# HOW DO APPLIED RESEARCHERS USE THE CAUSAL FOREST? A METHODOLOGICAL REVIEW OF A METHOD

**Patrick Rehill**
POLIS: The Centre for Social Policy and Research
Australian National University
Canberra, Australia
patrick.rehill@anu.edu.au

## ABSTRACT

This paper conducts a methodological review of papers using the causal forest machine learning method for flexibly estimating heterogeneous treatment effects. It examines 133 peer-reviewed papers. It shows that the emerging best practice relies heavily on the approach and tools created by the original authors of the causal forest such as their grf package and the approaches given by them in examples. Generally researchers use the causal forest on a relatively low-dimensional dataset relying on randomisation or observed controls to identify effects. There are several common ways to then communicate results — by mapping out the univariate distribution of individual-level treatment effect estimates, displaying variable importance results for the forest and graphing the distribution of treatment effects across covariates that are important either for theoretical reasons or because they have high variable importance. Some deviations from this common practice are interesting and deserve further development and use. Others are unnecessary or even harmful.

***Keywords*** Causal forest · Causal machine learning · Heterogeneous treatment effects · Methodological review

## 1 Introduction

Causal machine learning methods offer a flexible way to estimate heterogeneous treatment effects nonparametrically. This is particularly useful for high-dimensional datasets, those where drivers of heterogeneity are not known in advance and those with nonlinear effects. One of the most popular methods is the causal forest (Athey & Wager 2019, Athey et al. 2019). This is because the causal forest has in comparison to other estimators (such as other meta-learners (Künzel et al. 2019)) desirable statistical qualities — such as asymptotic normality and doubly robust estimation. It gives strong 'plug-and-play' performance for applied researchers who may not have much experience designing and tuning machine-learning models (Wager & Athey 2018). The most popular implementation is in the grf package (Athey et al. 2019) which includes some useful quality of life features like auto-ML via a tuning forest, automatic fitting of nuisance models when estimates are not provided, efficient error calculation without bootstrapping (via jackknifing), and tools to help understand results.

In the years since the method was proposed many applied papers have made use of it, however, the precise way they use it varies. The aim of this paper is to try and understand the state of applied work using the causal forest. This is a kind of methodological review but focused not on one field of research, but on one method. It examines peer reviewed papers indexed by Scopus and Web of Science that have used these methods in applied work and tries to make sense of just how they were used. This method is immensely promising for applied researchers, but it is only just beginning to be applied. That is why it is important to pay attention to the way the causal forest is being applied, highlight best practice, point out mistakes that have been made and suggest directions the development of the causal forest may take in the future. The review covers 133 peer-reviewed publications across many different fields. It focuses on three broad concerns — the identification of effects, the estimation of effects and the presentation of results.[1]

---

[1]In econometrics, how results of analysis with a method are presented generally gets much less attention than identification and estimation. However, the causal forest's main output (an often high-dimensional conditional average treatment effect distribution) is

The review finds that the causal forest has been used in many different ways across many different fields, however, most publications stick to the same basic design laid out by Athey, Wager and co-authors across several different papers (Wager & Athey 2018, Athey et al. 2019, Athey & Wager 2019, Athey & Imbens 2016). This approach has been encouraged by the excellent tooling in their widely used grf package. This approach generally involves fitting a causal forest via grf on observational or experimental data using somewhere around 2000-4000 trees and 10-20 covariates for each model. Often users trim the number of variables used in the final model by following the procedure laid out by Basu et al. (2018) and demonstrated for the causal forest by Athey & Wager (2019) (I call this the Basu Technique from hereon out). Identification is generally achieved either by randomisation or through controlling on observables. An average treatment effect (ATE) is then estimated using augmented inverse propensity weighting (AIPW) (Robins et al. 1994). Conditional average treatment effects (CATEs) are explored by some combination of plotting the distribution of individual estimates as a histogram, presenting variable importance (that is a depth-weighted count of the use of different variables in splitting), fitting a best linear projection (Semenova & Chernozhukov 2021, Tibshirani et al. 2021), or graphing treatment effects across values of important variables. Often (particularly on designs with observational data), this causal forest analysis is one part of a paper that uses more standard methods to estimate an average effect before using a causal forest to explore treatment effect heterogeneity or as a robustness check for their preferred method of studying heterogeneity. For example, it is not uncommon for a paper to use difference-in-differences to estimate an average treatment effect for the treated and then use a standard causal forest to understand heterogeneity.

While there is a lot of similarity in approaches, some authors are innovating; taking on methods developments from outside of the Athey and Wager body of work (e.g., Cockx et al. 2023, Osawa et al. 2023) or presenting results in innovative ways (e.g., Baum et al. 2017). However, not all these novel approaches are positive. Some changes seem to be redundant or even harmful. For example, several authors use cross-validation for hyperparameter tuning when there is already a tuning forest built into the grf causal forest that does tuning automatically (Guo et al. 2021). In all of these cases, there are reasons why the authors may have done this, but the papers do not make a point of justifying this approach. This suggests it is simply due to a misunderstanding of the causal forest implementation they are using. There are also cases where authors seem to misunderstand the method and do things that are outright counter-productive such as averaging together individual tree predictions rather than using the forest to provide kernel weights to a doubly robust estimator (Bittencourt & Albuquerque 2020) (thereby using the older method from Wager & Athey (2018) and ignoring the generalised random forest (Athey et al. 2019)).

The literature using the causal forest then is both rather conservative in the approach taken and also often unsuccessful when it tries to deviate from this conservative approach. That is not to say that it is best to stick to the current common practice. There are certainly innovations in the methods literature which will hopefully filter into applied work in time. There are also some successful innovations amongst these papers that show what is possible when the methods are used in a way that is both rigorous and creative.

## 2 A brief introduction to the causal forest

As this study will be talking a lot about the specificities of different causal forest designs, it is worth explaining what the causal forest is and how it works for any reader who may be unfamiliar. The causal forest is an approach to estimating CATEs based on the random forest — a popular predictive machine learning method (Breiman 2001). The standard causal forest has developed through several key methodological papers by Athey, Wager and co-authors (Athey & Imbens 2016, Wager & Athey 2018, Nie & Wager 2020, Athey et al. 2019).[2] Its estimates are asymptotically unbiased and normal.

The causal forest has several key steps — it removes confounding with nuisance models, it fits a model to cluster observations with treatment effect heterogeneity to form an adaptive kernel and then it estimates out treatment effects within each kernel bandwidth. Three models are used, a nuisance model that predicts treatment, a nuisance model that predicts outcome and finally a model of treatment effect heterogeneity. These nuisance models are generally random forests themselves (typically regression forests in grf) that can be used to partial out selection effects (similar to how they are used in double machine learning (Chernozhukov et al. 2018)). This partialing out is called local centering — essentially taking the residuals of the models as the treatment and outcome variables used in the final model. The final heterogeneity model is a random forest made up of causal trees (Athey & Imbens 2016). These differ from standard regression trees in two ways. Firstly they use a special splitting criterion designed to find treatment effect heterogeneity (estimated expected mean squared error). Secondly they are 'honest', this means each tree uses typically half of its sample to split the tree and the other half to make estimates for each terminal node. This prevents

---

so challenging to interpret on its own that presentation of results deserves a special focus here. We need some way to digest such a distribution into something that provides useful insights, and there is not a canonical way to do so (Rehill & Biddle 2023).

[2]There are many other variations on the causal forest that have been proposed for example by Lechner (2019) or Dandl et al. (2022). The review will show that they are used little in practice so for simplicity I do not discuss them in depth here.

over-fitting, giving asymptotically normal predictions out of each tree. The overall ensemble these fit into is trained as an R-Learner meta-learner (Nie & Wager 2020). This means it minimises the R-Loss function to estimate a heterogeneity model $\tilde{\tau}$. Here we see the local centering ($Y_i - m^*$, $W_i - e^*(X_i)$) which gives us the residuals (sometimes called pseudo-outcomes) that the final model is fit on. There is also a regularisation that is in this case, implicit in the structure of the causal forest (e.g. from tuning minimum node size $\Lambda_n\{\tau(\cdot)\}$).

$$\tilde{\tau}(\cdot) = \text{argmin}_\tau \left( \frac{1}{n} \sum_{i=1}^{n} [\{Y_i - m^*(X_i)\} - \{W_i - e^*(X_i)\}\tau(X_i)]^2 + \Lambda_n\{\tau(\cdot)\} \right)$$

This loss function allows for the tuning of the forest automatically via a smaller tuning forest that is fit as part of the causal forest function.

The last innovation of the causal forest is that rather than predicting directly with an average of trees like a standard random forest, the method simply uses the ensemble as an adaptive kernel, finding weights to calculate treatment effects for neighbourhoods in a high-dimensional covariate space. We find the weights for a covariate set $x$ by looking through each data-point $i$ and finding the proportion of trees that were not fit on $X_i$ (to prevent overfitting) for which $X_i$ and $x$ fall into the same leaf. This gives us the kernel function $h(x, X)$.

We use the nuisance models to estimate doubly robust scores $\hat{\Gamma}(X)$ (the standard approach is via Augmented Inverse Probability Weighting (Robins et al. 1994)), then we take a weighted average of these using the kernel function

$$\hat{\tau}(x) = \frac{\sum_{i=1}^{n} h(x, X_i)\hat{\Gamma}(X_i)}{\sum_{i=1}^{n} h(x, X_i)}.$$

Standard errors can be calculated via resampling, generally jackknifing which is much less computationally intensive for random forests than bootstrapping (Wager et al. 2014). This approach overall yields an asymptotically unbiased and normal estimator that can be fit easily on a researcher's personal computer and that does not require extensive manual hyperparameter tuning.
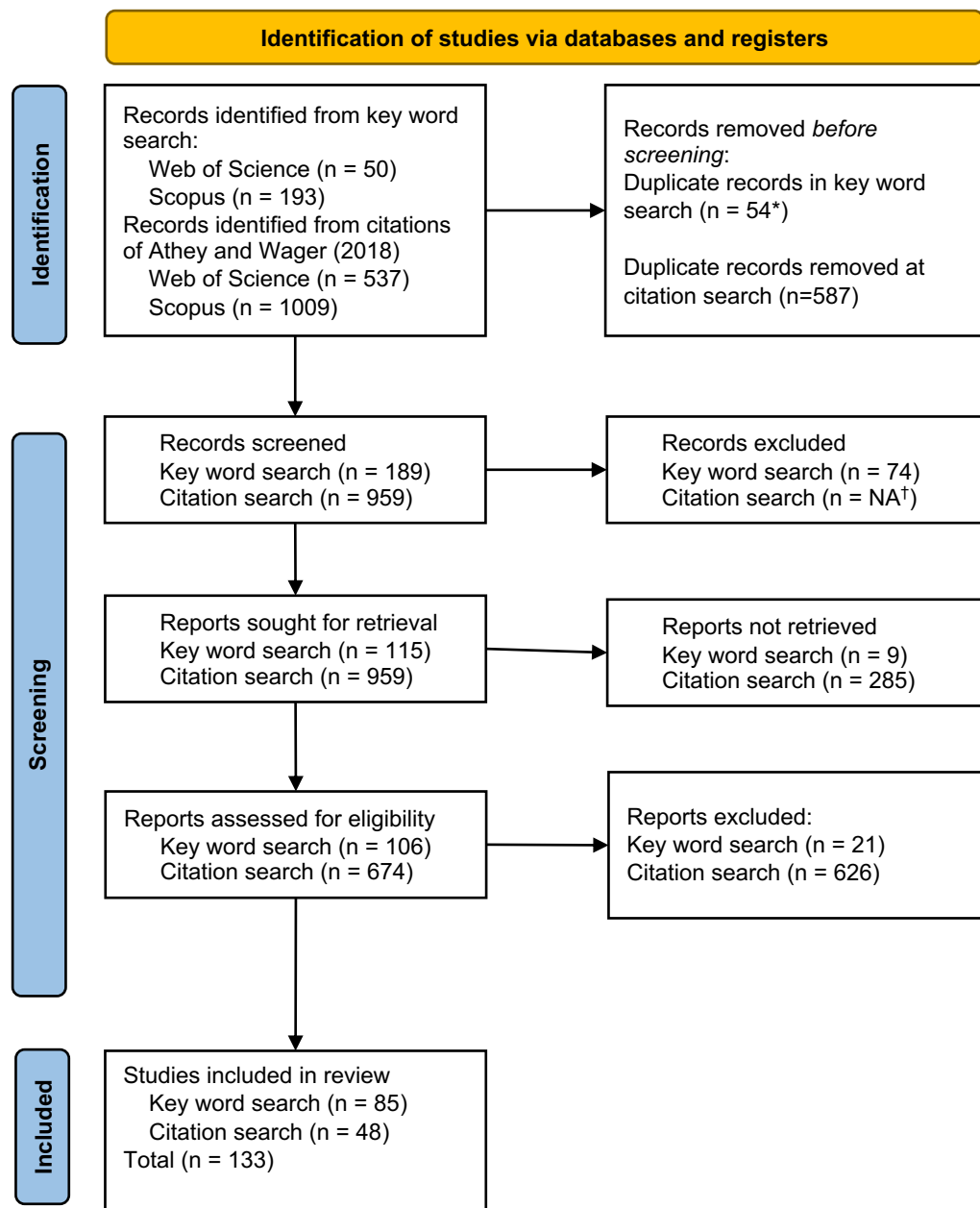
## 3  Study design

The inclusion criteria for this study were simple. The paper should use the causal forest method and it should be primarily an applied paper, not a methods one. The reason for this is simply that methods papers, even when they use non-synthetic data are fundamentally a different kind of paper with different aims, different methodologies, and different findings they want to communicate. For example many were primarily interested in comparing the causal forest against other methods and so emphasised the goodness-of-fit of the model rather than the actual insights that could be extracted.

The first step was searching Scopus and Web of Science using a keyword search for "causal forest". This yielded 197 unique articles. These were then screened based on their abstracts which yielded 107 articles that were then read in full. 85 of these were found to meet the inclusion criteria and were analysed in this study.

A broader search of any paper citing Wager & Athey (2018) on Scopus or Web of Science was then conducted. As any papers here that were not duplicates from the key word search would by definition not refer to the causal forest in their abstract, reading abstracts was not a feasible screening strategy. For this reason I attempted to retrieve all of these articles but rather than reading them in full, I first parsed them with a script. This searched the text for mentions of "causal forest" and provide the line of text around the mention for context so I could quickly assess why the Wager and Athey paper was cited. The vast majority of the papers cut here (keeping in mind these were papers that were not found in the key word search) simply made throw-away comments about Wager & Athey (2018) or included the citation in a larger point about causal machine learning or the performance of random forests without a specific mention of the method. The next most common reason for exclusion was that papers were primarily methodological. 62 of the papers from Scopus possibly met the inclusion criteria and were read in full. This approach was not used for the Web of Science papers as there were only 23 left at this stage (because this search was conducted after the Scopus search meaning any duplicates were assessed as part of the Scopus sample or key word search). For this reason, the Web of Science papers were read in full. Only 40 Scopus papers and eight Web of Science papers that had not already been picked up in the key word search met the inclusion criteria.

The citation search was conducted after the keyword search yielded 75 papers. Directly after the citation search, the keyword search results were updated and a further 10 papers had been published in that time that were not already included in the citation search. Figure 1 provides the full breakdown of this selection process. A full list of the papers selected and notes on their use of the causal forest is available in the web appendix.

**Identification of studies via databases and registers**

| | |
|---|---|
| **Identification** | Records identified from key word search:<br>    Web of Science (n = 50)<br>    Scopus (n = 193)<br>Records identified from citations of Athey and Wager (2018)<br>    Web of Science (n = 537)<br>    Scopus (n = 1009) |

Records removed *before screening*:
Duplicate records in key word search (n = 54*)

Duplicate records removed at citation search (n=587)

| | |
|---|---|
| **Screening** | Records screened<br>Key word search (n = 189)<br>Citation search (n = 959) |

Records excluded
Key word search (n = 74)
Citation search (n = NA†)

Reports sought for retrieval
Key word search (n = 115)
Citation search (n = 959)

Reports not retrieved
Key word search (n = 9)
Citation search (n = 285)

Reports assessed for eligibility
Key word search (n = 106)
Citation search (n = 674)

Reports excluded:
Key word search (n = 21)
Citation search (n = 626)

| | |
|---|---|
| **Included** | Studies included in review<br>    Key word search (n = 85)<br>    Citation search (n = 48)<br>Total (n = 133) |

* Six papers were duplicates when the keyword search was updated (these were already in the citation search. In the original search there were 48 duplicates.
† There was no abstract scan for the citation search papers.

Figure 1: Process for selecting studies into this review. Diagram from template by Page et al. (2021).

To provide context on the literature, I will briefly present some basic exploratory analysis on the dataset. Figure 2 plots the number of publications by field while Figure 4 plots the distribution of papers over time. These categorisations were made manually when reading the papers. It is clear that a few disciplines represent a large part of the causal forest literature, particularly economics and health research which together account for most of the papers in the sample. Use of the method has grown over time. It particularly took off after the grf package and generalised random forest paper in 2019.
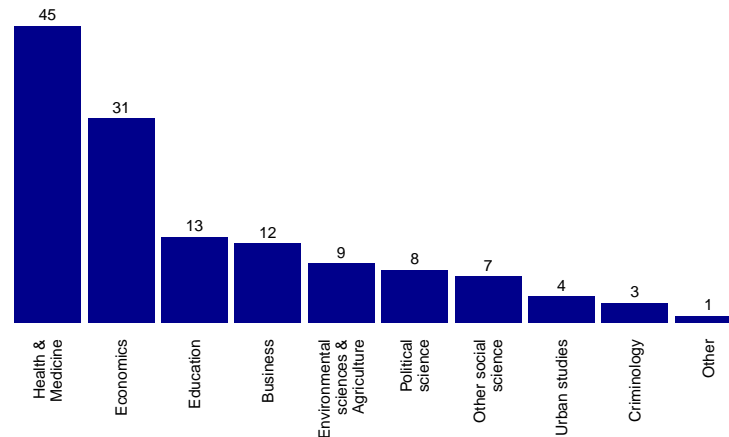
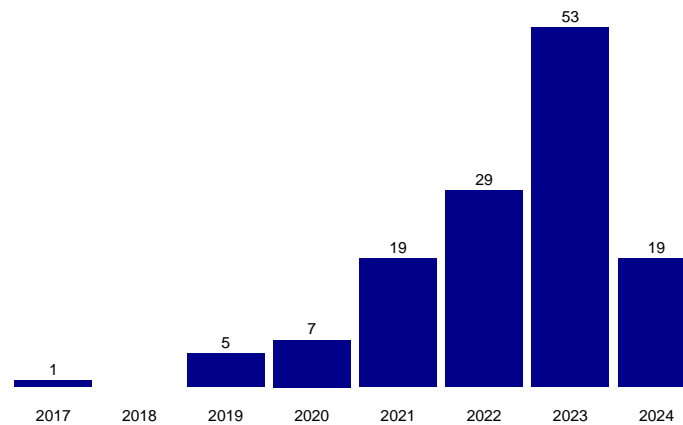Figure 2: Publications in study by their discipline



Figure 3: Publications in study by their discipline

## 4  Results - Identification and estimation approaches

### 4.1  Choice of software

Across the sample a clear majority of papers that detailed the software they used used the grf package by Athey et al. (2019) to fit their causal forest. In fact only seven papers that name the software they used did not use grf. The ubiquity of grf affects the approach to identification, estimation and presentation of results because some processes are easier than others with the tooling it offers and some have already been modelled in tutorials. Table 1 below shows the (non-exclusive) count of the packages used across different papers, i.e. if an approach combines two packages, it is counted once for each package. It is notable that despite the popularity of the EconML Python package in industry (Syrgkanis et al. 2021), it was only explicitly used by one paper in the sample. This is perhaps due to a language divide as only six paper used any Python packages at all (Kristjanpoller et al. 2021, Seitz et al. 2023) with the MCF package being used by three of these (Zhu 2023, Cockx et al. 2023, Hodler et al. 2023). Either because of the familiarity of academics with R or the quality of the grf tooling, it appears R is simply dominant in causal forest modelling. One paper used grf via the MLRtime Stata package.

While it is hard to say, my suspicion is that most of the papers that did not specify an implementation used the grf package as they often cited the Athey et al. (2019) without naming an implementation and they were generally using R where grf is the only popular implementation of a standard causal forest. In some cases with particularly with

| Name | Count |
|---|---|
| grf | 69 |
| mcf (Lechner 2019) | 3 |
| causalTree | 3 |
| EconML | 1 |
| tools4uplift | 1 |
| causalML | 1 |
| Original causal forest implementation (Athey & Wager 2019) | 1 |
| MLRtime | 1 |
| No paged stated | 54 |

Table 1: Count of packages used in the papers studied (some use multiple packges).

novel designs or papers which predate the grf release (e.g., Baum et al. 2017), it may be that the researchers simply implemented their own version or took an earlier implementation (like that in Athey's causalTree package[3]).

## 4.2 Identification strategies

Most papers in the study used either experimental (54) or control on observables (72) designs. However there were a number of papers that used interesting variations on these identification strategies.

### 4.2.1 Controlling on observables

It is interesting to see control-on-observables designs used so commonly, particularly in fields (like economics) where they are not generally favoured. Is there something about using a doubly robust estimator paired with powerful machine learning nuisance functions that allows us to identify causal effects without some kind of randomisation in a way we could not with simpler methods? Clearly many authors think there is and they are eager to use the nuisance models to identify effects. Whether they are right or wrong is unanswerable for specific applied questions due to the Fundamental Problem of Causal Inference (Holland 1986). More work needs to be done to understand the credibility of such designs. Certainly wider use of robustness checks and in particular placebo tests would be useful.

Some of the papers pursue unusual routes to identifying observational effects; these often seem redundant. For example Inoue et al. (2023) calculate propensity scores with logistic regression before matching observations and then running a causal forest analysis which includes a propensity score weighting step with different scores calculated by a presumably more powerful generalised random forest model. At best this is redundant and at worst it may lead to problems estimating standard errors using jackknifing as the same limitations that apply to bootstrapping likely also apply here (Abadie & Imbens 2006).

### 4.2.2 Experiments

It is notable that the causal forest from the grf package is being used so widely with experimental data. While the use of the causal forest on experimental data has been discussed in relation to other methods (the original causal forest (Wager & Athey 2018), the model-based forest (Dandl et al. 2022)), it has not been a focus of methodological discussion when using a forest with local centering. For most of the papers in this review, it is not clear whether oracle propensities were used, or whether propensities were estimated as if the data were observational. It is likely though that the latter is more the case as this is the default specification for the causal forest. This approach is likely the superior one as per Hirano et al. (2003), estimation of propensity scores with a nonparametric model for inverse probability weighting actually provides a more efficient estimator of the average treatment effect than using oracle probabilities. Analogously, we might expect this to apply to CATEs that are making the same AIPW estimation, just within a kernel bandwidth.

Three natural experiments or randomised controlled trials had data that was imperfectly randomised which provided some level of exogeneity to treatment but did not provide full identification (Brock & De Haas 2023, Allen et al. 2022, Habel et al. 2023, Hodula et al. 2023). Adjusting with controls was still necessary to get identification but the partial randomisation provided a level of confidence in identification. One paper used a data fusion design to combine experimental and observational data to improve identification (Kluger et al. 2022).

---

[3]https://github.com/susanathey/causalTree

### 4.2.3  Quasi-experimental

Given how prevalent quasi-experimental approaches are in economics, and how many of these papers are from that discipline, it is curious how few quasi-experimental designs were used. Of course there is no in-built support for many common quasi-experimental methods in grf, but there is a built in instrumental variable method — the instrumental forest (Athey et al. 2019, Tibshirani et al. 2021). Interestingly only one paper (Brooks et al. 2022) used this method.

Six papers used some kind of difference-in-differences design (Xue et al. 2023, Cui & Davis 2022, Wang et al. 2023, Wang 2022, Miao et al. 2023, Guo et al. 2021). Notably, many others used a standard difference-in-differences with a linear fixed effects model design for their analysis of ATE but used control-on-observables identification with a causal forest to explore heterogeneity.

Generally where papers used difference-in-differences for treatment effect heterogeneity they used the first difference between the last pre-treatment period and the first treatment period as an outcome and then modelled the relationship with a standard causal forest. Miao et al. (2023) term this design the First Differences Causal Forest. This approach helps to improve the credibility of the causal forest, at least in some disciplines where difference-in-differences is seen as a more credible design than simply controlling for observables. However, in all the papers that propose this method, the approach is not rigorously laid out. In particular, while there is testing for parallel trends in the main effects with a linear model, there is no attempt made in the papers to make sure trends are parallel for each CATE estimate. This is a problem because with machine learning of heterogeneous effects we no longer have just the one estimate we need to make this assumption for but a whole distribution of estimates.

Presumably for inference over the conditional average treatment effect on the treated (CATT) we would have to make a version of a parallel trends assumption for within each kernel bandwidth. More formally, where we target outcome $\Delta(W, X) \equiv Y(W, X)_t - Y(W, X)_{t-1}$ for the estimation of $\tau(x)$ we must assume that

$$\forall x, \forall X_i \in X, \mathbb{E}[\Delta_t(1, X_i) - \Delta_t(0, X_i)|X_i, W_i = 1] - \mathbb{E}[\Delta_t(1, X_i) - \Delta_t(0, X_i)|X_i, W_i = 0] = 0.$$

That is that in expectation potential outcomes are parallel for every data-point in the neighbourhood of $x$. This is obviously a much stronger assumption and one that is harder to test, but this difficulty does not mean we can afford to just ignore the assumption. A proper treatment of difference-in-differences and the causal forest is out of scope of this paper, but these unresolved problems should make applied researchers wary of naively using differenced outcomes as the target variable in their causal forest.

One more unorthodox approach inspired by difference in differences is the one proposed by Turjeman & Feinberg (2024), the Temporal Causal Forest. They have a sample which was all treated at one time with a data-breach of website users, but the sample joined the website at different times, so they can counterfactually model normal fall-off of users by looking at the change in user activity before the data breach versus after the breach happens. Essentially it is a difference-in-differences across time. This is an interesting approach at least for their specific research question and dataset. However, as with the more conventional first differences approaches it still requires an untested parallel trends assumption that may be violated for some neighbourhoods when calculating CATEs.

### 4.3  Choice of variables

In addition it is worth speaking about the way in which covariates are used. There is little distinction in the papers between the use of variables in the nuisance models and the use of those variables in the final stage model. Only one paper specifies a specific set of heterogeneity model variables that is different from the nuisance function variables. Several others specify different variables in a data-driven way, i.e. they use the Basu Technique to generate a subset of the nuisance function variables to use in their heterogeneity models. For the most part the same variables are generally being used for all three. This means there is little consideration of post-treatment moderators (Imai et al. 2010), perhaps because most of the data used is cross-sectional.

The number of variables used is also generally small with the median number being 17.[4] Figure 4 shows the distribution of covariate counts. While some designs use many variables, it seems obvious that for the most part, the ability for the causal forest to excel with very high-dimensional datasets (Gong et al. 2021) is not being used. This may be because they lack access to sufficiently high-dimensional data or because they do not see a large number of variables as being useful; they might have an existing theoretical framework that narrows down the variable set. In the final model, it may

---

[4]As an aside, the approach to sampling variables for use in the individual trees by default draws $\sqrt{p} + 20$ variables in the grf causal forest. This means for most papers in our sample, the causal forest is not so much a kind of modified random forest, but instead a modified bagged trees ensemble where different trees are fit on different subsets of the data but the full set of variables (Breiman 2001, 1996). This is not necessarily a problem, but it is worth keeping in mind that the decorrelation of trees via random variable selection is the key insight of the random forest and when this parameter is left at its default value, we may be hurting the out-of-sample fit of the ensemble.

also be that the problem of dealing extracting insights from CATEs calculated with a large number of variables is more trouble than it is worth.
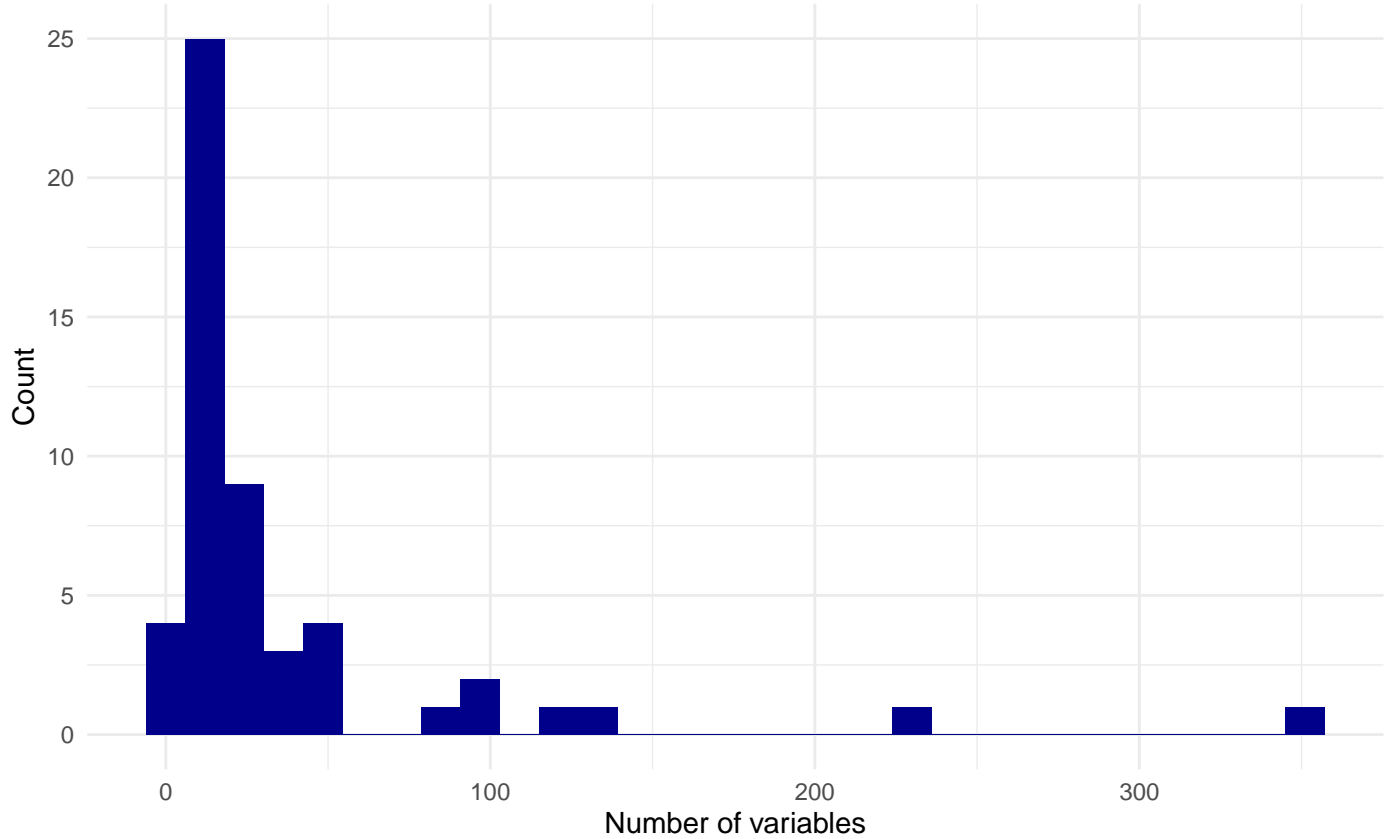


Figure 4: Number of $X$ variables used in each study for which the covariates are given (if separate variables are used across the three models e.g. if Basu Technique is applied, we take the highest count)

Related to this is the use of the Basu Technique for reducing the number of variables used to fit a causal forest. This approach aims to decrease the amount of splitting that is just due to noise to provide greater clarity in results. 14 papers employ this technique. Interestingly, they are not the studies with larger sets of covariates. In fact the papers using the Basu Technique had a median of 11 original variables compared to 18 for those that did not. The motivation here seems to be from Athey & Wager (2019) who use this technique in order to get a clearer heterogeneity model that offers clearer findings to researchers. There seems to be no clear guide as to when it is appropriate or not appropriate to use the Basu Technique and no methodological work on the improvements it may lead to in modelling either whether as just a regulariser or to make it easier to extract insights from a model.

### 4.4 Model specification choices

Interestingly, because most approaches use the stock grf, no papers rely on nuisance models that are not a random forest (though some use other kinds of models for matching before applying the causal forest and one uses a locally linear forest for propensity score estimation). This is not a bad choice, particularly given the small size of the data in most studies (both in width and length) and that grf is designed to easily fit these nuisance models under the hood. However, it is certainly a choice worth noting.

There is a little more diversity in the way random forests are tuned. A number of papers papers used their own manual cross-validation of hyperparameter values rather than relying on the grf tuning forest (Guo et al. 2021, Xu et al. 2021, Seitz et al. 2023, Verstraete et al. 2023, Inoue et al. 2023). It is unclear from the text of those papers whether this decision was made due to a mistrust of the tuning forest or whether the authors simply did not know about this functionality and relied on a more traditional ML fitting process instead.

| Method | Count |
|---|---|
| Graph across key variables | 58 |
| Variable importance | 35 |
| Graph individual CATEs in a histogram | 33 |
| Graph along quantiles | 26 |
| Single tree | 15 |
| Best Linear Projection | 12 |
| Derive policy | 11 |
| Rank individual CATEs | 9 |
| SHAP | 6 |
| Fit linear model on causal forest predictions | 7 |
| Graph across geography | 5 |
| Fit quantiles for individual CATEs then look at covariate distribution within quantile | 5 |
| Partial dependence plot | 3 |
| K-means clustering | 2 |

Table 2: Number of times a particular technique was used in the studied papers to communicate CATE results

The number of trees used varied widely from 300 up to 100,000 with a median of 4000 (for those that reported a number). 69 papers do not specify a number of trees used which suggests the default hyperparameter value in grf of 2000 was used. The number of trees in the random forest is an unusual hyperparameter in that unlike many others in machine learning, the number of trees should never cause overfitting — there is no trade-off (Breiman 2001). In expectation, more trees will be better than fewer and the main reason for choosing a smaller number is simply runtime. It is therefore better to err on the side of more trees than fewer. It is after all relatively inexpensive to fit more trees and papers should simply grow the ensemble until there is essentially no excess error (the estimate of Monte Carlo error from having a finite-sized ensemble).

### 4.5   Sample splitting

In some cases authors used the standard machine learning approach of splitting their sample into a training dataset and a dataset for estimation (analogous to a test dataset) (Seitz et al. 2023, Kianmehr et al. 2022, Serra-Garcia & Szech 2023). Because the causal forest always fits out-of-sample — that is it only predicts for data-points based on trees that were not fit on that data-point — this is redundant (Athey et al. 2019). In fact, in expectation, there will be no effect apart from shrinking sample size. The only reason this might be worth doing is if there is some sort of policy being derived from the forest for which one is estimating a value as estimating value in sample will probably overestimate due to over-fitting. A straight-forward example of where this is useful is Jakobsen et al. (2023) or Coleman et al. (2023), while Osawa et al. (2023) uses a more unusual approach without i.i.d. data between samples (validating on experimental data) which nonetheless is an interesting approach which could be useful for future policy learning work.

## 5   Results - Presentation of results

One serious challenge facing the causal forest is that even if there are useful patterns in results, it is not obvious that researchers will be able to identify and communicate these patterns ((Rehill & Biddle 2023). The reason heterogeneity can be challenging to understand is that the results are essentially a high-dimensional distribution of effects that must be simplified in some way to allow for human comprehension. This means that generally a causal forest needs to be explained with the use of some kind of additional model. This could be a simpler model fit on the same locally centered data or doubly robust scores as the causal forest like a best linear projection or a best causal tree (akin to interpretable AI in predictive modelling (Rudin 2019)). Alternatively, one can look at the predictions made by the model and try to explain them like through binned CATEs. Finally, one can look at variable importance which tries to reveal relationships by analysing the structure of the tree itself. The focus of this section is not on all the possible approaches but instead on discussing the practices in this sample where there are a few that are used widely.

There are three common ways of visualising results, most papers use one or more of these. Graphing of the overall distribution of individual treatment effects (individual CATEs were graphed in a histogram in 33 papers and a ranked plot of treatment effects was used to a similar end in 9), graphing of conditional treatments effects often with binning for continuous variables or a kernel regression (used in 58 papers) and showing variable importance (used in 35 papers). Table 7 shows how common these and other approaches are. Several other approaches were used only once. These are presented in an expanded version of this table in Appendix A.

## 5.1 Graphing individual treatment effects

A univariate plot graphing out the distribution of the individual CATEs can give a good sense of how much heterogeneity is in the data. While this is not always the case, it can be a good basic check on data and in some cases as in Guo et al. (2021) it is very expressive, laying out a pattern of treatment effects that is very useful in answering the research question. Graphing of ranked effects is a variation on this which shows the same information. While arguably the point estimates are a little harder to interpret here, such a plot does allow for visualisation of confidence intervals on the same plot ala Fukai et al. (2021). Another variation on this is the graphing of CATEs across quantiles of CATE estimates. This approach may violate the warning not to subset CATE estimates by outcomes from the grf documentation (Tibshirani et al. 2021), however it is widely done. One variation on this approach is that from Davis & Heller (2020) where the top quantile is compared to the other quantiles in order to get a heuristic of the amount of heterogeneity in a distribution (Mesplé-Somps & Nilsson 2023, Davis & Heller 2020). Two papers actually statistically tested the difference in quantile effects with a Wald test (Serra-Garcia & Szech 2023) or a Kruskai-Wallis test (Yu et al. 2024).

## 5.2 Variable importance

Variable importance is a simple metric but a useful one, and one that is easy to explain to a reader. It takes a depth weighted count of the splits on each variable for all trees in the forest (by default halving the weight of each layer compared to that above it before stopping counting at depth 4) and the normalises there to sum to 1 (Tibshirani et al. 2021). Many papers use it to drive their search for treatment effect heterogeneity where they lack a good theoretical framework for which variables to use. It can also help to check the robustness of the theoretical framework in the particular dataset being studied. That said it is certainly just a heuristic and is not as robust as techniques for variable importance in predictive modelling. These predictive metrics can better partial out the effects of variables on estimates because they have access to ground truth (Strobl et al. 2008).[5] It becomes less useful in cases where there are many variables because highly correlated features (Bénard & Josse 2023). This is because correlated features effectively split importance between them. Newer approaches propose more robust variable importance metrics, but at the cost of longer runtime and not having convenient implementations already available (Bénard & Josse 2023, Hines et al. 2023). Only three papers use an approach other than that from grf for variable importance (Davis & Heller 2020, Venkatasubramaniam et al. 2023) seemingly implementing their own method in both cases. Dylong & Uebelmesser (2024) use their own approach that is slightly different, they look at the difference in estimates for each quartile of a covariate which they call relevance.

## 5.3 Estimating CATEs across variable values

Graphing effects across variables can be done in a number of ways. There are many ways to choose which variables to estimate across (based on variable importance or based on theory) and they can be shown in many different ways (e.g. boxplots for binned or binary variables, scatter plots with a kernel regression line fit on predictions for continuous variables, or simply reporting subgroup estimates in text). While the standard approach was simply to graph CATEs as a function of a variable, there were many variations on this approach. Some papers used partial dependency plots brought across from the explainable AI (XAI) literature (Seitz et al. 2023, Xu et al. 2023, Esterzon et al. 2023) which if valid offer a better view of the conditional effect of a given variable. In practice this may not be appropriate given the nature of the causal forest model not itself making predictions but instead fitting a kernel for use in an AIPW estimator (Sverdrup 2021). A variation on this is the use of SHAP, a similar XAI method that is extremely expressive but again suffers from the same concerns about ignoring the model's purpose as an adaptive kernel. This is not to say that neither approach is appropriate, but rather there is no good theoretical work so far that establishes these XAI approaches are appropriate (Sverdrup 2021) (though it is worth noting that SHAP is built in to the EconML Python package (Oprescu et al. 2019) so is likely being widely used in industry applications).

An approach appropriate for some research questions was to graph CATEs for different areas on a map (Kluger et al. 2022, Yin et al. 2023, Deines et al. 2023). This seems to be quite a clear and expressive way to show treatment effect variation.

When estimating CATEs across pre-specified variables is taken to its extreme, the causal forest becomes superfluous (e.g., Esposti 2024). Essentially this is just a series of doubly robust ATE calculations for different subsamples for example different age brackets. This is easy without the causal forest. At most grf faciliates the easy estimation of doubly robust scores.

---

[5]It is worth noting that the mcf Python package which implements the modified causal forest (Lechner 2019) does implement a permutation variable importance. No papers here use this approach.

## 5.4  Deriving a single tree

15 papers derived a single expressive tree from the causal forest which they used in some way to communicate results. This is potentially a useful method which could capture interaction effects in CATEs in a way other ways of exploring CATEs do not. However, the usefulness of the tree depends on how it was derived. There is no clear best practice to derive a tree but there are many possible options whether a tree is selected based on some criterion from the forest (Wager 2018) or simply fit on the whole dataset separately. Interestingly, no paper seems to use the R-Loss based selection method recommended by Wager (2018). Instead it is common for a single tree to be fit independently or selected based on a tree-distance metric (Zhou, Zhan, Chen, Lin, Zhang, Zheng, Wang, Huang, Xu, Liao, Tian & Zhuang 2023, Zhou, Jian, Chen, Liu, Zhang, Fu, Li, Liang, Tian & Wu 2023). The aim of the tree distance metric is to select a central tree in an ensemble (Banerjee et al. 2012). This is certainly a valid approach to the problem of selecting a single tree where we lack good ground-truth data. It is hard to say which is better, this approach or minimising R-Loss (Wager 2018). While we obviously want a tree that performs well and so might reduce R-Loss, it is also a very noisy objective so whether this explicit optimisation actually gives a more 'representative' tree is not entirely clear. This question depends on what we actually want from a single tree (representativeness or fit) and how much we trust the intuition of Wager over these three papers. Finally it is possible to simply take a tree at random as representative (this seems to be what Amann & Rzepka (2023) does).

When fitting a separate tree it is possible to fit a tree separately on data as many papers do (e.g., Giannarakis et al. 2022, Rana & Miller 2019). However, it is also possible to distill a tree from the causal forest which may improve its stability and fit (Frosst & Hinton 2017). In fact the properties of the causal forest which give it good properties in an estimator (based on the insights of the semi-parametric estimation literature) may make it particularly amenable to distillation too (Dao et al. 2021).

## 5.5  Studying a policy derived from CATE results

Several papers also look at explicit targeting of treatments whether this is based on a 'black box' policy from the causal forest or a second interpretable policy learning model fit on the forest. In the first case, targeting is based directly on the CATE estimate even though that estimate is less interpretable while in the latter case the allocation follows a set of human interpretable rules (Athey & Wager 2020, Zhou, Athey & Wager 2023). Papers commonly followed the former approach, which is interesting given the policytree package in R interfaces well with grf; sharing many of the same authors (Sverdrup et al. 2020). The only paper to explicitly fit a policy tree was Cockx et al. (2023) which also looked at the effects of the 'black box' policy as well. Perhaps the policy tree is a poor fit for empirical researchers and may be more used in other types of work where direct prescriptions are more relevant.

Analysis of the black-box policy was more common with 11 papers doing some variation on this. This kind of analysis is useful particular where research is designed to directly inform practice (many of these papers were around changing treatment regimes in medicine). It gives a sense of the potential gains of incorporating the knowledge from the study. The use of a held-out test sample ala Osawa et al. (2023) makes sense here in a way it does not for other applications of the causal forest because the policy is in effect set based on the CATE estimates. We would expect the estimate of the policy value to overfit in-sample even though the estimates themselves are unbiased as causal effects. The way this learning is visualised varies too, but curves comparing outcomes under the 'treatment' of policy targeting and the standard regime like in Osawa et al. (2023) or the Qini curve used in Seitz et al. (2023) are good examples of how this information can be communicated. There are likely to be more insights on this that can be drawn from uplift modelling which tackles similar problems (Zhang, Li & Liu 2022).

Another version of this is Jakobsen et al. (2023) which uses a Rank-Weighted Average Treatment Effect (RATE) analysis (Yadlowsky et al. 2023). This is part of a fairly extensive effort to uncover treatment effect heterogeneity. In this case, the treatment is not really manipulable (COVID infection) so the direct utility of a policy is low, but the authors seem to find it useful for presenting their forest results.

## 5.6  Use of linear models to explore heterogeneity

12 papers relied on fitting a best linear projection (BLP) (Semenova & Chernozhukov 2021) to explain heterogeneous effects (Wang 2022, Zhang, Akinci & Qian 2022, Elamin 2023, Huber et al. 2022, Esterzon et al. 2023, Xu et al. 2021, Elek & Bíró 2021). A best linear projection regresses doubly robust scores onto a set of variables in a linear model. In grf, standard errors are calculated via resampling to adjust for having a dependent variable that is itself estimated from a model. While there are obviously limitations here (it can only measure linear effects where the rest of the approaches have been free of functional form limitations), it could be a useful way to hypothesis test results to get an output that is more easily interpretable to an audience more familiar with parametric regression outputs. Generally the BLP was used to projects the effects onto the same variables that were used to fit the causal forest in the first place.

However, there were a number of studies that did not project effects using the approach built-in to the grf package, but instead fit a linear model on CATE estimates directly directly as proposed by Nilsson et al. (2019) (Habel et al. 2023, Leite et al. 2022, Guo et al. 2021, Chen et al. 2020, Zhu 2022, Amann & Rzepka 2023). This is a bad idea for two main reasons. The first is that fitting a linear model on predictions rather than doubly robust scores means the model lacks double-robustness and is more likely to be biased. It is essentially a conditional average of a series of conditional averages (the predictions) the offer no guarantees about the properties of this conditional mean. The second problem is that it seems like in all these cases, the standard errors are calculated analytically as is standard for linear regression rather than with resampling. This does not meet the assumptions for valid analytical standard errors because doubly robust scores are outputs from another model (Wooldridge 2013). In order to get valid standard errors in this approach with looser assumptions, we need to estimate errors through resampling across all models like the grf BLP implementation does rather than analytically. Papers that fit their own models on predictions do not seem to do this. The result is a biased point estimate and likely over-confident standard errors. When fitting a model researchers should always use the procedure built into grf rather than attempting their own.

### 5.7 Using a clustering model

Finally, two papers clustered treatment effects with an unsupervised learning algorithm (K-means or K-means++) (Miao et al. 2023, Cockx et al. 2023). This could be a promising approach as it accounts for covariation between variables in a way that clustering based on a single tree does not. Hopefully this approach will be developed further in future papers such that its value across a wider domain of problems can be assessed.

This approach is related in a way to the approach that partitions treatment effects by quantiles and then calculates the sample averages of different characteristics for each quantile. This was often done in combination with calculating treatment effect quantiles.

## 6 Discussion

The causal forest literature is currently largely defined by an implicit best practice set out by the creators of the method, either in the way they have used the method in their own work (such as the use of the Basu Technique for dimensionality reduction) or the way in which they wrote the grf package (such as in the use of standard hyperparameter settings). This is not a bad thing at all — most of the criticisms I have made in this paper have been of papers stepping outside this practice — but it does mean there is for better or worse a methodological conservatism, at least in the applied space. As a side note, there are many methods papers that propose novel modifications to the causal forest, however these have yet to be adopted in the applied literature so far.

While this paper has been skeptical about some of the attempts authors have made to innovate on causal forest methods outside of the standard grf best practice, sensible innovations that are well-justified and build on a strong knowledge of the existing methods could offer substantial benefits for creating more flexible and more transparent methods in the future. Certainly on the former, new approaches to identification with the causal forest could dramatically improve the credibility of the methods in fields such as applied economics where control-on-observables is generally not viewed as a credible identification strategy and where experimental data is often scarce. For example, the work on difference-in-differences with the causal forest or the use of causal forest with regression discontinuity are areas that could be explored much more than they have been. In addition new approaches to estimation can open up new research questions that simply cannot be answered with the standard grf approach. For example, Cockx et al. (2023) studying policies derived from the causal forest is a good way to understand how the forest can add value in shaping actual policy interventions. Osawa et al. (2023) testing the policy on a second validation sample of experimental data is going to be useful in some contexts to test whether the value of the policy estimated in the policy learning step generalises.

On transparency, methods for understanding effects are still very much lacking, posing problems both for researchers who are trying to extract insights from their analysis and also, for those whose lives might be affected by decisions made on the basis of causal forest analysis (Rehill & Biddle 2023). One promising area here is in more robust methods for calculating variable importance (Bénard & Josse 2023, Hines et al. 2023), more akin to the way variable importance is calculated in predictive models than the heuristic implemented by Tibshirani et al. (2021). Another approach that could be valuable is to use SHAP values as the EconML package allows, however the theoretical work has not been done yet to establish this as a valid approach.

The papers in the study are all peer reviewed, applied research papers written largely by academic researchers, this has its benefit but it is also a limitation. There is a whole other area of activity where the causal forest and similar methods are being applied to solve problems in industry, often called uplift modelling problems in this context. It is also worth noting that by only looking at peer reviewed papers, in a relatively new and fast-moving field this paper may

have missed important advances that have so far only been made in papers published as preprints. Finally there are many similar causal machine learning methods where estimates are made differently but which have some of the same challenges as the causal forest, for example Bayesian Causal Forest and Bayesian Additive Regression Trees (Hahn et al. 2019) or causal meta-learners (Künzel et al. 2019). Chernozhukov et al. (2023) already has some interesting ideas (aside from the BLP) that could be more widely used with causal forest CATE distributions. There may be lessons to be learnt from this broader literature as well.

One potentially troubling factor in this sample is that many papers have quite low sample sizes. While the random forest performs well with smaller datasets compared to other machine learning approaches (Athey et al. 2019, Breiman 2001), it is still not clear that data-driven HTE analysis is appropriate in these settings. For starters there is real understanding of statistical power for kernel bandwidths in the causal forest. CATE estimates that use a small part of an already small sample are likely to be prone to type II error. Worse, the way the causal forest is often used does not just risk type II error in the way a parametric model might. Because the forest is effectively generating hypotheses, helping a researcher to understand the kinds of things they might want to test, it can be even more problematic. Many of the commonly used outputs like graphs along key variables and variable importance give no confidence intervals and so a researcher's understanding could be warped without their knowing. If they then go on to run tests based on this knowledge for example in picking subgroups for CATE estimation or in choosing projection variables for the BLP they have tailored their tests to potentially spurious effects in the data.

There is one last point worth making which relates to the task of assembling a review but also broader reproduceability and interpretability of research. Many researchers did not specify basic information about their use of the causal forest. What packages they used, the number of trees they used and the like. Unlike some more standard methods, the causal forest has several possible implementations all called the causal forest, it has hyperparameters that need to be picked and it is nondeterministic. There are a lot of choices that need to be made when modelling and a lot of information that should therefore be given in papers. Providing code is especially useful here. While it is always useful for researchers to make their code publicly accessible, the process of compiling this review has shown just how important that can be when using a causal forest. While some authors wrote detailed (and much appreciated) appendices on their specification, access to code is invaluable for understanding the exact approach a paper took. In many cases, the precise specification of a model was unclear from simply reading the published material.

# 7    Conclusion

The causal forest is a promising method that is being used more and more over time. This paper has sought to understand how this still nascent method is being used in order to reflect on and maybe even shape the emerging best-practice. The review shows the causal forest is largely used for research that achieves causal identification with randomisation or by controlling on observables, there is little in the way of quasi-experimental work using the method. Furthermore, the vast majority of work uses the grf package in R. There are a reasonable number of commonly used methods for presenting results and and a much larger set of methods that are used only rarely but which show — for better and for worse — that there is still some innovation happening in figuring out how to communicate insights from a CATE distribution.

No doubt the method will continue to evolve, but there appear to be a few key areas ripe for further work. Control-on-observables designs should do more to test for identification (for example assessing propensity score overlap which only a minority of papers did). Quasi-experimental variants of the causal forest may help to make identification more credible in some research. There is still a lot of work that could be done to improve the presentation of results, particularly useful would be the development of theoretically valid XAI approaches that account for the unusual process of getting causal forest estimates. Finally, it will be interesting to see whether larger datasets both in width and length begin to be used as researchers begin to embrace models that can better cope with these (and as they begin to get access to bigger datasets (Connelly et al. 2016)). The promise of better being able to understand treatment heterogeneity through causal machine learning is an exciting one, and many of the papers reviewed here show this promise beginning to be realised.

# References

Abadie, A. & Imbens, G. W. (2006), 'Large Sample Properties of Matching Estimators for Average Treatment Effects', *Econometrica* **74**(1), 235–267.
   **URL:** *http://doi.wiley.com/10.1111/j.1468-0262.2006.00655.x*

Allen, J., Carbo-Valverde, S., Chakravorti, S., Rodriguez-Fernandez, F. & Pinar Ardic, O. (2022), 'Assessing incentives to increase digital payment acceptance and usage: A machine learning approach', *PLOS ONE* **17**(11), e0276203.
   **URL:** *https://dx.plos.org/10.1371/journal.pone.0276203*

Amann, E. & Rzepka, S. (2023), 'The effect of goal-setting prompts in a blended learning environment—evidence from a field experiment', *Economics of Education Review* **92**, 102331.
**URL:** *https://linkinghub.elsevier.com/retrieve/pii/S0272775722001042*

Athey, S. & Imbens, G. (2016), 'Recursive partitioning for heterogeneous causal effects', *Proceedings of the National Academy of Sciences* **113**(27), 7353–7360. Publisher: National Acad Sciences.

Athey, S., Tibshirani, J. & Wager, S. (2019), 'Generalized random forests', *The Annals of Statistics* **47**(2), 1148–1178.
**URL:** *https://doi.org/10.1214/18-AOS1709*

Athey, S. & Wager, S. (2019), 'Estimating Treatment Effects with Causal Forests: An Application', *Observational Studies* **5**(2), 37–51.
**URL:** *https://muse.jhu.edu/article/793356*

Athey, S. & Wager, S. (2020), 'Policy Learning with Observational Data'. Number: arXiv:1702.02896 arXiv:1702.02896 [cs, econ, math, stat].
**URL:** *http://arxiv.org/abs/1702.02896*

Banerjee, M., Ding, Y. & Noone, A.-M. (2012), 'Identifying representative trees from ensembles', *Statistics in Medicine* **31**(15), 1601–1616.
**URL:** *https://onlinelibrary.wiley.com/doi/10.1002/sim.4492*

Basu, S., Kumbier, K., Brown, J. B. & Yu, B. (2018), 'Iterative random forests to discover predictive and stable high-order interactions', *Proceedings of the National Academy of Sciences* **115**(8), 1943.
**URL:** *http://www.pnas.org/content/115/8/1943.abstract*

Baum, A., Scarpa, J., Bruzelius, E., Tamler, R., Basu, S. & Faghmous, J. (2017), 'Targeting weight loss interventions to reduce cardiovascular complications of type 2 diabetes: a machine learning-based post-hoc analysis of heterogeneous treatment effects in the Look AHEAD trial', *The Lancet Diabetes & Endocrinology* **5**(10), 808–815.
**URL:** *https://linkinghub.elsevier.com/retrieve/pii/S2213858717301766*

Bittencourt, W. R. & Albuquerque, P. H. M. (2020), 'Evaluating company bankruptcies using causal forests', *Revista Contabilidade & Finanças* **31**(84), 542–559.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**(2), 123–140.
**URL:** *https://doi.org/10.1007/BF00058655*

Breiman, L. (2001), 'Random Forests', *Machine Learning* **45**(1), 5–32.
**URL:** *https://doi.org/10.1023/A:1010933404324*

Brock, J. M. & De Haas, R. (2023), 'Discriminatory Lending: Evidence from Bankers in the Lab', *American Economic Journal: Applied Economics* **15**(2), 31–68.
**URL:** *https://pubs.aeaweb.org/doi/10.1257/app.20210180*

Brooks, J. M., Chapman, C. G., Floyd, S. B., Chen, B. K., Thigpen, C. A. & Kissenberth, M. (2022), 'Assessing the ability of an instrumental variable causal forest algorithm to personalize treatment evidence using observational data: the case of early surgery for shoulder fracture', *BMC Medical Research Methodology* **22**(1), 190.
**URL:** *https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01663-0*

Bénard, C. & Josse, J. (2023), 'Variable importance for causal forests: breaking down the heterogeneity of treatment effects'. arXiv:2308.03369 [stat].
**URL:** *http://arxiv.org/abs/2308.03369*

Chen, Y., Lee, J.-Y., Sridhar, S. H., Mittal, V., McCallister, K. & Singal, A. G. (2020), 'Improving Cancer Outreach Effectiveness Through Targeting and Economic Assessments: Insights from a Randomized Field Experiment', *Journal of Marketing* **84**(3), 1–27.
**URL:** *http://journals.sagepub.com/doi/10.1177/0022242920913025*

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.
**URL:** *https://doi.org/10.1111/ectj.12097*

Chernozhukov, V., Demirer, M., Duflo, E. & Fernández-Val, I. (2023), 'Fisher-Schultz Lecture: Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments, with an Application to Immunization in India'. arXiv:1712.04802 [econ, math, stat].
**URL:** *http://arxiv.org/abs/1712.04802*

Cockx, B., Lechner, M. & Bollens, J. (2023), 'Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium', *Labour Economics* **80**, 102306.
**URL:** *https://linkinghub.elsevier.com/retrieve/pii/S0927537122001968*

Coleman, E. A., Schultz, B., Parker, A. R., Manyindo, J. & Mukuru, E. M. (2023), 'How Communities Benefit from Collaborative Governance: Experimental Evidence in Ugandan Oil and Gas', *Journal of Public Administration Research and Theory* **33**(4), 616–632.
**URL:** *https://academic.oup.com/jpart/article/33/4/616/6927134*

Connelly, R., Playford, C. J., Gayle, V. & Dibben, C. (2016), 'The role of administrative data in the big data revolution in social science research', *Social Science Research* **59**, 1–12.
**URL:** *https://linkinghub.elsevier.com/retrieve/pii/S0049089X1630206X*

Cui, Y. & Davis, A. M. (2022), 'Tax-Induced Inequalities in the Sharing Economy', *Management Science* **68**(10), 7202–7220.
**URL:** *https://pubsonline.informs.org/doi/10.1287/mnsc.2021.4277*

Dandl, S., Hothorn, T., Seibold, H., Sverdrup, E., Wager, S. & Zeileis, A. (2022), 'What Makes Forest-Based Heterogeneous Treatment Effect Estimators Work?'. Publisher: arXiv Version Number: 1.
**URL:** *https://arxiv.org/abs/2206.10323*

Dao, T., Kamath, G. M., Syrgkanis, V. & Mackey, L. (2021), Knowledge Distillation as Semiparametric Inference, *in* 'International Conference on Learning Representations'.

Davis, J. M. & Heller, S. B. (2020), 'Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs', *The Review of Economics and Statistics* **102**(4), 664–677.
**URL:** *https://direct.mit.edu/rest/article/102/4/664-677/96778*

Deines, J. M., Guan, K., Lopez, B., Zhou, Q., White, C. S., Wang, S. & Lobell, D. B. (2023), 'Recent cover crop adoption is associated with small maize and soybean yield losses in the United States', *Global Change Biology* **29**(3), 794–807.
**URL:** *https://onlinelibrary.wiley.com/doi/10.1111/gcb.16489*

Dylong, P. & Uebelmesser, S. (2024), 'Biased beliefs about immigration and economic concerns: Evidence from representative experiments', *Journal of Economic Behavior & Organization* **217**, 453–482.
**URL:** *https://linkinghub.elsevier.com/retrieve/pii/S0167268123004250*

Elamin, O. A. (2023), 'Overeducation wage penalty for university graduates: evidence from the MENA region using machine learning techniques', *Middle East Development Journal* **15**(1), 151–188.
**URL:** *https://www.tandfonline.com/doi/full/10.1080/17938120.2023.2167482*

Elek, P. & Bíró, A. (2021), 'Regional differences in diabetes across Europe – regression and causal forest analyses', *Economics & Human Biology* **40**, 100948.
**URL:** *https://linkinghub.elsevier.com/retrieve/pii/S1570677X20302185*

Esposti, R. (2024), 'Non-monetary motivations of the EU agri-environmental policy adoption. A causal forest approach', *Journal of Environmental Management* **352**, 119992.
**URL:** *https://linkinghub.elsevier.com/retrieve/pii/S0301479723027809*

Esterzon, E., Lemmens, A. & Van Den Bergh, B. (2023), 'Enhancing Donor Agency to Improve Charitable Giving: Strategies and Heterogeneity', *Journal of Marketing* **87**(4), 636–655.
**URL:** *http://journals.sagepub.com/doi/10.1177/00222429221148969*

Frosst, N. & Hinton, G. (2017), 'Distilling a Neural Network Into a Soft Decision Tree'. arXiv:1711.09784 [cs, stat].
**URL:** *http://arxiv.org/abs/1711.09784*

Fukai, T., Ichimura, H. & Kawata, K. (2021), 'Describing the impacts of COVID-19 on the labor market in Japan until June 2020', *The Japanese Economic Review* **72**(3), 439–470.
**URL:** *https://link.springer.com/10.1007/s42973-021-00081-z*

Giannarakis, G., Sitokonstantinou, V., Lorilla, R. S. & Kontoes, C. (2022), Towards assessing agricultural land suitability with causal machine learning, *in* '2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)', IEEE, New Orleans, LA, USA, pp. 1441–1451.
**URL:** *https://ieeexplore.ieee.org/document/9856946/*

Gong, X., Hu, M., Basu, M. & Zhao, L. (2021), 'Heterogeneous treatment effect analysis based on machine-learning methodology', *CPT: Pharmacometrics & Systems Pharmacology* **10**(11), 1433–1443.
**URL:** *https://ascpt.onlinelibrary.wiley.com/doi/10.1002/psp4.12715*

Guo, T., Sriram, S. & Manchanda, P. (2021), 'The Effect of Information Disclosure on Industry Payments to Physicians', *Journal of Marketing Research* **58**(1), 115–140.
**URL:** *http://journals.sagepub.com/doi/10.1177/0022243720972106*

Habel, J., Alavi, S. & Heinitz, N. (2023), 'Effective Implementation of Predictive Sales Analytics', *Journal of Marketing Research* p. 002224372211510.
  **URL:** *http://journals.sagepub.com/doi/10.1177/00222437221151039*

Hahn, P. R., Murray, J. S. & Carvalho, C. (2019), 'Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects', *arXiv:1706.09523 [stat]* . arXiv: 1706.09523.
  **URL:** *http://arxiv.org/abs/1706.09523*

Hines, O., Diaz-Ordaz, K. & Vansteelandt, S. (2023), 'Variable importance measures for heterogeneous causal effects'. arXiv:2204.06030 [stat].
  **URL:** *http://arxiv.org/abs/2204.06030*

Hirano, K., Imbens, G. W. & Ridder, G. (2003), 'Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score', *Econometrica* **71**(4), 1161–1189. Publisher: John Wiley & Sons, Ltd.
  **URL:** *https://doi.org/10.1111/1468-0262.00442*

Hodler, R., Lechner, M. & Raschky, P. A. (2023), 'Institutions and the resource curse: New insights from causal machine learning', *PLOS ONE* **18**(6), e0284968.
  **URL:** *https://dx.plos.org/10.1371/journal.pone.0284968*

Hodula, M., Melecký, M., Pfeifer, L. & Szabo, M. (2023), 'Cooling the mortgage loan market: The effect of borrower-based limits on new mortgage lending', *Journal of International Money and Finance* **132**, 102808.
  **URL:** *https://linkinghub.elsevier.com/retrieve/pii/S0261560623000098*

Holland, P. W. (1986), 'Statistics and causal inference', *Journal of the American statistical Association* **81**(396), 945–960. Publisher: Taylor & Francis.

Huber, M., Meier, J. & Wallimann, H. (2022), 'Business analytics meets artificial intelligence: Assessing the demand effects of discounts on Swiss train tickets', *Transportation Research Part B: Methodological* **163**, 22–39.
  **URL:** *https://linkinghub.elsevier.com/retrieve/pii/S0191261522001059*

Imai, K., Keele, L. & Tingley, D. (2010), 'A general approach to causal mediation analysis.', *Psychological Methods* **15**(4), 309–334.
  **URL:** *https://doi.apa.org/doi/10.1037/a0020761*

Inoue, K., Seeman, T. E., Horwich, T., Budoff, M. J. & Watson, K. E. (2023), 'Heterogeneity in the Association Between the Presence of Coronary Artery Calcium and Cardiovascular Events: A Machine-Learning Approach in the MESA Study', *Circulation* **147**(2), 132–141.
  **URL:** *https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.122.062626*

Jakobsen, K. D., O'Regan, E., Svalgaard, I. B. & Hviid, A. (2023), 'Machine learning identifies risk factors associated with long-term sick leave following COVID-19 in Danish population', *Communications Medicine* **3**(1), 188.
  **URL:** *https://www.nature.com/articles/s43856-023-00423-5*

Kianmehr, H., Guo, J., Lin, Y., Luo, J., Cushman, W., Shi, L., Fonseca, V. & Shao, H. (2022), 'A machine learning approach identifies modulators of heart failure hospitalization prevention among patients with type 2 diabetes: A revisit to the ACCORD trial', *Journal of Diabetes and its Complications* **36**(9), 108287.
  **URL:** *https://linkinghub.elsevier.com/retrieve/pii/S1056872722001994*

Kluger, D. M., Owen, A. B. & Lobell, D. B. (2022), 'Combining randomized field experiments with observational satellite data to assess the benefits of crop rotations on yields', *Environmental Research Letters* **17**(4), 044066.
  **URL:** *https://iopscience.iop.org/article/10.1088/1748-9326/ac6083*

Kristjanpoller, W., Michell, K. & Minutolo, M. C. (2021), 'A causal framework to determine the effectiveness of dynamic quarantine policy to mitigate COVID-19', *Applied Soft Computing* **104**, 107241.
  **URL:** *https://linkinghub.elsevier.com/retrieve/pii/S1568494621001642*

Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. (2019), 'Metalearners for estimating heterogeneous treatment effects using machine learning', *Proceedings of the National Academy of Sciences* **116**(10), 4156–4165.
  **URL:** *http://www.pnas.org/lookup/doi/10.1073/pnas.1804597116*

Lechner, M. (2019), Modified Causal Forests for Estimating Heterogeneous Causal Effects.
  **URL:** *https://arxiv.org/ftp/arxiv/papers/1812/1812.09487.pdf*

Leite, W. L., Kuang, H., Shen, Z., Chakraborty, N., Michailidis, G., D'Mello, S. & Xing, W. (2022), Heterogeneity of Treatment Effects of a Video Recommendation System for Algebra, *in* 'Proceedings of the Ninth ACM Conference on Learning @ Scale', ACM, New York City NY USA, pp. 12–23.
  **URL:** *https://dl.acm.org/doi/10.1145/3491140.3528275*

Mesplé-Somps, S. & Nilsson, B. (2023), 'Role models, aspirations and desire to migrate', *Journal of Economic Behavior & Organization* **212**, 819–839.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0167268123002196*

Miao, W., Deng, Y., Wang, W., Liu, Y. & Tang, C. S. (2023), 'The effects of surge pricing on driver behavior in the <span style="font-variant:small-caps;">ride-sharing</span> market: Evidence from a <span style="font-variant:small-caps;">quasi-experiment</span>', *Journal of Operations Management* **69**(5), 794–822.
URL: *https://onlinelibrary.wiley.com/doi/10.1002/joom.1223*

Nie, X. & Wager, S. (2020), 'Quasi-Oracle Estimation of Heterogeneous Treatment Effects', *arXiv:1712.04912 [econ, math, stat]* . arXiv: 1712.04912.
URL: *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.
URL: *http://link.springer.com/10.1007/s10654-019-00551-0*

Oprescu, M., Syrgkanis, V., Battocchi, K., Hei, M. & Lewis, G. (2019), EconML: A Machine Learning Library for Estimating Heterogeneous Treatment Effects, *in* '33rd Conference on Neural Information Processing Systems', Vancouver, Canada, p. 6.

Osawa, I., Goto, T., Kudo, D., Hayakawa, M., Yamakawa, K., Kushimoto, S., Foster, D. M., Kellum, J. A. & Doi, K. (2023), 'Targeted therapy using polymyxin B hemadsorption in patients with sepsis: a post-hoc analysis of the JSEPTIC-DIC study and the EUPHRATES trial', *Critical Care* **27**(1), 245.
URL: *https://ccforum.biomedcentral.com/articles/10.1186/s13054-023-04533-3*

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P. & Moher, D. (2021), 'The PRISMA 2020 statement: an updated guideline for reporting systematic reviews', *BMJ* p. n71.
URL: *https://www.bmj.com/lookup/doi/10.1136/bmj.n71*

Rana, P. & Miller, D. C. (2019), 'Machine learning to analyze the social-ecological impacts of natural resource policy: insights from community forest management in the Indian Himalaya', *Environmental Research Letters* **14**(2), 024008.
URL: *https://iopscience.iop.org/article/10.1088/1748-9326/aafa8f*

Rehill, P. & Biddle, N. (2023), 'Transparency challenges in policy evaluation with causal machine learning – improving usability and accountability'. arXiv:2310.13240 [cs, econ].
URL: *http://arxiv.org/abs/2310.13240*

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994), 'Estimation of Regression Coefficients When Some Regressors Are Not Always Observed', *Journal of the American Statistical Association* **89**(427), 846–866. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].
URL: *http://www.jstor.org/stable/2290910*

Rudin, C. (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence* **1**(5), 206–215.
URL: *https://doi.org/10.1038/s42256-019-0048-x*

Seitz, K. P., Spicer, A. B., Casey, J. D., Buell, K. G., Qian, E. T., Graham Linck, E. J., Driver, B. E., Self, W. H., Ginde, A. A., Trent, S. A., Gandotra, S., Smith, L. M., Page, D. B., Vonderhaar, D. J., West, J. R., Joffe, A. M., Doerschug, K. C., Hughes, C. G., Whitson, M. R., Prekker, M. E., Rice, T. W., Sinha, P., Semler, M. W. & Churpek, M. M. (2023), 'Individualized Treatment Effects of Bougie versus Stylet for Tracheal Intubation in Critical Illness', *American Journal of Respiratory and Critical Care Medicine* **207**(12), 1602–1611.
URL: *https://www.atsjournals.org/doi/10.1164/rccm.202209-1799OC*

Semenova, V. & Chernozhukov, V. (2021), 'Debiased machine learning of conditional average treatment effects and other causal functions', *The Econometrics Journal* **24**(2), 264–289.
URL: *https://doi.org/10.1093/ectj/utaa027*

Serra-Garcia, M. & Szech, N. (2023), 'Incentives and Defaults Can Increase COVID-19 Vaccine Intentions and Test Demand', *Management Science* **69**(2), 1037–1049.
URL: *https://pubsonline.informs.org/doi/10.1287/mnsc.2022.4405*

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. (2008), 'Conditional variable importance for random forests', *BMC Bioinformatics* **9**(1), 307.
URL: *https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-307*

Sverdrup, E. (2021), 'Add support for SHAPR'.
URL: *https://github.com/grf-labs/grf/issues/986*

Sverdrup, E., Kanodia, A., Zhou, Z., Athey, S. & Wager, S. (2020), 'policytree: Policy learning via doubly robust empirical welfare maximization over trees', *Journal of Open Source Software* **5**(50), 2232.
URL: *https://joss.theoj.org/papers/10.21105/joss.02232*

Syrgkanis, V., Lewis, G., Oprescu, M., Hei, M., Battocchi, K., Dillon, E., Pan, J., Wu, Y., Lo, P., Chen, H., Harinen, T. & Lee, J.-Y. (2021), Causal Inference and Machine Learning in Practice with EconML and CausalML: Industrial Use Cases at Microsoft, TripAdvisor, Uber, *in* 'Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining', ACM, Virtual Event Singapore, pp. 4072–4073.
URL: *https://dl.acm.org/doi/10.1145/3447548.3470792*

Tibshirani, J., Athey, S., Sverdrup, E. & Wager, S. (2021), 'Package 'grf''.
URL: *https://cran.r-project.org/web/packages/grf/grf.pdf*

Turjeman, D. & Feinberg, F. M. (2024), 'When the Data Are Out: Measuring Behavioral Changes Following a Data Breach', *Marketing Science* **43**(2), 440–461.
URL: *https://pubsonline.informs.org/doi/10.1287/mksc.2019.0208*

Venkatasubramaniam, A., Mateen, B. A., Shields, B. M., Hattersley, A. T., Jones, A. G., Vollmer, S. J. & Dennis, J. M. (2023), 'Comparison of causal forest and regression-based approaches to evaluate treatment effect heterogeneity: an application for type 2 diabetes precision medicine', *BMC Medical Informatics and Decision Making* **23**(1), 110.
URL: *https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-023-02207-2*

Verstraete, K., Gyselinck, I., Huts, H., Das, N., Topalovic, M., De Vos, M. & Janssens, W. (2023), 'Estimating individual treatment effects on COPD exacerbations by causal machine learning on randomised controlled trials', *Thorax* **78**(10), 983–989.
URL: *https://thorax.bmj.com/lookup/doi/10.1136/thorax-2022-219382*

Wager, S. (2018), 'Find the best tree in the random forest · Issue #281 · grf-labs/grf'.
URL: *https://github.com/grf-labs/grf/issues/281*

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. Publisher: Taylor & Francis.
URL: *https://doi.org/10.1080/01621459.2017.1319839*

Wager, S., Hastie, T. & Efron, B. (2014), 'Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife', *J. Mach. Learn. Res.* **15**(1), 1625–1651. Publisher: JMLR.org.

Wang, G. (2022), 'The Effect of Medicaid Expansion on Wait Time in the Emergency Department', *Management Science* **68**(9), 6648–6665.
URL: *https://pubsonline.informs.org/doi/10.1287/mnsc.2021.4239*

Wang, L., Wang, C. A. & Yao, X. (2023), 'Befriended to polarise? The impact of friend identity on review polarisation—A <span style="font-variant:small-caps;">quasi-experiment</span>', *Information Systems Journal* p. isj.12425.
URL: *https://onlinelibrary.wiley.com/doi/10.1111/isj.12425*

Wooldridge, J. M. (2013), *Introductory econometrics: a modern approach*, 5th ed edn, South-Western Cengage Learning, Mason, OH. OCLC: ocn827938223.

Xu, Y., Bechler, K., Callahan, A. & Shah, N. (2023), 'Principled estimation and evaluation of treatment effect heterogeneity: A case study application to dabigatran for patients with atrial fibrillation', *Journal of Biomedical Informatics* **143**, 104420.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S1532046423001417*

Xu, Y., Loftness, V. & Severnini, E. (2021), 'Using Machine Learning to Predict Retrofit Effects for a Commercial Building Portfolio', *Energies* **14**(14), 4334.
URL: *https://www.mdpi.com/1996-1073/14/14/4334*

Xue, Q., Wang, H. & Bai, C. (2023), 'Local green finance policies and corporate ESG performance', *International Review of Finance* **23**(4), 721–749.
URL: *https://onlinelibrary.wiley.com/doi/10.1111/irfi.12417*

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E. & Wager, S. (2023), 'Evaluating Treatment Prioritization Rules via Rank-Weighted Average Treatment Effects'. arXiv:2111.07966 [stat].
URL: *http://arxiv.org/abs/2111.07966*

Yin, W., Huo, W. & Lin, D. (2023), 'The effects of state coercion on voting outcome in protest movements: a causal forest approach', *Political Science Research and Methods* **11**(3), 645–653.

Yu, C., Yu, J. & Gao, D. (2024), 'Smart Cities and Greener Futures: Evidence from a Quasi-Natural Experiment in China's Smart City Construction', *Sustainability* **16**(2), 929.
URL: *https://www.mdpi.com/2071-1050/16/2/929*

Zhang, W., Li, J. & Liu, L. (2022), 'A Unified Survey of Treatment Effect Heterogeneity Modelling and Uplift Modelling', *ACM Computing Surveys* **54**(8), 1–36.
URL: *https://dl.acm.org/doi/10.1145/3466818*

Zhang, Z., Akinci, B. & Qian, S. (2022), 'Inferring heterogeneous treatment effects of work zones on crashes', *Accident Analysis & Prevention* **177**, 106811.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0001457522002469*

Zhou, H., Zhan, R., Chen, X., Lin, Y., Zhang, S., Zheng, H., Wang, X., Huang, M., Xu, C., Liao, X., Tian, T. & Zhuang, X. (2023), 'Targeting efficacy of spironolactone in patients with heart failure with preserved ejection fraction: the TOPCAT study', *ESC Heart Failure* **10**(1), 322–333.
URL: *https://onlinelibrary.wiley.com/doi/10.1002/ehf2.14068*

Zhou, Z., Athey, S. & Wager, S. (2023), 'Offline Multi-Action Policy Learning: Generalization and Optimization', *Operations Research* **71**(1), 148–183.
URL: *https://pubsonline.informs.org/doi/10.1287/opre.2022.2271*

Zhou, Z., Jian, B., Chen, X., Liu, M., Zhang, S., Fu, G., Li, G., Liang, M., Tian, T. & Wu, Z. (2023), 'Heterogeneous treatment effects of coronary artery bypass grafting in ischemic cardiomyopathy: A machine learning causal forest analysis', *The Journal of Thoracic and Cardiovascular Surgery* p. S0022522323007973.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0022522323007973*

Zhu, M. (2022), 'Living with offspring surely brings happiness to the elderly?—the heterogeneity in the effect of living arrangements on the life satisfaction of the Chinese elderly', *Procedia Computer Science* **214**, 359–366.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S1877050922018968*

Zhu, M. (2023), 'The Effect of Political Participation of Chinese Citizens on Government Satisfaction: Based on Modified Causal Forest', *Procedia Computer Science* **221**, 1044–1051.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S187705092300844X*

## Appendix A - Full table of presentation methods

| Method | Count |
|---|---|
| Graph across key variables | 58 |
| Variable importance | 35 |
| Graph ITEs in a histogram | 33 |
| Graph along quantiles | 26 |
| Single tree | 15 |
| Best Linear Projection | 12 |
| Derive policy | 11 |
| Rank ITEs | 9 |
| SHAP | 6 |
| Fit linear model on causal forest predictions | 7 |
| Graph across geography | 5 |
| Fit quantiles for ITE then look at covariate distribution within quantile | 5 |
| Partial dependence plot | 3 |
| K-means clustering | 2 |
| GAM model based on important variables | 1 |
| Graph predicted vs observed effects in future | 1 |
| Graph probability of negative outcome | 1 |
| PCA on most important variables | 1 |
| Plot across variable holding others fixed at median | 1 |
| Q-Score metric | 1 |
| Qini curve | 1 |
| Slope test | 1 |
| Rank-weighted Average Treatment Effect | 1 |

Table 3: Number of times a particular technique was used in the studied papers to communicate CATE results