KI - Künstliche Intelligenz
https://doi.org/10.1007/s13218-023-00827-4

TECHNICAL CONTRIBUTION

Predicting Individual Treatment Effects: Challenges and Opportunities for Machine Learning and Artificial Intelligence

Thomas Jaki[1,2] · Chi Chang[3] · Alena Kuhlemeier[4] · The Pooled Resource Open-Access ALS Clinical Trials Consortium · M. Lee Van Horn[4]

Received: 3 April 2023 / Accepted: 28 November 2023

Abstract
Personalized medicine seeks to identify the right treatment for the right patient at the right time. Predicting the treatment effect for an individual patient has the potential to transform treatment of patients and drastically improve patients outcomes. In this work, we illustrate the potential for ML and AI methods to yield useful predictions of individual treatment effects. Using the predicted individual treatment effects (PITE) framework which uses baseline covariates (features) to predict whether a treatment is expected to yield benefit for a given patient compared to an alternative intervention we provide an illustration of the potential of such approaches and provide a detailed discussion of opportunities for further research and open challenges when seeking to predict individual treatment effects.

Keywords BART · Heterogeneity in treatment effects · Personalized medicine · Predicted individual treatment effects

1 Introduction

Traditionally a clinical trial compares two treatments (experimental and control) based on an outcome (response) variable, *Y*. Data are gathered on an individual participant level, while decisions about treatment effectiveness are usually based on summaries (such as averages) of the individual data. Patients are, however, heterogeneous so that a patient's individual characteristics (e.g., gender, disease severity, genetics) can lead to a patient's personal treatment effect being markedly different from the average effect observed in trials. Analysis of (prespecified) subgroups can be used in the hope that an individual patient's treatment effect is closer to the average within the subgroup. As such subgroups are usually defined on a very limited set of characteristics, however, such approaches still do not fully utilize all patient characteristics that modify the treatment effect. Subgroup analysis has also been criticized, as many approaches lead to the identification of effects that often fail to replicate [1–3].

In recent years, researchers have increasingly been interested in developing methods to predict the treatment effect for individual patients based on all baseline covariates (features). The importance of individual treatment effects in randomized clinical trials are, for example, argued in Gadbury et al. [4] who proposed identifiable bounds for the proportion of patients in the population that responds favorably to one of the treatments using data from an unmatched 2 by 2 table and discuss the advantages to matching in a matched-pairs design. Dorresteijn et al. [5] predict treatment effects for individual patients and then evaluate the net benefit of making treatment decisions for individual patients based on a predicted absolute treatment effect. Van der Leeuw et al. [6] discusses an individual estimate of the absolute risk reduction in cardiovascular events given the specific combination of clinical characteristics of a patient, while Lamont et al.

Data used in the preparation of this article were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database. As such, the following organizations and individuals within the PRO-ACT Consortium contributed to the design and implementation of the PRO-ACT Database and/or provided data, but did not participate in the analysis of the data or the writing of this report: Neurological Clinical Research Institute, MGH; Northeast ALS Consortium; Novartis; Prize4Life; Regeneron Pharmaceuticals, Inc.; Sanofi; Teva Pharmaceutical Industries, Ltd.

✉ Thomas Jaki
thomas.jaki@ur.de

[1] University of Regensburg, Bajuwarenstraße 4, 93055 Regenburg, Germany

[2] University of Cambridge, Cambridge, UK

[3] Michigan State University, East Lansing, USA

[4] University of New Mexico, Albuquerque, USA

Published online: 22 January 2024

[7] define Predicted Individual Treatment Effects (PITE) and introduce the PITE framework on which we will base our discussion.

In order to obtain estimates for these individual treatment effects, often traditional statistical models, such as linear regression models are used. While some proposals for using modern machine learning (ML) approaches and artificial intelligence (AI) models, such as regression trees [8], artificial neural networks [9] and Gaussian processes [10], exist (e.g. [11, 12]) their use to date is mostly restricted to illustrative examples and few practical applications exist. In this work, we aim to illustrate the huge potential for ML and AI methods when predicting individual treatment effects and discuss open questions and opportunities for further research.

We will use the PITE framework [7], a simple intuitive framework, to illustrate the potential for ML and AI methods when predicting individual treatment effects. We use PITE as one member of a wide class of different methods that aim to estimate individual treatment effects. It has been chosen to show the huge potential of these methods due to it being simple and intuitive, yet able to encompass many ML methods within, not because it will be the best approach to take in all circumstances. Moreover, PITE has been used in at least two real applications in the past [13, 14]. The highlighted opportunities and challenges, however, exist for other approaches to predicting individual treatment effects as well.

2 The PITE Framework

In a clinical trial, we typically observe the outcome for a given patient only under either the experimental or the control condition. Potential outcomes [15–17] provide a powerful framework to overcome this and enable understanding of causal effects – even on an individual level. In a clinical trial, for example, each individual participant has a potential outcome under both the experimental treatment, $Y_{Ei,}$ and control, Y_{Ci}. The causal effect of an individual can then be defined as $Y_{Ei} - Y_{Ci}$. As the outcome is typically only observed under one treatment condition in a clinical trial, however, researchers typically estimate average treatment effects (ATE) defined as

$$ATE = E\left(Y_E\right) - E\left(Y_C\right)$$

where Y_E is the outcome when receiving the experimental treatment and Y_C under the control treatment, possibly accounting for covariates on a population level.

The PITE framework [7, 18] supposes that the outcome of an individual under a given treatment is a function of underlying characteristics so that we can capture some of the potential outcome by predicting this function. Specifically, it supposes that

$$Y_{Ti} = f_T\left(x_i\right) + \varepsilon_{Ti}$$

where T denotes the treatment (E for the experimental group and C for control), x_i is a vector of covariates for individual i, $\varepsilon_{\mathrm{Ti}} \sim \mathrm{N}(0, \sigma^2{}_{\mathrm{T}})$ is a patient-level random effect and $\mathrm{f}_{\mathrm{T}}(.)$ is an unknown function. Using an estimate of the unknown functions, $\widehat{f_T}\left(x_i\right)$, the predicted individual treatment effect (PITE) of a patient i is defined as

$$PITE_i = \hat{Y}_{Ei} - \hat{Y}_{Ci} = \hat{f}_E\left(x_i\right) - \hat{f}_C\left(x_i\right) \quad (1)$$

It therefore is an estimate of the individual treatment effect given the covariates and method of estimation of the underlying functions $\widehat{f_T}(.)$. Note that, even if the ATE equals zero, there will often be individuals who would be expected to benefit from the treatment while others would be expected to do better under control. As a consequence, PITE can be useful to help guide treatment decisions. Also note that PITEs are estimates of causal effects under conditions typical to those in a randomized controlled trial [19]. Finally it is worth pointing out that for the definition (1) to be valid the variability in the patient-level random effect does not have to be equal to yield unbiased estimates.

3 Machine Learning and PITE: An Illustration

To illustrate the PITE framework, we will consider Amyotrophic Lateral Sclerosis (ALS, also known as Motor Neuron Disease), a neurodegenerative disorder that affects motor neurons in the brain and spinal cord and will use the publicly available Pooled Resource Open-Access ALS Clinical Trials database (PRO-ACT, http://nctu.partners.org/ProACT) [20]. People with ALS have progressive weakness in voluntary muscle which affects movement of arms and legs but also impacts speech, swallowing and breathing. The PRO-ACT database includes complete information from close to 3000 patients with ALS who participated in 23 clinical trials of the drug Riluzole. The pooling of multiple randomized trials results in a large enough dataset to obtain predictions even for this rare disease yet due to unaccounted study to study differences, the findings presented here should be viewed as illustrative.

One of the outcome measures often used in ALS is the ALS Functional Rating Scale (ALSFRS) which comprises of a list of 10 different assessments of motor function each of which is scored 0 to 4 (4 = normal function and 0 = no function). The sum of the 10 assessment questions is the ALSFRS score and is measured repeatedly over time for each patient. Following Küffner et al. [20] we use the slope of the ALSFRS score from a repeated measures model for

each patient as the outcome of this study. We note that two other studies have used these data to investigate treatment effect heterogeneity [21, 22] using different estimators, both finding evidence for significant individual differences.

In line with [22] we focus on the 2,910 patients (1,766 on experimental treatments and 1,144 on control) who had complete data for 17 predetermined covariates, treatment condition, and the outcome. In this illustration, we will use Bayesian Additive Regression Trees (BART) [23] to estimate the unknown underlying patterns and use 1000 permutations for testing for the existence of heterogeneity. BART has the advantage over the linear model presented in [22] that it can predict well when higher order interactions of non-linearity are present, is robust to outliers in the data, and can handle high-dimensional data without overfitting.

Results To ensure that a PITE analysis is meaningful, we begin by testing for the presence of treatment effect heterogeneity using a permutation test [22]. Figure 1 confirms that there is strong evidence ($p < 0.001$) against the hypothesis of no treatment effect heterogeneity in the PITEs using BART as the predictive model. To be consistent with [22], we used the standard deviation of predicted PITE values to define the treatment effect heterogeneity. Interestingly, the standard deviation for the PITEs using BART is greater than those for the linear model reported in [22] suggesting that meaningful higher-order interactions or non-linearity are present. This allows for better prediction, and highlights one of the potential benefits of ML and AI methods in the context of PITE.

The predicted PITEs are highly variable (Fig. 2) and range from a very clear benefit of the experimental treatment to a clear benefit for control. Most strikingly, one can see that, despite having a small benefit on average, for a number of patients the PITE suggests that using the experimental treatment is actually notably worse than control. Figure 2 also clearly illustrates that, if the PITEs and their uncertainty would be used to make treatment decisions, clear recommendations (i.e. intervals not including zero) would arise for about 40% of patients showing the potential power of such approaches to transform patient care.

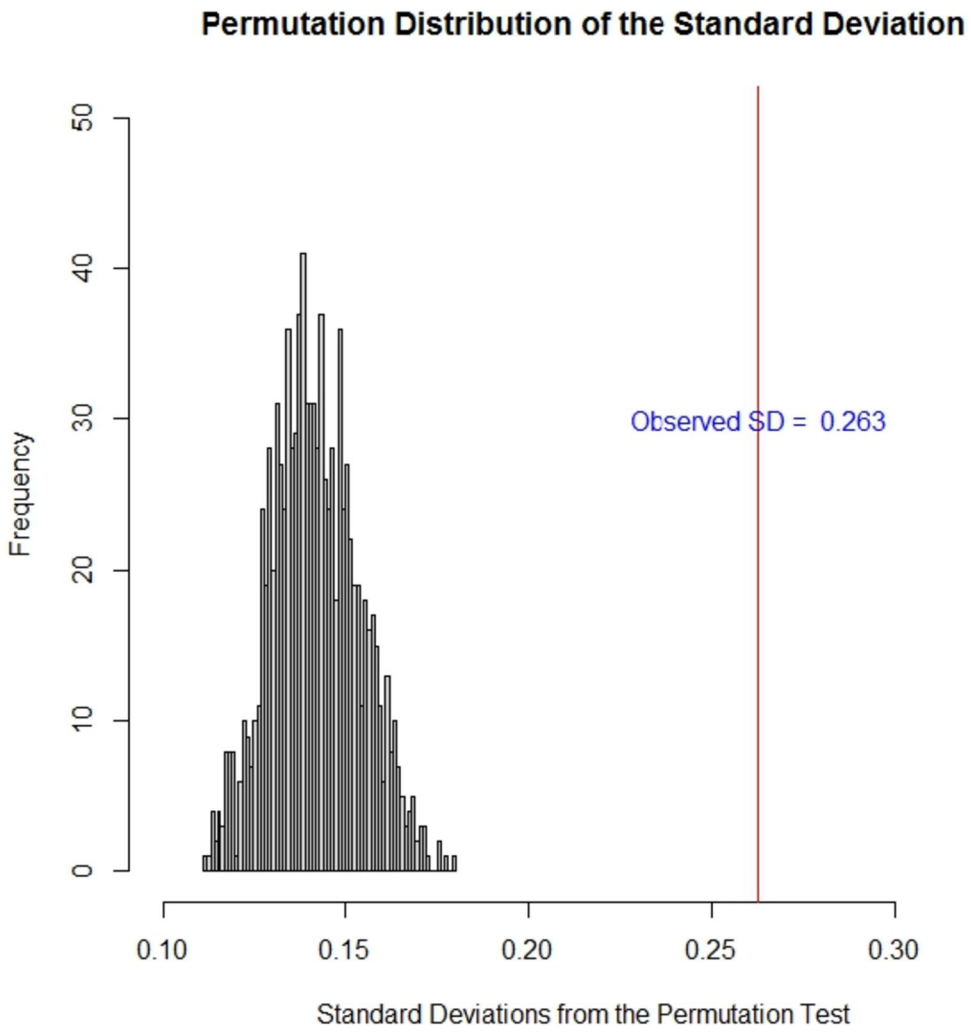

Fig. 1 Permutation distribution for the testing for the presence of treatment effect heterogeneity in the ALS dataset using BART as the predictive approach. Vertical red line shows the observed standard deviation resulting in a p-value < 0.001 of the hypothesis of no heterogeneity of effects

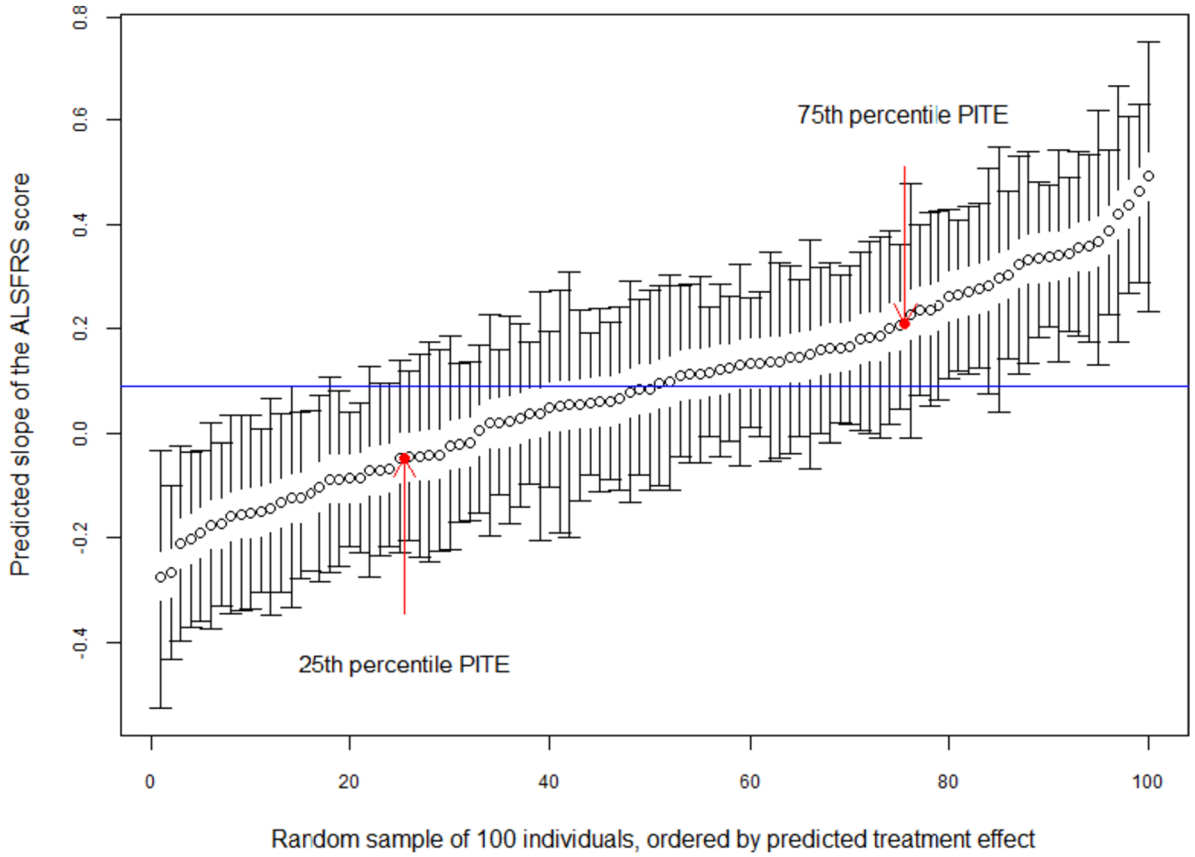

Fig. 2 Ordered predicted individual treatment effects together with their 80% intervals for 100 randomly selected patients in the ALS dataset. Horizontal line indicates the average treatment effect

A further step in the analyses might be to consider the importance of individual variables. We do not consider this step here as we simply wish to show the potential of these methods to inform treatment decisions rather than provide mechanistic insights for the particular application considered. Moreover we note that it often is to be expected that individual differences are due to accumulating small contributions of many factors rather than big effects of a limited number of factors limiting the utility of variable importance measures.

4 Challenges and Opportunities

The illustration above highlights the potential of PITEs to transform treatment of patients by using a patient's features to inform this patient's treatment and the flexibility of ML and AI methods make them particularly attractive to use when estimating PITEs. The illustration, however, also gives rise to a number of interesting challenges and under-researched areas, some of which we aim to put in the spotlight below.

Validation of PITE models It is widely acknowledged that it is essential for any prediction model to be validated [24, 25] and in an ideal setting, this would occur by evaluating the quality of the predictions in an independent dataset. While initial validation may us alternative approaches, such as sample splitting, independent validation provides much stronger support when PITEs are to be used for decisions in clinical practice.

In the context of personalized predictions based on separate models per treatment condition as in PITE, validation of each separate model, while undoubtedly important, does not automatically imply that the resulting PITEs are valid. To date, limited research has been undertaken to validate individual treatment effect predictions with some very recent suggestions for appropriate metrics being provided in [26].

Choice of prediction method From the construction of the PITEs in Eq. (1), it is clear that a necessary condition for them to be identical to the causal effect for an individual is that σ^2_T is zero for both groups and that the estimates $\hat{f}_T(.)$ are equal to the true underlying $f_T(.)$. In order to maximise the utility of PITE it is therefore paramount to predict the unknown functions, $f_T(.)$, as accurately as possible. In

general, the PITE framework can utilize any method that allows predicting outcomes on a patient level yet it is to be expected that some methods will perform better than others for a given true underlying data structure. A linear model in the covariates is expected to work very well if there are no non-linear relationships between the covariates and the outcome of interest and all relevant interactions are included in the model. At the same time, such a model will perform poorly in the presence of non-linear effects and higher order interactions which are not included.

In the case of simple linear models, it is well understood when these models are expected to perform well and different diagnostics have been developed to assess the appropriateness of these models. For many ML and AI methods, however, it is less clear if a particular method is suitable, robust and precise for specific underlying relationships and diagnostics to assess this are sometimes lacking. As a consequence, there is a need to develop more and better diagnostics that allow assessing if a particular predictive approach is appropriate.

Moreover, the fact that any predictive approach could be used for the construction of PITEs poses the yet unsolved question of which predictive approach is best for a given application. While one would expect ensemble methods [27] to perform well in general, a rigorous framework for choosing the best approach for a given setting is still missing.

Utility of PITE to guide treatment When developing a pharmacological intervention, a high level of evidence is required to show that the intervention is safe and beneficial to patients, and commonly agreed standards apply (e.g., use of two pivotal, well-powered studies in the confirmatory phase), no such standard exists to evaluate the potential added benefit of treatment guided by, for example, PITEs. In principle, randomised clinical trials can be used to evaluate the added benefit for an algorithm as recently done in [28] and some guidance on how to do so exists [29]. Frequently, however, prediction models are updated based on accumulating data so that such (potentially large) evaluations would be required repeatedly—every time a change occurs. Open questions that remain include: (i) when re-evaluation is necessary, and (ii) how to repeatedly assess the benefit of the algorithm if re-evaluation is needed.

Covariate selection When trying to best approximate the outcome of a particular patient using ML and AI methods one expects that inclusion of many covariates yields best results. At the same time, one expects that the inclusion of covariates that do not contribute meaningfully to the predictions may add noise to the prediction and, possibly more importantly, incur an unnecessary cost (both monetary and in terms of burden to the patient and/or staff) when collecting the data. Consequently, there remains a desire for the underlying models to be as parsimonious as possible, and a plethora of approaches exist that allow feature selection. In the context of PITE, however, where separate models are used to predict under each treatment condition, there is an opportunity to develop overarching feature selection methods that yield better PITE predictions than when selecting the features separately for each model.

A related area that deserves further attention anchors around the acceptability of these individual treatment predictions by the treating clinician. As one can expect resistance to use such a tool to decide between treatment options when the underlying rationale for the prediction is not well understood, explainable ML and AI methods have a particularly important role to play in this setting. We have also previously argued that utilizing a clinical advisory group to select covariates based on previous research and clinical knowledge can effectively reduce the number of covariates included and improve the acceptability of results [13].

Responsibility and liability One final point to raise involves the risk associated with making treatment decisions. Invariably the decision to favor one treatment over another will be incorrect for some patients which poses a question around the responsibility and (potentially) liability. While one can argue that this problem exists with all devices to support treatment decisions, the fact that ML and AI algorithms often are a black box to the medical professional amplifies the issue as the reason for a particular result of the AI approach is not apparent to the user. This question is further amplified when considering situations where the treating clinician deviates from the recommendation made by the algorithm. We believe that more work is urgently needed to clarify the responsibilities and liabilities of individual treatment predictions (and algorithms more generally) in the context of healthcare.

5 Discussion

ML and AI methods have a huge potential to transforming healthcare. In this work, we highlight predicting individual treatment effects as one area where ML and AI methods have a potentially large role to play as the flexibility of these methods implies that they are useful with no or limited assumptions about the underlying data structure. In our illustrative example we find that BART, as a representative of ML and AI approaches, showed greater individual differences than simpler linear prediction models, emphasizing the need to consider alternative flexible and robust prediction methods. Before a paradigm shift in which treatment decisions are informed by predictions for individual patients can take place and such modern approaches become widely used, however, and we believe a number of crucial areas deserve further attention. In this work, we have highlighted a few of the most important areas in need for further research and consensus.

Acknowledgements TJ received funding from the UK Medical Research Council (MC_UU_00002/14). For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The data are availabe from http://ncri1.partners.org/ProACT/.

References

1. Dmitrienko A, Muysers C, Fritsch A, Lipkovich I (2016) General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. J Biopharm Stat 26(1):71–98
2. Committee for Medical Products for Human Use (2019) Guideline on the investigation of subgroups in confirmatory clinical trials. European Medicines Agency. EMA/CHMP/539146/2013. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-subgroups-confirmatory-clinical-trials_en.pdf Accessed 27 Mar 2023
3. Wijn SR, Rovers MM, Le LH, Belias M, Hoogland J, IntHout J, Debray T, Reitsma JB (2019) Guidance from key organisations on exploring, confirming and interpreting subgroup effects of medical treatments: a scoping review. BMJ Open 9(8):e028751. https://doi.org/10.1136/bmjopen-2018-028751
4. Gadbury GL, Iyer HK, Albert JM (2004) Individual treatment effects in randomized trials with binary outcomes. J Stat Plan Inference 121:163–174
5. Dorresteijn JAN, Visseren FLJ, Braunwald PMRE, Wassink AMJ, Paynter NP, Steyerberg EW, der Graaf YV, Cook NR (2011) Estimating treatment effects for individual patients based on the results of randomised clinical trials. BMJ 343:d5888
6. van der Leeuw J, Ridker PM, van der Graaf Y, Visseren FLJ (2014) Personalized cardiovascular disease prevention by applying individualized prediction of treatment effects. Eur Heart J 35:837–843
7. Lamont AE, Lyons MD, Jaki TF, Stuart EA, Feaster D, Tharmaratnam K, Oberski D, Ishwaran H, Wilson DK, Horn MV (2016) Identification of predicted individual treatment effects (PITE) in randomized clinical trials. Stat Methods Med Res 27(1):142–157
8. Breiman L (2017) Classification and regression trees. Routledge
9. Haykin S (2008) Neural networks and learning machines, 3rd edn. Pearson
10. MacKay DJ (1998) Introduction to Gaussian processes. NATO ASI Ser F Comput Syst Sci 168:133–166
11. Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 113(523):1228–1242
12. Yoon J, Jordon J, Van Der Schaar M (2018) GANITE: Estimation of individualized treatment effects using generative adversarial nets. In: International Conference on learning representations
13. Kuhlemeier A, Desai Y, Tonigan A, Witkiewitz K, Jaki T, Hsiao YY, Chang C, Van Horn ML (2021) Applying methods for personalized medicine to the treatment of alcohol use disorder. J Consult Clin Psychol 89(4):288–300
14. Kuhlemeier A, Jaki T, Jimenez EY, Kong AS, Gill H, Chang C, Resnicow K, Wilson DK, Van Horn ML (2022) Individual differences in the effects of the ACTION-PAC intervention: an application of personalized medicine in the prevention and treatment of obesity. J Behav Med 45:211–226
15. Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66(5):688
16. Rubin DB (2005) Causal inference using potential outcomes. J Am Stat Assoc 100(469):322–331
17. Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81(396):945–960
18. Ballarini NS, Rosenkranz GK, Jaki T, König F, Posch M (2018) Subgroup identification in clinical trials via the predicted individual treatment effect. PLoS ONE 13(10):e0205971
19. Hoogland J, IntHout J, Belias M, Rovers MM, Riley RD, Harrell F Jr, Moons KG, Debray TP, Reitsma JB (2021) A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. Stat Med 40(26):5961–5981
20. Küffner R, Zach N, Norel R, Hawem J, Schoenfeld D, Wang L, Li G, Fang L, Mackey L, Hardiman O, Cudkowicz M, Sherman A, Ertaylan G, Grosse-Wentrup M, Hothorn T, van Ligtenberg J, Macke JH, Meyer T, Schölkopf B, Tran L, Vaughan R, Stolovitzky G, Leitner ML (2015) Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. Nat Biotechnol 33:51–57
21. Seibold H, Zeileis A, Hothorn T (2018) Individual treatment effect prediction for amyotrophic lateral sclerosis patients. Stat Methods Med Res 27:3104–3125
22. Chang C, Jaki T, Sadiq MS, Feaster D, Cole N, Lamont AE, Oberski D, Desai Y, Van Horn ML, Pooled Resource Open-Access ALS Clinical Trials Consortium (2021) A permutation test for assessing the presence of individual differences in treatment effects. Stat Methods Med Res 30(11):2369–2381
23. Chipman HA, George EI, McCulloch RE (2010) BART: Bayesian additive regression trees. Ann Appl Stat 6:266–298
24. Steyerberg EW, Harrell FE (2016) Prediction models need appropriate internal, internal–external, and external validation. J Clin Epidemiol 69:245–247
25. Eichler H, Koenig F, Arlett P, Enzmann H, Humphreys A, Pétavy F, Schwarzer-Daum B, Sepodes B, Vamvakas S, Rasie G (2020) Are novel, nonrandomized analytic methods fit for decision making? The need for prospective, controlled, and transparent validation. Clin Pharmacol Ther 107(4):773–779. https://doi.org/10.1002/cpt.1638
26. Efthimiou O, Hoogland J, Debray TP, Seo M, Furukawa TA, Egger M, White IR (2023) Measuring the performance of prediction models to personalize treatment choice. Stat Med 42(8):1188–1206
27. Opitz D, Maclin R (1999) Popular ensemble methods: an empirical study. J Artif Intell Res 11:169–198
28. Manz CR, Zhang Y, Chen K, Long Q, Small DS, Evans CN, Chivers C, Regli SH, Hanson CW, Bekelman JE, Braun J (2023) Long-term effect of machine learning-triggered behavioral nudges on serious illness conversations and end-of-life outcomes among patients with cancer: a randomized clinical trial. JAMA Oncol 9(3):414–418
29. US Food and Drug Administration (2016). Adaptive designs for Medical Devices. http://www.fda.gov/media/92671/download Accessed 24 Mar 2023.