Society for Causal Inference
EST. 2020

VANDERBILT UNIVERSITY

Evaluating Finite-Sample Properties of Machine Learning Approaches for Assessing Heterogeneity of Treatment Effect in Clinical Trials

L.LEVOIR[1], A. SPIEKER[1,2] and B. BLETTE[2]
1. Vanderbilt University, Nashville, Tennessee 2. Vanderbilt University Medical Center, Nashville, Tennessee

PART I: SETUP

Our goal:

Investigate the **finite-sample properties of popular methods** for estimation and inference of **individualized CATEs**

Through simulation, we will consider:

A range of scenarios and sample sizes (with focus on sample sizes that are more commonly used in clinical trials)

Our desired outcome:

Gain a better **understanding of when one can achieve valid CATE inference using RCT data in practice**.

KEY QUESTIONS

1) How **reliably** can we **detect HTE** in clinical trials?
2) **What sample size is necessary** to expect **valid performance** of different estimators?
3) **Given a sample size**, which method should be chosen for **better performance**?

The ATE exists when there is a mean difference between the treated and the control potential outcomes:

$$ATE = E\left[Y^{(1)} - Y^{(0)}\right] \overset{?}{=} E[Y|W=1] - E[Y|W=0]$$

Association does not imply causation in general, but the design of a RCT can make plausible a set of assumptions under which association and causation can align. We can use these assumptions to identify the CATE, the average treatment effect *conditional* on belonging to a subgroup defined by $\mathbf{x}$.

$$CATE(\mathbf{x}) = E[Y^{(1)} - Y^{(0)}|\mathbf{X} = \mathbf{x}]$$

Looking at a group with Lisa's covariate values
$\{X_1 = x_1, X_2 = x_2, \ldots\}$

Looking at someone with Lisa's covariate values
$\{X_1 = x_1, X_2 = x_2, \ldots\}$

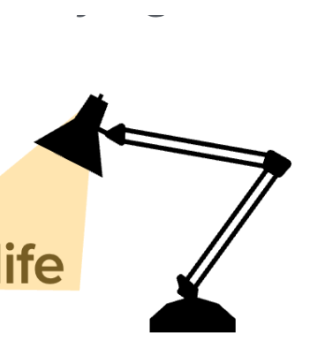
$$CATE_i = E[Y^{(1)} - Y^{(0)}|\mathbf{X} = \mathbf{x_i}]$$

Individualized CATE how we would expect the treatment to affect someone *LIKE* Lisa

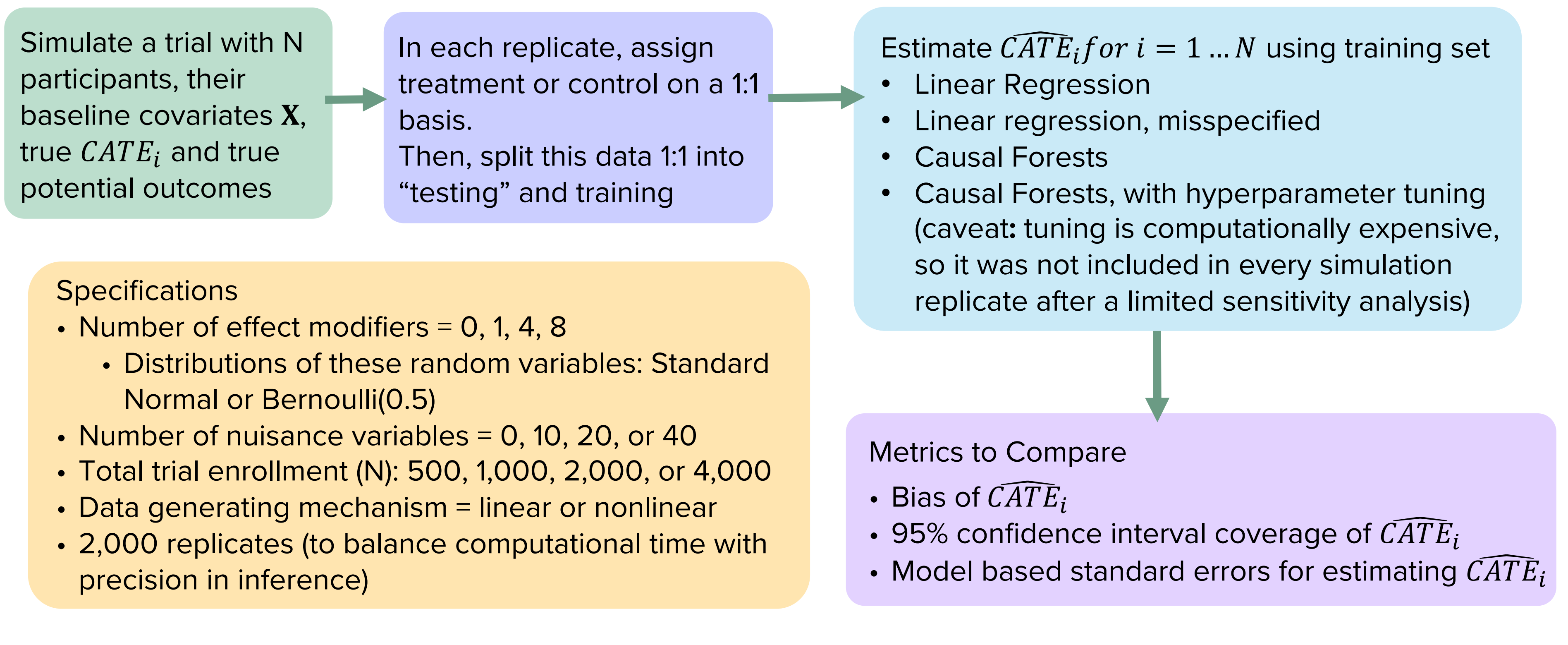
Various methods for estimation and inference of "individualized" CATEs have been proposed with good **asymptotic** properties, but the sample size required for good statistical properties may depend heavily on the underlying data.

Despite this, many practitioners are implementing these methods in clinical trials with smaller or moderate sample sizes where the performance of these methods is not clear.

"asymptopia" vs. real life



SIMULATION



Simulate a trial with N participants, their baseline covariates $\mathbf{X}$, true $CATE_i$ and true potential outcomes

In each replicate, assign treatment or control on a 1:1 basis.
Then, split this data 1:1 into "testing" and training

Estimate $\widehat{CATE}_i for\ i = 1 \ldots N$ using training set
- Linear Regression
- Linear regression, misspecified
- Causal Forests
- Causal Forests, with hyperparameter tuning (caveat: tuning is computationally expensive, so it was not included in every simulation replicate after a limited sensitivity analysis)

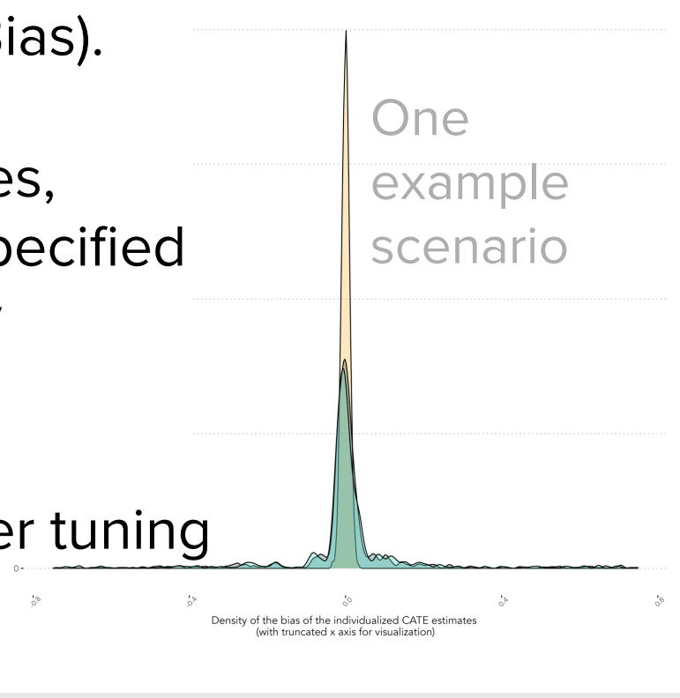
Specifications
- Number of effect modifiers = 0, 1, 4, 8
  - Distributions of these random variables: Standard Normal or Bernoulli(0.5)
- Number of nuisance variables = 0, 10, 20, or 40
- Total trial enrollment (N): 500, 1,000, 2,000, or 4,000
- Data generating mechanism = linear or nonlinear
- 2,000 replicates (to balance computational time with precision in inference)

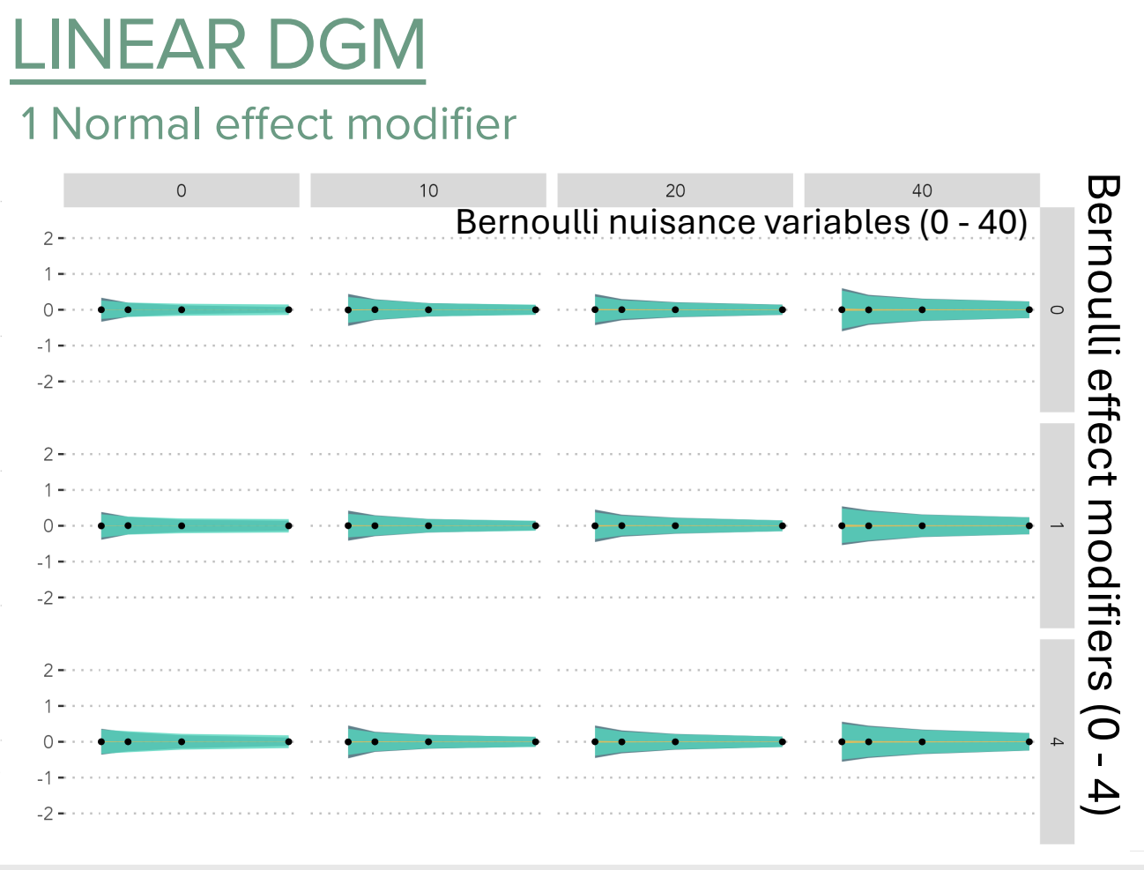
Metrics to Compare
- Bias of $\widehat{CATE}_i$
- 95% confidence interval coverage of $\widehat{CATE}_i$
- Model based standard errors for estimating $\widehat{CATE}_i$

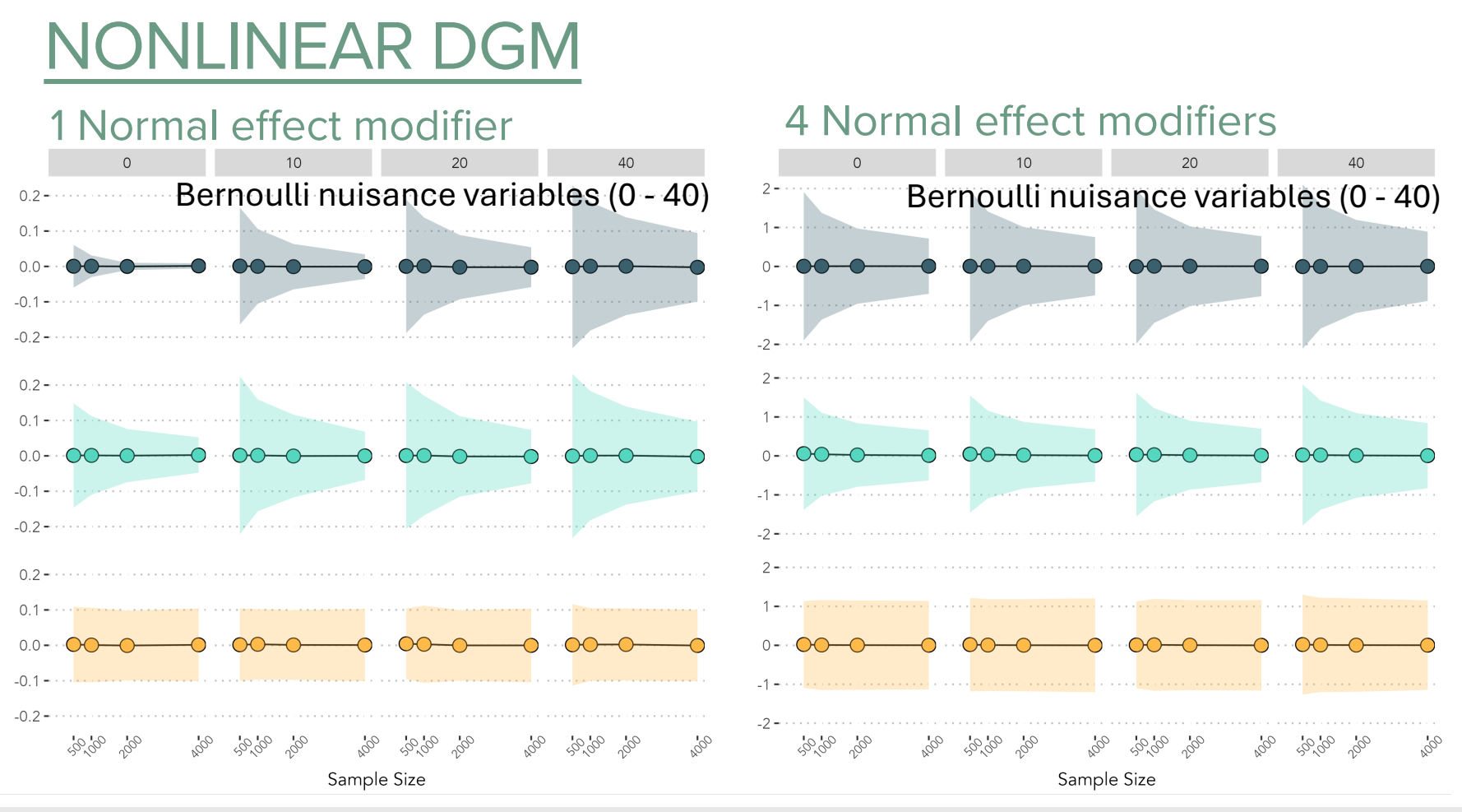
PART II: RESULTS

BIAS

- All methods evaluated are unbiased, but causal forests have *much* larger SE(Bias).
- As expected, as sample size increases, SE(Bias) decreases - except for misspecified linear regression which has relatively constant SE(Bias) width across N.
- Implementing limited hyper parameter tuning improves SE(Bias)

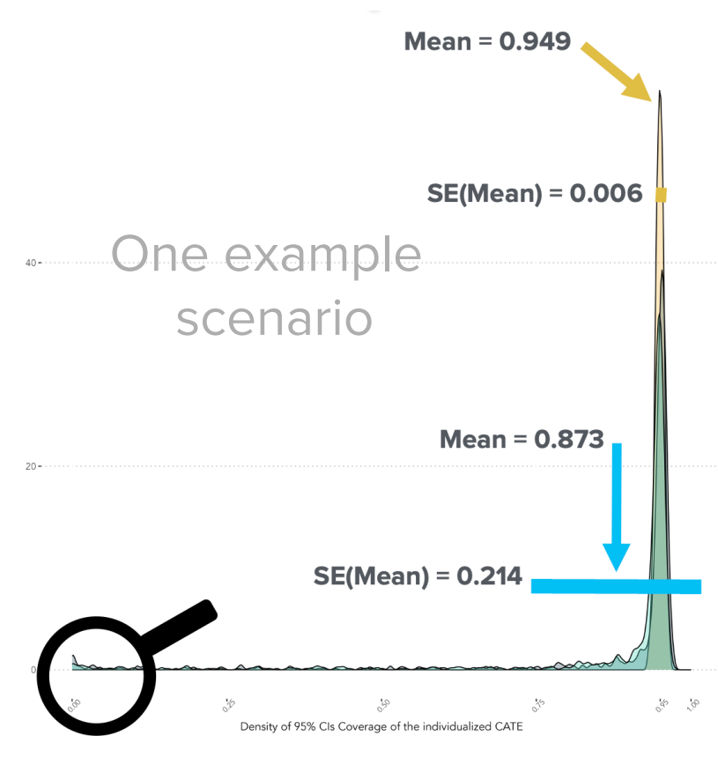

One example scenario



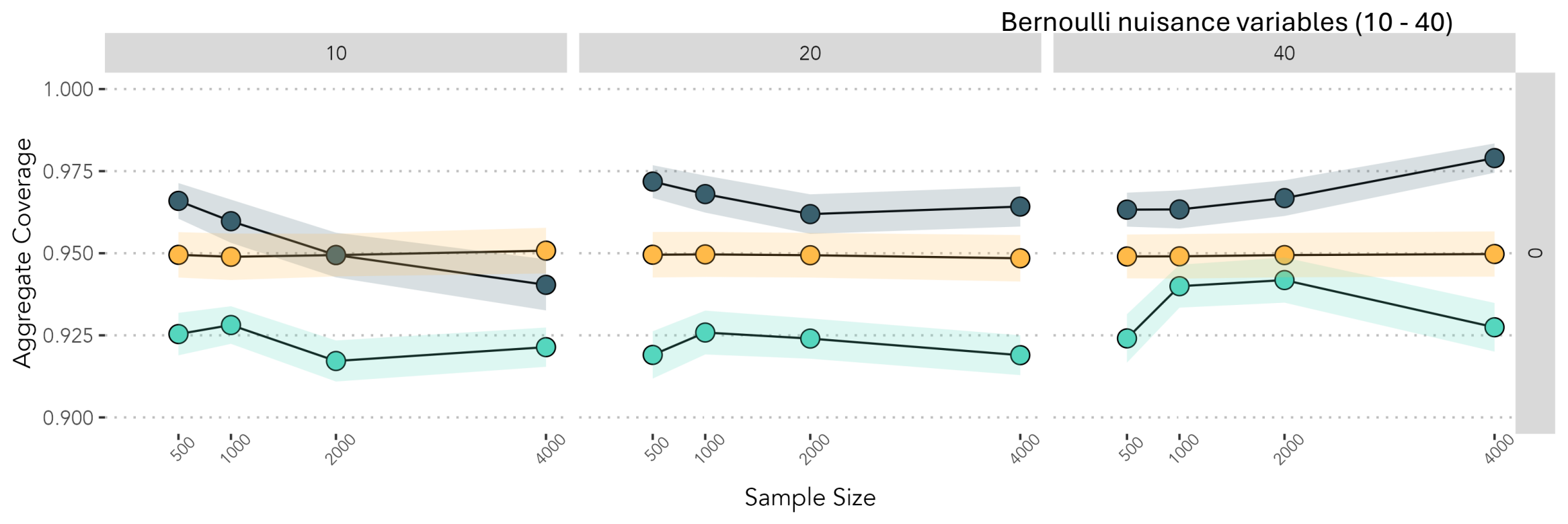


COVERAGE

- Under a linear DGM, linear regression is the clear choice in terms of SE(Bias) and 95% CI coverage.
- However, a linear DGM is likely not plausible in most real-life situations, so we should compare methods under a nonlinear DGM.
- If there is only 1 Normally distributed treatment effect modifier under a nonlinear DGM (top row, far right plot), causal forests with default settings are a good choice for 95% CI coverage.
- Implementing hyperparameter tuning improves causal forests 95% CI coverage proportion as compared to default causal forests.

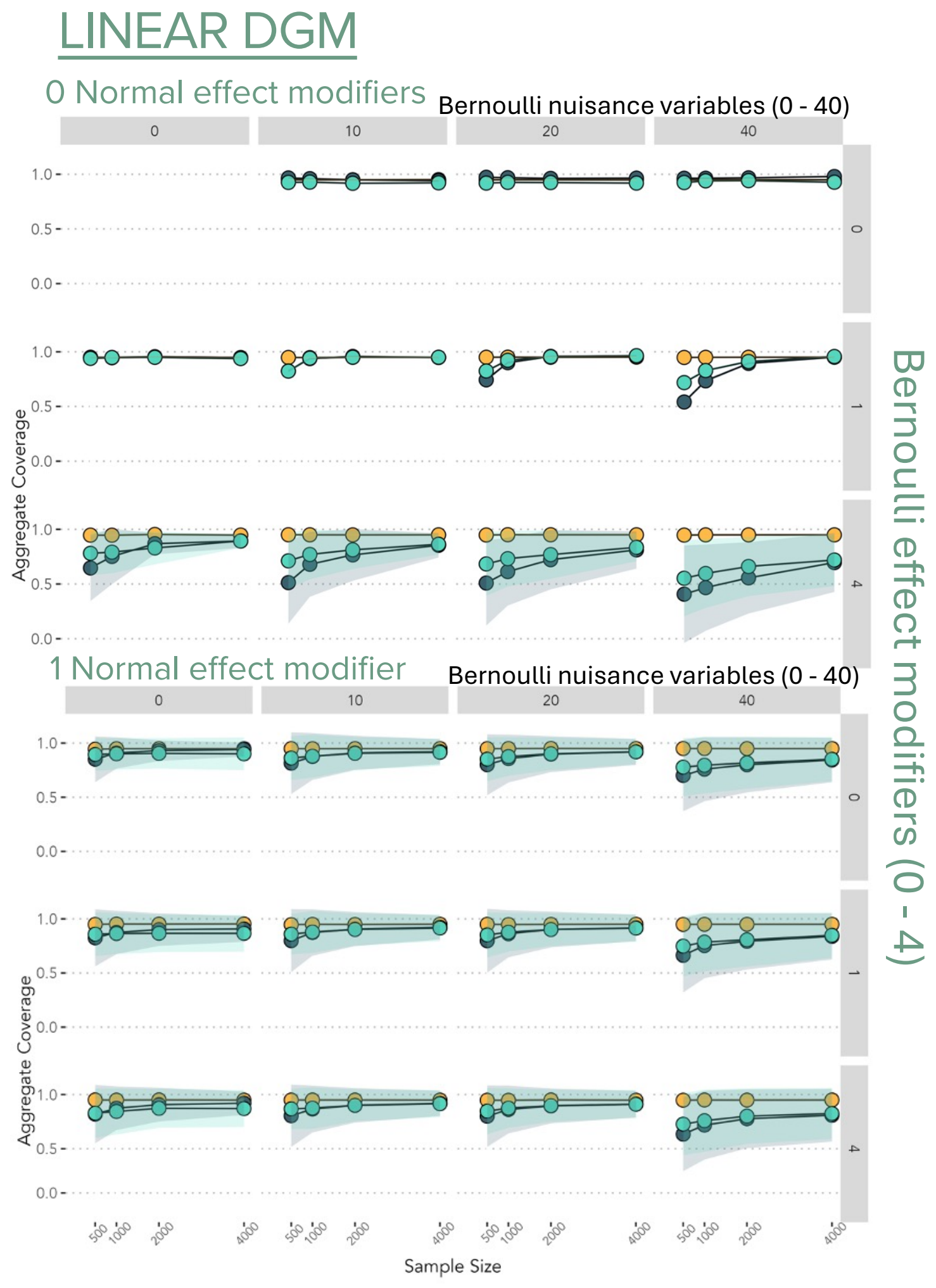

One example scenario

NULL SCENARIO: 0 EFFECT MODIFIERS



SCENARIOS WITH VARYING # EFFECT MODIFIERS

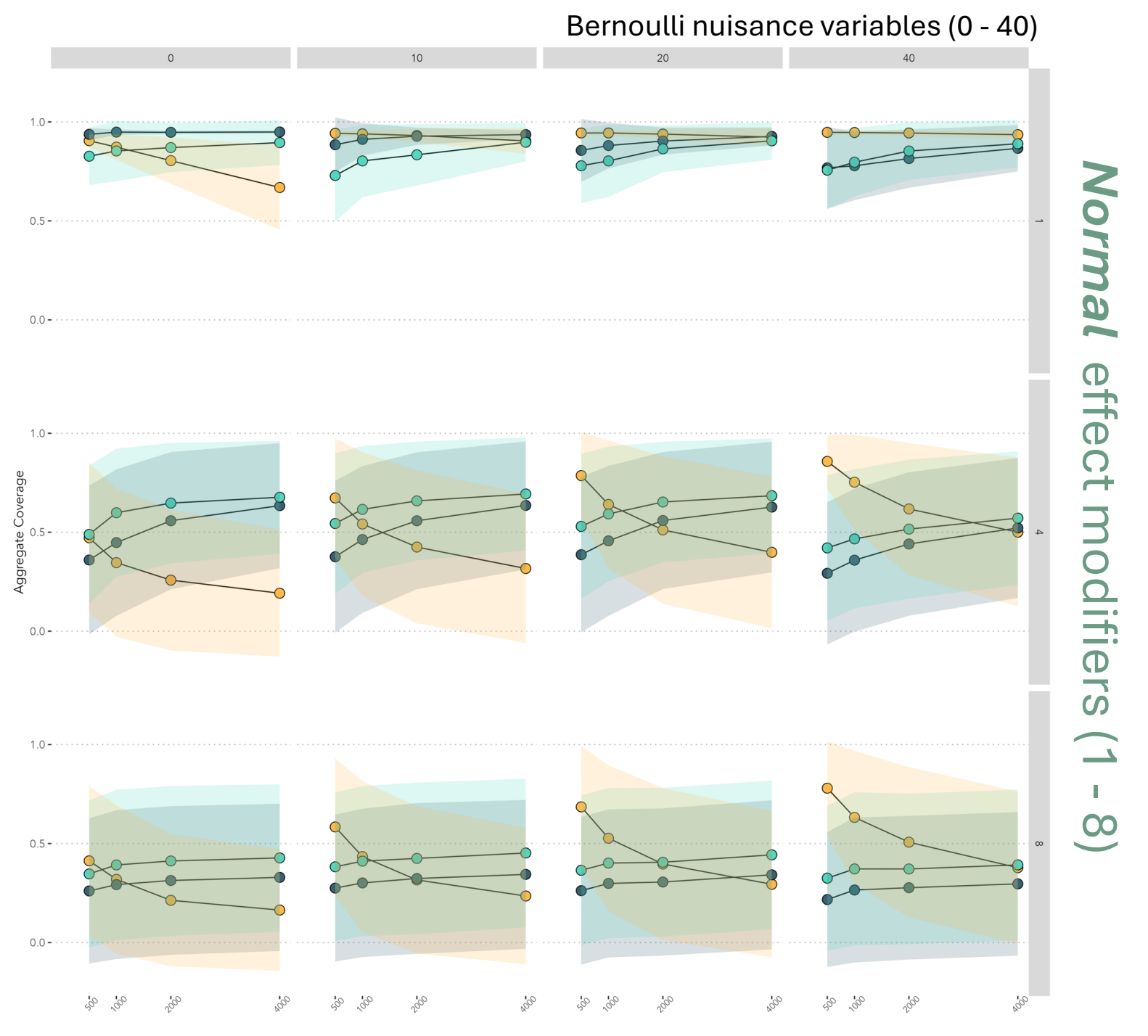
* Results not shown for all scenarios due to space constraints



NONLINEAR DGM

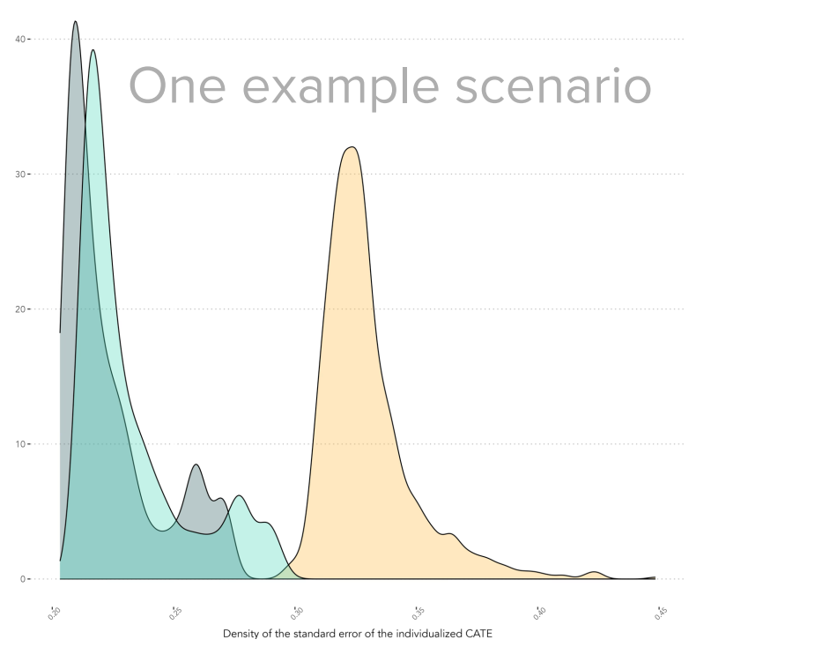
$$CATE_i = \sum_i f(X_i) \quad \text{where}$$

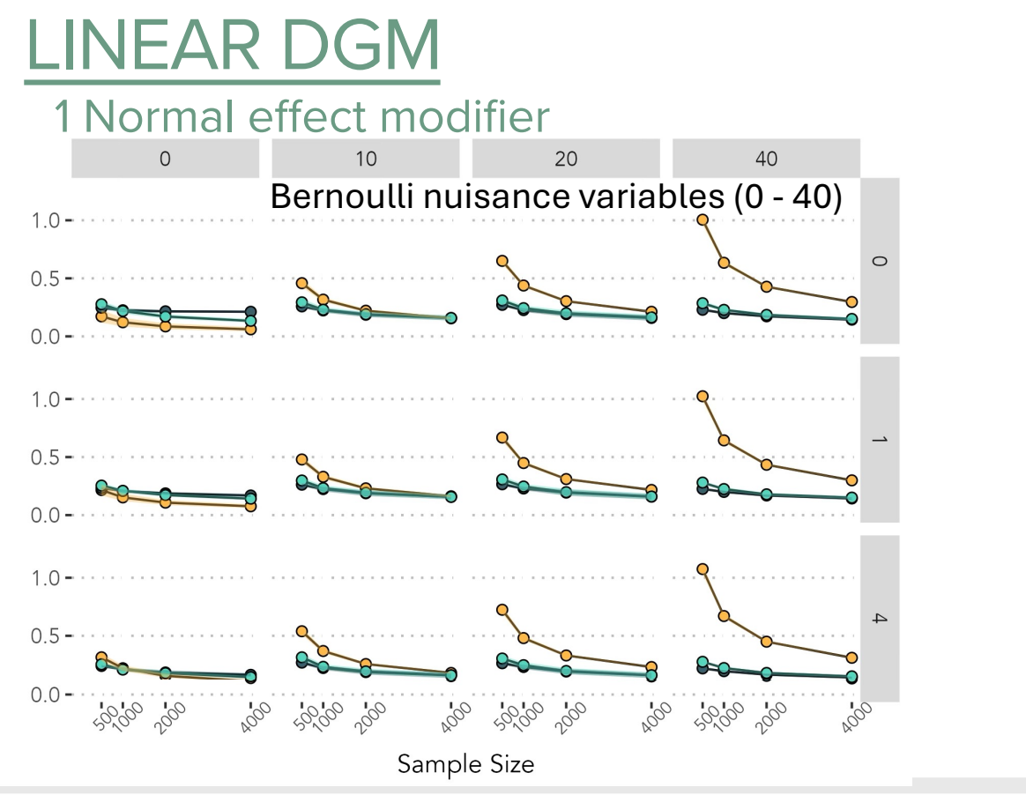
$$f(x) = 1 + \frac{1}{1+e^{-2(x-\frac{1}{4})}}$$

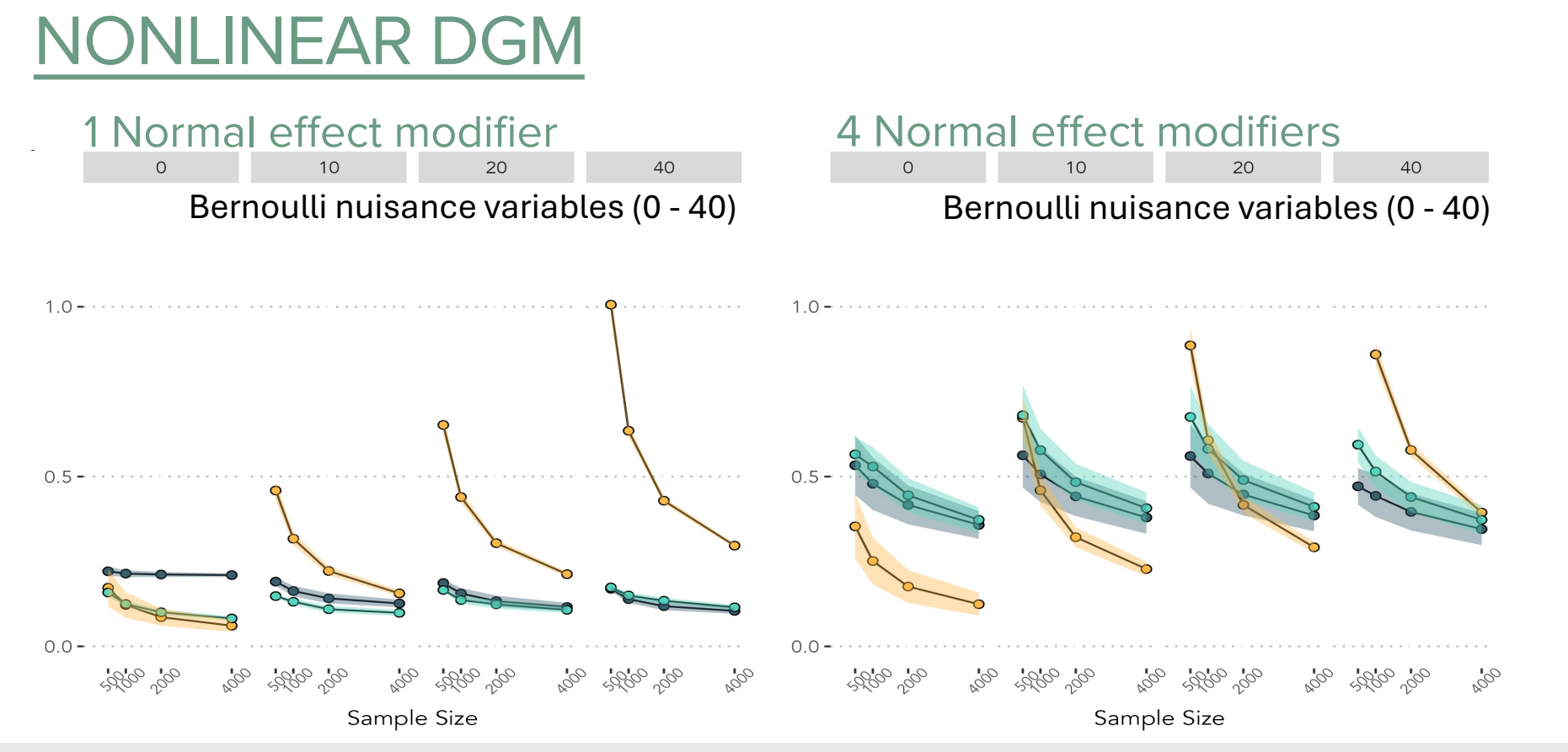


●	Causal Forests
●	Causal Forests, tuned
●	Linear Regression

STANDARD ERROR


One example scenario





References and more details

lisalevoir.github.io/projects

CONCLUSIONS

- Overall, these methods hold much promise but should be applied carefully in trials with modest sample sizes.
- Practitioners may want to consider how many treatment effect modifiers an expert could reasonably evaluate. If there are more than 4 and there are reasons to expect violations to assumptions of linearity, none of the methods considered seem to reliably deliver nominal CI coverage under the settings considered. This is concerning, since many practitioners seek to estimate individualized treatment effects with a moderate sample size and with models that consider many potential effect modifiers.
- We can expect linear regression to exhibit valid performance in all linear settings across sample sizes included in our study. We do notice higher standard errors which may be necessary to achieve nominal coverage.
- However, the plausibility of linearity is often dubious in practice. In the nonlinear settings considered in this study, the recommendation is less straightforward and will require the practitioner and study team to consider the sample size, number of covariates, and a forthright assessment of different potential subsets of covariates which may act as true effect modifiers.
- In the setting restricted to a smaller sample size, we cannot recommend casual forests without significant caveats. We can say that if the sample size is smaller (n = 500 to 1, 000), misspecified linear regression can outperform causal forest CI coverage *on average,* although the standard error bands overlap. Some of these limitations were outlined in the seminal work developing causal forests (Wager and Athey, 2018) and this is an avenue for further active research.
  - We can see the asymptotic nature of casual forests, but it's also clear that good performance doesn't happen until after N = 4,000 (which could be an issue for practitioners)