



# Treatment Effect Heterogeneity in Randomized Field Experiments: A Methodological Comparison and Public Policy Implications

Yixing Chen, Shrihari Sridhar , and Vikas Mittal

Online supplement: <https://doi.org/10.1177/07439156211032751>

Many public policy studies (Martin and Scott 2021) use randomized field experiments for drawing causal conclusions (e.g., Chen et al. 2020). A typical randomized field experiment involves a control group and a treatment group to which individual units (e.g., consumers, patients) are randomly assigned, after which an intervention (e.g., a marketing program) is implemented in the treatment group. To assess the efficacy of an intervention, researchers typically estimate the average treatment effect, which is computed as the mean difference in the outcome between the units in the treatment group and the control group. When applying the results of a randomized experiment, it is assumed that the treatment effect within the manipulated condition is the same for all the units assigned to the treatment condition. This may not always be the case, as the treatment might have differential causal effects on different subgroups (subgroup differences). More formally, this variation is called treatment effect heterogeneity. For example, the treatment effect may differ for men and women or for those who have different types of insurance coverage. By accounting for treatment effect heterogeneity, public policy researchers can get a better and more nuanced understanding of the efficacy of an intervention. Specifically, they can ascertain how the treatment effect may also vary across units based on characteristics that are not manipulated in the experiment. We discuss three prominent approaches to account for treatment effect heterogeneity: analysis of variance (ANOVA)/regression with covariates and moderators, a random-coefficients model, and causal forests (see Table 1). We illustrate each approach using a simulated version of the data in Chen et al. (2020). Our goal is to provide a practical understanding of each approach, with a focus on causal forests.

## Screening Completion for Liver Cancer: A Stylized Example

People at risk for liver cancer or hepatocellular carcinoma (HCC) should undergo semiannual screening to facilitate

early detection, which can be lifesaving. Due to low screening rates, health care institutions invest in patient outreach programs to encourage and increase screening rates among at-risk populations. To preserve confidentiality, we simulated the data for a hypothetical experiment that broadly follows the randomized experiment in Chen et al. (2020).

This article focuses on two randomized conditions: the control group (no-outreach condition [ $n = 600$ ]) and the treatment group (outreach condition [ $n = 600$ ]). In the control group, patients received visit-based HCC screening as recommended by primary or specialty care providers and were not contacted by anyone else. In the treatment group (outreach), patients were mailed a one-page letter describing (1) the risk of HCC in patients with cirrhosis, (2) the benefits and risks of HCC screening tests, (3) a summary of the screening procedure, and (4) a recommendation to the patient to make an appointment for an ultrasound.

The dependent variable was simulated as the probability of a patient completing the screening (a continuous variable ranging from 0 to 1 denoting screening probability). The basic dummy-variable regression equation used to estimate the average treatment effect (or main effect) of the outreach intervention is written as

$$\text{Screening}_i = \beta_0 + \beta_1 \text{Outreach}_i + \varepsilon_i. \quad (1)$$

In Equation 1,  $\text{Screening}_i$  is a continuous outcome variable that indicates the probability of the patient  $i$  completing the screening test.  $\text{Outreach}_i$  refers to the intervention, whether the patient received the outreach intervention (1 = yes; 0 = the patient was in the control group). The estimate of  $\beta_1$  provides the causal effect of the outreach intervention. A simple dummy-variable

Yixing Chen is Assistant Professor of Marketing, Mendoza College of Business, University of Notre Dame, USA (email: [ychen43@nd.edu](mailto:ychen43@nd.edu)). Shrihari Sridhar is Professor of Marketing and Joe Foster '56 Chair in Business Leadership, Mays Business School, Texas A&M University, USA (email: [ssridhar@mays.tamu.edu](mailto:ssridhar@mays.tamu.edu)). Vikas Mittal is J. Hugh Liedtke Professor of Marketing, Jones Graduate School of Business, Rice University, USA (email: [vmittal@rice.edu](mailto:vmittal@rice.edu)).

**Table 1.** A Methodological Comparison: ANOVA/Regression, Random-Coefficients Model, and Causal Forests.

	<b>ANOVA/Regression with Covariates and Moderators</b>	<b>Random-Coefficients Model</b>	<b>Causal Forests</b>
Goal of including additional heterogeneity-identifying factors	Hypothesis testing using covariates and moderators.	Separately estimate the slope for each subgroup as a deviation from the average treatment effect.	Separately estimate coefficients for individuals to capture complex interactions and detect unexpected heterogeneity.
Philosophy	<i>Top-down.</i> Compare mean of the outcome variable in the treatment group with control group, and use moderator variables to uncover subgroup differences.	<i>Somewhat top-down.</i> Compare mean of the outcome variable in the treatment group with control group and use random slopes to uncover subgroup differences.	<i>Bottom-up.</i> First obtain the treatment effect for every unit, and then relate individual-level treatment effect estimates to covariates to understand subgroup differences.
Theoretical equation	$Y_i = \beta_0 + \beta_1 W_i + \beta_2 X_i + \beta_3 W_i \times X_i + \varepsilon_i,$ where $Y$ = outcome variable, $W$ = treatment dummy, and $X$ is a vector of moderators.	$Y_{ij} = \beta_{0j} + \beta_{1j} W_{ij} + \varepsilon_{ij},$ $\beta_{0j} = \alpha_0 + \gamma_{0j},$ $\beta_{1j} = \alpha_1 + \gamma_{1j}.$ The data are stacked into two levels $i$ (patient, level 1) and $j$ (moderator, level 2), respectively; $\beta_{0j}$ is the random intercept, and $\beta_{1j}$ is the random slope. $\beta_{1j}$ can be specified as a function of other covariates $X$ to capture higher-order interactions.	$\hat{\tau}_i = \frac{\sum_{j=1}^n \alpha_j (Y_i - \hat{m}_i)(W_i - \hat{e}_i)}{\sum_{j=1}^n \alpha_j (W_i - \hat{e}_i)^2},$ where $\hat{\tau}_i$ is the heterogeneous treatment effect, $Y$ = outcome variable, $W$ = treatment dummy, $\alpha_i$ is patient-level weight, $\hat{e}_i$ refers to the estimates of the propensity score of receiving the treatment $e(x) = P[W_i   X_i = x]$ , and $\hat{m}_i$ refers to the estimates of the expected outcome marginalizing over treatment $m(x) = P[Y_i   X_i = x]$ .
Example equation	$\text{Screening}_i = \beta_0 + \beta_1 \text{Outreach}_i + \beta_2 \text{Female}_i + \beta_3 \text{Outreach}_i \times \text{Female}_i + \varepsilon_i.$ Screening is a continuous outcome variable that indicates the probability of the patient $i$ completing the screening test. Outreach <sub><math>i</math></sub> refers to whether the patient received the outreach intervention ( $1 = \text{yes}$ ; $0 = \text{the patient was in the control group}$ ), and Female <sub><math>i</math></sub> refers to the gender of the patient with male being the baseline ( $1 = \text{female}$ , $0 = \text{male}$ ).	$\text{Screening}_{ij} = \beta_{0j} + \beta_{1j} \text{Outreach}_{ij} + \varepsilon_{ij}$ $\beta_{0j} = \alpha_0 + \gamma_{0j}$ $\beta_{1j} = \alpha_1 + \gamma_{1j}.$ Screening is a continuous outcome variable that indicates the probability of the patient $i$ completing the screening test. Outreach <sub><math>i</math></sub> refers to whether the patient received the outreach intervention ( $1 = \text{yes}$ ; $0 = \text{the patient was in the control group}$ ).	$\hat{\tau}_i = \frac{\sum_{j=1}^n \alpha_j (\text{Screening}_i - \hat{m}_i)(\text{Outreach}_i - \hat{e}_i)^2}{\sum_{j=1}^n \alpha_j (\text{Outreach}_i - \hat{e}_i)^2},$ where Screening is a continuous outcome variable that indicates the probability of the patient $i$ completing the screening test. Outreach <sub><math>i</math></sub> refers to whether the patient received the outreach intervention ( $1 = \text{yes}$ ; $0 = \text{the patient was in the control group}$ ).

(continued)

**Table 1.** (continued)

	<b>ANOVA/Regression with Covariates and Moderators</b>		<b>Random-Coefficients Model</b>	<b>Causal Forests</b>
Subgroup-based treatment effects	Yes	Yes	Yes	Yes
Individual-level treatment effects	No	Possible (with longitudinal data)	Possible (with longitudinal data)	Yes
Stage when subgroups are identified	Before analysis (preregistered in trials) or during analysis	During model specification	During model specification	After estimating individual-level treatment effects
Ability to handle a large number of covariates and interactions	Low	Medium	Medium	High
Ability to detect unexpected heterogeneity	Low	Low	Low	High
Strengths	<ul style="list-style-type: none"> <li>• Simplicity</li> <li>• Ability to test hypotheses</li> </ul>	<ul style="list-style-type: none"> <li>• Ability to exploit the variation at different levels in the data</li> <li>• Treatment effect heterogeneity can be uncovered by the variance in the random slope term without specifying the interaction terms</li> </ul>	<ul style="list-style-type: none"> <li>• Ability to obtain individual-level treatment effects</li> <li>• Ability to capture complex interactions</li> <li>• Ability to detect unexpected heterogeneity</li> </ul>	
Weaknesses	<ul style="list-style-type: none"> <li>• Potential for high multicollinearity as the number of interactions increases</li> <li>• Model instability with increased covariates</li> <li>• Low ability to detect unexpected heterogeneity</li> </ul>	<ul style="list-style-type: none"> <li>• Potential for high multicollinearity as the number of interactions increases</li> <li>• Model convergence as the number of covariates and levels increases</li> <li>• Low ability to detect unexpected heterogeneity</li> </ul>	<ul style="list-style-type: none"> <li>• Output is difficult to interpret with respect to the sources of heterogeneity</li> </ul>	

regression (Column 1, Web Appendix Table A1) shows the average treatment effect is positive and statistically significant ( $\hat{\beta}_1 = .181, p < .001$ ). Practically speaking, the outreach intervention increases the screening probability by 18.1 percentage points compared with the control group (i.e., the mean difference in screening probability between the patients in the outreach condition [.439] and those in the control condition [.258]).<sup>1</sup> Stopping the analysis here assumes that the treatment effect is the same for all 600 patients assigned to the outreach condition.

The three approaches described in Table 1 can be used to determine if the benefit of outreach differs for subgroups such as for male versus female patients or patients with different types of insurance coverage. The three approaches are ANOVA/regression with covariates and moderators, random-coefficients model, and causal forests. The goal is to ascertain if the average treatment effect should be adjusted downward or upward in different subgroups. The first two approaches have been widely used in marketing and are described in the Web Appendix.

This commentary focuses on causal forests, a technique with a different conceptual focus. ANOVA/regression and random-coefficients models are *top-down* ways to think about treatment effect heterogeneity. They start by comparing the mean of the outcome variable in the treatment group with the mean of the outcome variable in the control group to get the average treatment effect. Next, they uncover how the main effect may vary among subgroups. ANOVA/regression uses interaction terms to dissect the main effect by various subgroups (e.g., gender, insurance type, a combination of gender and insurance type), while the random-coefficients models use a combination of data stacking, random slopes, and interactions to dissect the main effects.

Top-down approaches are feasible and efficient when the researcher is interested in testing how the main effect changes across a relatively small number of moderating conditions (e.g., two to five) or has a priori ideas about the moderators. However, in many field experiments, researchers may have several dozen subgroup variables and a relatively small sample size. For example, the researcher may have more than 100 variables from the electronic medical records and census data based on a patient's geographic location (e.g., household income, retail growth, unemployment rate). Including these covariates as moderators in the regression or random-coefficients model is not feasible as the model will run out of degrees of freedom. Moreover, the researcher may not have any basis to a priori specify theory-driven moderators in the model.

A *bottom-up approach* such as causal forests uncovers treatment effect heterogeneity with a large number of subgroup variables. This involves obtaining the treatment effect estimate for each unit in the sample and then relating individual-level

treatment effects to a variety of covariates to understand how high-treatment-effect individuals differ from low-treatment-effect individuals. Next, we describe the steps involved in a causal forests approach.

## Causal Forests: Key Steps

### Step 1. Obtaining Individual-Level Treatment Effects

To obtain the treatment effect for each of the 600 patients in the control group, the researcher needs an approach to estimate the lift in screening probability for every patient in the control group if they were instead placed in the treatment group. This is inherently an error-prone prediction, as a patient who was placed in the control group could not have simultaneously been placed in the treatment group. The researcher also needs an approach to estimate the lift in screening probability for every patient in the treatment group as if they were instead placed in the control group. This is again an error-prone prediction, because a patient who was placed in the treatment group could not have simultaneously been placed in the control group.

How can the researcher obtain the treatment effect for every patient in the sample? A starting point is to look at the mean difference in the outcome between the patients in the treatment group and those in the control group within a particular subgroup (e.g., women); that is,

$$\tau_i = \text{Avg}[\text{Screening}_{i,W=1} | \text{Female}_i = 1] - \text{Avg}[\text{Screening}_{i,W=0} | \text{Female}_i = 1], \quad (2)$$

where  $i$  indexes a patient,  $W$  represents the condition where  $W=1$  represents the outreach condition and  $W=0$  represents the control condition, and  $\text{Female}_i$  is a dummy variable indicating whether the patient is female (coded as 1) or male (coded as 0).

Drawing on this logic, the researcher can define a more fine-grained subgroup for every patient  $i$  by using more patient characteristics to split the sample into subgroups. If we have  $m$  covariates labeled  $X_1$  to  $X_m$ , we could write Equation 2 as

$$\tau_i = \text{Avg}[\text{Screening}_{i,W=1} | X_1 = c_1, X_2 = c_2, \dots, X_m = c_m] - \text{Avg}[\text{Screening}_{i,W=0} | X_1 = c_1, X_2 = c_2, \dots, X_m = c_m], \quad (3)$$

where  $i$  and  $W$  are as previously discussed. We define the subgroups by choosing a set of cutoff values  $c_1, c_2, \dots, c_m$  for  $X_1, X_2, \dots, X_m$ , respectively. Assuming  $X_1$  is gender (= 1 if female),  $X_2$  is insurance type (coded as 1, 2, ..., 10), and  $X_3$  is age (continuous variable from 21 to 90), an exemplary treatment effect for patient  $i$  could be obtained as

$$\tau_i = \text{Avg}[\text{Screening}_{i,W=1} | X_1 = 0, X_2 = 1, X_3 > 65] - \text{Avg}[\text{Screening}_{i,W=0} | X_1 = 0, X_2 = 1, X_3 > 65]. \quad (4)$$

Equation 4 calculates the mean difference in screening outcomes for male patients over 65 years old with Insurance Type 1 in the control group and their counterparts in the treatment group. Compared with Equation 2, Equation 4 allows us

<sup>1</sup> These simulated results follow Chen et al. (2020), where the average screening probability is 25% (45%) in the control (outreach) condition in Period 1.

to get a more fine-grained treatment effect estimate. Yet, as we add more covariates, we introduce more subjectivity into the choices of variables as well as cutoff values and might create many subgroups with zero observation. Therefore, the challenge is to choose the relevant covariates and associated cutoffs while maintaining the ability to capture the heterogeneity driven by the combination of covariates.

The causal forest algorithm addresses this challenge by forming data-driven subgroups. The algorithm splits the data into subgroups that share a similar profile of patient characteristics and uses the within-subgroup treatment effect as the estimate for any patient who belongs to the corresponding subgroup. To reduce bias, it uses one part of the sample to determine the subgroups and the other part to estimate the treatment effects. To reduce variance of the treatment effect estimates, it repeats the procedure over many random draws of the sample and averages the estimates. This algorithm is further developed into generalized random forests, where the trees are not used to compute the treatment effect estimates but to create individual-specific weights (Athey, Tibshirani, and Wager 2019). Concretely, for the set of independent and identically distributed patients, indexed  $i = 1, \dots, n$ , we observe the outcome of interest  $Y_i$  (screening probability), treatment assignment  $W_i$ , and a vector of patient characteristics  $X_i$  (e.g., gender and insurance type). The patient-level treatment effect estimate  $\hat{\tau}_i$  is

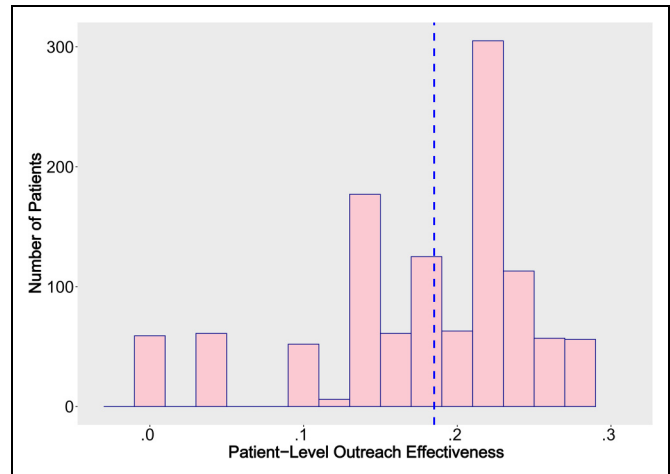
$$\hat{\tau}_i = \frac{\sum_{i=1}^n \alpha_i (Y_i - \hat{m}_i)(W_i - \hat{e}_i)}{\sum_{i=1}^n \alpha_i (W_i - \hat{e}_i)^2}, \quad (5)$$

where  $\alpha_i$  is patient-level weight,  $\hat{e}_i$  refers to the estimates of the propensity score of receiving the treatment  $e(x) = P[W_i|X_i = x]$ , and  $\hat{m}_i$  refers to the estimates of the expected outcome marginalizing over treatment  $m(x) = P[Y_i|X_i = x]$ . For details, see Athey, Tibshirani, and Wager (2019). The upshot is that the researcher can use the algorithm to obtain the treatment effect for every individual with valid confidence intervals.

As an illustration, Figure 1 shows the distribution of the patient-level treatment effect estimates,  $\hat{\tau}_i$ , based on the simulated data (i.e., 600 patients in the control condition and 600 patients in the treatment condition). First, we see that the average treatment effect is .185, which is close to the results from the main effect of ANOVA/regression model (.181, Column 1 of Web Appendix Table A1) and that of the random-coefficients models (.182, Column 2 of Web Appendix Table A4). Figure 1 also shows that the patient-level treatment effects vary substantially across patients from 0 percentage points to 31 percentage points.

### Step 2. Discovering Patterns in Treatment Effects Across Individuals

After collecting these treatment effect estimates, we can conduct a second-stage analysis to study how they vary by patient characteristics using a linear regression of  $\hat{\tau}_i$  on  $X_i$  or a subset of  $X_i$ . To conduct a more robust second-stage analysis, Athey and



**Figure 1.** Histogram of patient-level treatment effect estimates. Notes: Blue dotted line represents the average treatment effect.

Wager (2021) recommend an adjustment to  $\hat{\tau}_i$  to improve its precision and robustness to assumptions. For a given  $\hat{\tau}_i$ , the estimate  $\hat{\Gamma}_i$  is given as

$$\hat{\Gamma}_i = \hat{\tau}_i + \frac{W_i - \hat{e}_i}{\hat{e}_i(1 - \hat{e}_i)} \{Y_i - [\hat{m}_i + (W_i - \hat{e}_i)\hat{\tau}_i]\} \quad (6)$$

where  $\{Y_i, W_i, \hat{e}_i, \hat{m}_i\}$  are as previously defined. The estimate  $\hat{\Gamma}_i$  is considered doubly robust because it only requires either the propensity score (i.e.,  $e(x)$ ) or the expected outcome (i.e.,  $m(x)$ ) to be correctly specified. The doubly robust heterogeneous treatment effect estimates vary based on the gender (1 = female, 0 = male) and insurance type (Insurance Type 1 [1 = Insurance Type 1, 0 otherwise],... Insurance Type 10 [1 = Insurance Type 10, 0 otherwise]) as follows:

$$\begin{aligned} \hat{\Gamma}_i = & \beta_0 + \beta_1 \text{Female}_i + \beta_2 \text{InsuranceType2}_i + \beta_3 \text{InsuranceType3}_i \\ & + \beta_4 \text{InsuranceType4}_i + \beta_5 \text{InsuranceType5}_i \\ & + \beta_6 \text{InsuranceType6}_i + \beta_7 \text{InsuranceType7}_i \\ & + \beta_8 \text{InsuranceType8}_i + \beta_9 \text{InsuranceType9}_i \\ & + \beta_{10} \text{InsuranceType10}_i + \varepsilon_i. \end{aligned}$$

Table 2 shows the results from the second-stage analysis, which are largely consistent with those from the first two approaches. Female patients are more responsive to outreach than male patients, and outreach is less (more) effective among patients with Insurance Type 1 than those with Insurance Types 2, 3, 4, 5, 6, 7, and 8 (9 and 10). The second-stage analysis based on causal forests reveals the patterns regarding treatment effects within subgroups in a more formal way.

**Pros and cons of causal forests.** A key advantage of causal forests is the ability to uncover individual-level treatment effects with valid confidence intervals. It also systematically detects unexpected heterogeneity without (1) the need for a larger number of experimental conditions, (2) restrictions on the

**Table 2.** Sources of Treatment Effect Heterogeneity Based on Doubly Robust Estimates.

	Coef.	SE
Female (I = yes)	.073***	(.005)
Insurance Type 2	.026*	(.011)
Insurance Type 3	.118***	(.011)
Insurance Type 4	.023*	(.011)
Insurance Type 5	.056***	(.011)
Insurance Type 6	.103***	(.011)
Insurance Type 7	.078***	(.011)
Insurance Type 8	.037***	(.011)
Insurance Type 9	-.028*	(.012)
Insurance Type 10	-.142***	(.010)
Intercept	.121***	(.008)

\* $p < .05$ .\*\* $p < .01$ .\*\*\* $p < .001$ .

number of covariates or (3) limiting the nature and number of interactions among covariates.

In terms of cons, the output of causal forests is difficult to interpret with respect to the sources of heterogeneity. Causal forests is one of many estimators of heterogeneous treatment effects. Formal guidance for choosing the best estimator in a given context is lacking. This poses two challenges: First, relying on a single method might leave researchers too much freedom to make an arbitrary modeling choice. Second, each method may not perform very well in certain regions of the feature space. Researchers can compare multiple estimators of treatment effect heterogeneity and evaluate whether these estimators agree on the assignment of each individual to gain more confidence in the conclusions (Künzel, Walter, and Sekhon 2019).

## Conclusion

Causal forests is an emerging technique that accounts for treatment effect heterogeneity, in addition to an ANOVA/regression or random-coefficients model. As Table 1 shows, each approach

is slightly different, and no single approach is perfect for incorporating treatment effect heterogeneity. Our larger hope is that the public policy community embraces emerging approaches such as causal forests to provide more nuanced recommendations to policy makers in field experiments related to domains such as nutrition, educational programs, safety training evaluation, sustainability, and donation behavior, among others.

## Special Issue Guest Coeditors

Brennan Davis, Dhruv Grewal, and Steve Hamilton


## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with reor publication of /or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

## ORCID iD

Shrihari Sridhar  <https://orcid.org/0000-0002-6612-1871>

## References

- Athey, Susan, Julie Tibshirani, and Stefan Wager (2019), "Generalized Random Forests," *Annals of Statistics*, 47 (2), 1148–78.
- Athey, Susan and Stefan Wager (2021), "Policy Learning with Observational Data," *Econometrica*, 89 (1), 133–61.
- Chen, Yixing, Ju-Yeon Lee, Shrihari Sridhar, Vikas Mittal, Katharine McCallister, and Amit G. Singal (2020), "Improving Cancer Outreach Effectiveness Through Targeting and Economic Assessments: Insights from a Randomized Field Experiment," *Journal of Marketing*, 84 (3), 1–27.
- Künzel, Sören R., Simon J.S. Walter, and Jasjeet S. Sekhon (2019), "CausalToolbox—Estimator Stability for Heterogeneous Treatment Effects," *Observational Studies* 5, 105–17.
- Martin, Kelly D. and Maura L. Scott (2021), "A Strategic Vision for Rigor, Relevance, and Inclusivity," *Journal of Public Policy & Marketing*, 40 (1), 1–6.

Copyright of Journal of Public Policy & Marketing is the property of American Marketing Association and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.