How to detect heterogeneity in conjoint experiments[*]

Thomas S. Robinson[†] Raymond M. Duch[‡]

Abstract

Conjoint experiments are fast becoming one of the dominant experimental methods within the social sciences. Despite recent efforts to model heterogeneity within this type of experiment, the relationship between the conjoint design and lower-level causal estimands is underdeveloped. In this paper, we clarify how conjoint heterogeneity can be construed as a set of nested, causal parameters that correspond to the levels of the conjoint design. We then use this framework to propose a new estimation strategy, using machine learning, that better allows researchers to evaluate treatment effect heterogeneity. We also provide novel tools for classifying and analysing heterogeneity post-estimation using partitioning algorithms. Replicating two conjoint experiments, we demonstrate our theoretical argument, and show how this method helps estimate and detect substantive patterns of heterogeneity. To accompany this paper, we provide new a R package, **cjbart**, that allows researchers to model heterogeneity in their experimental conjoint data.

Keywords: Heterogeneity, conjoint, BART, AMCE, experiment

[*] This study was partially funded by the National Agency for Research and Development ANID, and we acknowledge the research support provided by FONDECYT 2020 grant number 1201397. The survey was conducted according to the University of Oxford's policy for human subjects research and approved by the University of Oxford Medical Sciences Interdivisional Research Ethics Committee (approval ID:R72328/RE001). Informed consent was obtained from each participant at the beginning of the survey. Replication files are available in the JOP Dataverse (https://dataverse.harvard.edu/dataverse/jop). The empirical analysis has been successfully replicated by the JOP replication analyst.

[†] Assistant Professor. Department of Methodology, London School of Economics and Political Science. Contact: Connaught House, 65 Aldwych, London, WC2B 4DS, UK. Email: t.robinson7@lse.ac.uk

[‡] Professor. Nuffield College, University of Oxford. Contact: Nuffield College, New Road, Oxford, OX1 1NF, UK. Email: raymond.duch@nuffield.ox.ac.uk. Phone: +44 (0)1865 278515

In the last decade, the number of papers per year that mention “conjoint experiments” has risen sixfold, from 110 articles published in 2010 to 600 published in 2020.[1] Conjoint designs offer researchers an efficient means of recovering multiple causal parameters across a wide range of research areas, including radical right voting (Chou et al. 2021), tax preferences (Ballard-Rosa et al. 2017), and contemporary drivers of migration (Spilker et al. 2020; Duch et al. 2020).

The predominant causal quantity estimated in conjoint experiments is the average marginal component effect (AMCE; Hainmueller et al. 2014), defined as “the effect of a particular attribute value of interest against another value of the same attribute while holding equal the joint distribution of the other attributes in the design, averaged over this distribution as well as the sampling distribution from the population" (Bansak et al. 2021, 29). While theoretically complex, the AMCE is easily estimated using conventional regression techniques, and allows researchers to isolate the average effect of attributes on the probability of choosing a profile.

By virtue of being an *average*, the AMCE may mask significant heterogeneity in subjects' behaviour. Researchers often want to know whether treatment effects differ depending on characteristics of the subjects who take part in their study, even though these sorts of analyses preclude causal interpretation since covariates are not randomised. For example, analysing heterogeneity can be useful in efforts to generalise treatment effects to populations of interest and for providing hints at possible causal mechanisms.

To estimate heterogeneity in AMCEs, studies typically present separate models for distinct sub-groups within the data.[2] Despite its simplicity, this subset approach is suboptimal. First, the strategy presumes that researchers have strong theoretical and empirical reasons to focus on specific sub-groups. Without such grounds, (repeated) subset analysis risks the

[1] Based on a Google Scholar keyword search for “conjoint experiment".

[2] Spilker et al. (2020), for example, conduct subgroup analyses on age, income, education and location.

inferential problems associated with multiple testing. Moreover, beyond convenient dichotomous splits in the data, subgroup analysis becomes unwieldy once researchers want to consider more complex groups of respondents. Second, directly interpreting subgroup differences across models can be misleading if each subgroup's preference differs over the reference level (Leeper et al. 2020). Third, subgroup analyses reduce the number of observations in each model, increasing uncertainty by preventing the models from "borrowing" shared variation between subsets of the data.

We propose a strategy for detecting and characterizing heterogeneity in these marginal effects, which addresses some of these limitations by exploiting the richness of the data generated in conjoint experiments. Recent methodological advances point to two specific causal concerns in conjoint experiments: causal *interaction* — estimating the efficacy of combinations of treatment variables (e.g. Ham et al. 2022; Goplerud et al. 2022) – and causal *moderation* – estimating variation in treatment effects across individuals or pre-treatment covariates (e.g. Zhirkov 2022). Our approach focuses on this latter concern, building on a growing corpus of work highlighting the utility of machine-learning methods in experimental settings (Hill 2011; Green and Kern 2012; Wager and Athey 2018; Künzel et al. 2019).

We make three novel contributions to the study of treatment effect heterogeneity in conjoint experiments. First, we clarify how lower-level causal quantities, i.e., subject-specific or conditional treatment effects, are situated within the structure of conjoint designs. We present a simple derivation of nested causal effects that disaggregates the AMCE to the level of the individual, round, and observation within the experiment.

Second, we leverage non-parametric machine learning estimators to estimate heterogeneity in conjoint treatment effects. We predict counterfactual treatment outcomes at the observation-level and aggregate these estimated effects to produce higher-level treatment effect estimates. Our non-parametric strategy is based on Bayesian Additive Regression

Trees (BART) (Hill 2011; Green and Kern 2012; Duch et al. 2020). Unlike subgroup analyses and other approaches that focus on modelling each individual separately (Zhirkov 2022), our approach leverages the full support of the data rather than relying on much smaller subsets of observations. We also provide variance estimators that exhibit good coverage, allowing researchers to quantify the uncertainty over these predicted effects.

Third, we characterize the extent and types of heterogeneity once we have estimated the nested causal quantities. We repurpose tools from the interpretable machine learning literature to measure how important different subject-level covariates are for partitioning the distribution of estimated individual-level marginal component effects (Ishwaran and Lu 2019). Our approach allows for bias-corrected estimates of the importance of variables, and thus to detect which variables are driving treatment effect heterogeneity.

We demonstrate our approach using data from the recent Duch et al. (2021) conjoint study of global preferences over COVID-19 vaccination policies. We also provide a new R package – *cjbart* – that implements our proposed method, allowing researchers to estimate and analyse treatment effect heterogeneity within conjoint experiments. This package is available on the Comprehensive R Archive Network (CRAN).

1 Heterogeneity in conjoint designs

Conjoint experiments allow for efficient estimation of multiple causal parameters that affect subjects' choices. Subjects are presented with profiles defined by a set of attributes. Each attribute has multiple values, or "levels", which are simultaneously randomised. Subjects then make a discrete choice over these profiles. Through repeated observation, researchers can estimate the marginal effects of each attribute-level (compared to some reference level) on subjects' choices.

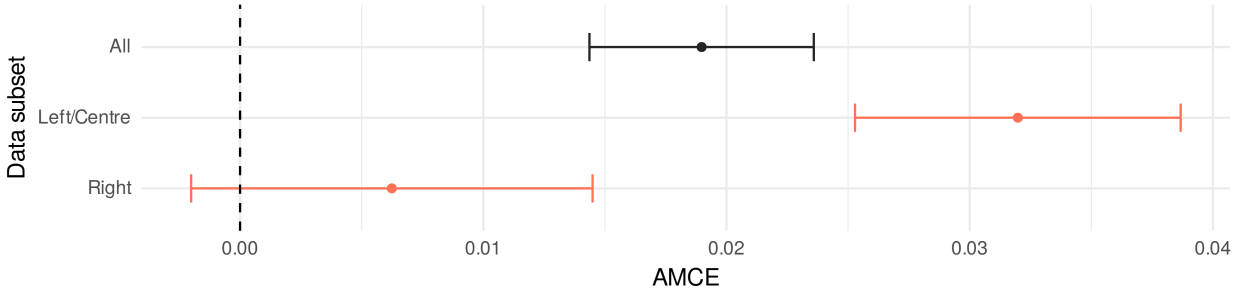
Typical conjoint estimands may, however, belie diversity in subjects' behavior. To illustrate the challenge facing scholars, consider a recent conjoint experiment conducted by Duch et al. (2021). This 13-country conjoint experiment asked subjects in each round to choose which of two profiles should be prioritized for a COVID-19 vaccine. In Figure 1a we replicate the AMCE and sub-group estimates for hypothetical profiles who had low incomes. On average, subjects were more likely to choose profiles that were labelled low-income relative to those on an average income, and subgroup analyses suggest the effect of this attribute-level is conditioned by subjects' own ideological stance.

However, Figure 1b suggests the narrative is not quite as simple as the subgroup analysis would suggest. Here we plot the selection probabilities (by colour) and densities (by height) for low-income profiles, conditioning on both ideological self-identification and the country of each subject in the Duch et al. (2021) study. The data comprising this plot covers 82,503 forced-choices made by the 15,536 participants in the study. While we do see a general trend that right-leaning subjects are less likely to prioritise low-income profiles, there is quite clear heterogeneity in this relationship *across* contexts. For some countries – like Brazil, Uganda, and India – the relationship is far less pronounced.

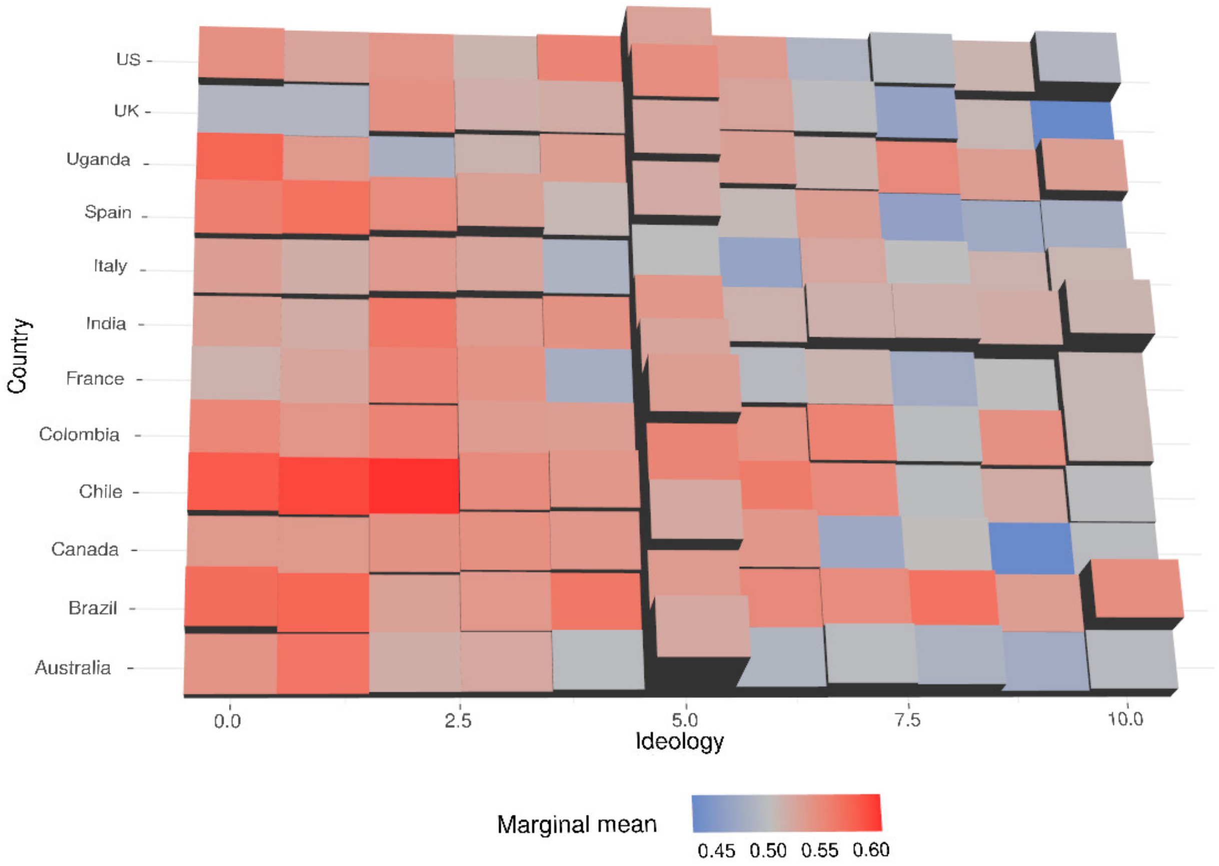
The subgroup strategy demonstrated in Figure 1a fails to capture this cross-country variation in part because it involves a more complicated, *a priori* specification of subgroups. Running separate models for left and right-identifying subjects in each country would entail estimating 26 separate models, each powered by far fewer observations. Even if this were feasible, this strategy would still omit variation within the dichotomous ideological splits. For example, subjects in Canada and the UK both exhibit notable variation *within* right- and left-leaning subjects respectively.

In order to understand heterogeneity in conjoint experiments, researchers should use methods that allow for the modelling of interactions between covariates and conjoint attributes without imposing *a priori* functional form on these relationships. Given the rich data generated by the conjoint design, we should allow the model itself to find interactions between randomised conjoint attributes and subjects' characteristics.

Figure 1. Impact of respondents' ideology on choosing to prioritise vaccinating low-income profiles



(a) AMCE estimates for the "Lowest 20% income-level" attribute-level, estimated on the full data and subsets containing Left/Centre and Right-leaning subjects respectively.



(b) Proportion of profiles selected (marginal mean) that contain the "Lowest 20% income-level" attribute-level, by subjects' ideology and country. The height of the bars reflect the number of observations in each cell.

The remainder of this section outlines a series of lower-level causal estimands that relate to the multi-level structure of the conjoint design and that allow us to model heterogeneous treatment effects. We initially restrict our focus to cases where there is complete randomisation of values in the conjoint experiment.[3] This assumption simplifies the analysis and estimation of the causal parameters, and is the typical design employed by researchers in practice. In Section 2.4, we demonstrate how our strategy can incorporate non-uniform distributions of attribute-levels following the insights of de la Cuesta et al. (2022).

1.1 Nested causal quantities in conjoint designs

Suppose N individuals (indexed by i) choose between J profiles across K rounds of the experiment. Within each round of the experiment, we randomly assign attribute-levels across L attributes for each profile (Hainmueller et al. 2014). Having run the experiment, the researcher faces a data structure with $N \times J \times K$ rows and $L + X$ columns (where X are any covariates observed for each subject), from which causal parameters of interest can be estimated.

The most common parameter estimated from this design is the **average marginal component effect** (AMCE). This estimand reflects the overall effect of a specific attribute-level on the probability of choosing a profile (compared to some baseline reference level), after accounting for the possible effects of the other attributes in the design. To account for these other effects, the parameter is averaged over the effect variations caused by these other attributes.

With complete randomisation of the attributes, and adapting the notation set by Hainmueller et al. (2014), we define the potential outcome for a profile shown to a respondent

[3] In other words, where the probability of assigning each attribute-level is constant within each attribute and independent of the values of other attributes.

in the experiment as the (non-parametric) function:

$$Y_{ijk}(t_l, T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k}) = g\Bigg(\mathcal{S}_i(t_l, T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k}), \mathcal{R}_{ik}(t_l, T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k}), \mathcal{P}_{ijk}(t_l, T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k})\Bigg),$$

where t_l is the value of the lth attribute shown to individual i, in profile j, of round k of the experiment, $T_{ijk[-l]}$ is the vector of values for the remaining attributes in the same profile, and $\boldsymbol{T}_{i[-j]k}$ is the unordered set of possible treatment vectors.[4] $\mathcal{S}_i$, $\mathcal{R}_{ik}$, and $\mathcal{P}_{ijk}$ are respondent-, round-, and profile-level random components of this function.

Using the defined potential outcomes, the AMCE can be expressed as:

$$\tau_l = \mathbb{E}\big[Y_{ijk}(t_l = l_1, T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k}) - Y_{ijk}(t_l = l_0, T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k})\big].$$

By definition, the AMCE captures the central tendency of subjects' behavior with respect to each attribute of the design. Often, however, researchers are interested in whether these effects differ dependent on subject characteristics or the context of the experiment. As others have noted, the AMCE can be disaggregated into more granular causal quantities of interest (Hainmueller et al. 2014; Abramson et al. 2020; Zhirkov 2022). Here we formalise this logic with respect to the structure of the data generating process itself.

First, we disaggregate the AMCE into N individual-level effects by conditioning the AMCE estimand on the individual-level random component of our model:

$$\tau_{il} = \mathbb{E}\big[Y_{ijk}(t_l = l_1, T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k}) - Y_{ijk}(t_l = l_0, T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k}) | \mathcal{S}_i\big].$$

This lower-level parameter is the **individual-level marginal effect** (IMCE), and reflects the change in probability *for a specific subject* i of choosing a profile given an attribute-level (compared to some reference category) averaged over the effects of all other attributes. The estimand is similar to subgroup analysis of AMCEs – what Hainmueller et al. (2014) call conditional AMCEs. Unlike that specification, rather than subsetting the data along a

[4] In Appendix A, we relax this assumption. Note also, for the sake of completeness, that $T_{ijkl} = t_l$.

vector of covariates, we subset based on the subject identifier and therefore consider the conditional effect based on all of subject i's characteristics. [5]

The IMCE is substantively useful because it allows researchers to inspect heterogeneity in the treatment effects derived from conjoint experiments (Abramson et al. 2020), and is commensurate with more general heterogeneous effect estimation strategies (Künzel et al. 2019). By recovering a vector of individual-level estimates, researchers can compare how non-randomised aspects of the data (i.e. subjects' characteristics) correspond to the magnitude and direction of the individual-level predicted effects.

In turn, the IMCE can be decomposed over the repeated observations taken for that individual (i.e. the choices over profiles subjects make across multiple rounds of the conjoint experiment). This decomposition can be split into two steps since subjects typically see $J \geq 2$ profiles per round.[6] First, therefore, we can disaggregate the **round-level marginal component effect** (RMCE):

$$\tau_{ikl} = \mathbb{E}\big[Y_{ijk}(t_l = l_1, T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k}) - Y_{ijk}(t_l = l_0, T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k}) | \mathcal{S}_i, \mathcal{R}_{ik}\big].$$

This estimand reflects the effect of a component within a specific round (k) of the experiment for a given individual. Under the conventional no carryover assumption, this random component should be mean zero and evidence to the contrary may suggest this assumption has been violated.

Finally, the RMCE can be further decomposed into an **observation-level marginal component effect** (OMCE) by conditioning on the profile-level random component:

$$\tau_{ijkl} = \mathbb{E}\big[Y_{ijk}(t_l = l_1, T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k}) - Y_{ijk}(t_l = l_0, T_{ijk[-l]}, \boldsymbol{T}_{i[-j]k}) | \mathcal{S}_i, \mathcal{R}_{ik}, \mathcal{P}_{ijk}\big].$$

This profile-level random component contains three aspects of the conjoint design: the

[5] Despite subsetting on the identifier, the IMCE may be moderated by covariate features that generalise across subjects. Our estimation strategy in Section 2 allows for this moderation to be discovered.

[6] Though we note it is possible to have single-profile conjoint rounds.

randomised values of the L-1 attributes, plus the randomised order of those attributes (as is typical in conjoint experiments), and finally the order of the two profiles (whether a profile appears on the left or right in a two-profile design). The expectation is therefore taken over the fundamental uncertainty in the outcome, uncertainty that one would expect under the assumption of a super-population inference framework. In other words, the OMCE captures what a specific individual would have done, in exactly that vignette, at that point in the experiment, varying *only* the conjoint attribute of interest. Across multiple (hypothetical) realisations we might nevertheless expect some variation in choosing that profile, given the probabilistic nature of that choice.

From an applied perspective, the informativeness of the OMCE is limited given the granularity of the estimand. That said, it serves a useful statistical purpose given the more general, nested relationship between the OMCE, RMCE, IMCE, and AMCE. In particular, by the law of iterated expectations $\tau_l = \mathbb{E}_i\Big[\mathbb{E}_k\big[\mathbb{E}_j[\tau_{ijkl}]\big]\Big]$.[7]Assuming there are no carry-over effects across rounds, the OMCE can be thought of as an independent draw from the individual-level distribution. The individual-level marginal effect can therefore be estimated by aggregating OMCEs (as we discuss in Section 2.1).

2 Estimating the IMCE

Estimating lower-level estimands provides specific leverage over questions about the heterogeneity of these effects. The most effective level of analysis is the individual-level, since we can analyse how the IMCE varies dependent on characteristics of the subjects. Therefore, we propose a three step strategy to recover estimates of the IMCEs.

First, we model the relationship between the forced-choice outcome, conjoint attribute-levels, and subject-level covariates. This allows us to estimate some function that captures the potentially heterogeneous relationship between the conjoint attributes and subjects'

[7] Subscripts under the expectation symbol indicate over what level the conditional means are taken. Table A1 in the Appendix illustrates this relationship from a data perspective.

characteristics when making choices in the experiment. Second, we use the trained model to predict counterfactual outcomes at the observation-level from which we can estimate OMCEs. Third, following the nested logic outlined in Section 1, we aggregate these OMCE estimates to the level of the individual in order to recover estimates of the IMCEs.

It is worth noting that researchers could use any number of possible estimators to model subject-level heterogeneity in the first step. We provide a specific implementation in this paper and accompanying software that uses Bayesian Additive Regression Trees (BART) (Chipman et al. 2010), but other researchers may wish to pursue alternative types of models. To that extent, the general approach detailed here can be considered a meta-strategy for estimating individual-level marginal effects in conjoint designs. In Appendix E, for example, we demonstrate our method using causal forests instead of BART (Athey et al. 2019).

One key benefit of this meta-strategy is that *all* data is included in the model when estimating the relationship between observed covariates, attribute-level assignments, and the conjoint outcome. This feature is in contrast to both subgroup analysis (where effects are modelled using only a smaller number of individuals who share a covariate value) and more recent approaches that recommend running separate models for each respondent (Zhirkov 2022).[8] Particularly when modelling each individual separately, constraints on experimental survey length may lead to large imprecision in the estimates. In our proposed method, the model leverages the full support of the data, across all observations, to discover covariate interactions that modify the causal effect at the individual-level. In Section 4 we demonstrate the comparative performance of our method compared to a subset-modelling strategy.

[8] Our strategy is a form of data-adaptive subgroup analysis, as the predicted outcomes are determined by observations closest to the datapoint after recursive partitioning of the full data. Unlike conventional subgroup analyses, however, tree-based approaches also use the data to *find* the most informative clusters, rather than relying on researchers to specify these *a priori*.

Moreover, by using machine learning, this method improves the analysis of potential heterogeneity in two ways. First, it reduces researcher degrees of freedom to arbitrarily run many subgroup analyses, which we would expect to inflate the chances of false positive discoveries. Second, it enables the identification of more complex relationships between variables. Common to many machine-learning methods, the model itself (rather than the researcher) determines the final functional form of the relationship between the supplied predictor variables and the outcome.

2.1 Parameter estimation

Step 1 In the first step, we use BART to model potential heterogeneity in the observed experimental data defined as:

$$P(Y_{ijk} = 1 | T_{ijk}, X_i) = f(T_{ijk}, X_i) \approx \hat{f}(T_{ijk}, X_i),$$

where Y_{ijk} is the observed binary outcome, T_{ijk} is the vector of treatment assignments across the L attributes, and X_i is the vector of covariate information for subject i considering profile j in round k of the experiment. f is some unknown true data generating process, and $\hat{f}$ is an estimate of that function.

BART is a tree-based supervised machine learning strategy that models the response surface by *summing* the predictions of many constrained individual tree models – recursive splits of the data into ever more homogeneous groups (Chipman et al. 2010). Appendix B provides a more detailed description of the BART algorithm. In short, there are two major difference between BART and other tree-based methods like random forests. First, the final prediction is not the average across a set of trees. Instead each tree is a "weak learner" that seeks to explain only the *residual* variance in the outcome not explained by the $T - 1$ other trees. In that sense, the constituent trees in the BART forest work together to predict the full outcome (rather than all trying to predict the same outcome entirely). Second, BART models include random variables as parameters, allowing draws to be taken from the posterior. This feature entails convenient Bayesian properties that allow us to recover variance estimates at the IMCE level, which we discuss below.

In addition to these advantage, we also use BART because the models are relatively robust to the choice of tuning parameters (He et al. 2019), as discussed in Appendix B. These priors are set partially with respect to the observed data, and the default parameters identified by Chipman et al. (2010) are known to perform well across data contexts (Kapelner and Bleich 2016). Cross-validation can be used to improve model performance further, if necessary.

To estimate the BART model, we supply a matrix of "training" data at the observation-level. The training data are simply the results of the conjoint experiment. Each row reflects a profile within a round shown to a specific subject. The matrix columns comprise the observed individual decision (0 or 1) regarding that profile; the assigned attribute-levels for each of the L attributes in the vignette (which vary within individuals); and covariate columns that are invariant at the individual-level. During training, the BART algorithm iterates through the trees in the model, many times over, updating the model parameters to minimize the error between a vector of predictions $\hat{Y}$ and the observed outcomes Y.[9]

Step 2 Using the final trained model ($\hat{f}$), we predict counterfactual outcomes by altering the value of attribute-levels in the conjoint data. Specifically, to recover a vector of OMCE estimates of attribute-level l_1, we take z draws from the predicted posterior using a "test" matrix which is identical to the training dataset, except each element in the column corresponding to attribute l is set to the value l_1.[10] We then repeat this process, except

[9] We use a probit-specific version of BART that better handles the binary outcome typical of this type of discrete-choice design. The probit outcomes are transformed back to probabilities prior to the computation of OMCEs.

[10] In our software implementation, $z = 1000$. These draws are taken using a Gibbs Sampler, obtained through a Monte Carlo Markov Chain (MCMC) backfitting algorithm. Chipman et al. (2010) show that, with sufficient burn-in, these sequential draws converge to the posterior of the true data generating process (p.275). Users can assess convergence using Geweke's convergence diagnostic test available in the **BART** `R` package (see §4.5, Sparapani et al. 2021).

the value of this column is now set to l_0, the reference category. This process yields two separate matrices of dimensions $z \times N$, which approximate the posterior distribution for each observation for two separate attribute values respectively (l_1 and l_0). Subtracting these two matrices yields a single matrix of predicted OMCE estimates – z per observation. To recover a parameter estimate of the OMCE, we simply average these z predictions for each observation to yield a vector of observation-level effects:

$$\text{OMCE} = \hat{\tau}_{ijkl} = \frac{1}{z}\left(\hat{f}(T_{ijkl} = l_1, T_{ijk[-l]}, X_i) - \hat{f}(T_{ijkl} = l_0, T_{ijk[-l]}, X_i)\right).$$

Step 3 Finally, consistent with the logic outlined in Section 1, the IMCE estimates can then be calculated by averaging the OMCEs for each individual i:

$$\text{IMCE} = \hat{\tau}_{il} = \frac{1}{J \times K}\sum^{K}\sum^{J}\hat{\tau}_{ijkl}.$$

Uncertainty estimation We also use the $z \times N$ matrix of predicted OMCEs from the BART model to estimate the uncertainty both at the observation and individual level. Since our estimating strategy is Bayesian, we implement a credible interval approach to capture the parameter uncertainty. We take the $1-\alpha$ posterior interval of the OMCE-level predictions. To aggregate this interval to the IMCE level, we concatenate the posterior draws for each OMCE estimate, and take the $\alpha/2$ and $(1-\alpha)/2$ quantiles. Given that the posterior distribution is a random variable, this credible interval indicates the central $1-\alpha$ proportion of the probability mass for the parameter's posterior.

2.2 Simulation tests of the estimation strategy

Using Monte Carlo simulations, we find that our method effectively detects IMCE heterogeneity caused by heterogeneous *preferences*. We simulate a full conjoint experiment in which subjects make choices between two profiles. Each profile contains three conjoint attributes that are randomly assigned one of two values: $A_1 = \{a, b\}, A_2 = \{c, d\}, A_3 = \{e, f\}$. To induce heterogeneity, we define subjects' preferences over attribute levels as a function of two individual-level covariates varying this relationship across attributes. The first covariate c_1 is a binary variable drawn from a binomial distribution of size 1 with probability 0.5; the second covariate c_2 is a continuous variable drawn from a uniform distribution with bounds [-1,1].

We define the change in utility as a result of observing the second level for each attribute as follows:

$$\Delta U_{A_1} \sim \begin{cases} \mathcal{N}(\mu = 1, \sigma = 1), & \text{if } c_1 = 1 \\ \mathcal{N}(\mu = -1, \sigma = 1), & \text{otherwise.} \end{cases}$$

$$\Delta U_{A_2} \sim \mathcal{N}(\mu = |c_2 - 0.2|, \sigma = 1)$$

$$\Delta U_{A_3} \sim \mathcal{N}(\mu = 0, \sigma = 0.5)$$

We then simulate the conjoint experiment run on 500 subjects, for 5 rounds each, in which individuals choose between 2 profiles. For each observation, we calculate the utility for subject i given profile j in round k as:

$$U_{ijk} = \mathbb{I}(A_1 = b) \times \Delta U_{A_1} + \mathbb{I}(A_2 = d) \times \Delta U_{A_2} + \mathbb{I}(A_3 = f) \times \Delta U_{A_3} + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, 0.0005)$ adds a small amount of noise to each utility calculation (to prevent exact draws). For each round j that subject i sees, the profile that yields the higher change in utility is “chosen" ($Y = 1$), and the other is not ($Y = 0$). This mimics the technical dependence between observations that forms the basis of the discrete choice design.

Given this specification, the BART estimation strategy should predict heterogeneous IMCEs for the first two attributes (A1 and A2) but not for the last attribute (A3). Since tree-based ML methods operate by partitioning the data, our strategy should easily identify the dichotomous IMCE relationship with c_1. The IMCEs for A1 should be positive when

$c_1 = 1$, but negative when $c_2 = 0$. We should observe no correlations between c_1 and A2. The covariate c_2 poses a harder challenge for our estimation strategy for two reasons. First, subdivision of the data cannot perfectly partition the IMCEs since the covariate is continuous. Second, the defined relationship is more complex and asymmetric over the covariate's range. The strongest positive effects should occur for *negative* values, and the weakest effects when $c_2 = 0.2$. We anticipate no correlation between c_2 and attribute A1 or A3.[11]

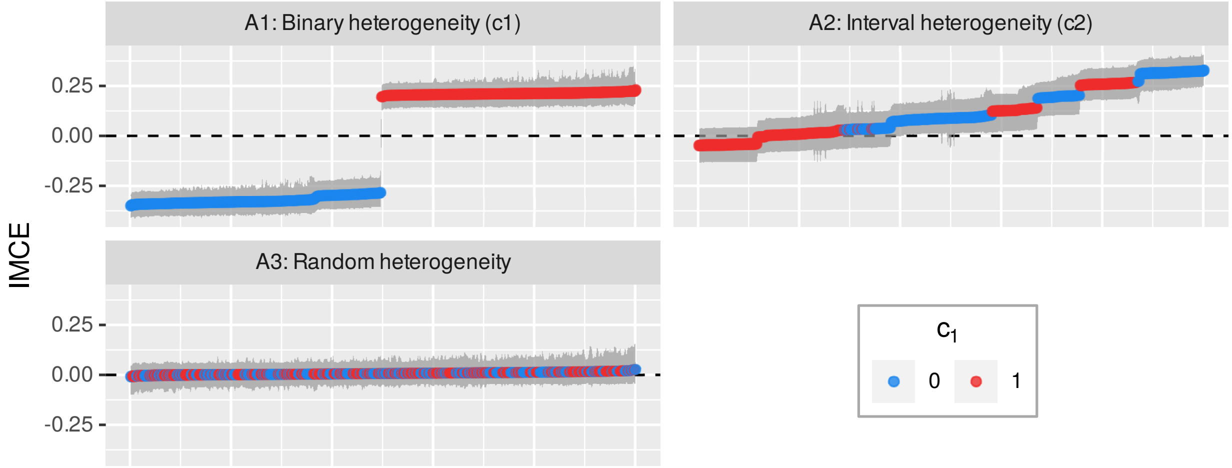
Figure 2 demonstrates the results of this experiment, colouring predicted IMCEs by the values of c_1. Our strategy effectively discovers heterogeneous IMCEs when the heterogeneity over preferences is a function of a binary variable – the positive and negative preferences perfectly correspond to the values of this covariate. Conversely, in the third facet, there is no indication of heterogeneity in the predicted IMCEs nor correlation between c_1 and the size of effects.

Over 100 simulations of this experiment, we observe an almost perfect correlation between c_1 and A1 ($\bar{r} = 0.998$), but negligible correlations between the same covariate and A2 and A3 ($\bar{r} = 0.004$ and -0.003 respectively). Similarly, the correlation between c_2 and A2 is substantive but, as expected, the magnitude is moderated by the non-linear and asymmetric relationship imposed ($\bar{r} = -0.557$). Again, there are negligible correlations for A1 and A3 ($\bar{r} = 0.000$ and 0.074 respectively).

In Figure G1 in the Appendix, we demonstrate that the heterogeneous IMCEs for A2 correlate with the continuous covariate c_2, as expected. Under a conventional, subsetting strategy, the analyst would likely also note that conditional AMCEs for A2 do not covary with c_1. However, subsetting based on c_1 would not indicate that there is substantial heterogeneity to the marginal component effect. We conjecture that as the complexity of

[11] In Appendix C4 and E1 we replicate this exercise using the Zhirkov (2022) OLS and Athey et al. (2019) causal forest methods, respectively.

Figure 2. Detecting heterogeneity in IMCEs using simulated conjoint data derived from preferences over profiles



Point estimates of the IMCEs for 500 subjects shown with 95% credible intervals (described in Section 2.1)

the covariance between covariates and IMCEs increases, it becomes harder for the analyst to adequately pre-specify models that would be capable of detecting this heterogeneity.

We extend this discussion of the simulated performance of our method in the Appendix. In Section C1 we demonstrate that the estimation method exhibits good predictive accuracy when IMCEs themselves are simulated across DGPs of varying form and complexity. We also find that our variance estimation strategy exhibits good coverage (Section C2). Finally, we test whether RMCEs can be used to detect whether effects are serially correlated by round (a violation of a conjoint experiment's assumptions) in Section C3.

2.3 Applied test of BART-estimated AMCEs

Under the various conjoint design assumptions, parameter estimates of the AMCEs from a linear probability model (LPM) are unbiased (Hainmueller et al. 2014). In Section 1, moreover, we note that the AMCE estimand can be considered the average of the IMCEs across subjects. Therefore, if our our estimation strategy is performing well, we expect that averaging the BART IMCEs will be very similar, if not the same as, the unbiased AMCEs estimated from a LPM.

As an applied sense check, we test this expectation empirically using data from two conjoint experiments. First, we analyse the archetypal conjoint experiment by Hainmueller et al. (2014) where U.S. subjects made a series of forced-choices between two profiles describing potential immigrants, indicating which they would prefer to admit. The attributes presented in the profiles reflected traits hypothesized to matter in immigration decision making, including the migrant's profession, country of origin, and language skills. Second, we analyse the previously discussed COVID-19 vaccine conjoint by Duch et al. (2021).

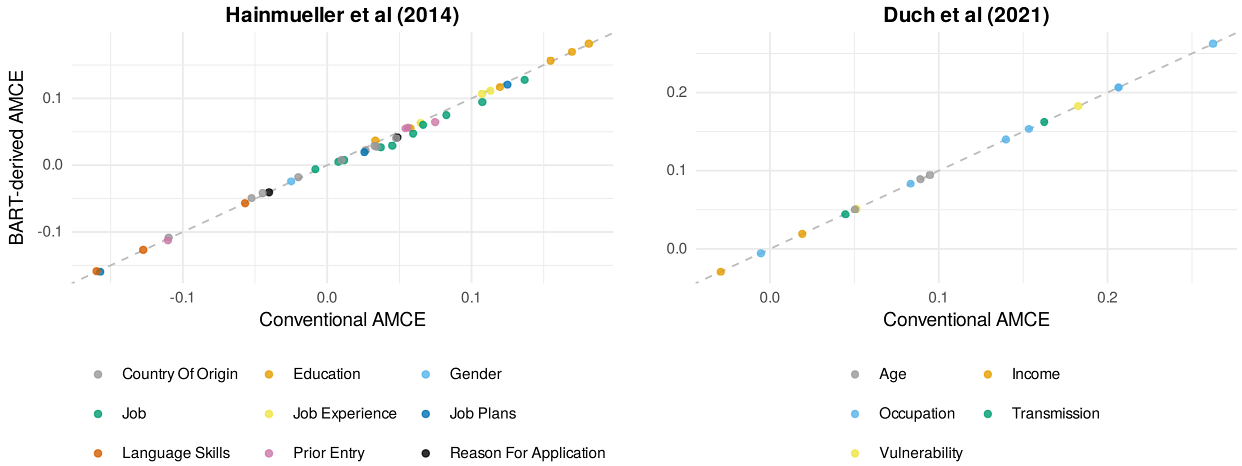
Figure 3 plots the point estimates of each (non-reference) attribute-level using our BART strategy and those of the conventional LPM approach, for both datasets. In both cases, and for every parameter, we see that the predicted effects are very similar. These results are strong *prima facie* evidence that the BART model is appropriately estimating the response surface: the individual-level effects do, in practise, aggregate correctly to the AMCE.[12]

2.4 Non-independent randomisation of attribute-levels

So far, we have assumed that attributes are completely and independently randomised, which is by far the most common type of conjoint design in practise (de la Cuesta et al. 2022). However, as others have noted, it is possible and informative to consider non-uniform distributions of profiles that better correspond to real-world profile distributions (Hainmueller et al. 2014; Bansak et al. 2021; de la Cuesta et al. 2022). This adaptation is also possible at the individual level, and we define the population-weighted quantity of interest as the “population-IMCE" (pIMCE).

[12] In Appendix D, we provide further estimation details for the Hainmueller et al. (2014) data, and Appendix Tables D1 and D3 report the LPM and **cjbart** coefficient estimates as well as the percentage differences between them.

Figure 3. Comparison of conventional GLM-derived AMCE to AMCEs recovered from the BART estimated IMCEs



We approach this challenge as a post-hoc exploratory analysis of existing conjoint data, similar to the model-based approach discussed by de la Cuesta et al. (2022). As before, we first use the observed experimental data to train a BART model. Adapting the previous strategy, we then predict a full set of counterfactual potential outcomes for *every* combination of the $L-1$ attributes in the design, for each subject (holding constant the individual-level covariates), and estimate the corresponding OMCEs by setting the lth attribute to l_1 and l_0 respectively.

We then recover the pIMCE by taking a weighted average of the predicted OMCEs, using researcher-specified marginal probabilities for the $L-1$ attributes. In effect, this marginalizes the IMCE over the profile distributions at the individual-level. The weight for a specific partial profile (ignoring the Lth attribute) is calculated as the product of the marginal probabilities for every other attribute-level in the profile:

$$w_{T_{ijk[-l]}} = P(T_{ijk[-l]}) = \prod_{l' \neq l} P(T_{ijkl'}).$$

Since our BART strategy takes z draws from the posterior, we calculate the weighted sum over the OMCEs for each draw separately, and then take the average over these predictions to generate our pIMCE estimate:

$$\text{pIMCE}_{il} = \mathbb{E}_z \Bigg[\sum_{T_{ijk[-l]} \in \boldsymbol{T}_{ijk[-l]}} (\hat{\tau}_{ijkl} \times w_{T_{ijk[-l]}}) \Bigg],$$

where the subscript z indexes draws from the model posterior, and $\boldsymbol{T}_{ijk[-l]}$ is the set of possible attribute-level combinations across the $L - 1$ other attributes.[13]

While this adaptation is relatively straightforward from a theoretical perspective, it comes at a computational cost. As the number of attributes (and attribute-levels) increases, the number of potential outcomes that need to be predicted inflates rapidly. Compared to the standard strategy, the number of predictions increases by the factorial of the number of levels for the $L - 1$ other attributes in the design. Researchers will want to narrow their analysis to specific population profiles, otherwise the computational demands will quickly become infeasible. We present an example of estimating pIMCEs in Appendix F.

3 Comparing Sources of Heterogeneity

A particular attraction of heterogeneous effects estimation is that we are able to examine whether treatment effects differ at the individual-level. To date, however, researchers have lacked principled methods of characterising this heterogeneity. In this section, we propose two tools researchers can use to systematically recover indicators of *which* covariates are driving heterogeneity in the marginal treatment effects and the interactions between variables. Both tools rely on tree-based learning methods to group the predicted IMCEs based on covariate information. In general, tree-based modelling approaches are well suited to this type of problem since they work by partitioning observations into clusters where the differences in outcomes between members of the same cluster are as small as possible

[13] Similar to the standard strategy, we also recover credible interval uncertainty estimates by taking the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles over the weighted distributions.

(Breiman et al. 1984).

We first introduce a standardised variable importance (VIMP) measure that summarises how well different covariates predict each distribution of IMCEs. This measure can be used to explore the potential sources of heterogeneity in the marginal component effects systematically across all attributes in the experiment. Second, we show how single regression trees can be used to better inspect the determinants of heterogeneity for specific attribute-levels of interest. This second step builds on the VIMP analysis by using the tree's decision rules to identify clusters, defined by subject covariates, that best define this heterogeneity. For each cluster, researchers can recover the conditional marginal component effect and thus analyse the extent of heterogeneity in the treatment effects.

3.1 Random forest variable importance

Our first tool summarises which covariates matter for predicting differences in the IMCE distributions for all attribute-levels in a conjoint experiment. We use random forests to estimate the relationship between the predicted IMCEs and subject-level covariates. For each attribute-level, we train a random forest to model the heterogeneity in the predicted IMCE distribution using subjects' covariate information as the predictor variables. Once each model has been trained, we recover variable importance measures (VIMPs) – a common form of model analysis for tree-based methods – to understand which covariate dimensions are most useful for partitioning the data. In turn, these variables can drive subsequent analyses which we present in Section 3.2.

In general, VIMP measures work by measuring the degradation in model performance when noise is added to a predictor variable. A larger drop in performance is indicative that the variable in question is more important for predicting the outcome. For our purposes, we use VIMP scores to measure how well the included subject covariates predict each vector of IMCEs. Higher importance scores suggest that partitioning the IMCEs on these variables is informative.[14]

The importance of subject-level covariates may differ dependent on the specific attribute-level in question. We therefore recover separate VIMP scores for each combination of attribute-level and subject covariate, allowing us to plot a heatmap of variable importance across the design as a whole. In Section 4 we demonstrate how this schedule of VIMP scores can be analysed to understand what drives heterogeneity for each attribute-level in the conjoint experiment.

3.2 Single decision tree partitioning

The random forest VIMP tool compares how well subject-level covariates predict each IMCE distribution. Given its reliance on random forests, however, it is less useful for substantively interpreting the partitioned IMCE space. The final model contains many trees, where each individual tree only considers a random subset of variables and a bootstrap sample of the data. We therefore propose a complementary tool that fits a single decision tree on an attribute-level of interest. Like the random forest model, the single-tree model recursively partitions the vector of IMCEs using a matrix of covariate information. Unlike the random forest method, since only one model is fit the individual splitting rules from this tree can be directly interpreted and used to inspect the heterogeneity in the IMCEs.[15]

Single tree models typically fit many splits to the data, making interpretation difficult. This feature reflects the inherent trade-off in machine learning methods between model complexity and the risk of mispredicting observations. In other words, a more complex tree may reduce prediction error (in training) but the incurred complexity reduces the variance of the model (leading to overfitting). Therefore, to ensure the tree is interpretable, we

[14] We use the Breiman-Cutler approach, which randomly permutes the predictor variable and measures the standardised difference in prediction error when using the original data compared to this permuted data. Taking advantage of recent developments in VIMP theory, and noting earlier critiques of bias in VIMP measures (Strobl et al. 2007), we recover bias-corrected variance estimates of these VIMP scores using delete-d jackknife estimation (Ishwaran and Lu 2019).

[15] A similar strategy has been pursued by Hahn et al. (2020).

follow the convention of “pruning" the fit model. Since the partitioning is recursive and “greedy", earlier splits in the tree are those that provide the greatest leverage over differentiating observations. By removing later splits, pruning has the effect of paring back the cluster definitions (i.e. the combination of decision rules) to a more parsimonious level.

In practice, trees are pruned by setting a complexity parameter. In the case of continuous outcomes, this determines how much of an increase in the overall R^2 of the model is needed in order for a split to be kept in the model. In our experience, a complexity parameter of about 0.02-0.04 is sufficient to constrain the tree’s depth to an interpretable level – about two or three degrees of partitioning.

Post-pruning, researchers can use the fit model to describe the underlying heterogeneity in the IMCE distribution. The terminal nodes reflect the conditional average marginal component effects defined by the splitting rules in the tree. This is similar to estimating marginal component effects for specific subgroups. Unlike manual subset analyses, however, the clusters are *discovered* during model fitting. This feature is particularly useful since the tree, splitting sequentially on multiple variables, may define complex groups. For example, it may find a stronger effect for young *and* ideologically left-leaning subjects, compared to those who are left-leaning but older. We illustrate this approach in the next section.

Comparing across methods While both the VIMP and decision tree methods share a similar partitioning logic, these methods could yield different insights: the VIMP analysis may, for example, highlight an additional feature not used in the single decision tree. These methods are complementary, and aim to extract as much information from the experimental data as possible. If, and where, inconsistencies do arise, that can be a signal that the researcher does not have strong or robust evidence of the sources of heterogeneity.

4 Heterogeneity in a multi-national conjoint experiment

In this section, we consider an application of the framework and estimation strategy outlined in Sections 1 and 2. We analyse heterogeneity in a very large conjoint experiment that encompasses a diverse group of subjects surveyed from 13 countries, and then compare our approach to a recent alternative strategy proposed in Zhirkov (2022).

Detecting heterogeneous effects The conjoint data are from the Duch et al. (2021) multi-national study on COVID-19 vaccine prioritization. Subjects choose which of two hypothetical individuals should be given priority for a COVID-19 vaccine. Each profile displays five attributes – the recipients' vulnerability to the virus, likely transmission of the virus, income, occupation, and age – and all levels are completely randomly assigned. Subjects make a total of 8 choices in the experiment. The experiment also recorded *subjects*' country of origin, age, gender, ideology, income, education, hesitancy over vaccination, and measures of their willingness to pay for a vaccine.

The original study, using subgroup analysis, finds consistent AMCEs across all the countries surveyed. Nevertheless, it is reasonable to suspect that these AMCEs may mask heterogeneity with respect to individual-level covariates. This experiment is particularly suited to a study of heterogeneous effects, since with approximately 250,000 observations in total and harmonised covariate information across countries, there is ample data to model complex relationships. We train a BART model on all five conjoint attributes and the set of covariate information for each profile using **cjbart**, using all observations from the 13 countries surveyed in the experiment. We recover a schedule of IMCE estimates for each attribute-level.

With multiple covariates, however, systematically identifying the drivers of heterogeneity is difficult. This is particularly acute in the case of conjoint experiments where we have separate IMCE vectors for each attribute-level, which means researchers are faced with a dense schedule of predicted effects. We address this challenge by using the tree-based measure of variable importance, as discussed in Section 3.1.

We use our proposed VIMP tool as the first step in identifying plausible sources of heterogeneity in the schedule of IMCEs estimated from the Duch et al. (2021). The method estimates a standardised importance score for each combination of the 10 covariates and 16 attribute-levels in the conjoint design. Figure 4 provides a graphical summary of how well each covariate predicts the attribute-levels in the Duch et al. (2021) conjoint. Clearly, the country of a respondent is a highly predictive factor across most attribute-levels in the model. This is perhaps unsurprising, given the diversity of contexts considered and differing levels of COVID-19 infections at the point the experiment was fielded.

Most interestingly, some subject-level variables appear to condition IMCEs for specific attributes. For example, while subjects' age is not a particularly important predictor of heterogeneity across most attributes, it is very predictive when considering the age of the potential vaccine recipient. In general, this suggests that whether one is willing to prioritise individuals based on age may well be driven by one's own age (which we explore in more detail below), and second that this is perhaps most important for the 65 year old label where the risks of COVID-19 are more severe. Similarly, ideology appears particularly important when partitioning the IMCEs related to the potential vaccine recipient's income. This result accords with conventional expectations about the relationship between political ideology and service provision, and highlights that one's own ideological position appears to predict how willing one is to prioritise those on low incomes.

Given the results from the VIMP summary measure, we can use a single pruned decision tree (as described in Section 3.2) to inspect this heterogeneity in more detail. On the basis of the variable importance heatmap in Figure 4, for example, we would expect that subjects' age is used to partition the IMCE vectors for prioritising vaccine recipients of different ages.

Figure 4. Variable importance matrix having estimated separate random forest models on each attribute-level in the model. Higher values indicate variables that were more important in terms of predicting the estimated IMCE distribution

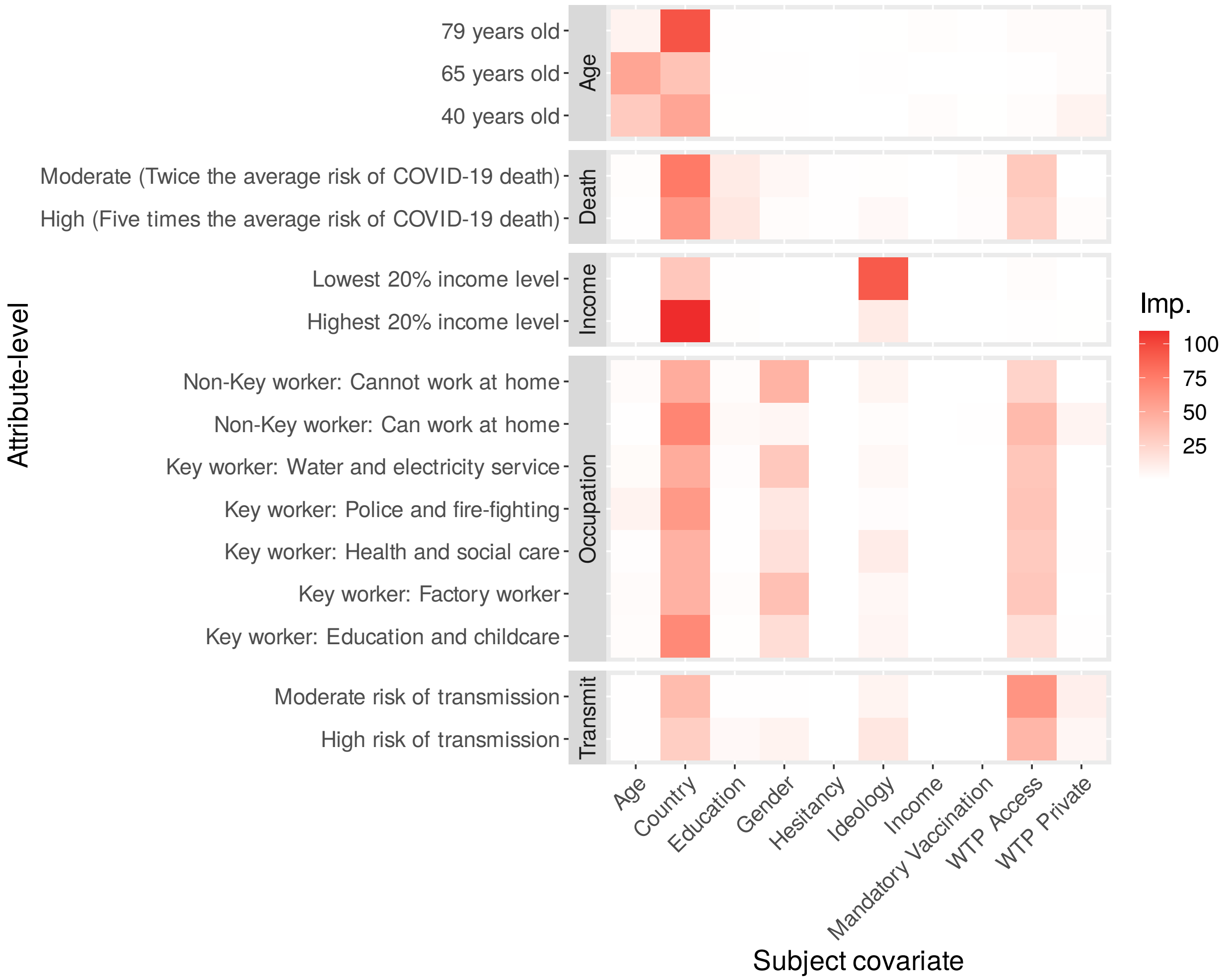
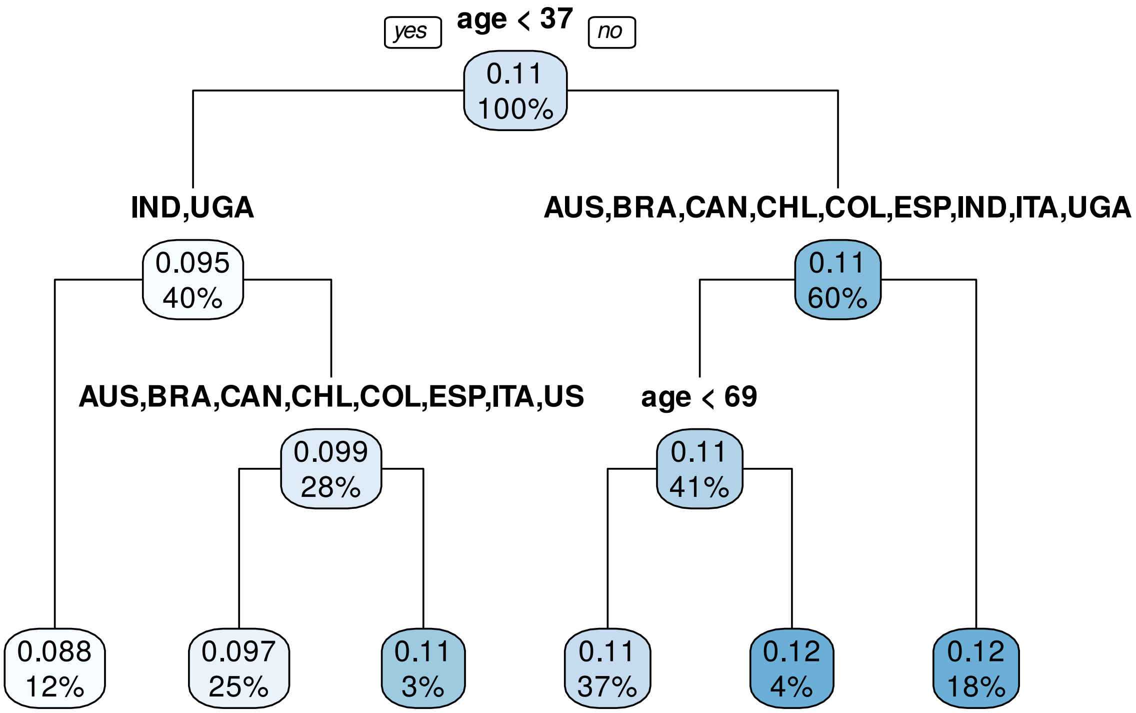


Figure 5 presents a single decision tree for the IMCEs related to prioritising vaccines for “65 year olds". Note first that the split confirms the VIMP analysis results in Figure 4 that identify subject’s age as an important source of heterogeneity for this attribute-level: older subjects (over the age of 37) exhibit a predicted average marginal effect (0.11) that is about 20 percent larger than younger subjects. Notably, moreover, this partitioning strategy captures more complex interactions between covariates.The smallest IMCEs are defined by younger subjects (< 37) in India and Uganda. Conversely, the strongest effects

Figure 5. Pruned decision-tree of predicted IMCEs for prioritising vaccines for those "65 years old", using subject-level covariate information to partition the vector of individual-level effects.

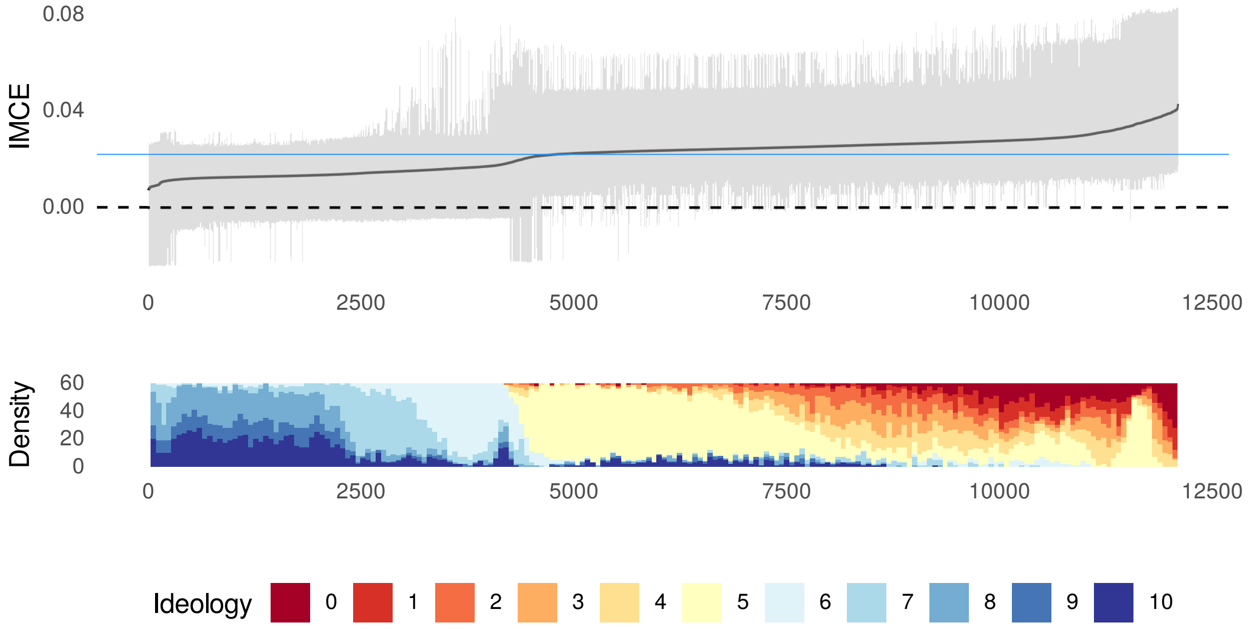


are for those subjects older than 37 resident in the UK, US, and France (countries with older-aged populations), and those resident in other countries who are above the age of 69 (and thus closest in age to the profile age).

Finally, we demonstrate one further way of summarizing these results visually by plotting the full ordered distribution of IMCEs for a given variable against the corresponding distribution of a covariate. Figure 4 suggests that subjects' ideology is an important predictor of IMCEs for the income-related attribute-levels in the conjoint experiment. In Figure 6, therefore, we visualize this particular relationship by plotting the IMCEs against a histogram of subjects' self-reported ideological position.

As Figure 6 shows, there is quite clear and distinct heterogeneity. Smaller IMCEs

Figure 6. Comparison of IMCEs for the "Lowest 20% income level" attribute-level ordered from smallest to largest and corresponding histogram of individuals' self-reported ideology.



The grey ribbon indicates 95% credible intervals for the IMCEs, and the blue line in the top panel indicates the estimated AMCE

(around the 0.01 mark) are individuals whose ideology is right-leaning (at or above 6 on a 0-10 scale). In contrast, larger IMCEs are predicted for those who are typically more left-leaning. Clearly, however, ideology does not play a perfect role. Within these two portions of the distribution, varying degrees of ideology are more uniformly distributed, and at the very right of the IMCE distribution other factors appear to drive a further uptick in the predicted IMCE, to approximately four times the effect size of right-leaning subjects.[16]

[16] To check for overfitting, we re-estimated these models using smaller random subsets of the data. Appendix Figure G2 demonstrates that despite fewer observations these models also identified similar correlations between the income IMCEs and subjects' ideology, with left-leaning subjects typically having higher AMCEs on average (and *vice versa*).While evidence of overfitting is low, there is some evidence of sensitivity to the training batch: the fourth and fifth random subsets have higher predicted effects overall, and the distribution of IMCEs spikes upwards more for batch 3 at the left ideological extreme compared to the other training batches. Separately, in Appendix Figure G3, we show an example from the same model of an attribute-level where there is no apparent correlation between ideology and the substantial heterogeneity observed in the IMCEs.

Comparison to OLS-based approach To demonstrate the comparative performance of our approach, we also estimate IMCEs using an alternative strategy proposed recently by Zhirkov (2022). In short, this method estimates separate OLS regression models for each subject separately. The resultant coefficients are unbiased estimates of the same IMCE quantity we outline in Section 1.

Our method finds a strong correlation between individuals' ideology and the predicted IMCEs for the low income attribute-level of the Duch et al. (2021) experiment. Under the Zhirkov (2022) OLS strategy, we expect to see a similar result – both in terms of the distribution of IMCEs and its correlation with individuals' self-reported ideology.

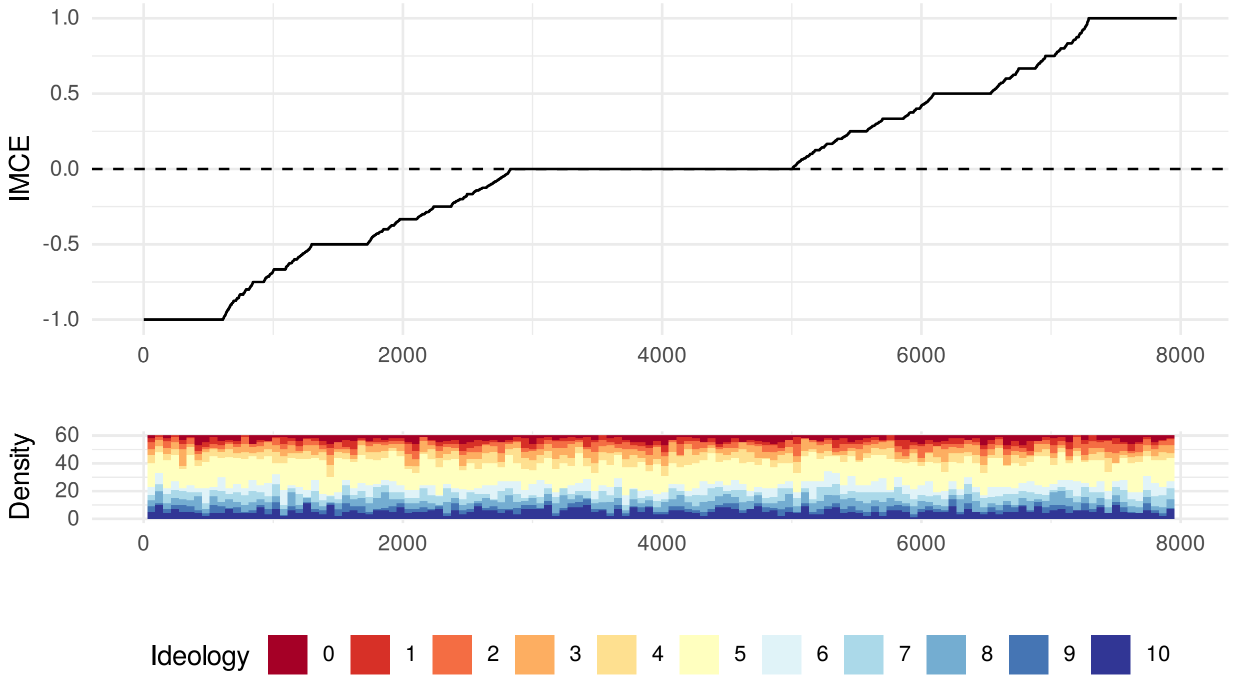
Two practical features of the regression approach complicate this analysis using OLS. Since each subject completed eight rounds of the conjoint experiment (a number we think is quite typical for a conjoint design), each model has only 16 observations (2 profiles per round) and thus the individual models will be imprecise. Zhirkov (2022) acknowledges this limitation, and notes that the OLS approach requires subjects to rate closer to 30 profiles in total. While this large number of activities may be feasible in principle, we rarely see this number of profiles in practice.

Moreover, even if the number of observations approaches 30, Zhirkov (2022) recommends using interval rating scales rather than the binary, forced-choice outcome. While many conjoint experiments implement both rating and forced-choice scales of measurement, we believe the forced-choice outcome is the most interesting aspect. It allows us to think of the effects directly in terms of marginal probabilities, and thus to consider the behaviour of subjects (a choice of candidate) rather than just an attitude (the subjects' rating of two candidates).[17]

[17] This is consistent with the Bansak et al. (2022) finding that the estimated AMCEs from forced-choices of political candidates map well to actual election outcomes. Moreover, in Section C4 of the Appendix we present simulation evidence that even when we adapt designs to meet this requirement, heterogeneity in preferences is less well detected using interval rating scales.

Figure 7 displays the ordered distribution of estimated IMCEs using this OLS strategy, plot against a histogram of individuals' self-reported ideology. The OLS approach yields 5,369 IMCE estimates outside of the range of possible changes in probability. We exclude these estimates from our analysis, leading to a 34 percent reduction in the number of IMCEs we can inspect.[18] We do not observe the same correlation as in our BART estimation. The correlation coefficient between the IMCEs and ideology in the OLS case is negligible and statistically insignificant ($r = -0.01,\ p = 0.20$) compared to a strong correlation with respect to BART ($r = -0.75,\ p < 0.001$). The OLS strategy does not seem to have modelled the data well: The distribution of IMCEs is symmetric centred on zero with tails that contain implausibly large effects.

Figure 7. Comparison of estimated IMCEs using the OLS method proposed in Zhirkov (2022), on the "Lowest 20% income level" attribute-level within Duch et al. (2021)



While these are not ideal conditions for the Zhirkov (2022) approach, our vaccine ex-

[18] Of these individuals, only 5935 uncertainty estimates were parametrically recoverable.

periment resembles a typical conjoint design with 16 observations per individual. Our OLS comparison confirms Zhirkov's (2022) recommendation that the OLS method should only be implemented for conjoints with at least 30 observations per individual. An advantage of our ML-based approach is that it leverages all observations in the data and hence our estimation strategy is less reliant on having many observations per experimental subject.

Perhaps most importantly, our approach is able to detect and capture how subject covariate information modifies the size and direction of these marginal component effects. The OLS method rests on the fact that this heterogeneity is implicitly detected when the marginal effects are modelled for each individual separately. In our proposed method, since the trees in the BART model can identify interactive effects between the supplied covariates and the attribute-levels, it can model these effect modifiers. The result, in this case, is that our method identifies the correlation between subjects' ideology and their treatment of low-income vaccine recipients in a way that the OLS strategy does not.

5 Discussion and Conclusion

The attraction of conjoint experiments is a rich data generating process that allows us to tease out the choice characteristics that shape individuals' decision making. Conjoint experiments are fast becoming one of the dominant methods within the social sciences. Alongside this rise in use, a rich methodological literature is developing that explores how advances in conjoint estimation can enhance its informative value (Jenke et al. 2021; Ham et al. 2022; Goplerud et al. 2022; de la Cuesta et al. 2022).

We make a small contribution to this wider development, by clarifying how the conjoint design relates to the structure of the data collected, and how we can leverage the nature of this data generation to estimate heterogeneous treatment effects across conjoint attributes. Heterogeneity can be characterized in terms of a set of nested, causal estimands that correspond to the repeated observations across individuals, rounds, and profiles of the conjoint design. Using machine learning tools, we show how to estimate heterogeneous treatment effects in the conjoint design using the potential outcomes framework. Our strategy allows researchers to assess treatment effect heterogeneity in a straightforward and flexible manner.

We suggest that machine learning is particularly useful given its ability to identify more complicated relationships between predictor variables without the need for researchers to specify these *a priori*. By reducing researcher degrees of freedom, our proposed general method provides a more robust means of analysing heterogeneity compared to *ad hoc* subgroup analyses. Moreover, since our estimation strategy leverages all observations in the modelling stage, our method has greater statistical power than approaches that rely on estimating separate subset models.

As a consequence, when researchers use forced-choice outcomes, have relatively few observations per subject, or many substantively important values per attribute, we believe the BART method is more appropriate than OLS-based alternatives. There are, however, a smaller subset of conjoint designs – where the outcome is a rating, there are many profile-observations per subject, and few substantively important values per attribute – when either OLS or BART strategies are appropriate. In these cases, researchers should consider the trade offs that both methods entail.

Therefore, notwithstanding its advantages, there are also limitations to our estimation strategy. Principally, our BART modelling strategy assumes that conditional on the observed covariates, outcomes depend only on the assigned treatment values. In other words, two individuals with identical covariate profiles and the same attribute-level assignments would get assigned the same predicted OMCEs. This is, in part, a limitation of the underlying BART algorithm, with limited development of cluster-specific estimation. Future research may wish to implement recent advances in random intercept modelling to better capture these latent effects (see Tan et al. 2018).

More generally, and as with many ML methods, overfitting the training data can lead to poor out-of-sample generalisability. As we have pointed to, researchers can assess these issues by, for example, estimating separate ML models on subsets of the data to ensure the findings replicate. Even when the model is not overfit, these methods can be quite sensitive to the data: particularly with smaller training sets. Moreover, ML models can be sensitive to the choice of hyperparameter values. As we note earlier, we chose BART because of its greater resilience to these issues: BART predictions are relatively stable over hyperparameter choices, and the Bayesian priors provide strong regularisation to prevent overfitting (Chipman et al. 2010; Hill et al. 2020). Other ML implementations, for example causal forests, offer separate tuning and validation algorithms, and we recommend researchers take these procedures seriously.

To accompany this paper, we provide a new `R` package, **cjbart**, that allows researchers to use our method on their experimental conjoint data. However, our proposed meta-strategy could be used with many other forms of modelling. For example, researchers may wish to use random forests or neural networks instead, and we demonstrate one such alternative example in Section E of the Appendix.

Finally, generating individual-level estimates of treatment effects is only half the battle. Once researchers recover these individual-level estimates, the challenge is to identify the most significant sources of heterogeneous treatment effects. We provide two complementary tools that help researchers make sense of the estimated distribution of individual-level effects. We demonstrate how VIMP measures can be used to summarise which variables are most important for predicting heterogeneity in the IMCEs. We then show how single regression tree models can be used to partition IMCE distributions into clusters, where the decision rules provide information about which covariates define those clusters. This paper also shows how these results can be visualized to aid analysis.

References

Abramson, Scott F. , Korhan Kocak, Asya Magazinnik, and Anton Strezhnev (2020, July). Improving preference elicitation in conjoint designs using machine learning for heterogeneous effects.

Athey, Susan , Julie Tibshirani, and Stefan Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Ballard-Rosa, Cameron , Lucy Martin, and Kenneth Scheve (2017). The structure of american income tax policy preferences. *The Journal of Politics 79*(1), 1–16.

Bansak, Kirk , Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto (2021). *Conjoint Survey Experiments*. Cambridge University Press.

Bansak, Kirk , Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto (2022). Using conjoint experiments to analyze election outcomes: The essential role of the average marginal component effect. *Political Analysis*, 1–19.

Breiman, L. , J. Friedman, C.J. Stone, and R.A. Olshen (1984). *Classification and Regression Trees*. Taylor & Francis.

Chipman, Hugh A. , Edward I. George, and Robert E. McCulloch (2010). Bart: Bayesian additive regression trees. *Annals of Applied Statistics 4*(1), 266–298.

Chou, Winston , Rafaela Dancygier, Naoki Egami, and Amaney A. Jamal (2021). Competing for loyalists? how party positioning affects populist radical right voting. *Comparative Political Studies 54*(12), 2226–2260.

de la Cuesta, Brandon , Naoki Egami, and Kosuke Imai (2022). Improving the external validity of conjoint analysis: The essential role of profile distribution. *Political Analysis 30*(1), 19–45.

Duch, Raymond , Denise Laroze, Thomas Robinson, and Pablo Beramendi (2020). Multimodes for detecting experimental measurement error. *Political Analysis 28*(2), 263–283.

Duch, Raymond , Laurence S. J. Roope, Mara Violato, Matias Fuentes Becerra, Thomas S. Robinson, Jean-Francois Bonnefon, Jorge Friedman, Peter John Loewen, Pavan Mamidi, Alessia Melegaro, Mariana Blanco, Juan Vargas, Julia Seither, Paolo Candio, Ana Gibertoni Cruz, Xinyang Hua, Adrian Barnett, and Philip M. Clarke (2021). Citizens from 13 countries share similar preferences for covid-19 vaccine allocation priorities. *Proceedings of the National Academy of Sciences 118*(38).

Duch, Raymond M , Denise Laroze, Constantin Reinprecht, and Thomas S Robinson (2020). Nativist policy: the comparative effects of trumpian politics on migration decisions. *Political Science Research and Methods*, 1–17.

Goplerud, Max , Kosuke Imai, and Nicole E. Pashley (2022). Estimating heterogeneous causal effects of high-dimensional treatments: Application to conjoint analysis.

Green, Donald P. and Holger L. Kern (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public Opinion Quarterly 76*(3), 491–511.

Hahn, P. Richard , Jared S. Murray, and Carlos M. Carvalho (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis 15*(3), 965 – 2020.

Hainmueller, Jens , Daniel J. Hopkins, and Teppei Yamamoto (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis 22*(1), 1–30.

Ham, Dae Woong , Kosuke Imai, and Lucas Janson (2022). Using machine learning to test causal hypotheses in conjoint analysis. *arXiv preprint arXiv:2201.08343*.

He, Jingyu , Saar Yalov, and P. Richard Hahn (2019, 16–18 Apr). Xbart: Accelerated bayesian additive regression trees. In K. Chaudhuri and M. Sugiyama (Eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, Volume 89 of *Proceedings of Machine Learning Research*, pp. 1130–1138. PMLR.

Hill, Jennifer , Antonio Linero, and Jared Murray (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application 7*(1).

Hill, Jennifer L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics 20*(1), 217–240.

Ishwaran, Hemant and Min Lu (2019). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in medicine 38*(4), 558–582.

Jenke, Libby , Kirk Bansak, Jens Hainmueller, and Dominik Hangartner (2021). Using eye-tracking to understand decision-making in conjoint experiments. *Political Analysis 29*(1), 75–101.

Kapelner, Adam and Justin Bleich (2016). bartmachine: Machine learning with bayesian additive regression trees. *Journal of Statistical Software 70*(4), 1–40.

Künzel, Sören R. , Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences 116*(10), 4156–4165.

Leeper, Thomas J. , Sara B. Hobolt, and James Tilley (2020). Measuring subgroup preferences in conjoint experiments. *Political Analysis 28*(2), 207–221.

Sparapani, Rodney , Charles Spanbauer, and Robert McCulloch (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software 97*(1), 1–66.

Spilker, Gabriele , Vally Koubi, and Tobias Böhmelt (2020, 07). Attitudes of urban residents towards environmental migration in kenya and vietnam. *Nature Climate Change 10*.

Strobl, Carolin , Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics 8*(1), 1–21.

Tan, Yaoyuan Vincent , Carol AC Flannagan, and Michael R Elliott (2018). Predicting human-driving behavior to help driverless vehicles drive: random intercept bayesian additive regression trees. *Statistics and Its Interface*, 557—572.

Wager, Stefan and Susan Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Zhirkov, Kirill (2022). Estimating and using individual marginal component effects from conjoint experiments. *Political Analysis 30*(2), 236–249.

Bibliographical Statement

Thomas Robinson is an Assistant Professor at the Department of Methodology, London School of Economics and Political Science, UK, WC2A 2AE.

Raymond Duch is the co-founder and Director of the Centre for Experimental Social Sciences (CESS) at Nuffield College, University of Oxford, UK, OX1 1NF.

Acknowledgements

We would like to thank Mats Ahrenshop, Thomas Plümper, and Kosuke Imai, as well as participants at the Asian PolMeth 2023 Annual Meeting (Singapore) for their feedback on drafts of this paper. We acknowledge the Health Economics Research Centre at the University of Oxford Nuffield Department of Population Health for their support of the CANDOUR project. We also thank our anonymous reviewers for their feedback.