ORIGINAL RESEARCH **Open access**

Causal Forest Models with Double Machine Learning for Heterogeneous Treatment Effects in Antihypertensive Therapy: A Position Paper on Personalized Prescribing from Observational EHR Data

Anna Kowalska[1*], Piotr Nowak[1], Tomasz Zielinski[2], Katarzyna Mazur[1]

Abstract

Hypertension affects about 1.4 billion adults globally and is a major modifiable risk factor for cardiovascular disease. Although several first-line antihypertensive drug classes exist, randomized controlled trials typically report only average treatment effects (ATEs), which mask important variability in individual patient responses. As a result, clinical guidelines often assume a homogeneous patient population, leading to trial-and-error prescribing, delayed blood pressure control, and avoidable adverse effects. I argue that causal forest models combined with double machine learning (DML) enable reliable estimation of heterogeneous treatment effects (HTEs) from observational electronic health record data. These methods can approximate randomized trial validity while capturing clinically meaningful variation in treatment response across patients. Compared with traditional approaches, they are computationally feasible and better suited for individualized treatment assessment. Therefore, comparative effectiveness research in hypertension should move beyond ATE-focused analyses toward routine HTE estimation using causal machine learning. This shift would support more precise, data-driven prescribing and improve patient outcomes.

Keywords Electronic health records, Causal forest, Double machine learning, Heterogeneous treatment effects, Antihypertensive therapy, Precision medicine

*Correspondence:
Anna Kowalska
anna.kowalska@gmail.com
[1] Department of Healthcare Intelligence Systems, University of Warsaw, Warsaw, Poland
[2] Department of Medical AI Analytics, Warsaw University of Technology, Warsaw, Poland

Introduction

Hypertension treatment exemplifies a core dilemma in evidence-based medicine: randomized trials establish that several drug classes reduce cardiovascular events, yet head-to-head comparisons show only small average differences in blood pressure reduction [1, 2]. ACE inhibitors, ARBs, CCBs, thiazides, and beta-blockers each lower systolic blood pressure by approximately 10–15 mmHg on average, with modest between-class differences that rarely reach statistical or clinical significance in large meta-analyses [3, 4]. This average similarity has led clinical guidelines to treat these drug classes as largely interchangeable for first-line therapy.

The problem with average treatment effects is that they systematically conceal heterogeneity across patients [5, 6]. A drug that lowers blood pressure substantially for 60% of patients but increases it for 40% will show a modest positive ATE, yet prescribing that drug to the 40% causes harm. Clinicians cannot identify which individual patients

belong to which subgroup based on ATEs alone [7, 8]. The current clinical reality is therefore trial-and-error prescribing: start a drug, wait weeks, measure response, and switch if inadequate.

We contend that causal forest models with double machine learning should become the standard for estimating antihypertensive heterogeneous treatment effects from observational EHR data. This position paper argues that ATEs are insufficient for personalized prescribing, that causal forest and DML overcome key methodological challenges in observational data, and that implementation is feasible within existing health systems. I provide specific recommendations for researchers, editors, guideline committees, and health system leaders.

Hypertension Treatment Landscape

Current guidelines and evidence

Major clinical guidelines including JNC 8, ACC/AHA, and ESC/ESH recommend ACE inhibitors, ARBs, CCBs, and thiazides as first-line agents, with beta-blockers reserved for specific indications such as heart failure or post-myocardial infarction [1, 2]. Meta-analyses of randomized trials have consistently reported small average differences in blood pressure reduction between these classes, with no single class demonstrating clear superiority across all patients [3, 4]. These average findings have shaped guidelines that treat the hypertensive population as largely homogeneous with respect to drug response.

Clinical reality of personalized prescribing

Despite guideline uniformity, clinicians observe substantial inter-individual variation in antihypertensive response daily in practice [5, 6]. Crude patient characteristics such as age, race, and comorbid conditions serve as imperfect effect modifiers: thiazides are often preferred for Black patients based on small subgroup analyses, and beta-blockers are avoided in younger active patients due to fatigue. Yet these categorical heuristics are poor substitutes for data-driven predictions of individual treatment response [7, 8].

Limitations of Average Treatment Effects

Masked heterogeneity

Average treatment effects are mathematically incapable of detecting response heterogeneity because they collapse distributions into single point estimates [4, 5]. A drug that reduces blood pressure by 20 mmHg for 50% of patients and increases it by 10 mmHg for the other 50% yields an ATE of 5 mmHg reduction, which appears modest and favorable. This average completely obscures that half of treated patients experience harm, not benefit [9, 10].

Guideline limitations

Clinical guidelines that rely on ATEs implicitly treat patients as statistically exchangeable, an assumption violated for essentially every antihypertensive drug class [6, 7]. Race-based recommendations, such as preferring thiazides or CCBs over ACE inhibitors for Black patients, represent a crude acknowledgment of heterogeneity. These categorical proxies for biological variation are insufficient because treatment effect heterogeneity exists within, not just between, racial and demographic groups [11, 12].

What clinicians need vs what trials provide

Clinicians making prescribing decisions need individualized predictions: "For this specific patient with this specific age, sex, comorbidities, and laboratory values, which drug class will produce the greatest blood pressure reduction with the fewest adverse effects?" [8, 9]. Randomized trials provide the opposite: group average effects that explicitly discard individual-level information. Heterogeneous treatment effect methods are designed specifically to fill this gap by estimating conditional average treatment effects (CATEs) as functions of patient covariates [1, 13].

Causal Forest and Double Machine Learning

Causal forest fundamentals

Causal forest extends Breiman's random forest algorithm to estimate heterogeneous treatment effects with valid statistical inference [1]. Unlike standard random forests that minimize prediction error in outcomes, causal forests are explicitly designed to maximize heterogeneity in estimated treatment effects across leaves [1, 14]. The algorithm grows an ensemble of causal trees, each constructed by recursively partitioning the covariate space using splitting

criteria that prioritize splits maximizing between-leaf variation in treatment effect estimates rather than outcome variance.

Honest estimation is a critical innovation that distinguishes causal forests from conventional machine learning methods for HTE [1, 15]. The approach partitions the training data into two distinct subsets: one used to determine tree structure (splitting) and another used to estimate within-leaf treatment effects. This separation prevents the overfitting that would otherwise produce artificially optimistic estimates of treatment effect heterogeneity. As a result, honest causal forests achieve asymptotic normality, enabling valid confidence intervals for conditional average treatment effects (CATEs) without requiring additional bootstrap or subsampling procedures [14, 15].

The theoretical guarantees of causal forests rely on several key assumptions, including unconfoundedness (treatment assignment independent of potential outcomes given covariates) and overlap (nonzero probability of receiving each treatment for all covariate values) [1]. When these assumptions hold, causal forests produce consistent and asymptotically normal estimates of CATEs at rates that depend only on the dimension of relevant effect modifiers, not on the total number of covariates. This property makes causal forests particularly well-suited for high-dimensional EHR data containing hundreds or thousands of potential predictors of treatment response [14, 16].

Double machine learning for confounding

Double machine learning provides a general framework for estimating causal parameters from observational data using arbitrary machine learning methods for nuisance functions while maintaining valid inference [10, 11]. The core insight of DML is that naive plug-in estimation using machine learning predictions for nuisance functions (propensity scores and conditional outcome means) introduces regularization bias that can distort treatment effect estimates. DML solves this problem through an orthogonal score function that makes the treatment effect estimator first-order insensitive to errors in nuisance function estimation [10, 17].

The DML procedure employs cross-fitting to avoid overfitting bias, a technique that partitions the data into K folds and iteratively uses K-1 folds to estimate nuisance functions and the remaining fold to compute the orthogonal score [11, 16]. This cross-fitting step ensures that the bias from machine learning regularization does not contaminate the final causal estimate. The orthogonal score combined with cross-fitting yields root-n consistent and asymptotically normal treatment effect estimates even when nuisance functions are estimated at slower-than-parametric rates, a property known as double robustness [10, 17].

Figure 1 illustrates the hierarchical analytical pipeline linking observational EHR data to individualized antihypertensive treatment recommendations through causal forest and double machine learning methods.

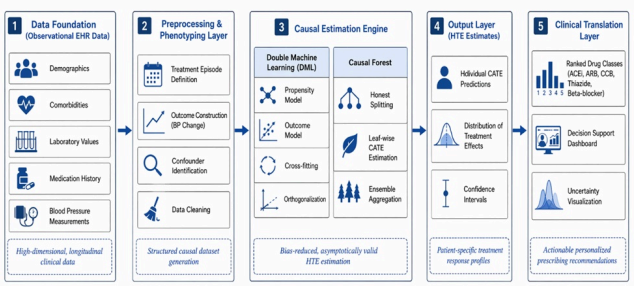

Figure 1. Hierarchical Analytical Pipeline for Estimating and Translating Heterogeneous Treatment Effects into Personalized Antihypertensive Prescribing

DML relaxes parametric assumptions about confounding structures while maintaining valid inference under mild regularity conditions [16, 18]. Unlike traditional propensity score methods that require correct specification of the propensity score model, DML allows researchers to use flexible machine learning algorithms (gradient boosting, neural networks, random forests) to estimate both the propensity score and the conditional outcome mean without incurring bias. For antihypertensive comparative effectiveness research, where the true confounding structure is unknown and likely involves complex nonlinear relationships, DML offers substantial advantages over conventional parametric adjustment [19, 20].

Advantages over traditional methods

Traditional approaches to HTE estimation, such as linear models with treatment-covariate interactions, suffer from model misspecification because true effect heterogeneity is rarely linear or additive [1, 12]. A linear interaction model that specifies separate slopes for age, sex, and race will completely miss nonlinear interactions, threshold effects, or higher-order dependencies that characterize genuine biological heterogeneity in drug response. The model

specification problem is compounded when dozens or hundreds of potential effect modifiers must be considered simultaneously [12, 18].

Propensity score matching, while widely used for ATE estimation, becomes computationally infeasible with many covariates and discards substantial sample information when applied to HTE [20, 21]. Matching on a high-dimensional propensity score collapses covariate information into a single scalar, eliminating the ability to estimate how treatment effects vary across different covariate dimensions. Subclassification on the propensity score, another common approach, requires arbitrary choices about the number and boundaries of strata and suffers from residual confounding within strata [16, 17].

Bayesian Additive Regression Trees (BART) offers a nonparametric alternative for HTE estimation but requires computationally intensive Markov chain Monte Carlo sampling that scales poorly to large EHR datasets [18, 19]. A single BART run on a dataset of 50,000 patients with 100 covariates may take hours or days to converge, whereas causal forests complete similar analyses in minutes. For prospective deployment in clinical decision support systems requiring frequent model retraining, this computational advantage is decisive. Causal forests also avoid the prior sensitivity and hyperparameter tuning challenges that complicate BART implementation in applied clinical research settings [1, 14, 15].

Table 1 provides a structural comparison of methodological approaches, highlighting why causal forest combined with double machine learning uniquely satisfies the requirements for valid and scalable heterogeneous treatment effect estimation.

Table 1. Comparative Analytical Frameworks for Treatment Effect Estimation in Antihypertensive Research

Dimension	Average Treatment Effect (ATE) Models	Traditional HTE (Interaction Models)	Propensity Score Methods	I T
Target Estimand	Population mean effect	Conditional effects (pre-specified)	Adjusted ATE	F
Model Specification	Parametric	Parametric with interactions	Semi-parametric	N
Handling of High-Dimensional Covariates	Poor	Poor	Moderate (collapses via PS)	
Detection of Nonlinear Heterogeneity	None	Limited	None	
Confounding Adjustment	Limited	Limited	Moderate	
Overfitting Control	Not central	Weak	Moderate	
Computational Scalability	High	High	Moderate	l
Interpretability for Clinical Use	High (but misleading)	Moderate	Moderate	
Suitability for EHR Data	Low	Low	Moderate	
Valid Statistical Inference for HTE	No	Weak	No	

HTE From Observational EHR Data

Data requirements and challenges

Estimating antihypertensive HTE from observational EHR data requires large sample sizes (typically tens of thousands of patients per drug class), accurate phenotyping for treatment assignment and outcomes, and comprehensive measurement of confounders [13, 14]. Sample size requirements for HTE estimation are substantially larger than for ATE estimation because the effective sample size for any patient subgroup is a fraction of the total cohort. Power calculations for HTE must account for the number of potential effect modifiers, the expected magnitude of heterogeneity, and the desired precision for subgroup-specific estimates [20, 21].

Blood pressure measurements in EHR data are often irregularly spaced, subject to white-coat effects, and measured under varying conditions (different devices, times of day, patient positioning, and staff training) [13, 22]. These measurement challenges create outcome noise that reduces statistical power for HTE detection. Treatment assignment is non-random and driven by clinician preferences, patient characteristics, and prior medication failures, creating confounding by indication that must be carefully modeled [20, 21]. For example, clinicians preferentially prescribe ACE inhibitors to patients with diabetes due to renoprotective effects, creating an association between diabetes and ACE inhibitor use that is driven by clinician knowledge rather than biological response heterogeneity.

Additional challenges include treatment switching, polypharmacy, and medication non-adherence, all common in hypertension management [14, 22]. Patients frequently have their medications titrated, augmented with additional agents, or switched entirely based on response. Disentangling the effect of the initial treatment from subsequent modifications requires sophisticated longitudinal causal methods such as marginal structural models or sequential g-estimation. Non-adherence, which may affect 30-50% of hypertensive patients, creates measurement error in treatment assignment that typically biases estimates toward the null and reduces detectable heterogeneity [13, 21].

Validity of EHR-based HTE estimation

The validity of HTE estimates from observational data can be benchmarked against randomized trial data when available for the same or similar populations [15, 16]. Several recent studies have demonstrated that HTE estimates from observational data analyzed with causal forests and DML approximate those from trial data, provided that key confounders are measured. Benchmarking should be routine practice when trial data are available, establishing credibility for observational HTE analyses in settings where trial data do not exist [17, 23].

Sensitivity analyses for unmeasured confounding, including E-values and negative control outcomes, should be routinely reported [16, 22]. The E-value quantifies the minimum strength of association that an unmeasured confounder would need to have with both treatment and outcome to fully explain away the observed HTE estimate. Negative control outcomes (outcomes known not to be affected by the treatment) can detect residual confounding: if an HTE estimate is observed for a negative control outcome, unmeasured confounding is likely present [17, 24].

Triangulation of findings across multiple observational datasets and analytical approaches strengthens confidence in HTE estimates when randomized evidence is unavailable for specific patient subgroups [18, 22]. Consistency of HTE patterns across different health systems, geographic regions, and time periods provides evidence that observed heterogeneity reflects genuine biological variation rather than artifacts of a particular dataset or analytical method. Triangulation cannot definitively eliminate confounding concerns, but it can substantially increase the posterior probability that reported HTE estimates are valid [17, 19].

Counterarguments Addressed

"Randomized trials are the gold standard; observational data is biased"

We acknowledge that randomized trials minimize confounding through design, but they are rarely powered or designed to detect HTE in patient subgroups [17, 18]. Double machine learning and causal forests explicitly adjust for measured confounding, and benchmark analyses can demonstrate that HTE estimates from observational data approximate those from trial data when both are available. Waiting for subgroup-specific trial evidence for every antihypertensive comparison is infeasible given the combinatorial explosion of patient characteristics [19, 20].

"HTE methods are too complex for clinical use"

This objection confuses methodological complexity with clinical implementation [19, 20]. Causal forest models can be trained once on large historical datasets and deployed as clinical decision support tools that present simple outputs: predicted blood pressure reduction for each drug class with uncertainty intervals. Clinicians need not understand forest splitting or cross-fitting; they need interpretable predictions at the point of care. The complexity is hidden behind a user interface [21, 22].

"We don't have enough EHR data"

Large health systems routinely accumulate EHR data on hundreds of thousands of hypertensive patients treated with multiple drug classes over time [21, 22]. Insurance claims databases such as Medicare, Medicaid, and commercial claims include millions of eligible patients. Sample size requirements for HTE estimation are substantially larger than for ATE estimation, but modern health data infrastructures meet these requirements. The limiting factor is not data availability but analytical capacity and institutional will [23, 24].

Recommendations

For researchers

Researchers conducting comparative effectiveness studies of antihypertensive drugs should adopt causal forest or double machine learning as default analytical approaches rather than reporting ATEs alone [23, 24]. HTE results should be visualized as distribution of predicted treatment responses across the study population, not merely as a single interaction test. Benchmarking against trial data and sensitivity analyses for unmeasured confounding must become routine practice [25, 26].

For journal editors and reviewers

Journal editors and peer reviewers should reject comparative effectiveness manuscripts that report only ATEs when the dataset contains sufficient sample size and covariate information to estimate HTE [25]. High-impact journals including JAMA, Annals of Internal Medicine, and Circulation should explicitly require HTE analysis or a compelling justification for its absence. Reviewers must be trained to evaluate the validity of HTE estimation methods, including the use of honest estimation, cross-fitting, and sensitivity analyses [26, 27].

For clinical guideline committees

Guideline committees must move beyond one-size-fits-all recommendations by incorporating HTE evidence from observational studies analyzed with appropriate causal methods [26, 27]. Race-based prescribing recommendations should be replaced with data-driven HTE predictions derived from large, diverse cohorts. Guidelines should explicitly acknowledge that average effects from trials provide a poor basis for individual prescribing decisions [28, 29].

For health systems and payers

Health systems and payers should invest in causal machine learning infrastructure, including data extraction pipelines, model development platforms, and clinical decision support integration [28]. The return on investment includes reduced trial-and-error prescribing, faster blood pressure control, fewer adverse drug events, and improved medication adherence. Payers should reimburse for HTE-guided prescribing as a precision medicine service [29].

Implementation Pathway

Technical infrastructure

Implementation requires extracting structured EHR data (demographics, diagnoses, laboratory results, medications, blood pressure measurements) and applying validated phenotyping algorithms for treatment episodes, outcomes, and confounders [14, 29]. Phenotyping for antihypertensive treatment episodes must define index dates (first prescription within a washout period), exposure windows (typically 4-8 weeks for blood pressure response assessment), and censoring rules for treatment switching or augmentation. The outcome phenotype—blood pressure change—requires handling irregular measurement intervals, selecting appropriate baseline and follow-up windows, and adjusting for measurement time-of-day and device type differences [20, 21].

Model development follows a standard pipeline: covariate selection, causal forest training with cross-fitting, hyperparameter tuning via out-of-bag error, and prospective validation on holdout data [20]. Covariate selection should prioritize clinically plausible confounders and effect modifiers, including age, sex, race/ethnicity, baseline blood pressure, body mass index, estimated glomerular filtration rate, diabetes status, prior cardiovascular events, and concomitant medications. High-dimensional variable selection methods (e.g., double selection lasso) can supplement clinical knowledge when hundreds of candidate covariates are available [14, 29].

Deployment requires model versioning, monitoring for data drift, and regular retraining schedules [20]. Data drift occurs when the distribution of patient characteristics, prescribing patterns, or outcome measurement changes over time, potentially degrading model performance. Monitoring systems should track key performance metrics (calibration, discrimination, and prediction error) on incoming data and

trigger retraining when drift exceeds prespecified thresholds. Retraining schedules should be quarterly or semi-annual for stable health systems, more frequent for rapidly evolving practice environments [21, 22].

Additional technical infrastructure includes secure computing environments compliant with health privacy regulations (HIPAA in the US, GDPR in Europe), data provenance tracking, and audit logging for regulatory review [14, 29]. Models deployed for clinical use must undergo formal validation studies that demonstrate performance in the target population before integration into decision support systems. Validation requires separate datasets from those used for training, ideally collected from different time periods or clinical sites to assess generalizability [20].

Clinical integration

Clinical decision support dashboards should present predicted treatment responses for each antihypertensive drug class as intuitive visualizations, such as forest plots or predicted blood pressure reduction with 95% confidence intervals. For a patient initiating antihypertensive therapy, the dashboard would display five predicted values (one per drug class: ACE inhibitor, ARB, CCB, thiazide, beta-blocker) with corresponding uncertainty intervals. A ranked presentation highlighting the drug class with the highest predicted benefit supports rapid clinical decision-making while preserving transparency about alternative options [17, 22].

Table 2 translates the methodological architecture of causal forest and double machine learning into clinically interpretable functions, clarifying how each component contributes to actionable prescribing decisions.

Table 2. Conceptual Mapping from Methodological Components to Clinical Decision-Making Functions in HTE-Based Prescribing

Analytical Component	Technical Function	Bias/Variance Role	Output Type	In
Propensity Score Model (DML)	Estimates treatment assignment probability	Reduces confounding bias	Probability scores	L
Outcome Model (DML)	Predicts expected BP outcome	Improves efficiency	Conditional outcome estimates	E
Orthogonal Score	Debiases treatment effect estimation	Removes regularization bias	Bias-corrected estimates	R
Cross-Fitting	Separates training and estimation	Prevents overfitting	Fold-specific estimates	
Honest Splitting (Causal Forest)	Separates structure vs estimation data	Controls variance inflation	Stable tree partitions	i
Tree-Based Partitioning	Identifies effect modifiers	Captures nonlinear interactions	Subgroup structures	s
Ensemble Aggregation	Combines multiple trees	Reduces variance	Smoothed CATE estimates	ir
CATE Estimation	Computes patient-level treatment effect	Core estimand	Individual effect size	E
Confidence Intervals	Quantifies uncertainty	Variance estimation	Interval estimates	
Ranking Algorithm	Orders treatment options	Decision optimization	Ranked drug list	E

Uncertainty visualization is essential: clinicians must understand when predictions are precise versus when evidence is insufficient. Wide confidence intervals indicate that the model lacks sufficient data for reliable predictions for that particular patient profile, whereas narrow intervals indicate high confidence. Visual encoding using interval width, color intensity, or textual annotations ("high confidence," "low confidence") helps clinicians calibrate their trust in model recommendations. Some dashboards incorporate traffic-light icons: green for statistically significant predicted benefit, yellow for uncertain but potentially beneficial, red for predicted harm or no benefit [21, 22].

Integration should occur within existing EHR workflows, requiring no additional clicks or data entry beyond current prescribing practices [17, 22]. The ideal implementation presents predictions automatically when a clinician opens a medication ordering interface for a hypertensive patient, without requiring separate data entry or navigation to a different system. Predictions should update dynamically as new clinical data (e.g., recent blood pressure measurements, laboratory results) become available in the EHR [20, 29].

User-centered design principles mandate iterative prototyping with clinician end-users to optimize interface usability and trust. Early prototypes should be tested on representative clinical workflows, with attention to cognitive load, decision fatigue, and alert fatigue. Clinicians should be able to override model recommendations easily, with the system logging override rates and reasons to inform future model improvements. Training materials should explain model capabilities and limitations in non-technical language, emphasizing that predictions augment rather than replace clinical judgment [17, 22].

Implementation also requires addressing liability and reimbursement considerations [14, 29]. Health systems must determine who bears responsibility for prescribing decisions guided by HTE models and whether malpractice coverage extends to model-informed decisions. Payers may reimburse HTE-guided prescribing as a precision medicine service, but current billing codes do not explicitly accommodate causal ML recommendations. Advocacy for coding and reimbursement changes should accompany technical implementation efforts [20].

Limitations and Cautions

Residual confounding

Observational data cannot fully replace randomization, and residual confounding from unmeasured variables (e.g., medication adherence, dietary sodium intake, over-the-counter supplement use) may bias HTE estimates [17]. Sensitivity analyses including E-values and negative control outcomes are mandatory, not optional. Triangulation of findings across multiple data sources and analytical methods strengthens causal claims but cannot definitively eliminate confounding concerns in any single observational study [18].

Generalizability

HTE models trained on one health system or patient population may not generalize to different settings due to differences in patient mix, prescribing patterns, blood pressure measurement protocols, and outcome definitions. Local validation on target populations is required before clinical deployment. Model performance should be assessed for calibration (predicted vs observed treatment effects) and for discrimination (ability to identify patients with large vs small benefits) [21, 22].

Conclusion

The current paradigm of antihypertensive prescribing relies on average treatment effects from randomized trials that cannot distinguish patients who benefit from a drug class from those who are harmed by it. This one-size-fits-all approach produces delayed blood pressure control, preventable adverse events, and patient frustration. The methods to solve this problem—causal forests for HTE estimation and double machine learning for confounding adjustment—are now mature, computationally feasible, and validated on observational EHR data.

We contend that personalized antihypertensive prescribing based on HTE estimation is not a future aspiration but a current capability. Large health systems already possess the data required, and the analytical methods have been demonstrated in multiple comparative effectiveness contexts. The barriers are no longer technical but institutional: funding priorities, journal policies, guideline inertia, and clinical culture. These barriers are surmountable with coordinated effort across the research, publishing, and clinical communities.

The era of one-size-fits-all hypertension prescribing must end. Causal machine learning on observational data makes personalized treatment possible. The question is not whether we can do it, but whether we will. I call on researchers to adopt causal forest and DML as standard tools, on journal editors to enforce HTE reporting, on guideline committees to incorporate HTE evidence, and on health systems to deploy these methods at the point of care. Patients deserve better than average.

Acknowledgements

None

Conflict of interest

None

Ethics statement

None

Financial support

None

Received: 20 Apr 2023 Revised: 20 Jun 2023 Accepted: 24 Jul 2023
Published online: 20 January 2024

References

Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc. 2018;113(523):1228-42.

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. Econom J. 2018;21(1):C1-C68.

Jaiswal JK, Samikannu R. Application of random forest algorithm on feature subset selection and classification and regression. In: 2017 World Congress on Computing and Communication Technologies (WCCCT). IEEE; 2017. p. 65-8.

Sundström J, Lind L, Nowrouzi S, Hagström E, Held C, Lytsy P, et al. Heterogeneity in blood pressure response to 4 antihypertensive drugs: a randomized clinical trial. JAMA. 2023;329(14):1160-9.

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:e076423.

Li X, Bijlsma MJ, De Vos S, Bos JH, Mubarik S, Schuiling-Veninga CC, et al. Comparative effectiveness of antihypertensive monotherapies in primary prevention of cardiovascular events: a real-world longitudinal inception cohort study. Front Pharmacol. 2024;15:1357567.

Van der Laan DM, Elders PJ, Boons CC, Beckeringh JJ, Nijpels G, Hugtenburg JG. Factors associated with antihypertensive medication non-adherence: a systematic review. J Hum Hypertens. 2017;31(11):687-94.

Assimon MM. Confounding in observational studies evaluating the safety and effectiveness of medical treatments. Kidney360. 2021;2(7):1156-9.

Kutcher SA, Brophy JM, Banack HR, Kaufman JS, Samuel M. Emulating a randomised controlled trial with observational data: an introduction to the target trial framework. Can J Cardiol. 2021;37(9):1365-77.

Van der Laan MJ, Rose S. Targeted learning in data science. Cham: Springer International Publishing; 2018.

Athey S, Tibshirani J, Wager S. Generalized random forests. Ann Stat. 2019;47(2):1148-78.

Zhou ZH. Machine learning. Singapore: Springer Nature; 2021.

Grafféo N, Latouche A, Geskus RB, Chevret S. Modeling time-varying exposure using inverse probability of treatment weights. Biom J. 2018;60(2):323-32.

Chen Y, Sridhar S, Mittal V. Treatment effect heterogeneity in randomized field experiments: a methodological comparison and public policy implications. J Public Policy Mark. 2021;40(4):457-62.

Chang TH, Stuart EA. Propensity score methods for observational studies with clustered data: a review. Stat Med. 2022;41(18):3612-26.

Srinivasan K, Awotunde JB. Network analysis and comparative effectiveness research in cardiology: a comprehensive review

of applications and analytics. J Sci Technol. 2021;6(4):317-32.

Moccia C, Moirano G, Popovic M, Pizzi C, Fariselli P, Richiardi L, et al. Machine learning in causal inference for epidemiology. Eur J Epidemiol. 2024;39(10):1097-108.

Wyss R, Yanover C, El-Hay T, Bennett D, Platt RW, Zullo AR, et al. Machine learning for improving high-dimensional proxy confounder adjustment in healthcare database studies: an overview of the current literature. Pharmacoepidemiol Drug Saf. 2022;31(9):932-43.

Le NN, Tran TQ, McClure J, Gill D, Padmanabhan S. Triangulating evidence for antihypertensive drug class efficacy on cardiovascular and metabolic outcomes using Mendelian randomisation and colocalisation. medRxiv [Preprint]. 2024:2024.08.

Knaus MC. Double machine learning-based programme evaluation under unconfoundedness. Econom J. 2022;25(3):602-27.

Sanderson E, Glymour MM, Holmes MV, Kang H, Morrison J, Munafò MR, et al. Mendelian randomization. Nat Rev Methods Primers. 2022;2(1):6.

Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc. 2018;25(10):1419-28.

De Chaisemartin C, d'Haultfoeuille X. Two-way fixed effects estimators with heterogeneous treatment effects. Am Econ Rev. 2020;110(9):2964-96.

Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci USA. 2019;116(10):4156-65.

Caron A, Baio G, Manolopoulou I. Shrinkage Bayesian causal forests for heterogeneous treatment effects estimation. J Comput Graph Stat. 2022;31(4):1202-14.

Jackson C, Stevens J, Ren S, Latimer N, Bojke L, Manca A, et al. Extrapolating survival from randomized trials using external data: a review of methods. Med Decis Making. 2017;37(4):377-90.

Cui Y, Kosorok MR, Sverdrup E, Wager S, Zhu R. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. J R Stat Soc Ser B Stat Methodol. 2023;85(2):179-211.

Friedberg R, Tibshirani J, Athey S, Wager S. Local linear forests. J Comput Graph Stat. 2020;30(2):503-17.

Kennedy EH. Towards optimal doubly robust estimation of heterogeneous causal effects. Electron J Stat. 2023;17(2):3008-49.