ECONOMETRICA

JOURNAL OF THE ECONOMETRIC SOCIETY

An International Society for the Advancement of Economic Theory in its Relation to Statistics and Mathematics

https://www.econometricsociety.org/

Econometrica, Vol. 93, No. 4 (July, 2025), 1165–1170

A COMMENT ON:
"Fisher–Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India" by Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val

KOSUKE IMAI
Department of Government and Department of Statistics, Harvard University

MICHAEL LINGZHI LI
Technology and Operations Management, Harvard Business School

The copyright to this Article is held by the Econometric Society. It may be downloaded, printed and reproduced only for educational or research purposes, including use in course packs. No downloading or copying may be done for any commercial purpose without the explicit permission of the Econometric Society. For such commercial purposes contact the Office of the Econometric Society (contact information may be found at the website http://www.econometricsociety.org or in the back cover of *Econometrica*). This statement must be included on all copies of this Article that are made available electronically or in any other format.

Econometrica, Vol. 93, No. 4 (July, 2025), 1165–1170

A COMMENT ON: *"Fisher–Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India" by Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val*

KOSUKE IMAI
Department of Government and Department of Statistics, Harvard University

MICHAEL LINGZHI LI
Technology and Operations Management, Harvard Business School

We examine the split-sample robust inference (SSRI) methodology introduced by Chernozhukov, Demirer, Duflo, and Fernandez-Val for quantifying uncertainty in heterogeneous treatment effect estimates produced by machine learning (ML) models. Although SSRI properly accounts for the additional variability due to sample splitting, its computational cost becomes prohibitive with complex ML models. We propose an alternative approach based on randomization inference (RI) that preserves the broad applicability of SSRI while eliminating the need for repeated sample splitting. Leveraging cross-fitting and design-based inference, the RI procedure yields valid confidence intervals with substantially reduced computational burden. Simulation studies demonstrate that the RI method preserves the statistical efficiency of SSRI while scaling to much larger applications and more complex settings.

KEYWORDS: Heterogeneous treatment effects, machine learning, randomization inference, confidence intervals, GATES.

1. INTRODUCTION

WE CONGRATULATE THE AUTHORS—Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val (hereafter CDDF)—for their excellent contribution to the growing literature on the estimation of heterogeneous treatment effects. CDDF's proposed approach is both methodologically innovative and practically relevant. Their method also comes with three important advantages. First, it can incorporate essentially any machine learning (ML) model. Second, the proposed methodology does not require that the ML model accurately estimates the conditional average treatment effect (CATE). Finally, it quantifies uncertainty including the one that arises from data splitting used to avoid overfitting the ML model. We believe that these attractive features will enable applied researchers to effectively but safely explore heterogeneous treatment effects in randomized experiments while leveraging the power of modern ML algorithms.

In this commentary, we focus on CDDF's key contribution—the development of uncertainty quantification methodology. The authors correctly point out that while data splitting is a standard technique to avoid overfitting ML models, the resulting causal estimates can vary substantially across different random splits of the data. In practice, this leads to an undesirable consequence: one may obtain different empirical conclusions, depending on the random seed. This practical problem motivates the need to quantify the uncertainty that arises from data splitting as well as the estimation uncertainty conditional on the data split.

Kosuke Imai: imai@harvard.edu
Michael Lingzhi Li: mili@hbs.edu

© 2025 The Econometric Society https://doi.org/10.3982/ECTA22261

As a solution to this problem, CDDF propose the split-sample robust inference (SSRI) methodology that requires randomly splitting the data many times. Then, for each split, researchers obtain an estimated causal quantity of interest and its conditional confidence interval. The authors show that a valid unconditional confidence interval, which accounts for the randomness of data splitting, can be constructed by simply using the median of lower conditional confidence bounds and that of upper conditional confidence bounds.

2. AN ALTERNATIVE RANDOMIZATION-INFERENCE APPROACH

One potential limitation empirical researchers may face when applying the SSRI method in practice is that repeated data splitting increases computational burden. Since fitting a complex ML model to even a single data split can be computationally expensive, repeating this process many times is not desirable. Moreover, ML models often require parameter tuning, which in itself may require another data splitting and incur additional computational cost. Given CDDF's recommendation that researchers split the data 250 times and fit a ML model to each split, the SSRI methodology can be, in practice, quite computationally intensive.

In a separate paper, we developed an alternative, randomization-inference (RI) methodology that addresses this remaining challenge (Imai and Li (2025)). Like the SSRI method, the RI methodology is applicable to a generic ML model regardless of whether it estimates the CATE well. Indeed, this alternative approach essentially requires only three randomization assumptions—random assignment of treatment, random sampling of units, and random splitting of data. This is consistent with the design-based approach pioneered by Neyman (1923). Unlike the SSRI method, however, the RI approach does not require repeated data splitting and hence is computationally more efficient.

Here, we briefly discuss the RI methodology while referring the readers to Imai and Li (2025) for details. For simplicity, suppose that we have K sorted groups of equal size and the data are randomly divided into $L \geq 2$ splits of equal size. To both avoid overfitting and gain statistical efficiency, the RI method uses the standard cross-fitting procedure. That is, every split, denoted by $\mathcal{L}_\ell$, is used once as the evaluation data to estimate the GATES $\hat{\gamma}_k^{(\ell)}$, while the remaining $L-1$ splits, denoted by $\mathcal{L}_{-\ell}$, are used to estimate the ML proxy $S^{(-\ell)}$. This process is repeated for each split, and the final estimate is given by the following simple average across all splits, that is,

$$\hat{\gamma} = \frac{1}{L}\sum_{\ell=1}^{L} \hat{\gamma}_k^{(\ell)}. \tag{1}$$

Using CDDF's notation whenever possible, a natural estimator of the GATE for the kth group based on the ℓth split (evaluation data) is given by

$$\hat{\gamma}_k^{(\ell)} = \frac{K}{N_1^{(\ell)}} \sum_{i\in\mathcal{L}_\ell} Y_i D_i \hat{f}_k^{(-\ell)}(Z_i) - \frac{K}{N_0^{(\ell)}} \sum_{i\in\mathcal{L}_\ell} Y_i (1-D_i) \hat{f}_k^{(-\ell)}(Z_i), \tag{2}$$

where $N_d^{(\ell)} = \sum_{i\in\mathcal{L}_\ell} \mathbf{1}\{D_i = d\}$ for $d = 0, 1$ denotes the treatment/control group size in the evaluation data, $\hat{f}_k^{(-\ell)}(Z_i) = \mathbf{1}\{S^{(-\ell)}(Z_i) \geq \hat{c}_k^{(\ell)}\} - \mathbf{1}\{S^{(-\ell)}(Z_i) \geq \hat{c}_{k-1}^{(\ell)}\}$ is the sorted group indicator variable based on the ML proxy obtained from the remaining splits, and $\hat{c}_k^{(\ell)} = \inf\{c \in \mathbb{R} : \sum_{i\in\mathcal{L}_\ell} \mathbf{1}\{S^{(-\ell)}(Z_i) > c\} \leq |L_\ell| k/K\}$ is the estimated kth quantile for the evaluation data.

Equation (2) resembles the standard estimator for evaluating the expected utility of estimated individualized treatment rules (ITR) represented by $\hat{f}_k^{(-\ell)}(Z_i)$. This relation enables us to leverage the result from the experimental evaluation of ITR (Imai and Li (2023)). In particular, Theorem 5 of Imai and Li (2025) presents the finite sample bias and variance of the estimator given in Equation (1) for the following estimand, which is a variant of the GATES:

$$\gamma_k = \mathbb{E}\big\{\mathbb{E}\big(Y_i(1) - Y_i(0) | \hat{c}_k \leq S(Z_i) \leq \hat{c}_{k+1}\big)\big\}. \tag{3}$$

In this equation, the inner expectation is taken over the distribution of $\{Y_i(1), Y_i(0), Z_i\}$ that fall within the sorted group based on a training data set, while the outer expectation is taken over the random sampling of training data. Based on this result, one can construct the asymptotic confidence interval. Importantly, while the resulting confidence interval accounts for both estimation and data splitting uncertainties, it only requires a single round of data splitting into L splits.

To derive the variance expression, we must account for the fact that the cross-fitting procedure uses the same data to both train the ML algorithm and estimate the GATES. We apply the useful result of Nadeau and Bengio (2000, Lemma 1), which implies the following:

$$\mathbb{V}(\hat{\gamma}_k) = \mathbb{V}\big(\hat{\gamma}_k^{(\ell)}\big) - \frac{L-1}{L}\mathbb{E}\big(V_k^2\big), \tag{4}$$

for each sorted group k where $V_k^2 = \sum_{\ell=1}^{L}(\hat{\gamma}_k^{(\ell)} - \hat{\gamma}_k)^2/(L-1)$ is the sample variance of split-specific GATE estimate. While the first term of Equation (4) denotes the variance of GATE estimate based on a single split, the second term represents the efficiency gain due to cross-fitting. Thus, the efficiency gain due to cross-fitting is large when the split-specific GATE estimate has a large variance across splits.

Finally, we note that CDDF's classification analysis of CLAN can also be conducted using this estimator by replacing the outcome Y_i with a quantity of interest $g(Y_i, Z_i)$. CLAN compares the two groups that are predicted to benefit most and least from the treatment, that is, γ_K and γ_1. This means that we need to derive the covariance between $\hat{\gamma}_K$ and $\hat{\gamma}_1$. Theorem 7 of Imai and Li (2025) derives this covariance as a part of the nonparametric test of treatment effect heterogeneity.

2.1. *Comparing the Finite-Sample Performance Through Simulation Study*

We compare the finite-sample performance of the SSRI confidence interval with that of the RI confidence interval. We compute the latter using the open-source software package evalITR (Li, Imai, Li, and Yang (2023)). Unlike the simulation study presented in CDDF, which assumes a linear interactive model, we randomly selected three data generating processes out of the 77 settings used for the 2016 Atlantic Causal Inference Conference Competition (Dorie, Hill, Shalit, Scott, and Cervone (2019)), including one linear, polynomial, and step-change treatment effect model (No. 13, 28, and 66). These three simulation settings cover various levels of treatment effect heterogeneity and different functional forms. We use the empirical distribution of covariates in the full sample ($n = 4802$) as their population distribution. See Imai and Li (2025) for additional details about the simulation setup.

We focus on the properties of confidence intervals for GATES with $K = 5$ groups using a total sample size of 100, 500, and 2500. For all simulation experiments, we train LASSO

to estimate the CATE. We choose LASSO here because other more complex ML algorithms such as BART and Causal Forest are computationally too expensive for repeated sample splitting required by the SSRI method. For the SSRI method, we follow CDDF's empirical application and utilize a 67%/33% split and set the number of repeated data splitting to 250. The results presented below are based on 1000 simulated data sets (the true values are approximated by averaging over 10,000 simulated data sets).

To compare the SSRI method with the RI method on an equal footing, we present the results based on the mean metric. While CDDF recommend the use of the median metric, they also note that the mean metric enjoys the same properties as the median (Section 4.1 Estimation and Inference: The Generic Targets, below Lemma 4.1). Indeed, we find that the results based on the median metric are essentially identical. Following the advice of CDDF (Section 4.4: Confidence Intervals for Median Parameter with Multiple Splits, below Definition 4.4), we further test a 80%/20% split so that the auxiliary data set is much larger than the main data set, making it more likely to satisfy one of the SSRI's key assumptions (R3). Finally, for the sake of direct comparison, we set the number of splits to $L = 3$ and $L = 5$ for the RI method, implying that the split between training and evaluation is 67%/33% and 80%/20%, respectively. In both cases, the SSRI method was approximately 40 times more computationally intensive than the RI method.

Figures 1 and 2 present the results for 67%/33% and 80%/20% splits, respectively. For direct comparison with RI, we use the mean point estimate of SSRI, rather than the

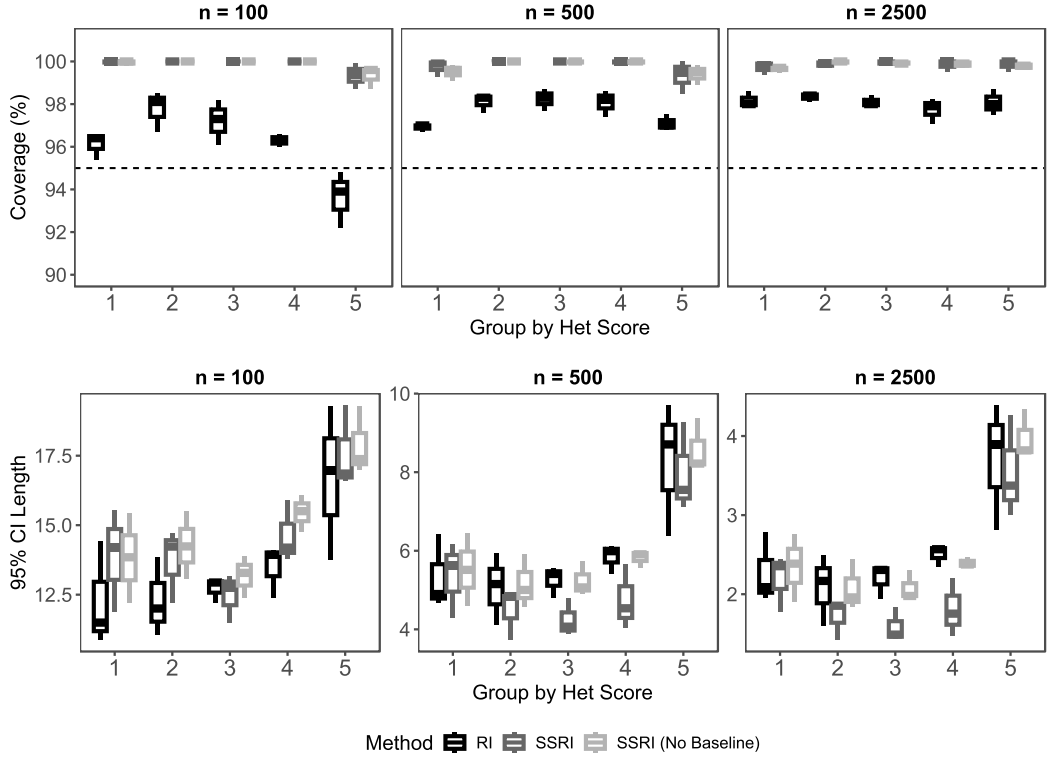

FIGURE 1.—Empirical coverage (upper panel) and average length (lower panel) of 95% confidence intervals for CDDF's Split-Sample Robust Inference (SSRI) confidence intervals (gray) and our Randomization-Inference (RI) confidence intervals (black) under the 67%/33% split scenario. For direct comparison with the RI confidence intervals, we also include the SSRI confidence intervals without baseline estimation (light gray).

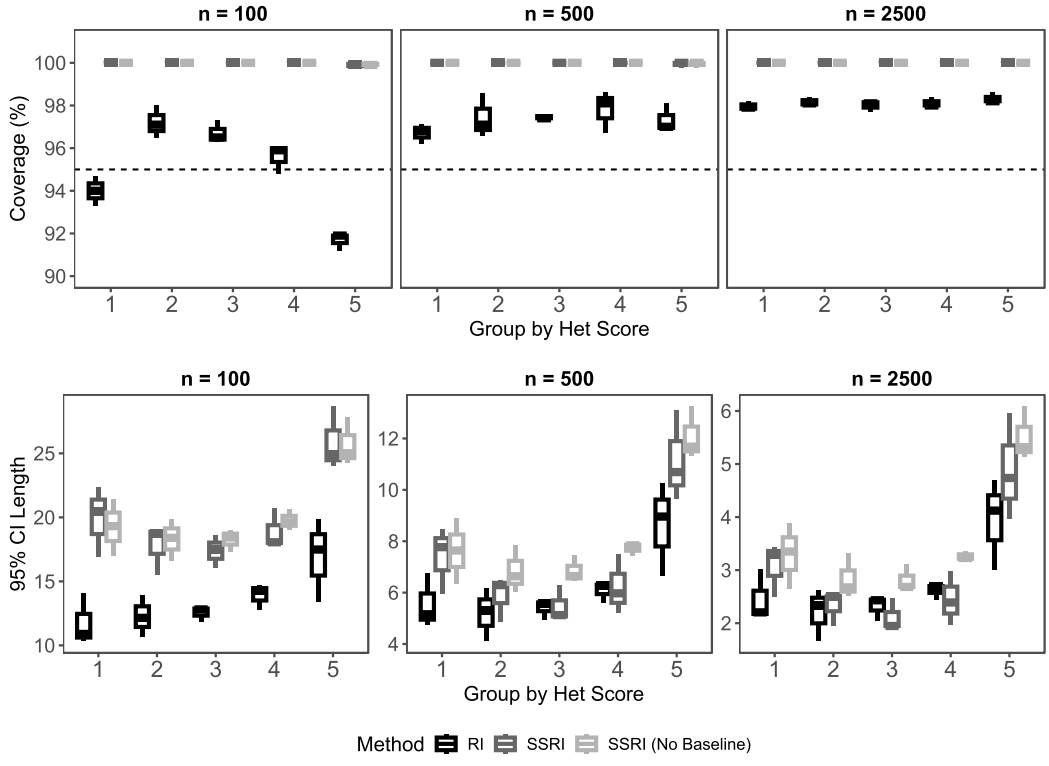

FIGURE 2.—Empirical coverage (upper panel) and average length (lower panel) of 95% confidence intervals for CDDF's Split-Sample Robust Inference (SSRI) confidence intervals (gray) and our Randomization-Inference (RI) confidence intervals (black) under the 80%/20% split scenario. For direct comparison with the RI confidence intervals, we also include the SSRI confidence intervals without baseline estimation (light gray).

estimate based on median as recommended by CCDF. The upper panels of these figures present the empirical coverage of 95% confidence intervals based on the RI (black) and SSRI (gray) methods. Given that the SSRI method utilizes another ML model for estimating the baseline conditional average function, we also examine the performance of SSRI without baseline estimation (light gray) for direct comparison with RI. We find that the SSRI confidence interval (with or without baseline estimation) is conservative across sorted groups in all sample sizes, leading to nearly 100% coverage in every case.

Although it is difficult to pinpoint the exact source of this conservativeness, we hypothesize that it is due to the use of Markov's inequality in CCDF's theoretical analysis. While the inequality is tight for a general random variable, it may be conservative in realistic data generating settings including those of the ACIC data competition setup used in our paper. In contrast, while the RI confidence interval also tends to be conservative, it exhibits a smaller degree of overcoverage than the SSRI confidence interval.

The conservativeness of the SSRI method leads to wider confidence intervals. The bottom panels of Figures 1 and 2 show that the average length of the RI confidence interval is generally shorter than the SSRI method when no extra ML baseline estimation is used. In theory, the extra information about the estimator distribution obtained from SSRI's repeated sampling procedure should help construct a shorter confidence interval than RI. In practice, however, the conservativeness of the SSRI estimator appears to largely erase this benefit in our simulations.

The baseline estimation helps shrink the SSRI confidence intervals, suggesting that a similar strategy might also improve the RI confidence intervals. For a large sample size, the average length of the SSRI confidence intervals with baseline estimation is roughly comparable to that of the RI confidence intervals without. Finally, increasing the relative size of training data does not significantly affect the RI confidence intervals, but greatly widens the SSRI confidence intervals.

3. CONCLUDING REMARKS

We commend CDDF for proposing novel ways to explore heterogeneous effects with new causal quantities such as GATES and CLAN while developing a statistical method to quantify estimation uncertainty. We expect their work to have significant impacts on both applied empirical research and methodological literature. In this commentary, we discussed an alternative, randomization-inference approach to the same problem. We hope that this alternative perspective inspires further methodological development. Finally, many interesting open problems remain. For example, future work should consider the data-driven selection of cutoffs used to define sorted groups as well as the exploration of heterogeneous treatment effects in adaptive experiments.

REFERENCES

DORIE, VINCENT, JENNIFER HILL, URI SHALIT, MARC SCOTT, AND DAN CERVONE (2019): "Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned From a Data Analysis Competition," *Statistical Science*, 34 (1), 43–68. [1167]

IMAI, KOSUKE, AND MICHAEL LINGZHI LI (2023): "Experimental Evaluation of Individualized Treatment Rules," *Journal of the American Statistical Association*, 118 (541), 242–256. [1167]

IMAI, KOSUKE, AND MICHAEL LINZHI LI (2025): "Statistical Inference for Heterogeneous Treatment Effects Discovered by Generic Machine Learning in Randomized Experiments," *Journal of Business & Economic Statistics*, 43 (1), 256–268. [1166,1167]

LI, MICHAEL LINGZHI, KOSUKE IMAI, JIALU LI, AND XIAOLONG YANG (2023): "evalITR: Evaluating Individualized Treatment Rules,", https://CRAN.R-project.org/package=evalITR. Version 1.0.0. [1167]

NADEAU, CLAUDE, AND YOSHUA BENGIO (2000): "Inference for the Generalization Error," in *Advances in Neural Information Processing Systems*, 307–313. [1167]

NEYMAN, JERZY (1923): "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9. (Translated in 1990)," *Statistical Science*, 5, 465–480. [1166]

Co-editor Guido W. Imbens handled this manuscript as an invited Fisher–Schultz lecture. The invitation to deliver the Fisher–Schultz lecture is also an invitation to publish a suitable version of the lecture in Econometrica.

Manuscript received 15 September, 2023; final version accepted 29 July, 2024; available online 7 April, 2025.

The replication package for this paper is available at https://doi.org/10.5281/zenodo.12789158. The Journal checked the data and codes included in the package for their ability to reproduce the results in the paper and approved online appendices.