Received: 25 August 2021 | Revised: 15 October 2021 | Accepted: 20 October 2021

DOI: 10.1002/psp4.12761

COMMENTARY

Machine learning for tumor growth inhibition: Interpretable predictive models for transparency and reproducibility

Andreas D. Meid[1] | **Alexander Gerharz**[2] | **Andreas Groll**[2]

[1]Department of Clinical Pharmacology and Pharmacoepidemiology, University of Heidelberg, Heidelberg, Germany

[2]Department of Statistics, TU Dortmund University, Dortmund, Germany

Correspondence
Andreas D. Meid, Department of Clinical Pharmacology and Pharmacoepidemiology, University of Heidelberg, Im Neuenheimer Feld 410, 69120 Heidelberg, Germany.
Email: andreas.meid@med.uni-heidelberg.de

Funding information
No funding was received for this work

Machine learning (ML) has recently enriched the possibilities for predicting clinical events using prognostic and/or predictive variables. Clinical pharmacology also benefits from more accurate but often increasingly complex ML models. For such models to be accepted in practice, it is essential to make them reproducible for the analyzing scientist and interpretable for the clinician. When reproducing an ML work, we exemplarily show how methods of interpretable ML can support to meet these prerequisites.

Machine learning (ML) has become increasingly important in recent years, both in medicine and in the field of clinical pharmacology.[1] This applies to early phases of drug development and complex analyses of typical clinical trial data. ML methods promise a higher predictive value, which may result from the flexible analysis of nonlinear correlations or higher-order interactions.[2] At the same time, there is a desire on the clinical side that the path to good prediction should be comprehensible and transparent.[3] If one now intends to reproduce such a predictive model, it is immensely important to clarify the model's ML process and how predictions are obtained. This is exactly what we have approached when reproducing the results of the previously published original work of Chan et al.[2] In particular, we addressed three fundamental questions: (1) what is needed for reproducible ML analyses, (2) how can these methods be described in a transparent and interpretable way, and (3) what else can be predicted from the original work?

In the original manuscript, Chan et al. introduced a modeling platform based on the OAK study, which allows to compare different ML methods for predicting overall survival.[2] In brief, the OAK study included patients with previously treated non-small cell lung cancer into a randomized, open-label, phase III trial in which they were randomly assigned to receive either atezolizumab or docetaxel once every 3 weeks.[4] The source data are accessible with an adequate request to a clinical trials portal (www.vivli.org) and the original work also provides rudimentary analysis code. Random seeds are set in the analysis code, which is essential for ML methods (e.g., random forests[5]) in terms of reproducibility. With available software (versions) and helpful answers upon contacting the authors, many prerequisites for reproducible research seemed to be given. However, the practical situation was complicated by the fact that all baseline variables of the clinical trial were available, but not the variables generated during the follow-up, which would have had to be modeled first in a separate preprocessing step severely limiting reproducibility (e.g., individual drug exposures retrospectively derived by

population pharmacokinetic modeling or parameters from tumor growth models relying on measurements after the study baseline). However, it is this particular situation that further emphasizes the importance of interpretable machine learning (IML): now we are all the more interested in which variables (with different availability) contribute to what extent and in which way to the prediction.

If we want to describe ML procedures in an interpretable fashion, this is generally possible from two perspectives. First, the procedure can be monitored during the modeling process itself. For example, this could mean to describe how hyperparameters are determined. In their supplementary materials, Chan et al. show coefficient solution paths together with the optimal estimated coefficients selected by the tuning parameter λ in least absolute shrinkage and selection operator (Lasso) regression or by the optimal boosting step number in boosting regression.[2] It is also conceivable from this perspective to visualize, for example, how the number of split candidates of a random forest is determined. However, these metrics can only hardly be interpreted in a clinical sense; nevertheless, they can of course be reassuring if they match with the original analysis when reproducing this analysis. The second perspective, on the other hand, looks at the influence of predictor variables on the prediction itself. Whereas new methods for interpretability are published at breakneck speed,[3] we restrict ourselves here first for illustration to two catchy and intuitive methods, the variable importance (synonym: feature importance) and the partial dependence plots (synonym: marginal means, predictive margins, and marginal effects).

Variable importance (VI) indicates the relevance of a single predictor variable for the overall accuracy of the prediction (or vice versa the prediction error). This importance metric can be determined independently of the type of the (ML) model and thus also allows for comparing different methods directly. The idea behind this is that the prediction error increases after permutation of the values in the predictor variable.[5] It is straightforward that the prediction error will increase more for an important variable than for less important ones if the relationship between predictor and outcome is broken down. Figure 1a shows which variables in which order contribute how much to the prediction for the Lasso regression, boosting, and random survival forests examined as examples adapted from the original manuscript.[2] In principle, a cutoff in variable importance could also be used for preselecting predictors for another subsequent model, however, this is not done here. For our comparison to the original manuscript, it is also interesting to see which variables in our set of predictor variables are now predictively important. For example, the top six predictor variables of the random survival forests are clinically well-known and expected variables to influence prognosis in lung cancer.

This VI plot illustrates the extent, but not the direction, in which the predictor influences the prediction. This marginal effect of a predictor on the outcome can be determined via the partial dependence plot[6] from a previously fitted model. Although linear models yield linear relationships, it can be more complex for nonlinear or nonparametric methods, such as random forests. For the top eight variables from the random survival forest, Figure 1b shows how the outcome prediction is influenced by alternating the influence variable on the *x*-axis. In addition to the functional relationship (curve shape), this also shows the strength in predicting the outcome. Consistent trajectories are desirable when pursuing reproducible results, which must also appear plausible from a clinical point of view. In a clinically plausible manner, higher risks resulted from histologic classification as squamous, higher burden of disease (ECOG status [Eastern Cooperative Oncology Group]), more metastatic sites, higher levels of lactate dehydrogenase (LDH), or lower albumin levels. All these considerations of predictive values can describe ML methods more transparently and be additionally considered during reproduction—but they are not basic requirements (such as, e.g., the use of identical software algorithms and random seeds).

"The ultimate aim of artificial intelligence (and ML) is prediction,"[2] this thought has become particularly clear to us once again during reproduction. The relevant question is what one wants to achieve with a predictive model. It can be regarded as a good proof-of-principle, if tumor-growth-metrics (derived under treatment during the follow-up) are also predictive for overall survival (in the follow-up), in addition to effect measures for those treatments. This is well-demonstrated by the original paper. In practice, however, the clinician would like to know (based on patient information at baseline only) what the prognosis of a patient would be under the various available treatment options. This is particularly interesting if different patients respond differently to available treatment options (i.e., if so-called heterogeneous treatment effects [HTEs] exist). This means that there are treatment effect modulators that determine how individual responses to available treatment options T may differ.[7] Consequently, we are not only interested in the average treatment effect (population mean), but in the conditional average treatment effect (CATE) in the individual patient given his or her covariates Z. Because we can only observe the outcome under one treatment in a patient, we consider the predicted potential outcomes Y^* to estimate CATEs. In formula notation for our two treatments, this results in the following:

$$\widehat{\mathrm{CATE}}_i = E\left(Y_i^* | Z = z_i, T = \text{atezolizumab}\right) - E\left(Y_i^* | Z = z_i, T = \text{docetaxel}\right). \quad (1)$$

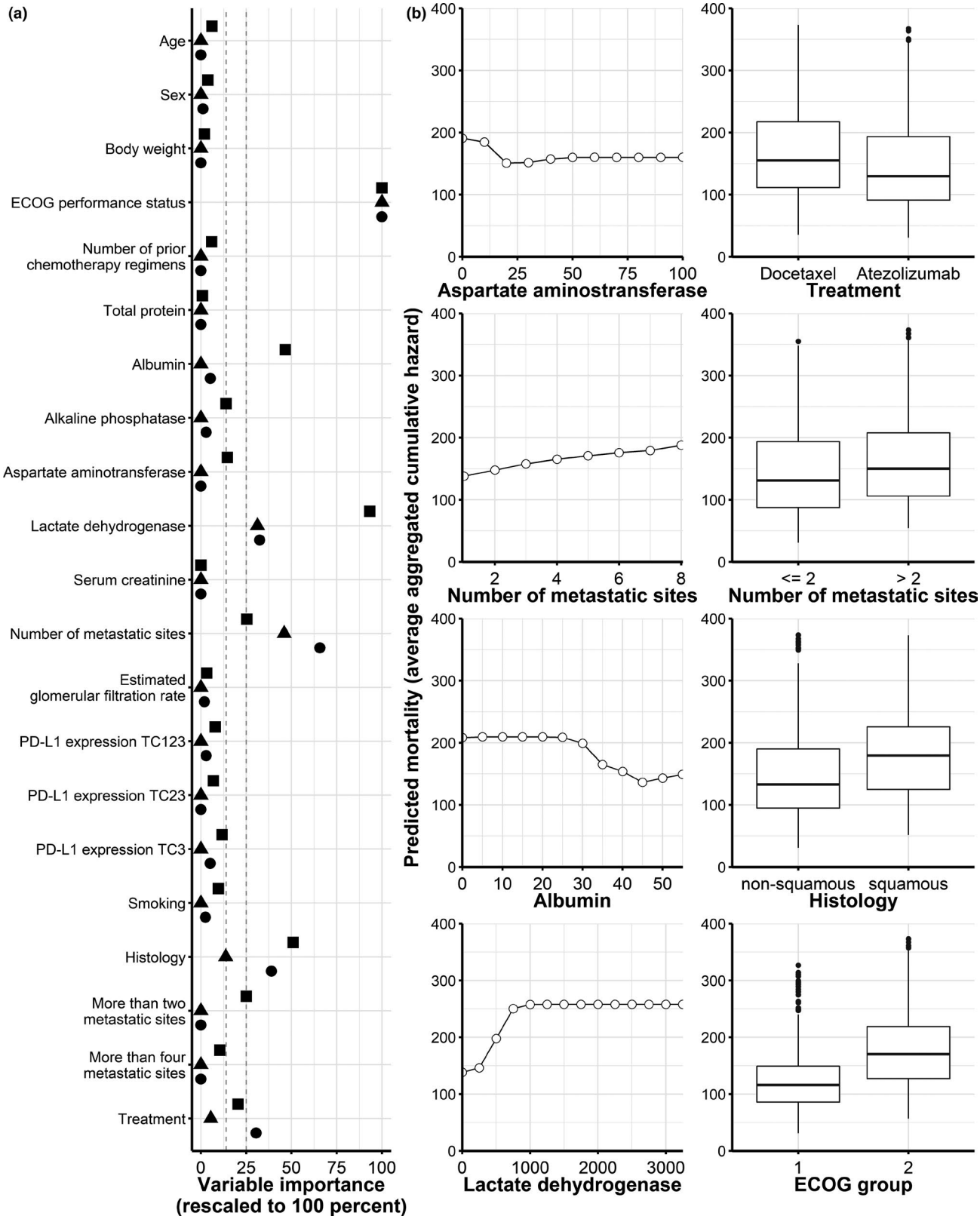

FIGURE 1 (a) Variable importance from random survival forest (square: ■), boosting Cox regression (circle: ●), and least absolute shrinkage and selection operator (Lasso) Cox regression (triangle: ▲; scaled to 100% of most relevant variable, respectively). Of note, permutation variable importance considered both categorical and continuous variables. Vertical dashed lines indicate the top six and top eight predictors in the random survival forest, respectively. (b) Partial dependence plots of the top eight predictors from the random survival forest. Of note, the same packages and functions were used as in the original manuscript of Chan et al.[2] Abbreviations: ECOG, Eastern Cooperative Oncology Group; TC123/TC23/TC3, Group indicators for patients with programmed death-ligand 1 (PD-L1)-expression of at least 1%/5%/50%

Among the many possible modeling methods,[8] the so-called split model approach appears intuitively understandable. One develops a prognostic model M in each of the two subgroups of patients under atezolizumab and docetaxel ($M1$ and $M2$, respectively). For each patient under any allocation, the CATE can then be determined from the difference in predictions of $\widehat{\text{CATE}}_i = Y_i^*(M1) - Y_i^*(M2)$, where Y_i^* represents the individual risk for mortality. Equation 1 thus shows the individual differences in expected risks associated with the

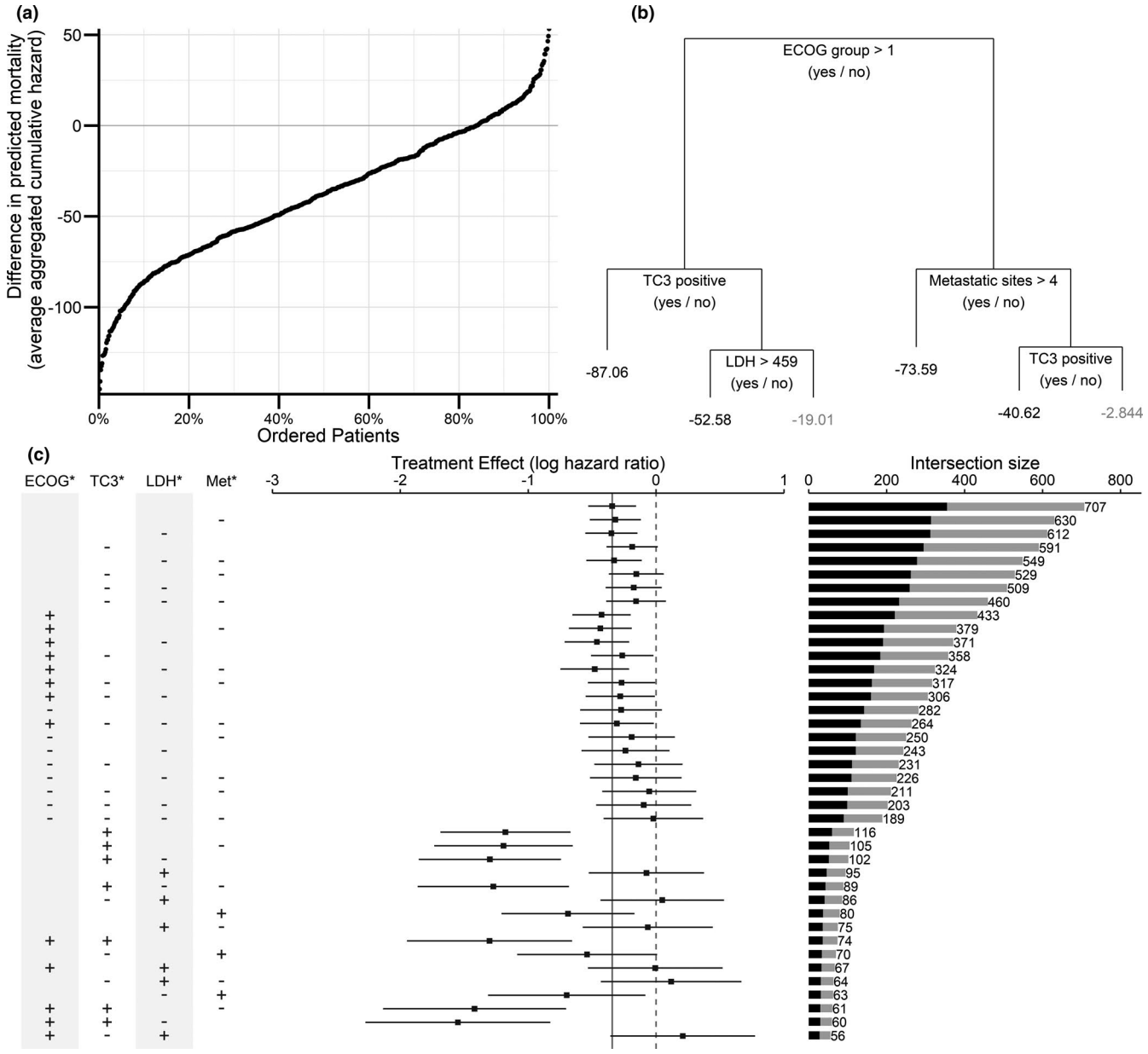

FIGURE 2 (a) Waterfall plot of heterogeneous treatment effects (HTEs)[9] as individual differences for the predicted mortality to atezolizumab or docetaxel in a random survival forest. In particular, the predicted mortality is the difference in the average aggregated cumulative hazard (i.e., values lower than zero indicate individual benefit for atezolizumab). (b) Surrogate model ("fit-the-fit") as a regression tree of the most influential predictors for this outcome. Predicted individual responses below the median are highlighted in gray. (c) Modified upset plot[10] to visualize subgroup effects in dependence of most of those influential predictors for HTEs. On the left, situations for subgroup generation are indicated by plus and minus for the states of a binary variable, whereas the absence allows both options. Subgroup effects are represented by a forest plot, in which a solid line indicates the overall treatment effect for the comparison of atezolizumab versus docetaxel (i.e., values lower than zero indicate benefit for atezolizumab). When considering sample sizes in subgroups on the right, atezolizumab patients are indicated in black. Abbreviations: ECOG*, binary indicator for at least level 2 performance according to Eastern Cooperative Oncology Group; TC3*, binary indicator for programmed death-ligand 1 (PD-L1)-expression of at least 50%; LDH*, binary indicator for lactate dehydrogenase (LDH) levels of at least 459 [units/L]; Met*, binary indicator for at least four metastatic sites

two treatments-positive values of CATE illustrate higher risks under atezolizumab (estimated with M1), whereas negative values reflect a higher risk under docetaxel (estimated with M2). A modified waterfall plot[9] can clarify the distribution and also make the prediction for an individual patient therein apparent to the clinician (Figure 2a).

In addition, HTE modeling is also an ideal showcase to demonstrate how the complex CATE estimates can be made interpretable. For this purpose, a so-called surrogate model can be fitted downstream, which considers the estimated CATEs as the new dependent variable and the set of original predictors of the respective patients as independent variables. Naturally, "fit-the-fit" is a synonymously used term. For this purpose, one can, for example, develop a regression tree and use it to identify important variables (with associated cutoffs) that trigger the complexly determined individual probability of success (CATE) under atezolizumab and docetaxel (Figure 2b). We see that atezolizumab is generally superior on average and for the clear majority of patients. The few explanatory variables identified from a (pruned) regression tree primarily reveal how much a patient in the respective tree branch will benefit from atezolizumab. If these variables are taken to define subgroups, for example, a differential prognosis depending on them may help to classify the outcome clinically. A modified upset plot[10] visualizes these relationships in the form of a familiar forest plot (Figure 2b). In patients without high programmed death-ligand 1 (PD-L1)-expression, moderate LDH levels, and less than five metastatic sites, atezolizumab appears to be less beneficial than in the reverse case, where the benefit of atezolizumab appears to be increased for a higher ECOG performance status indicating a larger burden of disease. It should also be noted that all of the presented procedures were used to illustrate the ML models; performance measures for internal and external validity were outside the scope of this commentary.

As a conclusion, it is fundamentally important to make ML methods interpretable in order to enhance their comprehensibility and thus acceptance in clinical practice, but also to facilitate reproducibility as a confirmatory step along the way. In our re-analysis of a manuscript based on clinical trial data, we have clarified basic principles and exemplified which methods can be used for each purpose. It would be desirable if such methods were used routinely. This is a fundamental step toward the implementation of such procedures, for example, as decision support in clinical practice with all the legal and ethical implications. With the constantly growing number of possible methods of IML, we like to highlight that further developments can be particularly useful if they are developed together with clinicians, are intuitively understandable, and can be applied independently of the model's nature.

ACKNOWLEDGMENTS

This publication is based on research using data from data contributor Hoffmann-La Roche that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this publication.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

ORCID

Andreas D. Meid https://orcid.org/0000-0003-3537-3205

REFERENCES

1. Wang Y, Zhu H, Madabushi R, Liu Q, Huang SM, Zineh I. Model-Informed drug development: current US regulatory practice and future considerations. *Clin Pharmacol Ther*. 2019;105(4):899-911.
2. Chan P, Zhou X, Wang N, Liu Q, Bruno R, Jin JY. Application of machine learning for tumor growth inhibition - overall survival modeling platform. *CPT Pharmacometrics Syst Pharmacol*. 2021;10(1):59-66.
3. Molnar C. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/; 2019.
4. Rittmeyer A, Barlesi F, Waterkamp D, et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *Lancet*. 2017;389(10066):255-265.
5. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
6. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232.
7. Meid AD, Ruff C, Wirbka L, et al. Using the causal inference framework to support individualized drug treatment decisions based on observational healthcare data. *Clin Epidemiol*. 2020;12:1223-1234.
8. Rekkas A, Paulus JK, Raman G, et al. Predictive approaches to heterogeneous treatment effects: a scoping review. *BMC Med Res Methodol*. 2020;20(1):264.
9. Gewandter JS, McDermott MP, He H, et al. Demonstrating heterogeneity of treatment effects among patients: an overlooked but important step toward precision medicine. *Clin Pharmacol Ther*. 2019;106(1):204-210.
10. Ballarini NM, Chiu Y-D, König F, Posch M, Jaki T. A critical review of graphics for subgroup analyses in clinical trials. *Pharm Stat*. 2020;19(5):541-560.