI (N = 549) 2. FACTT (N = 1000) 3. SAILS (N = 745)	1. PEEP management 2. Fluid management 3. Rosuvastatin	90-day mortality	OR	Unsupervised learning: 1) K-means clustering 2. Partitioning around medoids 3. Hierarchical clustering 4. Spectral clustering 5. Latent class analysis Supervised learning: 1. Model-based recursive partitioning 2. Causal forest 3. X-learner with Random Forest 4. X-learner with BART	1. No 2. No 3. No	Yes
Furukawa	2020	Psychiatry	SUN☺D (N = 1,549)	1. Sertraline only 2. Mirtazapine only 3. Sertraline + Mirtazapine	Depression remission at week 9	Difference in outcomes	1. Penalized regression (LASSO, ridge) 2. SVM 3. Neural network	Yes	Yes
Shepherd-Banigan	2020	Geriatrics	HI-FIVES (N = 241)	Caregiver education intervention	1. Number of days the veteran was not at home due to medical reason 2.Caregiver depressive symptoms at 12 month	Difference in outcomes	1. Model-based recursive partitioning 2. mCART 3. Random forest	Yes	Yes
Solnick	2020	Sociology	Original RCT (N = 3,277)	A clinical vignette was presented to the participant with a picture of the emergency department physician.	A composite of participant's confidence and satisfaction with the physician	Difference in outcomes	BART	No	No
Foster	2019	Psychiatry	TADS trial (N = 439)	1. Placebo 2. Cognitive-behavioral therapy (CBT) 3. Fluoxetine (FLX) 4. CBT and FLX	CDRS-R	Difference in outcomes	Model-based random forest	Yes	Yes

(*Continued*)

Table 1. Continued

Author	Year	Field	Trial (sample size)	Treatment	Outcome	Causal estimand[a]	Method/Base-learner	Significant ATE reported in the original RCT?	HTE identified?
Scarpa	2019	Cardiovascular	SPRINT (N = 9,361)	Intensive BP control	CV event	HR	Causal forest	Yes	Yes
Furukawa	2018	Psychiatry	1. Trial by Keller et al (N = 681) 2. REVAMP Trial (N = 296) 3. Trial by Schramm et al (N = 59)	1. Cognitive-behavioral analysis system of psychotherapy 2. Antidepressants 3. Combination	1. Depression severity 2. Dropout for any reason	Difference in outcomes, OR	Penalized regression	1. Yes 2. No 3. No	Yes
Baum	2017	Cardiovascular	Look Ahead (N = 5,145)	Weight loss intervention	CV event	RD	Causal forest	No	Yes

BART, Bayesian additive regression trees; BP, blood pressure; CDRS-R, Children's Depression Rating Scale-Revised; CT, computed tomography; CV event, cardiovascular event; CXR, chest X-rays; HR, hazard ratio; IL-6, interleukin-6; LASSO, least absolute shrinkage and selection operator; MACE, major adverse cardiovascular event; mCART, multivariate classification and regression tree; ML, machine learning; MMSE, mini-mental state examination; PCS, physical component score; PEEP, positive end-expiratory pressure; RD, risk difference; RR, risk ratio; SVM, support vector machine; WOMAC, Western Ontario and McMaster Universities Osteoarthritis Index; XGBoost, eXtreme Gradient Boosting.

[a] ''Difference in outcomes'' means that the authors investigated the difference in continuous outcomes (specified in ''outcome'' column) between treatment and control groups.

compared the performance of mixedBART and BART using root mean square error of the outcome as a metric in simulations [46]. In another example, Conzuelo Rodriguez et al compared the magnitude of bias of predicted CATE when using doubly robust learners compared to generalized linear models or flexible parametric models with spline functions [41]. In most of the 14 included studies, simulation data was first used to compare the performance of the newly proposed algorithm with pre-existing algorithms, and then an RCT dataset was used to show how the new algorithm could be applied to the real data. A total of 18 RCT datasets were used in 15 studies, and these datasets were from various fields including cardiology, endocrinology, and psychiatry.

3.2. Overview of methods for HTEs assessment

In Tables 3 and 4, we provided a summary of the identified methods for assessing HTE, outlining the way models work, and highlighting their respective strengths and limitations. In the Supplementary Method, we described two algorithms that are most commonly used in the studies identified: penalized regression and causal forest. Additionally, we discussed the Bayesian causal forest algorithm to highlight that, despite being a tree-based method for estimating CATE, it is not the Bayesian counterpart of the causal forest algorithm. We then introduced the general metalearner framework. Subsequently, we explained how to evaluate the calibration of these algorithms.

3.3. Implementation

To demonstrate how each algorithm works to identify HTEs, we provide R code in simulated data. In this implementation, we simulated a hypothetical RCT with 10,000 participants to investigate the effect of intensive blood pressure management on cardiovascular outcomes. Each individual has been attributed a potential outcome, either from the intervention (Y_1) or the placebo (Y_0); that is, Y_1 equals to the observed Y and Y_0 are not observed for the intervention group whereas Y_0 equals to the observed Y and Y_1 are not observed for the placebo group. Outcomes were labeled as 0 in the absence of events and one when events occurred. The treatment effect was calculated by contrasting these potential outcomes, where $\tau = Y_1 - Y_0$. Our data incorporated four baseline covariates, including age (continuous), systolic blood pressure (continuous), estimated glomerular filtration rate (eGFR; continuous), and statin use (binary). We simulated two scenarios of HTEs by setting different treatment effects based on eGFR values and statin use: (i) linear interaction between eGFR and treatment and (ii) nonlinear interaction between eGFR and treatment (ie, eGFR interacted with treatment only between 45 and 90 mL/min/1.73 m^2, and no interaction for eGFR <45 and ≥ 90 mL/min/1.73 m^2). Our code implements penalized regression, causal forest, Bayesian causal forest, and metalearners. The code is available online (https://github.com/Koinoue/HTE_review), and can be used for future implementation of the algorithms.

4. Discussion

In this ScR, we searched for published studies that applied existing or developed new methods for assessing HTEs in RCT data. Although topics and algorithms varied across studies, cardiology was the most popular field of application, and the causal forest was the most frequently applied model in health-care literature. We then outlined the identified algorithms, elucidating their architecture and highlighting their advantages and limitations. For example, penalized regression efficiently selects features and is less computationally intense while causal forest and Bayesian causal forest are less prone to misspecification of the nonlinear complex interaction. Regarding the metalearner framework, S-learner and T-learner are simple approaches while X-learner, DR-learner, and R-learner are particularly effective in scenarios where the sample size of

Table 2. Characteristics of studies that developed ML methods for the HTE assessment

Author	Year	Field	Trial used e.g. illustration (sample size)	Treatment	Outcome	Causal estimand	Methodological contribution
Doubleday	2022	Diabetes	DURABLE trial (N = 1,498)	Twice-daily insulin vs once-daily basal insulin	Change in HbA1c from baseline to week 24	Difference in outcomes	Proposed risk-controlled individual treatment rule (rcITR) estimation using decision tree/random forest
Montoya	2022	Psychiatry	Correctional Intervention for People with Mental Illness "Intervetions" trial (N = 441)	CBT	Recidivism	RD	Provided optimal dynamic treatment rule framework using SuperLearner
Conzuelo Rodriguez	2021	Pregnancy	EAGeR Trial (N = 1,228)	Low-dose aspirin	Live birth	RD	Compared performance between generalized linear models and DR-learner (using SuperLearner)
Du	2021	Cardiovascular	SOLVD-T (N = 2,569)	Enalapril	Time to hospitalization/ death	Difference in survival time	Proposed constrained Lasso approach
Fazzari	2021	Neurology	AADDOPT-2 (N = 569)	Acupuncture	12-week chronic pain	RD	Proposed virtual twin method
Guo	2021	Nutrition	1. Almond consumption trial (N = 68) 2. Avocado consumption trial (N = 108)	Almond/avocado consumption	Composition of GI microbiota and host characteristics	RD	Proposed Multiple Outcome Treatment Effect Forests (MOTEF)
Li	2021	Sociology	Jobs dataset (N = 2,915)	Employment program	Trainee earnings	RD	Proposed causal optimal transport model
Spanbauer	2021	Diabetes Infection	1. TBSI trial (N = 255) 2. ACTG175 Study (N = 1,762)	1. Knowledge/ motivation intervention 2. Didanosine/ azidothymidine treatment	1. HbA1c change from baseline 2. Relative change of CD4 T-cell count	RD	Extended BART by incorporating random effects and clustering for repeated measures (mixedBART)
Chen	2020	Psychiatry	STAR*D (N = 2,555)	SSRIs	Depressive symptoms (HAM-D sum score, QIDS, WSAS, and CGI)	Difference in outcomes	Proposed integrated learning framework using multi-layer neural network
Henderson	2018	Cardiovascular	1. SOLVD-T (N = 2,569) 2. SOLVD-P (N = 4,228)	ACE inhibitor	Time until death/ hospitalization	Difference in expected log-survival	Proposed Bayesian accelerated failure time models
Seibold	2018	Neurology	PRO-ACT (N = 3,306)	Riluzole	Survival time and the ALSFRS at 6 months	Difference in survival time	Proposed personalized models using model-based random forest in time-to-event data
Zhu	2017	Neurology	CATIE-AD (N = 213)	Atypical antipsychotics	Minimal improvement on the CGI scale at 12 weeks	OR	Proposed weighted random forests
Lipkovich	2016	Infection Hematology	Two RCT-datasets, name not specified (N = 470, N = 599)	1. Novel treatment for sepsis 2. Experimental therapy for hematological malignancy	1. All-cause survival at 28 days 2. Overall survival	RD, HR	Illustrated subgroup identification based on 1) Differential effect search (SIDES) 2) Virtual twins' method (VT) 3) Outcome-weighted learning (OWL)
Shen	2016	Cardiovascular	AVID (N = 1,016)	Defibrillator	Two-year Mortality	RD	Proposed Bayesian tree based latent variable model
Weiss	2015	Gastrointestinal	Primary biliary cirrhosis dataset (N = 288)	D-penicillamine	Three-year survival	RD	Compared performance between logistic regression models and AdaBoost

ALSFRS, ALS functional rating scale; BART, Bayesian additive regression trees; CBT, cognitive behavioral therapy; CGI, clinical global impression scale; DR-learner, doubly robust learner; HAM-D, Hamilton depression rating scale; HR, hazard ratio; ML, machine learning; QIDS, quick inventory of depressive symptomatology; OR, odds ratio; RD, risk difference; RR, risk ratio; SSRI, selective serotonin reuptake inhibitors; WSAS, work and social functioning.

Table 3. Summary concept of commonly used algorithms to assess heterogeneous treatment effect

Algorithm	Method paper	Brief description	Strengths	Limitations
Penalized regression (LASSO, ridge, elastic net)	Imai and Ratkovic. 2013 [55]	LASSO is a penalized regression model that shrinks regression coefficients, aiming to maximize predictive performance in new samples. It also performs variable selection, by completely removing some predictors from the model. Ridge is similar to LASSO but does not perform variable selection. Elastic net combines LASSO and ridge. Treatment-covariate interactions can be included in all penalized regression models, to model CATE.	-Feature selection -Simplicity	-Cannot account for interaction by covariates if not prespecified -Difficulty of detecting interactions across a high-dimensional set of covariates.
Causal tree/ causal forest[a]	Wager and Athey. 2018 [56]	Causal forest, a forest-based algorithm, splits samples to maximize the variance in treatment effect estimates across leaves (defined by individual characteristics), employing the R-learner framework to minimize loss function. It adopts an 'honest' estimation approach by using different subsamples for growing trees and estimating CATE, ensuring independence between tree structure and effect estimation.	-Nonparametric (less prone to misspecification of the nonlinear complex interaction) -Embedded estimation of uncertainty	-Computational intensity
Bayesian additive regression trees/ Bayesian causal forest[a]	Hahn et al 2020 [57]	Bayesian causal forest applies two Bayesian additive regression trees functions to evaluate the HTEs. This algorithm calculates the sum of base trees to predict outcomes, and updates the trees to minimize the residual iteratively (MCMC). The framework avoids overfitting and reduces ''regularization-induced confounding'' (which occurs particularly in observational studies).	-Nonparametric (less prone to misspecification of the nonlinear complex interaction) -Embedded estimation of uncertainty	-Computational intensity

CATE, conditional average treatment effect; LASSO, least absolute shrinkage and selection operator; ML, machine learning; MCMC, Markov chain Monte Carlo.

[a] Strengths and limitations of R-learner can also be applied in these methods while Bayesian Causal Forest takes some approaches to consider nonoverlap regions of covariates.

the treatment group is much larger than the other or the covariate distribution is imbalanced which often occurs in observational studies. We provided R code using simulated data; the code can be used to implement the various algorithms, exemplify their use, and facilitate the uptake of these methods in future epidemiological research.

Traditionally, the focus of RCT designs has been on the estimation of ATE in the target population. In some clinical specialties, such as cardiology, multiple studies applied ML-based methods to RCT datasets, indicating the increased interest in HTEs assessment using ML-based methods. Such HTEs assessment via CATE estimation also allows us to create treatment strategies that prioritize individuals who are expected to receive benefit from the treatment (''high-benefit approach'') rather than treating individuals at high risk of developing the outcome (''high-risk approach'') [12,63]. Thus, when designing future RCTs, researchers may prespecify the HTEs assessment methods in the protocol, and include a sufficient set of known or suspected effect modifiers in the study to enrich HTEs assessment. Meanwhile, assessing HTEs via effect modeling typically requires large samples. Given that interaction effects equal to ATE require a sample size four times larger [64], ability to assess HTEs would be limited if we use a single RCT with a small sample, as observed in several studies in our review. In such scenarios, individual participant data metaanalysis could be a viable option to overcome this limitation if data are available [65,66].

There are three important points to note. First, some algorithms assess HTEs on an absolute scale (eg, risk difference, change in score, etc.), but not on a relative scale (eg, risk ratio, odds ratio, hazard ratio). Although assessing HTEs on an absolute scale may be more relevant from a public health perspective—partially because the estimated effects need to be weighed against the harms and costs of the treatment [67,68]—assessing HTEs on a relative scale is also important, particularly when distinguishing between prognostic factors and effect measure modifiers. Indeed, the PATH statement recommends reporting treatment effects in both absolute and relative scale [8]. When researchers want

Table 4. Metalearner framework to assess heterogeneous treatment effect

Algorithm	Method paper	Brief description	Strengths	Limitations
S-learner	Hill et al 2011 [58]	It predicts outcome under treatment and control using base learners that model the interaction between treatment and covariates.	-Simple -Perform better than T-learner when the treatment effect is simple or even zero	-It is not directly optimized to estimate the treatment effect -Risks exclusion of treatment variable from the model when using methods with variable selection as base learners. -Unstable when treatment and control groups are imbalanced.
T-learner	Athey and Imbens. 2016 [59]	It predicts outcomes separately under treatment and control using base learners, and then subsequently calculates the difference in their expected outcome values.	-Simple -Explicit modeling in treatment and control groups, separately - Perform better than S-learner when the treatment effect is strongly heterogeneous	-It is not directly optimized to estimate the treatment effects -Unstable when treatment and control groups are imbalanced.
X-learner, DR-learner	Künzel et al 2019, Kennedy. 2023 [60,61]	It estimates treatment effects on treated patients and on untreated patients using the difference between observed outcomes and estimated counterfactuals for each group. It incorporates propensity score weights to address scenarios with imbalances in covariate distribution (which often occur in observational studies). DR-learner is a similar form of X-learner using a doubly robust estimator instead of propensity score weights.	-Directly estimates heterogenous treatment effects. -Particularly effective in scenarios with imbalanced designs (which often occur in observational studies).	-Complex -Unstable in the presence of extreme propensity scores. -Vulnerable to model misspecification of the propensity scores.
R-learner	Nie and Wager. 2021 [62]	It calculates propensity scores of the exposure and marginal outcomes, calculates the residuals of treatment and outcome, and then minimizes the loss function defined by these residuals. R-Learner requires ML that incorporate some form of regularization for minimizing the loss function.	-Uses different subsamples to estimate the nuisance parameters and to predict CATE by constructing a direct loss function on it.	-Complex -Unstable in the presence of extreme propensity scores. -Vulnerable to model misspecification of the propensity scores.

CATE, conditional average treatment effect; ML, machine learning.

to assess heterogeneity on a relative scale, they may want to use approaches involving the calculation of potential outcomes under treatment and control to obtain such estimands for each individual. Second, once these models are built, it is crucial to check model performance. One thing that complicates this assessment is that a model may be good at predicting absolute outcomes but may nevertheless fail in predicting treatment benefit [69,70]. Thus, assessing model calibration and discrimination, as in simple prediction models, is not enough. Moreover, while S-learner and T-learner predict outcomes among treated and untreated, some models such as causal forest and Bayesian causal forest directly predict CATE which further complicates this assessment. Recently, a range of methods and measures was developed specifically for assessing performance of models for predicting CATE [71–74]. It is also crucial that such an evaluation avoids issues related to overfitting. One way to do this is via using resampling methods (eg, bootstrapping) or data splitting methods such as k-fold cross-validation [73]. While we need more comprehensive discussion on how to evaluate the comparative performance of each method, a standard checklist would be helpful for authors to report these analyses in future applications. Lastly, our review only covers effect modeling approach to assess HTEs. In general, the effect modeling approach is prone to several pitfalls such as overfitting, low statistical power, and multiplicity owing to using multiple treatment interactions [75]. Furthermore, it sometimes lacks sufficient prior knowledge of critical effect modifiers. When these issues cannot be avoided despite employing some statistical approaches such as penalization, consideration of the risk modeling approach—another useful approach for analyzing HTEs in RCT data—is recommended.

Our study has several limitations. First, our review focused on the applications to RCTs and did not consider observational studies. Focusing on randomized datasets helped simplify and clarify the differences among the existing approaches. Given the recent development of approaches

to identify HTEs in observational studies [76,77], future work is needed targeting observational studies in addition to RCTs. Second, our review has covered the literature up to January 2023 only, while the number of publications in ML-based assessments of HTEs in health-care literature has been increasing steadily and some methods (eg, model-based forests) have been newly developed and extended after 2023. However, such is inevitable for any review of a hot and rapidly developing topic, and we believe we were able to cover essential methods. Third, several studies were excluded from this review because they focused on the identification of subgroups rather than CATE estimation. One example is the paper that applied interaction trees to identify qualitative interactions (a type of interaction where, one treatment is better than the other for some subgroups of patients, whereas the reverse is true for other subgroups) in the study population [78]. Such methods are more suitable when the primary objective is to identify specific subgroups with large (or small) treatment effects, rather than CATE. Lastly, our study did not aim to compare model performance across algorithms. While the causal forest algorithm was most frequently applied, our results do not necessarily suggest that the algorithm performs better than the others. Further simulation studies and prospective studies are required to externally validate each algorithm's performance (including the performance of algorithm-based prioritization of treatment) and assess their comparative strengths and limitations.

Due to the increasing availability of statistical and ML methods for assessing treatment effects at the individual level, epidemiologists should carefully select algorithms based on the causal estimands of interest, the performance of the algorithms, and the practical application perspectives.

Ethics statement

Not required.
Consent to participate: Not applicable.
Consent to publish: Not applicable.

Transparency statement

The corresponding author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

CRediT authorship contribution statement

Kosuke Inoue: Writing − review & editing, Writing − original draft, Supervision, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Motohiko Adomi:** Writing − review & editing, Writing − original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Orestis Efthimiou:** Writing − review & editing, Writing − original draft, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Toshiaki Komura:** Writing − review & editing, Writing − original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kenji Omae:** Writing − review & editing, Writing − original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Akira Onishi:** Writing − review & editing, Writing − original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yusuke Tsutsumi:** Writing − review & editing, Writing − original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Tomoko Fujii:** Writing − review & editing, Writing − original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Naoki Kondo:** Writing − review & editing, Writing − original draft, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Toshi A. Furukawa:** Writing − review & editing, Writing − original draft, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Data availability

All data, protocols, and statistical codes are available either through the manuscript and supplementary materials, the Open Science Framework (https://osf.io/3fqgh/), or GitHub (https://github.com/Koinoue/HTE_review).

Declaration of competing interest

All authors have completed the ICMJE uniform disclosure format (available on request from the corresponding author). M.A. received financial supports from Kyoto University related to this work. There are no competing interests for any other author.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2024.111538.

References

[1] Varadhan R, Segal JB, Boyd CM, Wu AW, Weiss CO. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol 2013;66(8):818−25. https://doi.org/10.1016/j.jclinepi.2013.02.009.

[2] Inoue K, Athey S, Baicker K, Tsugawa Y. Heterogeneous effects of Medicaid coverage on cardiovascular risk factors: secondary analysis of randomized controlled trial. BMJ 2024;386:e079377. https://doi.org/10.1136/bmj-2024-079377.

[3] Baum A, Scarpa J, Bruzelius E, Tamler R, Basu S, Faghmous J. Targeting weight loss interventions to reduce cardiovascular complications of type 2 diabetes: a machine learning-based post-hoc analysis of heterogeneous treatment effects in the Look AHEAD trial. Lancet Diabetes Endocrinol 2017;5(10):808−15. https://doi.org/10.1016/S2213-8587(17)30176-6.

[4] Kiyohara K, Kondo N, Iwami T, Yano Y, Nishiyama A, Node K, et al. Heterogeneous effects of intensive glycemic and blood pressure on cardiovascular events among diabetes by living arrangements. J Am Heart Assoc 2024;13(13):e033860. https://doi.org/10.1161/JAHA.123.033860.

[5] Desai RJ, Glynn RJ, Solomon SD, Claggett B, Wang SV, Vaduganathan M. Individualized treatment effect prediction with machine learning — salient considerations. NEJM Evid 2024;3(4):EVIDoa2300041. https://doi.org/10.1056/EVIDoa2300041.

[6] Angus DC, Chang CCH. Heterogeneity of treatment effect: estimating how the effects of interventions vary across individuals. JAMA 2021;326(22):2312−3. https://doi.org/10.1001/jama.2021.20552.

[7] Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. Int J Epidemiol 2016;45(6):2184−93. https://doi.org/10.1093/ije/dyw125.

[8] Kent DM, Paulus JK, van Klaveren D, D'Agostino R, Goodman S, Hayward R, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. Ann Intern Med 2020;172(1):35−45. https://doi.org/10.7326/M18-3667.

[9] Tipton E. Beyond generalization of the ATE: designing randomized trials to understand treatment effect heterogeneity. J R Stat Soc Ser A Stat Soc 2021;184(2):504−21. https://doi.org/10.1111/rssa.12629.

[10] Lipkovich I, Dmitrienko A, D'Agostino BR Sr. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat Med 2017;36(1):136−96. https://doi.org/10.1002/sim.7064.

[11] van Klaveren D, Balan TA, Steyerberg EW, Kent DM. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. J Clin Epidemiol 2019;114:72−83. https://doi.org/10.1016/j.jclinepi.2019.05.029.

[12] Inoue K, Athey S, Tsugawa Y. Machine-learning-based high-benefit approach versus conventional high-risk approach in blood pressure management. Int J Epidemiol 2023;52(4):1243−56. https://doi.org/10.1093/ije/dyad037.

[13] Li Z, Chen J, Laber E, Liu F, Baumgartner R. Optimal treatment regimes: a review and empirical comparison. Int Stat Rev 2023;91(3):427−63. https://doi.org/10.1111/insr.12536.

[14] 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. https://doi.org/10.1002/sim.7623.

[15] Tibshirani J, Athey S, Friedberg R, Hadad V, Hirshberg D, Miner L, et al. Grf: generalized random forests. Available at: https://cran.r-project.org/web/packages/grf/. Accessed October 22, 2023.

[16] McCulloch R, Sparapani R, Gramacy R, Pratola M, Spanbauer C, Plummer M, et al. BART: bayesian additive regression trees. Available at: https://cran.r-project.org/web/packages/BART/. Accessed October 22, 2023.

[17] Hu A. Heterogeneous treatment effects analysis for social scientists: a review. Soc Sci Res 2023;109:102810. https://doi.org/10.1016/j.ssresearch.2022.102810.

[18] Ling Y, Upadhyaya P, Chen L, Jiang X, Kim Y. Emulate randomized clinical trials using heterogeneous treatment effect estimation for personalized treatments: methodology review and benchmark. J Biomed Inform 2023;137:104256. https://doi.org/10.1016/j.jbi.2022.104256.

[19] Lipkovich I, Svensson D, Ratitch B, Dmitrienko A. Modern approaches for evaluating treatment effect heterogeneity from clinical trials and observational data. Stat Med 2024;43:10167. https://doi.org/10.1002/sim.10167.

[20] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467−73. https://doi.org/10.7326/M18-0850.

[21] Inoue K, Adomi M, Efthimiou O, Komura T, Omae K, Onishi A, et al. Machine learning approaches to identify individualized treatment effect in randomized controlled trial: a scoping review. Available at: https://osf.io/3fqgh/. Accessed October 22, 2023.

[22] Duan T, Rajpurkar P, Laird D, Ng AY, Basu S. Clinical value of predicting individual treatment effects for intensive blood pressure therapy: a machine learning experiment to estimate treatment effects from randomized trial data. Circ Cardiovasc Qual Outcomes 2019;12(3):e005010. https://doi.org/10.1161/CIRCOUTCOMES.118.005010.

[23] Edward JA, Josey K, Bahn G, Caplan L, Reusch JEB, Reaven P, et al. Heterogeneous treatment effects of intensive glycemic control on major adverse cardiovascular events in the ACCORD and VADT trials: a machine-learning analysis. Cardiovasc Diabetol 2022;21(1):58. https://doi.org/10.1186/s12933-022-01496-7.

[24] Falet JPR, Durso-Finley J, Nichyporuk B, Schroeter J, Bovis F, Sormani MP, et al. Estimating individual treatment effect on disability progression in multiple sclerosis using deep learning. Nat Commun 2022;13(1):5645. https://doi.org/10.1038/s41467-022-33269-x.

[25] Kianmehr H, Guo J, Lin Y, Luo J, Cushman W, Shi L, et al. A machine learning approach identifies modulators of heart failure hospitalization prevention among patients with type 2 diabetes: a revisit to the ACCORD trial. J Diabetes Complications 2022;36(9):108287. https://doi.org/10.1016/j.jdiacomp.2022.108287.

[26] Oikonomou EK, Spatz ES, Suchard MA, Khera R. Individualising intensive systolic blood pressure reduction in hypertension using computational trial phenomaps and machine learning: a post-hoc analysis of randomised clinical trials. Lancet Digit Health 2022;4(11):e796−805. https://doi.org/10.1016/S2589-7500(22)00170-4.

[27] Sadique Z, Grieve R, Diaz-Ordaz K, Mouncey P, Lamontagne F, O'Neill S. A machine-learning approach for estimating subgroup- and individual-level treatment effects: an illustration using the 65 trial. Med Decis Making 2022;42(7):923−36. https://doi.org/10.1177/0272989X221100717.

[28] Hu L, Lin JY, Sigel K, Kale M. Estimating heterogeneous survival treatment effects of lung cancer screening approaches: a causal machine learning analysis. Ann Epidemiol 2021;62:36−42. https://doi.org/10.1016/j.annepidem.2021.06.008.

[29] Jiang X, Nelson AE, Cleveland RJ, Beavers DP, Schwartz TA, Arbeeva L, et al. Precision medicine approach to develop and internally validate optimal exercise and weight-loss treatments for overweight and obese adults with knee osteoarthritis: data from a single-center randomized trial. Arthritis Care Res 2021;73(5):693−701. https://doi.org/10.1002/acr.24179.

[30] Kessler RC, Furukawa TA, Kato T, Luedtke A, Petukhova M, Sadikova E, et al. An individualized treatment rule to optimize probability of remission by continuation, switching, or combining antidepressant medications after failing a first-line antidepressant in a two-stage randomized trial. Psychol Med 2022;52(15):3371−80. https://doi.org/10.1017/S0033291721000027.

[31] Raghavan S, Josey K, Bahn G, Reda D, Basu S, Berkowitz SA, et al. Generalizability of heterogeneous treatment effects based on causal forests applied to two randomized clinical trials of intensive glycemic control. Ann Epidemiol 2022;65:101−8. https://doi.org/10.1016/j.annepidem.2021.07.003.

[32] Sinha P, Spicer A, Delucchi KL, McAuley DF, Calfee CS, Churpek MM. Comparison of machine learning clustering algorithms for detecting heterogeneity of treatment effect in acute respiratory distress syndrome: a secondary analysis of three randomised controlled trials. EBioMedicine 2021;74:103697. https://doi.org/10.1016/j.ebiom.2021.103697.

[33] Furukawa TA, Debray TPA, Akechi T, Yamada M, Kato T, Seo M, et al. Can personalized treatment prediction improve the outcomes, compared with the group average approach, in a randomized trial? Developing and validating a multivariable prediction model in a pragmatic megatrial of acute treatment for major depression. J Affect Disord 2020;274:690−7. https://doi.org/10.1016/j.jad.2020.05.141.

[34] Shepherd-Banigan M, Smith VA, Lindquist JH, Cary MP, Miller KEM, Chapman JG, et al. Identifying treatment effects of an informal

caregiver education intervention to increase days in the community and decrease caregiver distress: a machine-learning secondary analysis of subgroup effects in the HI-FIVES randomized clinical trial. Trials 2020;21(1):189. https://doi.org/10.1186/s13063-020-4113-x.

[35] Solnick RE, Peyton K, Kraft-Todd G, Safdar B. Effect of physician gender and race on simulated patients' ratings and confidence in their physicians: a randomized trial. JAMA Netw Open 2020;3(2): e1920511. https://doi.org/10.1001/jamanetworkopen.2019.20511.

[36] Foster S, Mohler-Kuo M, Tay L, Hothorn T, Seibold H. Estimating patient-specific treatment advantages in the 'treatment for adolescents with depression study.'. J Psychiatr Res 2019;112:61−70. https://doi.org/10.1016/j.jpsychires.2019.02.021.

[37] Scarpa J, Bruzelius E, Doupe P, Le M, Faghmous J, Baum A. Assessment of risk of harm associated with intensive blood pressure management among patients with hypertension who smoke: a secondary analysis of the systolic blood pressure intervention trial. JAMA Netw Open 2019; 2(3):e190005. https://doi.org/10.1001/jamanetworkopen.2019.0005.

[38] Furukawa TA, Efthimiou O, Weitz ES, Cipriani A, Keller MB, Kocsis JH, et al. Cognitive-behavioral analysis system of psychotherapy, drug, or their combination for persistent depressive disorder: personalizing the treatment choice using individual participant data network metaregression. Psychother Psychosom 2018;87(3): 140−53. https://doi.org/10.1159/000489227.

[39] Doubleday K, Zhou J, Zhou H, Fu H. Risk controlled decision trees and random forests for precision Medicine. Stat Med 2022;41(4): 719−35. https://doi.org/10.1002/sim.9253.

[40] Montoya LM, Van Der Laan MJ, Luedtke AR, Skeem JL, Coyle JR, Petersen ML. The optimal dynamic treatment rule superlearner: considerations, performance, and application to criminal justice interventions. Int J Biostat 2023;19(1):217−38. https://doi.org/10.1515/ijb-2020-0127.

[41] Conzuelo Rodriguez G, Bodnar LM, Brooks MM, Wahed A, Kennedy EH, Schisterman E, et al. Performance evaluation of parametric and nonparametric methods when assessing effect measure modification. Am J Epidemiol 2022;191(1):198−207. https://doi.org/10.1093/aje/kwab220.

[42] Du Y, Chen H, Varadhan R. Lasso estimation of hierarchical interactions for analyzing heterogeneity of treatment effect. Stat Med 2021; 40(25):5417−33. https://doi.org/10.1002/sim.9132.

[43] Fazzari MJ, Kim MY. Subgroup discovery in non-inferiority trials. Stat Med 2021;40(24):5174−87. https://doi.org/10.1002/sim.9118.

[44] Guo B, Holscher HD, Auvil LS, Welge ME, Bushell CB, Novotny JA, et al. Estimating heterogeneous treatment effect on multivariate responses using random forests. Stat Biosci 2021;15: 545−61. https://doi.org/10.1007/s12561-021-09310-w.

[45] Li Q, Wang Z, Liu S, Li G, Xu G. Causal optimal transport for treatment effect estimation. IEEE Trans Neural Netw Learn Syst 2023; 34(8):4083−95. https://doi.org/10.1109/TNNLS.2021.3118542.

[46] Spanbauer C, Sparapani R. Nonparametric machine learning for precision medicine with longitudinal clinical trials and Bayesian additive regression trees with mixed models. Stat Med 2021;40(11):2665−91. https://doi.org/10.1002/sim.8924.

[47] Chen Y, Zeng D, Xu T, Wang Y. Representation learning for integrating multi-domain outcomes to optimize individualized treatments. Adv Neural Inf Process Syst 2020;33:17976−86.

[48] Henderson NC, Louis TA, Rosner GL, Varadhan R. Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. Biostatistics 2020;21(1): 50−68. https://doi.org/10.1093/biostatistics/kxy028.

[49] Seibold H, Zeileis A, Hothorn T. Individual treatment effect prediction for amyotrophic lateral sclerosis patients. Stat Methods Med Res 2018; 27(10):3104−25. https://doi.org/10.1177/0962280217693034.

[50] Zhu K, Huang Y, Zhou XH. Tree-based ensemble methods for individualized treatment rules. Biostat Epidemiol 2018;2(1):61−83. https://doi.org/10.1080/24709360.2018.1435608.

[51] Shen C, Hu Y, Li X, Wang Y, Chen P, Buxton AE. Identification of subpopulations with distinct treatment benefit rate using the Bayesian tree. Biom J 2016;58(6):1357−75. https://doi.org/10.1002/bimj.201500180.

[52] Weiss J, Kuusisto F, Boyd K, Liu J, Page D. Machine learning for treatment assignment: improving individualized risk attribution. AMIA Annu Symp Proc 2015;2015:1306−15.

[53] SPRINT Research Group, Wright JT Jr, Williamson JD, Whelton PK, Snyder JK, Sink KM. A randomized trial of intensive versus standard blood-pressure control. N Engl J Med 2015;373(22):2103−16. https://doi.org/10.1056/NEJMoa1511939.

[54] ACCORD Study Group, Cushman WC, Evans GW, Byington RP, Goff DC Jr, Grimm RH Jr. Effects of intensive blood-pressure control in type 2 diabetes mellitus. N Engl J Med 2010;362(17):1575−85. https://doi.org/10.1056/NEJMoa1001286.

[55] Imai K, Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. Ann Appl Stat 2013;7(1):443−70. https://doi.org/10.1214/12-AOAS593.

[56] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 2018;113(523): 1228−42. https://doi.org/10.1080/01621459.2017.1319839.

[57] Hahn PR, Murray JS, Carvalho CM. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). Bayesian Anal 2020;15(3). https://doi.org/10.1214/19-BA1195.

[58] Hill J, Linero A, Murray J. Bayesian additive regression trees: a review and look forward. Annu Rev Stat Its Appl 2020;7(1):251−78. https://doi.org/10.1146/annurev-statistics-031219-041110.

[59] Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proc Natl Acad Sci 2016;113(27):7353−60. https://doi.org/10.1073/pnas.1510489113.

[60] Künzel 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. https://doi.org/10.1073/pnas.1804597116.

[61] Kennedy EH. Towards optimal doubly robust estimation of heterogeneous causal effects. Electron J Stat 2023;17(2):3008−49. https://doi.org/10.1214/23-EJS2157.

[62] Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. Biometrika 2021;108(2):299−319. https://doi.org/10.1093/biomet/asaa076.

[63] Rose G. Sick individuals and sick populations. Int J Epidemiol 2001; 30(3):427−32. https://doi.org/10.1093/ije/30.3.427.

[64] Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses;: power and sample size for the interaction test. J Clin Epidemiol 2004;57(3):229−36. https://doi.org/10.1016/j.jclinepi.2003.08.009.

[65] Debray TPA, Moons KGM, van Valkenhoef G, Efthimiou O, Hummel N, Groenwold RHH, et al. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. Res Synth Methods 2015;6(4):293−309. https://doi.org/10.1002/jrsm.1160.

[66] Seo M, White IR, Furukawa TA, Imai H, Valgimigli M, Egger M, et al. Comparing methods for estimating patient-specific treatment effects in individual patient data meta-analysis. Stat Med 2021;40(6): 1553−73. https://doi.org/10.1002/sim.8859.

[67] Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004; 82(4):661−87. https://doi.org/10.1111/j.0887-378X.2004.00327.x.

[68] VanderWeele TJ, Knol MJ. A tutorial on interaction. Epidemiol Methods 2014;3(1):33−72. https://doi.org/10.1515/em-2013-0005.

[69] Rolling CA, Yang Y. Model selection for estimating treatment effects. J R Stat Soc Ser B Stat Methodol 2014;76(4):749−69. https://doi.org/10.1111/rssb.12043.

[70] Zhao Y, Fang X, Simchi-Levi D. Uplift modeling with multiple treatments and general response types. 2017. Available at: http://arxiv.org/abs/1705.08492. Accessed July 19, 2024.

[71] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W. Double/debiased/neyman machine learning of treatment effects. Am Econ Rev 2017;107(5):261−5. https://doi.org/10.1257/aer.p20171038.

[72] Van Klaveren D, Steyerberg EW, Serruys PW, Kent DM. The proposed 'concordance-statistic for benefit' provided a useful metric when

modeling heterogeneous treatment effects. J Clin Epidemiol 2018;94:59−68. https://doi.org/10.1016/j.jclinepi.2017.10.021.

[73] Efthimiou O, Hoogland J, Debray TPA, Seo M, Furukawa TA, Egger M, et al. Measuring the performance of prediction models to personalize treatment choice. Stat Med 2023;42(8):1188−206. https://doi.org/10.1002/sim.9665.

[74] Maas CCHM, Kent DM, Hughes MC, Dekker R, Lingsma HF, van Klaveren D. Performance metrics for models designed to predict treatment effect. BMC Med Res Methodol 2023;23(1):165. https://doi.org/10.1186/s12874-023-01974-w.

[75] Kent DM, van Klaveren D, Paulus JK, D'Agostino R, Goodman S, Hayward R, et al. The PATH statement explanation and elaboration document. Ann Intern Med 2020;172(1):W1−25. https://doi.org/10.7326/M18-3668.

[76] Robertson SE, Leith A, Schmid CH, Dahabreh IJ. Assessing heterogeneity of treatment effects in observational studies. Am J Epidemiol 2021;190(6):1088−100. https://doi.org/10.1093/aje/kwaa235.

[77] Segal JB, Varadhan R, Groenwold RHH, Li X, Nomura K, Kaplan S, et al. Assessing heterogeneity of treatment effect in real-world data. Ann Intern Med 2023;176(4):536−44. https://doi.org/10.7326/M22-1510.

[78] Maruo K, Furukawa TA, Noma H, Imai H, Ikeda K, Yamawaki S. Qualitative treatment-subgroup interactions in the antidepressant treatment of major depression: application of QUINT to individual participant data from seven placebo-controlled randomized controlled trials. Pers Med Psychiatry 2020;21-22:100054. https://doi.org/10.1016/j.pmip.2019.100054.