Using causal forests to assess heterogeneity in cost-effectiveness analysis

Carl Bonander | Mikael Svensson

School of Public Health and Community Medicine, Institute of Medicine, University of Gothenburg, Gothenburg, Sweden

Correspondence
Carl Bonander, Sahlgrenska Academy, PO Box 463, SE-405 30 Gothenburg, Sweden.
Email: carl.bonander@gu.se

Funding information
Familjen Kamprads Stiftelse, Grant/Award Number: 20180122;
Forskningsrådet för hälsa, arbetsliv och välfärd, Grant/Award Number: 2017-00414

Abstract
We develop a method for data-driven estimation and analysis of heterogeneity in cost-effectiveness analyses (CEA) with experimental or observational individual-level data. Our implementation uses causal forests and cross-fitted augmented inverse probability weighted learning to estimate heterogeneity in incremental outcomes, costs and net monetary benefits, as well as other parameters relevant to CEA. We also show how the results can be visualized in relevant ways for the analysis of heterogeneity in CEA, such as using individual-level cost effectiveness planes. Using a simulated dataset and an R package implementing our methods, we show how the approach can be used to estimate the average cost-effectiveness in the entire sample or in sub-populations, explore and analyze the heterogeneity in incremental outcomes, costs and net monetary benefits (and their determinants), and learn policy rules from the data.

KEYWORDS
causal forest, cost-effectiveness analysis, machine learning, stratified analysis, treatment heterogeneity

JEL CLASSIFICATION
C1, I1

1 | INTRODUCTION

The standard cost-effectiveness analysis (CEA) in health care settings, which compares the mean difference in health benefits over the mean difference in costs between alternative treatments, is increasingly being used to inform decisions on reimbursement and guidelines of pharmaceuticals and/or health care services in many jurisdictions across the world (Chalkidou et al., 2009; ISPOR, 2013). However, the focus on differences in average costs and health outcomes can mask important heterogeneity regarding costs and health outcomes across patient subgroups. Patient heterogeneity, that is, the variability in costs and/or health outcomes that is related to characteristics of patients, can exist in sociodemographic as well as biological characteristics (Kohli-Lynch & Briggs, 2019).

If CEA studies would routinely assess and report heterogeneity, it would provide a possibility for improving health and efficiency outcomes by policy making that targets implementation or reimbursement strategies toward subgroups...
with the most advantageous cost-effectiveness outcomes (e.g., Basu, 2011; Basu & Meltzer, 2007; Coyle et al., 2003; Espinoza et al., 2014a, 2014b; Sculpher, 2008). As an illustration, imagine a new treatment with an average incremental cost per quality-adjusted life year (QALY)—that is, an incremental cost-effectiveness ratio (ICER)—of 45,000 USD per QALY. The patient population consists of two equally sized patient subgroups called A and B, for example, based on disease severity, where the ICER is 30,000 USD per QALY in subgroup A and 60,000 USD per QALY in subgroup B. If we assume that the decision maker reimburses treatments if the ICER is below 50,000 USD per QALY (the cost-effectiveness threshold), by restricting reimbursement to subgroup A, it is possible to increase the total value (health) in the health-care system. The difference in expected net benefit per population member of adopting a subgroup-specific strategy over a “one-size-fits-all” approach is sometimes referred to as the static value of heterogeneity (Espinoza et al., 2014a). These ideas can also be extended to the value of making decisions at the individual level, as outlined Basu and Meltzer (2007).

The literature demonstrating the potential value of assessing heterogeneity in CEA is also reflected in government and scientific guidelines on best-practice CEA. For example, the Second Panel on Cost-Effectiveness in Health and Medicine argued that when evidence of heterogeneity exists, subgroup-specific CEA results should be presented alongside average results (Sanders et al., 2016). A similar view can be found in many governmental recommendations on CEA reference cases (e.g., Ramaekers et al., 2013). However, the governmental guidelines are typically not informative in terms of how subgroup analyses should be carried out in practice so to provide the best input for decision making.

The lack of details about the “when and how” of assessing heterogeneity in CEA is perhaps not surprising given the many potential problems and pitfalls involved. For instance, subgroup analyses give rise to multiple hypothesis testing, with an increased risk of false positive findings unless multiplicity is corrected for, which may in turn reduce statistical power (Savin, 1984). Another concern with conventional subgroup analyses is that it often involves manual, ad hoc searches for subgroups where the treatment is particularly effective. While pre-registration can prevent such “fishing expeditions”, it limits our possibility to learn from the data.

In this paper, we suggest a way to address the challenges of how to carry out such subgroup analyses by proposing a data-driven method for identification and visualization of heterogeneity in the cost-effectiveness across patient subgroups. Our approach extends the causal forest methodology developed by Athey et al. (2019) and Wager and Athey (2018) to what we call CEA forests, which have several potential advantages over conventional approaches for assessing heterogeneity in cost-effectiveness. In our view, its main advantages follow directly from the causal forest methodology, including that it (i) can learn complex patterns of heterogeneity from the data without explicit assumptions on functional form and (ii) automates the identification of subgroups. In addition to this, we suggest a number of ways to visualize the results from CEA forests, such as individual-level cost-effectiveness planes (see e.g., Basu, 2014), that can be used to explore and describe heterogeneity in incremental outcomes, costs and net monetary benefits. We also show how to construct subgroup-specific cost-effectiveness acceptability curves and cost-effectiveness planes based on a single CEA forest without having to perform the computationally expensive task of re-estimating the forest model. We expect that the suggested tools will enable credible and transparent assessment of heterogeneity in the context of CEA by automating the process of selection of relevant subgroups and estimating of the subgroup results. To facilitate and simplify their implementation, we implement our methods in a freely available package for R (CEAforests, available at: https://github.com/bonander/CEAforests).

The rest of the paper is organized as follows. In Section 2, we provide an overview of the causal forest methodology. In Section 3, we introduce our CEA forest algorithm. In Section 4, we demonstrate how CEA forests can be used to assess both average cost-effectiveness as well as heterogeneity across subgroups. In Section 5, we show how to estimate policy trees that can be used to automatically identify which subgroups should be offered the treatment in order to maximize the expected net monetary benefit per population member. Finally, in Section 6, we conclude by discussing avenues for future research and potential improvements to our approach.

2 | GENERAL METHODOLOGY

2.1 | Definitions and setting

We focus on a setting with a binary treatment \( W_i \) and where the analyst has access to individual-level data on each subject’s health outcome \( Y_i \) and (healthcare) cost \( C_i \) from either a randomized trial or an observational study. For
simplicity, we assume that the sample is representative of the target population for a reimbursement decision (sampling weights can, however, be incorporated into the framework to transport the results to external target populations (Tibshirani et al., 2020; Westreich et al., 2017)).

Following the potential outcomes framework, we define the potential outcome under the treatment and control states as \( Y_i(1) \) and \( Y_i(0) \), respectively, and the individual-level causal effect as \( \tau_i = Y_i(1) - Y_i(0) \). Throughout, we assume unconfoundedness conditional on observed characteristics \( X_i \), such that \( \{Y_i(1), Y_i(0)\} \perp W_i|X_i, \) that is, that the treatment is independent of the potential outcomes after adjustment for observable characteristics (such as age, sex, blood pressure, baseline health, etc). We impose the same assumption on healthcare costs, that is, that \( \{C_i(1), C_i(0)\} \perp W_i|X_i \). These assumptions hold naturally under random treatment assignment even without conditioning on \( X_i \). Jointly, they convey that we assume that estimates for incremental outcomes and costs are unconfounded after adjusting for observable characteristics. Generally, the variables required to achieve unconfoundedness may differ for health outcomes and for healthcare costs. In this case, the variable set \( X \) can be thought of as the union of the two sets of variables that are needed to achieve unconfoundedness in both. We also assume that there is sufficient overlap in covariate distributions between treatment groups such that \( P(W_i = w|X = x) > 0 \) for all \( w \) and \( x \). This assumption rules out settings where a particular covariate combination perfectly predicts either treatment or control status (e.g., when only men are eligible for treatment but sex is included in the covariate set). Further, we assume that counterfactual consistency holds such that \( Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0) \) and \( C_i = W_i C_i(1) + (1 - W_i) C_i(0) \). That is, we assume that the observed outcomes and costs among individuals in the treatment group correspond to their potential outcome under the treated state, and vice versa for controls. This assumption rules out settings with spillover effects.

Our methods rely heavily on the causal forest methodology introduced by Wager and Athey (2018) and further developed by Athey et al. (2019), which is a machine learning method that can be used to estimate heterogeneous causal effect functions under the assumptions invoked above. In our view, the main strength associated with causal forests are that they provide a strategy for learning patterns of heterogeneity from the data that requires little researcher input compared to conventional subgroup analyses. For instance, one does not have to decide which cutoffs to use to investigate heterogeneity in effects along continuous covariates, nor do we need to specify the functional form of the heterogeneity; the causal forest algorithm is designed to automate this search and find the most appropriate model that best describes the effect heterogeneity in the data.

We also note that the individual-level treatment effect estimates from a causal forest can be used to estimate average effects in the entire sample, as well as in user-defined subgroups, akin to conventional CEA estimation. The methodology proposed below can therefore be used for this task as well. While the added value of using causal forests compared to conventional analyses is less apparent in that case, we find it appealing that one can easily produce and report conventional CEA results without having to turn to another estimation method. Indeed, we believe that most investigators would want to know what the average result is before investigating heterogeneity in cost-effectiveness. In Section 4, we therefore show how to produce standard cost-effectiveness planes and cost-effectiveness acceptability curves to visualize uncertainty in the average estimates, alongside tools for visualizing and investigating patterns of heterogeneity.

Before moving on, we also want to highlight that the estimates from causal forests tend to exhibit small sample bias. Their variance estimates and associated confidence intervals are also only motivated asymptotically—in our case, we need to assume that the sample is large enough to achieve a normal sampling distribution for incremental costs and outcomes. Causal forests may therefore perform poorly in studies with small samples. Skewed outcomes, such as costs and QALYs, may exacerbate these issues. In Appendix S3, we report a simulation study that tests the finite sample performance of CEA forests on data generated to mimic a typical cost-effectiveness study (with skewed outcome and cost variables). As expected, we find that causal forests perform well when applied to data from relatively large samples. Our simulations also reaffirm that the method suffers from small sample bias. While we cannot provide an exact minimum sample size beyond which these errors become negligible as it depends partly on the data generating process (e.g., the degree of skewness in the outcomes), our simulations indicate that around 250–500 individuals may be sufficient. The results may be severely biased if the method is applied to smaller datasets, and we would therefore not recommend using causal forests for the purpose of CEA in small-sample studies (a Bayesian alternative may then be more appropriate, as we return to in Section 6).
2.2 Causal forests

A causal forest can be seen as a modified random forest (Wager & Athey, 2018). The original random forest algorithm averages the results from $B$ noisy regression trees into a more stable prediction model, with the aim of predicting an outcome $Y_i$ based on covariates $X_i$ (Breiman, 2001). In that case, each regression tree is created by first drawing a random subsample of the data. The individuals in the subsample are then split into subgroups (“leaves”) based on covariate values, where the splits are defined by minimizing some loss function such as the root mean squared error. This process is then repeated and averaged over $B$ trees, resulting in a random forest that can be used to predict $Y_i$ for a particular covariate combination $X_i = x$.

Expressed intuitively, the main difference between a causal forest and a random forest is that a causal forest consists of causal trees (Athey & Imbens, 2016), which estimate conditional average treatment effects (CATE) at the leaves of the trees instead of predicting the outcome. The CATE can be defined as a function of covariates, such as $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$. Confounding, if present, is adjusted for using residualization following Robinson (1988) and Nie and Wager (2019). In practice, this is achieved by first estimating a regression forest that predicts the conditional mean outcome $\mu(x) = \mathbb{E}[Y_i|X_i = x]$ and another regression forest estimating the conditional propensity score $e(x) = \mathbb{P}[W_i = 1|X_i = x]$. The predictions from these forests are then used to locally center (residualize) $Y_i$ and $W_i$ before training the causal forest. The forest-based estimates for $\tau(x)$ are then given by

$$\hat{\tau}(x) = \frac{\sum_{i=1}^{n} \alpha_i(x)(Y_i - \hat{\mu}^{(-i)}(X_i))(W_i - \hat{e}^{(-i)}(X_i))}{\sum_{i=1}^{n} \alpha_i(x)(W_i - \hat{e}^{(-i)}(X_i))^2},$$

(1)

where the superscript $(−i)$ is used to denote out-of-bag predictions (i.e., predictions obtained from the subset of trees where individual $i$ was not used to determine the splits), and the weights $\alpha_i(x)$ (which are non-negative and sum to one) measure how often the individual $i$ falls in the same leaf as the test point $x$ in the causal forest (Athey et al., 2019). More precisely, the causal forest algorithm proceeds in a fashion similar to random forests in that it repeatedly takes random subsamples of the data to train $B$ noisy causal trees for aggregation. However, there are some important differences worth noting with respect to how effects are estimated and the splits are determined by the causal forest algorithm. To avoid overfitting and ensure valid statistical inference, the trees in a causal forest must be “honest” in the sense that they never use the same information to determine the splits and estimate effects, which is achieved by splitting each subsample into two parts where one is used to determine the splits and the other to estimate the effects (Athey & Imbens, 2016). This is also known as cross-fitting, which is an essential component for valid statistical inference based on machine learning algorithms (Chernozhukov et al., 2017). While this may sound like an inefficient use of the available data, we note that every observation is used for either splitting or estimation in some iteration provided that the forest contains a sufficiently large number of trees. Internally, the splits in each tree are decided such that the accuracy of the estimates improves as fast as possible using a gradient-based algorithm detailed in Athey et al. (2019).

The complexity of the model is also controlled by a set of hyperparameters, such as the subsampling fraction, the minimum number of individuals in each leaf and the maximum imbalance in the ratio of treatment and control units allowed in each split. The grf package for R allows for automatic tuning of these parameters via cross-validation (Tibshirani et al., 2020), which avoids the need for manual specification searches. In practice, the only parameter that needs to be specified manually is the number of trees to grow ($B$), where the general recommendation is to grow enough trees such that the size of the random noise induced by the subsampling scheme is ignorable and the results remain stable across several forests fit to the same data. Another practical benefit to causal forests is that they perform an implicit variable selection that prunes irrelevant covariates, given that variables that do not modify the effect size will rarely be used to determine the splits within each tree.

Finally, the fully grown forest can then be used to estimate treatment effects for individuals with a particular (exact) covariate combination or for subpopulations defined by covariates (e.g., older men with a body mass index above 30) using Equation (1). These estimates are central to our methodology below, as they allow us to effectively explore and visualize the heterogeneity in the data.

Beyond the individual-level estimates, we are also interested in structural parameters for the entire sample and for various subpopulations. Following Chernozhukov et al. (2017, 2018), our inferential procedures for structural parameters (e.g., average effects) rely on cross-fitted augmented inverse propensity learning, a de-biasing technique required for valid statistical inference after machine learning estimation. Specifically, we construct the following scores:
\[ \hat{\Gamma}_i = \hat{\epsilon}^{(-i)}(X_i) + \frac{(Y_i - \hat{\mu}^{(-i)}(1, X_i)) W_i}{e^{(-i)}(X_i)} - \frac{(Y_i - \hat{\mu}^{(-i)}(0, X_i))(1 - W_i)}{1 - e^{(-i)}(X_i)} \]  

(2)

where \( \hat{\epsilon}^{(-i)}(X_i) \) is an out-of-bag conditional treatment effect estimate for individual \( i \) based on the causal forest; \( \hat{\mu}^{(-i)}(1, X_i) = \hat{\mu}^{(-i)}(X_i) + \hat{\epsilon}^{(-i)}(X_i) \hat{\epsilon}^{(-i)}(X_i) \) is an (out-of-bag) estimate of the conditional mean function \( E[Y_i(1)|X_i = x] \) under the treatment state, where \( \hat{\mu}^{(-i)}(X_i) \) is an estimate of \( [Y_i|X_i = x] \) obtained using the outcome regression forest (i.e., estimated before the causal forest to residualize the outcome) and \( e^{(-i)}(X_i) \) is an estimate of \( P[W_i = 1|X_i = x] \) given by the propensity score for treatment regression forest, and \( \hat{\mu}^{(-i)}(0, X_i) = \hat{\mu}^{(-i)}(X_i) - (1 - e^{(-i)}(X_i)) \hat{\epsilon}^{(-i)}(X_i) \) is an analogous estimate for the conditional mean under the control state. The resulting scores, \( \hat{\Gamma}_i \), are equivalent to those used in the doubly robust augmented inverse propensity score estimator described by Robins et al. (1995), and we therefore refer to them as doubly robust scores below. As shown by Chernozhukov et al. (2017), the sample mean of \( \hat{\Gamma}_i \) is a consistent estimator for the ATE under unconfoundedness, with asymptotically valid standard errors given by \( \hat{\sigma}/\sqrt{n} \), where

\[ \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{N} \left( \hat{\Gamma}_i - \hat{\Gamma}_i \right)^2, \]  

(3)

CATEs and their standard errors can be estimated in a similar fashion by taking averages over subgroups. Below, we use the doubly robust scores to estimate cost-effectiveness acceptability curves (CEAC), train policy decision trees, and conduct inference for the determinants of heterogeneity.

3 | COST-EFFECTIVENESS FORESTS

In this section, we discuss an extension to the causal forest methodology to accommodate CEA. We propose an algorithm that consists of a combination of several causal forests that, when used for this purpose, we refer to as a CEA forest. In essence, a CEA forest takes individual-level data on the health outcome and stores the corresponding out-of-bag estimates \( e^{(-i)}(X_i) \). Otherwise, use known propensities (e.g., 0.5 in a 1:1 randomized trial).

1. If the treatment propensity is unknown (e.g., in an observational study), train a regression forest using the treatment (\( W \)) as the outcome to estimate propensity scores for treatment and store the corresponding out-of-bag estimates \( e^{(-i)}(X_i) \). Otherwise, use known propensities (e.g., 0.5 in a 1:1 randomized trial).
2. Train two regression forests using \( Y_i \) and \( C_i \) and store the estimates of each respective conditional mean function and store the estimated \( \hat{\mu}^{(-i)}(X_i) \).
3. Using the estimates from Steps 1 and 2 to locally center \( Y_i, C_i \) and \( W_i \), train the following three causal forests and store the estimated \( \hat{\epsilon}^{(-i)}(X_i) \) and their associated variance:
   a. An incremental outcome forest using \( Y \) as the outcome to estimate the function \( \Delta Y(x) = E[Y_i(1) - Y_i(0)|X_i = x] \).
   b. An incremental cost forest using \( C \) as the outcome to estimate \( \Delta C(x) = E[C_i(1) - C_i(0)|X_i = x] \).
   c. A net monetary benefit forest using \( Y_i \lambda - C_i \) as the outcome to estimate the function \( B(x, \lambda) = E[\Delta Y \lambda - \Delta C|X_i = x] \).
4. Calculate the doubly robust scores for incremental outcomes \( \hat{\Gamma}_i^{\Delta Y} \) and costs \( \hat{\Gamma}_i^{\Delta C} \) by plugging the appropriate estimates from Steps 1–3 into Equation (2).

In practice, we use the causal_forest and regression_forest functions from the grf package for R to estimate each forest (using the tune.parameters option to automatically tune all hyperparameters to the data via cross-validation) (Tibshirani et al., 2020). As noted above, the only option that requires manual input is the number of trees used to grow the forest, which only affects the size of the Monte Carlo errors introduced by the subsampling scheme. While the minimally appropriate number of trees depends on the data, somewhere around 100,000 trees seems to be a common choice in applications (Athey et al., 2019; Cui et al., 2020). Ultimately, we recommend running sensitivity analyses with different random number seeds to ensure that the results are stable enough to produce the same key conclusions in each iteration. If not, it is necessary to increase the number of trees further.
The proposed algorithm can be used to explore patient heterogeneity in incremental costs and health outcomes separately using the results from Step 3a and 3b. Doing so provides a possibility to gain insights into whether the heterogeneity in cost-effectiveness is driven by variation in \( \Delta Y(x) \) or \( \Delta C(x) \) (or both). However, our primary focus lies in the combination of these two functions. Our algorithm also estimates a *net monetary benefit* forest by using \( Y_i - \lambda - C_i \) as the outcome (Step 3c) (Hoch et al., 2002), where \( \lambda \) is the WTP per one-unit increase in \( Y \). (If one prefers to express the results in terms of net health benefits, we note that the net monetary benefit forest can be replaced with a net health benefit forest by using \( Y_i - C_i/\lambda \) as the outcome [Hoch et al., 2002]). The estimated CATEs from these forests can then be directly interpreted as the conditional net monetary benefit (\( B \)) given \( X_i \), that is, \( B(x, \lambda) = E[\Delta Y \lambda - \Delta C|X_i = x] \) (see Appendix S1 for proof). In practice, however, we only use the net monetary benefit forest for plotting individual-level estimates and for obtaining variance estimates for the individual-level predictions of the net monetary benefit. For structural parameter estimation, we allow for greater post-estimation flexibility by instead combining the doubly robust scores from the outcome and cost forests as follows:

\[
\hat{\Gamma}^B_i = \hat{\Gamma}^{\Delta Y}_i \lambda - \hat{\Gamma}^{\Delta C}_i
\]

where \( \hat{\Gamma}^{\Delta Y}_i \) is the doubly robust scores in Equation (2) based on the outcome forest and \( \hat{\Gamma}^{\Delta C}_i \) are analogously defined scores from the cost forest (see Appendix S2 for more detailed definitions). These scores are equivalent in expectation to the corresponding scores derived directly from the net monetary benefit forest (see Appendix S2 for proof), but allows us to change the WTP for subsequent analyses (e.g., cost-effectiveness acceptability curves) without having to re-estimate the net monetary benefit forest. We obtain standard errors for structural parameter estimates based on \( \hat{\Gamma}^{\Delta Y}_i \), \( \hat{\Gamma}^{\Delta C}_i \) and \( \hat{\Gamma}^B_i \) using the same method as in Equation (4).

### 3.1 Quantifying uncertainty

We note that the bootstrapping methods often used in conventional CEA estimation with individual-level data, where the entire sample is resampled with replacement, are not appropriate for causal forests. This is because they introduce bias into the honesty subsampling scheme required for consistent estimation due to duplication of individuals. Growing new forests within each bootstrap resample would also be very inefficient from a computational perspective. Instead, Athey et al. (2019) rely on subsampling and a method called the bootstrap of little bags to estimate the variance for \( \hat{\tau}(x) \) at almost no computational cost, and provide evidence for the consistency and asymptotic normality of forest-based effect estimates. In general, we should therefore be able to rely on the central limit theorem in large samples and for asymptotically normal quantities such as average incremental outcomes and costs (Nixon et al., 2010). However, non-linear combinations of these quantities, for example, ICER (\( \Delta C/\Delta Y \)), are non-normally distributed even in large samples. In this case, Fieller’s theorem can be used to produce confidence intervals assuming a bivariate normal distribution for \( \Delta Y \) and \( \Delta C \), but this approach may sometimes result in unbounded or imaginary interval limits (e.g., when \( \Delta Y \) does not differ significantly from zero) (Willan & O’Brien, 1996). In Appendix S3, we provide simulation evidence showing that bootstrapping the doubly robust scores with replacement after training the forest performs well for constructing confidence intervals for ICERs provided that the sampling distribution for the incremental outcomes and cost estimates is normal at sample size \( n \). This approach can therefore serve as an alternative method of inference for parameters where there is no closed-form solution for the variance. As we show below, bootstrapping pairs of doubly robust scores for outcomes and costs (\( \hat{\Gamma}^{\Delta Y}_i \), \( \hat{\Gamma}^{\Delta C}_i \)) also serves fast and convenient way for constructing cost-effectiveness planes based on a CEA forest.

### 4 APPLYING CEA FORESTS

In this section, we provide suggestions for how the results from CEA forests can be analyzed and communicated. We propose a simple workflow containing several tools that we find to be useful for (i) estimating the average cost-effectiveness in the entire sample or in subgroups, (ii) exploring and analyzing the heterogeneity in incremental outcomes, costs and net monetary benefits (and their determinants), and (iii) learning policy rules from the data. Each method presented below can be thought of as a post-estimation method that can be applied to a trained CEA forest. We
implement the CEA forest algorithm and the proposed post-estimation methods in an R package (CEAforests, available at: https://github.com/bonander/CEAforests).

We use simulated data in our demonstrations below to enable replication. The dataset is intended to mimic a typical individual-level CEA study. It contains 500 individuals, three correlated covariates (a hypothetical risk score for disease progression [Risk_score], severe disease subtype [Severe_disease], and health-related quality of life at baseline [HRQoL]), a binary treatment variable $W_i$, a cost variable $C_i$ intended to emulate healthcare costs in USD and an outcome $Y_i$ that is intended to depict QALYs at follow-up. The data is generated so that the treatment effect is confounded by disease subtype, and that $\Delta Y$ and $\Delta C$ exhibit heterogeneity depending on the variable sets {risk score, disease subtype} and {HRQoL, risk score}, respectively (we provide further details on the data generating process in Appendix S3). Throughout, we use 50,000 USD as the WTP per QALY (“threshold value”) unless otherwise noted. We begin by training a CEA forest using this data. We tune all tunable hyperparameters via the grf package, and grow 50,000 trees per forest. Code that enables creation of the simulated data and replication of the analyses presented are available online as supplementary materials to this paper.

### 4.1 Cost-effectiveness results for the entire sample and user-specified subgroups

After training the CEA forest, the first natural step is to estimate the average cost-effectiveness in the population that the sample represents, that is, to perform a conventional CEA analysis. We use the doubly robust scores from Step 4 in the CEA forest algorithm to compute the statistics in Table 1A for the entire sample. Beyond average incremental outcomes, costs and net monetary benefits, these scores can also be used to estimate the ICER. In Table 1, we present confidence intervals (CIs) based on Fieller's theorem for this estimate, but as noted above, one can also bootstrap pairs of the doubly robust scores ($\Gamma^{\Delta Y}_i$, $\Gamma^{\Delta C}_i$) to obtain CIs for ICERs (both methods provide almost identical results in our simulated data, see online supplementary files for details). More importantly, the bootstrapped pairs can be used to construct conventional cost-effectiveness planes for the entire sample, or specified subgroups, based a single CEA forest (Figure 1). Overall, the results show that treatment is very close to our WTP threshold for the average individual.

The same CEA forest can also be used to estimate average effects for user-specified subpopulations, similar to conventional subgroup analyses but without the need for re-estimation of the model. In Table 1, we have arbitrarily selected two subgroups based on progression risk scores to highlight that the CEA forest methodology can incorporate manual subgroup analyses. The estimates indicate that the treatment is (barely) cost-effective on average (Table 1A), and that there is fairly strong evidence that it is cost-effective for patients with progression risk scores > 50 (Table 1B). The control treatment dominates the new treatment on average among individuals with risk scores below this point. However, it is not obvious that 50 is the optimal cut point for a targeted reimbursement policy, and one might also have to consider subgroups based on multiple variables to identify the best available policy. In Section 5, we therefore suggest a data-driven post-estimation procedure for identifying the optimal covariates and cut points that maximize the net monetary benefit per population member based on the individual-level estimates from a CEA forest.

Assuming the sample is sufficiently large for asymptotic normality assumptions to hold, one can also estimate the (Bayesian) probability that the data are consistent with a positive net monetary benefit for a specified subpopulation via the inverse normal cumulative distribution function (Nixon et al., 2010). This estimate is presented in Table 1 for the entire sample and for the chosen subpopulations. By extension, the WTP threshold ($\lambda$ in Equation (4)) can be varied to produce CEAC for the entire sample or for specified subpopulations based on a single CEA forest at almost no computational expense (Figure 2).

### 4.2 Describing heterogeneity in cost-effectiveness

Now turning to the heterogeneity analysis, we suggest starting with descriptive analysis of the results. For instance, we suggest plotting the individual (out-of-bag) estimates for incremental costs and outcomes in a cost-effectiveness plane. This results in an individual-level cost-effectiveness plane (cf. Basu, 2014; Basu & Meltzer, 2007), which can—besides the fact that we are viewing the variation in individual-level effect estimates rather than the variation in average effect estimates under uncertainty—be interpreted as a conventional cost-effectiveness plane. More precisely, the individual-level cost-effectiveness plane can be interpreted as the joint distribution of (estimated) individual-level differences in potential outcomes with respect to costs (on the y-axis) and health outcomes (on the x-axis). We present an example
using our simulated data in Figure 3. Based on this analysis, we see that the majority of the individual-level treatment effects are located in the northeast quadrant (i.e., new treatment improves health but is more costly), and that 45% of the sample has a positive net monetary benefit estimate.

To describe the data further, we stratify the sample into two groups based on their individual-level net monetary benefit estimate (negative or null [“not cost-effective”], or positive [“cost-effective”]). We then examine covariate balance between these groups a standard descriptive table. The resulting table gives a quick overview of how the covariate patterns differ between individuals for whom treatment is determined to be cost-effective versus ineffective or undetermined according to the CEA forest (Table 2). In our example data, we can directly see that the treatment appears to be more cost-effective for patients with a severe disease subtype and lower HRQoL at baseline.
4.3 Analyzing the determinants of heterogeneity

In some situations, the analyst may want to learn more about the causes of heterogeneity in multivariable settings with correlated covariates. In this case, it may be useful to estimate an ordinary least squares (OLS) model to explain the variation in effect size:

\[
\hat{\Gamma}_i = \alpha + \sum_{j=1}^{J} (\beta_j)X_{ji} + \epsilon_i
\]  

**Table 2** Descriptive statistics by subgroup based on individual-level net monetary benefit estimates (WTP = 50,000 USD)

|                          | Not cost-effective (n = 273) | Cost-effective (n = 227) |
|--------------------------|------------------------------|--------------------------|
| Severe disease subtype (n [%]) | 96 (35.2)                  | 168 (74.0)               |
| HRQoL (mean [SD])        | 0.45 (0.20)                 | 0.17 (0.14)              |
| Risk score (mean [SD])   | 47.49 (6.65)                | 53.61 (8.48)             |

Abbreviations: HRQoL, health-related quality of life; WTP, willingness to pay.
which provides the best linear projection of the effect size given $X_i$ (Nilsson et al., 2019; Semenova & Chernozhukov, 2020). The results can be interpreted as those from a standard OLS model with one key difference; the coefficients $\beta_j$ estimate the marginal (partial) effects of the covariates on the effect size. We note that the supplied covariate matrix can, but does not necessarily have to be, the same as the covariate matrix supplied to the forest algorithm.

We apply this method to estimate the determinants of heterogeneity in incremental outcomes, costs and net monetary benefits in our simulated data to illustrate the practical utility of the approach. The results can be presented as a conventional regression table (Table 3), which provides an overview of the strength and direction of the partial effect of each covariate on the effect size. In our simulated data, we find that HRQoL is not associated with variations in incremental outcomes after conditioning on the other covariates, which is consistent with the data generating process. The same is true for disease subtype and incremental costs. Finally, we can see that all three variables are associated with variations in cost-effectiveness.

4.4 | Generic plotting methods

A major benefit to methods that estimate individual-level effects is that they allow for visualization using generic plotting methods (Nilsson et al., 2019). Figure 4, panels A–C provides examples using histograms of the estimated incremental outcomes, costs and net monetary benefits to visualize the variation in estimated effect sizes in the data. Scatter plots can be used to explore how the effect sizes vary with respect to a single covariate (Figure 4, panels D–F). Wager and Athey (2018) use line plots to plot the predicted variation in effect size (and associated confidence intervals) with respect to a covariate while keeping all other variables in the data constant at their mean (Figure 4, panels G–I).

5 | POLICY LEARNING USING CEA FORESTS

The potential value of subgroup-specific treatment recommendations or reimbursement decisions has been conceptually demonstrated in a number of previous studies (cf. Basu & Meltzer, 2007; Coyle et al., 2003; Espinoza et al., 2014a). While subgroups can be identified manually by exploring the results from CEA forests as demonstrated above, one may also want to consider a data-driven search for optimal policy rules to identify the optimal policy based on the individual-level estimates from a CEA forest.

Suppose that we want to learn for which subgroups the new treatment should be recommended in order to maximize the net monetary benefit in the population. This is equivalent to finding the subgroups that maximize the static value of heterogeneity (Espinoza et al., 2014a). As shown by Athey and Wager (2019), such policy rules can be efficiently learned from the doubly robust scores by estimating a decision tree that maximizes the net monetary benefit per population member:

$$\hat{\pi} = \arg\max_{\pi \in \Pi} \left\{ \frac{1}{n} \sum_{i=1}^{n} (2\pi(X_i) - 1) \left( \hat{i}^{\Delta Y}_i - \hat{i}^{\Delta C}_i \right) : \pi \in \Pi \right\}, \quad (6)$$

where the goal is to learn an efficient policy $\pi$ from the policy class $\Pi$, which in our case is a finite-depth policy tree that maps an individual’s characteristics to a binary decision (to reimburse, or not reimburse), that is, $\pi : x \rightarrow \{0, 1\}$.

Our implementation relies on the policytree package for R (Zhou et al., 2020), which uses an exhaustive search to find the optimal tree of a pre-specified depth, where the depth determines the amount of subgroups to allow for. Figure 5 shows two suggested policies with different maximum tree depths, where we have supplied the three covariates in our dataset for consideration in the policy trees for illustration. The simplest tree suggests a policy based on HRQoL, whereas the larger tree also considers additional subgroups based on progression risk scores in combination with HRQoL. The policy trees provide estimates of the optimal cut points along these variables given the individual-level estimates from the CEA forest. In practice, it may be necessary to round the cut points or move them manually to ensure the feasibility of the policy.

We also note that the covariates supplied to the policy tree do not need to be the same as those included in the forest. Preferably, they should be characteristics that can be measured objectively and can be legally and ethically used for decision-making within the given healthcare context. In practice, it is also important to consider the costs of
implementing a targeted policy (Espinoza et al., 2014a). These costs are not considered by the policy tree algorithm as defined above, and must either be taken into account afterward (e.g., by incorporating them into Equations (7) and (8) below) or assumed to be ignorable (e.g., by only using data that would have been collected independent of the chosen policy).

5.1 Estimating the expected welfare gain from targeted reimbursement policies

The average welfare gain per population member under the suggested reimbursement policy, as opposed to giving everyone the control treatment (e.g., treatment as usual), can be estimated using:

$$\hat{E}[(\Delta Y_{i\lambda} - \Delta C_i)\pi(X_i)] = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\Gamma}_i^{\Delta Y_{i\lambda}} - \hat{\Gamma}_i^{\Delta C} \right) \hat{\pi}(X_i) = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\Gamma}_i^{B} \hat{\pi}(X_i) - \hat{\Gamma}_i^{B} \right), \quad (7)$$

where $\hat{\pi}(X_i)$ is a policy indicator coded as 1 if individual $i$ has a covariate combination that is assigned to the new (experimental) treatment by the learned policy and 0 if they are assigned to the control treatment.

The expected welfare gain from the suggested policy compared to giving everyone the new treatment can also be estimated as follows:

$$\hat{E}[(\Delta Y_{\lambda} - \Delta C) \times \pi(X_i)] - \hat{E}[(\Delta Y_{i\lambda} - \Delta C)] = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\Gamma}_i^{B} \hat{\pi}(X_i) - \hat{\Gamma}_i^{B} \right), \quad (8)$$

which can be used to assess whether a personalized treatment policy is sufficiently cost-effective to be motivated compared to a scenario in which everyone gets the new treatment. Standard errors for these estimates can be obtained.
**Figure 4** Histograms of the estimated effects (a–c), unconditional heterogeneity plots (d–f), and conditional heterogeneity plots (g–i) illustrating the heterogeneity in incremental outcomes, costs and net monetary benefits (NMB) in the entire sample (a–c) or the heterogeneity in effects with respect to health-related quality of life (HRQoL; d–i) using simulated data. The scatter plots in (d)–(f) illustrate the unconditional heterogeneity (including the effects of variables correlated with HRQoL), and the line plots in (g)–(i) illustrate the impact of HRQoL on the effect size keeping all other variables in the data constant at their mean.

**Figure 5** Depth-1 and depth-2 policies based on a cost-effectiveness analyses forest estimated using the policytree package for R. The optimal depth-1 policy suggested by the algorithm is to reimburse the new (experimental) treatment for only patients with health-related quality of life (HRQoL) ≤0.25 at baseline (43.8% of the population), which would result in an estimated welfare gain per population member of 24,535 USD (95% CI: 13,063, 36,008) compared to not reimbursing the new treatment for any group. The average net monetary benefit in the population is 363, so the added value from using the learned policy compared to reimbursing the new treatment for all patients would be 24,535−363 = 24,172 (95% CI: 11,826, 36,518) per population member. The depth-2 policy is more complex, and allows for a more inclusive HRQoL threshold for patients with higher progression risk scores, and would therefore cover roughly 50% of the population. The estimated average welfare gain of applying this policy would be 30,868 USD (95% CI: 18,296, 43,441).
We have presented a data-driven approach for assessing heterogeneity in CEA, which we believe can serve as an (at least partially) automated way to implement previous conceptual work on heterogeneity analysis in this context (Basu & Meltzer, 2007; Coyle et al., 2003; Espinoza et al., 2014a). Automated methods can serve as a complement to clinically informed subgroup analyses, especially if there are complex, multivariable patterns of heterogeneity in the data that are difficult to capture using a limited set of pre-defined subgroups (Athey & Imbens, 2019). For instance, CEA forests may help identify previously unknown non-linear patterns in cost-effectiveness alongside continuous variables and can easily incorporate several variables at once. These complexities can technically be incorporated into manually specified models and in subgroup analyses (Nilsson et al., 2019), but conventional practice is typically to perform subgroup analyses either along a single variable or to divide the data based on multiple binary variables (e.g., age group and sex; VanderWeele et al., 2019). While such simplifications may serve as sufficient input for policy decisions, they may also result in a loss of important information (VanderWeele et al., 2019). Nonetheless, it is important to recognize that there is a trade-off to be made between technical rigor and the utility of the output for decision-making. Data-driven approaches may help find hidden patterns in the data that can serve as input for drafting a policy, but there may be many other external factors (that are unknown to the model) that need to be taken into account before it is finalized. Thus, we do not believe that CEA forests (or any other data-driven method) should be seen as a replacement for conventional methods—rather, we view them as complements. Indeed, the framework presented above can be seen as both automated and manual; the CEA forest performs an automated search to create a prediction model for individual-level cost-effectiveness, and these estimates can then be explored either manually (using descriptive tables, plots or regression models) or automatically (using policy trees).

We note that causal forests are not the only available alternative for data-driven heterogeneity analysis (Athey & Imbens, 2019). We chose to focus on this approach because it inherits many appealing traits shared by tree-based machine learning methods; it is free of functional form assumptions, and is data-driven and requires limited researcher input and tuning compared to other machine learning methods. It also has a built-in cross-fitting algorithm, which simplifies statistical inference based on machine learning estimators (Athey & Imbens, 2019; Chernozhukov et al., 2018). To facilitate implementation, we have also developed an R package that implements our proposed methods (CEAforests, available at: https://github.com/bonander/CEAforests). The package works primarily as a wrapper for the grf package (Tibshirani et al., 2020), but is specifically tailored to produce the analyses and outputs presented in this paper.

While our proposed method serves as one way to perform data-driven heterogeneity analysis in the context of CEA, we believe that the rapidly growing literature on causal machine learning methods holds great promise for further refinement (Athey & Imbens, 2019). Importantly, a key drawback to causal forests is that they tend to perform poorly in small samples (Appendix S3), which currently limits the application of CEA forests to studies with large samples. A Bayesian alternative to causal forests has shown better finite sample performance than the original approach (Hahn et al., 2020), and it would therefore be of considerable interest to extend the methods discussed in this paper to incorporate Bayesian causal forests. We also note that causal forests may suffer from so-called “edge effects” (Athey et al., 2019), where the slope of the heterogeneity estimates may taper of as they reach the edges of the observed covariate space (even if the true data generating process keeps changing at that point). Hence, extrapolation beyond the range of the observed data should be done with due caution. The method may also fail to adequately model smooth patterns of heterogeneity (e.g., linear and quadratic shapes) in small samples (Athey & Imbens, 2019). In that case, a modified approach that relies on local linear regression may be a more appropriate alternative to standard causal forests (Friedberg et al., 2019). A limitation to the policy learning approach is that it is not designed to automatically find the optimal number of subgroups for targeted reimbursement policies. An extension to this idea that also identifies the optimal complexity of a policy while considering transaction costs could potentially increase the practical utility of the method for decision-making. In addition, we have focused solely on analyses of individual-level data, such as, from a trial and/or observational study. It would be of considerable interest to explore the potential benefits of combining our approach with decision-analytic models to be able to capture the long-term perspective that is often lacking in studies that rely solely on individual-level data. It would also be interesting to assess the value of using causal forests for...
estimating the dynamic value of heterogeneity (Espinoza et al., 2014a). Finally, we note that our paper has focused on
the case with a binary treatment and where the analyst is interested in intention-to-treat estimates. Extensions can
easily be made to incorporate instrumental variables and continuous treatment variables, as these features have already
been developed for causal forests (Athey et al., 2019).

ACKNOWLEDGMENTS
Carl Bonander was supported by a research grant from the Swedish Research Council for Health, Working Life and
Welfare (grant number: 2017-00414). MS was supported by a grant from Familjen Kamprads Stiftelse (grant number:
20180122).

CONFLICT OF INTEREST
None to declare.

DATA AVAILABILITY STATEMENT
The data that supports the findings of this study are available in the supplementary material of this article.

ORCID
Carl Bonander https://orcid.org/0000-0002-1189-9950
Mikael Svensson https://orcid.org/0000-0003-1113-7478

REFERENCES
Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences of the
United States of America, 113(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113
Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. Annual Review of Economics, 11(1),
685–725. https://doi.org/10.1146/annurev‐economics‐080217‐053433
Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. Annals of Statistics, 47(2), 1148–1178. https://doi.org/10.1214/18‐
AOS1709
Athey, S., & Wager, S. (2019). Efficient policy learning. arXiv:1702.02896v5.
Baker, R., Chilton, S., Donaldson, C., Jones‐Lee, M., Lancsar, E., Mason, H., Metcalf, H., Pennington, M., & Wildman, J. (2011). Searchers vs
surveyors in estimating the monetary value of a QALY: Resolving a nasty dilemma for NICE. Health Economics Policy and Law, 6,
435–447.
Basu, A. (2011). Economics of individualization in comparative effectiveness research and a basis for a patient-centered health care. Journal of Health Economics, 30(3), 549–559. https://doi.org/10.1016/j.jhealeco.2011.03.004
Basu, A. (2014). Estimating person-centered treatment (PeT) effects using instrumental variables: An application to evaluating prostate cancer treatments. Journal of Applied Economics, 29(4), 671–691. https://doi.org/10.1002/jae.2343
Basu, A., & Meltzer, D. (2007). Value of information on preference heterogeneity and individualized care. Medical Decision Making, 27(2),
112–127. https://doi.org/10.1177/0272989x06297393
Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
Chalkidou, K., Tunis, S., Lopert, R., Rochaix, L., Sawicki, P. T., Nasser, M., & Xerri, B. (2009). Comparative effectiveness research and
evidence-based health policy: Experience from four countries. The Milbank Quarterly, 87(2), 339–367. https://doi.org/10.1111/j.1468-
0009.2009.00560.x
Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/Debiased/Neyman machine learning of
treatment effects. American Economic Review, 107(5), 261–265. https://doi.org/10.1257/aer.p20171038
Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning
for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097
Coyle, D., Buxton, M. J., & O’Brien, B. J. (2003). Stratified cost-effectiveness analysis: A framework for establishing efficient limited use
criteria. Health Economics, 12(5), 421–427. https://doi.org/10.1002/hec.788
Cui, Y., Kosorok, M. R., Wager, S., & Zhu, R. (2020). Estimating heterogenous treatment effects with right-censored data via causal survival
forests. ArXiv:2001.09887v1.
Espinoza, M. A., Manca, A., Claxton, K., & Sculpher, M. J. (2014a). The value of heterogeneity for cost-effectiveness subgroup Analysis.
Medical Decision Making, 34(8), 951–964. https://doi.org/10.1177/0272989X14538705
Espinoza, M. A., Sculpher, M. J., Manca, A., & Basu, A. (2014b). Analysing heterogeneity to support decision making. In A. J. Culyer (Ed.),
Encyclopedia of health economics (pp. 71–76). San Diego, CA: Elsevier.
Friedberg, R., Tibshirani, J., Athey, S., & Wager, S. (2019). Local linear forests. arXiv:1807.11408v3.
Hahn, P. R., Murray, J. S., & Carvalho, C. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and
heterogenous effects. arXiv:1706.09523v4.
Hoch, J. S., Briggs, A. H., & Willan, A. R. (2002). Something old, something new, something borrowed, something blue: A framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics, 11*(5), 415–430. https://doi.org/10.1002/hec.678

ISPOR. (2013). *International Society for Pharmacoeconomics and Outcomes Research: Pharmacoeconomic guidelines around the world*. Retrieved from [http://www.ispor.org/peguidelines/index.asp](http://www.ispor.org/peguidelines/index.asp)

Kohli-Lynch, C. N., & Briggs, A. H. (2019). Heterogeneity in cost-effectiveness analysis. *Oxford Research Encyclopedia of Economics and Finance*. https://doi.org/10.1093/acrefore/9780190625979.013.101

Nie, X., & Wager, S. (2019). Quasi-oracle estimation of heterogenous treatment effects. *ArXiv:1712.04912*. Retrieved from [http://arxiv.org/abs/1712.04912](http://arxiv.org/abs/1712.04912)

Nilsson, A., Bonander, C., Strömberg, U., & Björk, J. (2019). Assessing heterogeneous effects and their determinants via estimation of potential outcomes. *European Journal of Epidemiology, 34*(9), 823–835. https://doi.org/10.1007/s10654-019-00551-0

Nixon, R. M., Wonderling, D., & Grieve, R. D. (2010). Non-parametric methods for cost-effectiveness analysis: The central limit theorem and the bootstrap compared. *Health Economics, 19*(3), 316–333. https://doi.org/10.1002/hec.1477

Ramaekers, B. L. T., Joore, M. A., & Grutters, J. P. C. (2013). How should we deal with patient heterogeneity in economic evaluation: A systematic review of national pharmacoeconomic guidelines. *Value in Health, 16*(5), 855–862. https://doi.org/10.1016/j.jval.2013.02.013

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association, 90*(429), 106–121. https://doi.org/10.1080/01621459.1995.10476493

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica, 56*(4), 931–954. https://doi.org/10.2307/1912705

Sanders, G. D., Neumann, P. J., Basu, A., Brock, D. W., Feeny, D., Krahn, M., Kuntz, K. M., Meltzer, D. O., Owens, D. K., Prosser, L. A., Salomon, J. A., Sculpher, M. J., Trikalinos, T. A., Russell, L. B., Siegel, J. E., & Ganiats, T. G. (2016). Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses. *Journal of the American Medical Association, 316*(10), 1093–1103. https://doi.org/10.1001/jama.2016.12195

Savin, N. E. (1984). Chapter 14 Multiple hypothesis testing, *Handbook of Econometrics, 2*, 827–879.

Sculpher, M. (2008). Subgroups and heterogeneity in cost-effectiveness analysis. *Pharmacoconomics, 26*(9), 799–806.

Semenova, V., & Chernozhukov, V. (2020). Debiased machine learning of conditional average treatment effects and other causal functions. *arXiv:1702.06240v6*.

Tibshirani, J., Athey, S., Friedberg, R., Hadad, V., Hirschberg, D., Miner, L., Sverdrup, E., Wager, S., & Wright, M. (2020). grf: generalized random forest. *R package version 1.0.1*. Retrieved from [https://github.com/grf-labs/grf](https://github.com/grf-labs/grf)

VanderWeele, T. J., Luedtke, A. R., van der Laan, M. J., & Kessler, R. C. (2019). Selecting optimal subgroups for treatment using many covariates. *Epidemiology, 30*(3), 334–341. https://doi.org/10.1097/ede.0000000000000991

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., & Cole, S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology, 186*(8), 1010–1014. https://doi.org/10.1093/aje/kwx164

Willan, A. R., & O’Brien, B. J. (1996). Confidence intervals for cost-effectiveness ratios: An application of Fieller’s theorem. *Health Economics, 5*(4), 297–305. https://doi.org/10.1002/(sic)1099-1050

Zhou, Z., Athey, S., Wager, S., Kanodia, A., & Sverdrup, E. (2020). Multi-action treatment effect estimation and exact tree search (beta): policytree. *R package version 0.1.0*. Retrieved from [https://github.com/grf-labs/policytree](https://github.com/grf-labs/policytree)

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

---

**How to cite this article:** Bonander, C., & Svensson, M. (2021). Using causal forests to assess heterogeneity in cost-effectiveness analysis. *Health Economics, 30*(8), 1818–1832. [https://doi.org/10.1002/hec.4263](https://doi.org/10.1002/hec.4263)