CEDIL LESSONS LEARNED PAPER 3

Machine learning for impact evaluation in CEDIL-funded studies: an ex ante lesson learning paper

Alex Lewin, Karla Diaz-Ordaz, Chris Bonell, James Hargreaves, and Edoardo Masset

UKaid from the British people

Centre of Excellence for Development Impact and Learning

Oxford Policy Management

LIDC London International Development Centre

About CEDIL

The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by the UK government through UK Aid. The mission of the Centre is to test innovative methodologies in evaluation and evidence synthesis and promote evidence-informed development. CEDIL-supported projects fall into three programmes of work: evaluating complex interventions, enhancing evidence transferability, and increasing evidence use.

CEDIL Lessons Learned Paper

The CEDIL Lessons Learned series shares learning from across the CEDIL programme.

About this working paper

Suggested citation: Lewin, A., Diaz-Ordaz, K., Bonell, C., Hargreaves, J. & Masset, E. (2023) 'Machine learning for impact evaluation in CEDIL-funded studies: an ex ante lesson learning paper'. CEDIL Lessons Learned Paper 3, CEDIL, Oxford. Available at https://doi.org/10.51744/LLP3

Cover design: PhilDoesDesign

Cover images: (clockwise from top left) Oxfam International; UNAMID; Alex Tilley, WorldFish; CPT Palestine, all shared on flickr under Creative Commons licences.

Copyright: © 2023 This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Machine learning for impact evaluation in CEDIL-funded studies: an ex ante lesson learning paper

Authors

Alex Lewin

Karla Diaz-Ordaz

Chris Bonell

James Hargreaves

Edoardo Masset

Institute

London School of Hygiene and Tropical Medicine

Contents

1. Introduction

The Centre of Excellence for Development Impact and Learning (CEDIL) has recently funded several studies that use machine learning methods to enhance the inferences made from impact evaluations. These studies focus on assessing the impact of complex development interventions, which can be expected to have impacts in different domains, possibly over an extended period of time. These studies therefore involve study participants being followed up at multiple time-points after the intervention, and typically collect large numbers of variables at each follow-up. The hope is that machine learning approaches can uncover new insights into the variation in responses to interventions, and possible causal mechanisms, which in turn can highlight potential areas that policy and planning can focus on.

This paper describes these studies using machine learning methods, gives an overview of the common aims and methodological approaches used in impact evaluations, and highlights some lessons and important caveats. The scope of the paper is limited to the use of supervised learning models in the estimation of heterogeneous treatment effects and the investigation of potential mediating mechanisms, since these are the areas of focus for the CEDIL-funded studies discussed here. We do not consider other machine learning methods, such as unsupervised learning or reinforcement learning, though these may be used in other areas of impact evaluation.

In the field of statistical modelling, the popularity of machine learning methods has grown hugely in recent years. Innovations in modelling, in particular greater flexibility and the ability to accommodate large numbers of variables, have enabled researchers to be more ambitious in modelling relations between input variables and outcomes/responses. The availability of "big data" has encouraged this, though the real gains from flexible machine learning approaches come from data complexity rather than size (in particular, non-linear relationships and interactions between variables). Additionally, the availability of powerful software with many contributed packages has made these models accessible to a wide user-base.

In the context of modelling treatment effects and mechanisms, there are two main benefits of statistical machine learning models as compared with traditional statistical approaches: (1) flexible modelling and (2) the use of regularisation or latent structure to accommodate large numbers of variables.

Flexible modelling (often misleadingly called non-parametric modelling) allows for modelling associations between variables without making strong assumptions, such as linearity. Flexible models are in fact highly parameterised, so that the model can closely track non-linear patterns and interactions in the data. Such models can be so complex that it can be difficult to interpret the relations between variables. For this reason, machine learning traditionally focused on predicting outcome variables as a black box function of input variables – a task in which it succeeds very well. However, statistical models with latent structure and which make use of interpretable machine learning approaches give us a way to gain insight into the most important input variables. This is essential when seeking to make causal inference, where we want to understand which variables are impacting the development of relevant outcomes. The benefit of flexible modelling in impact evaluation is the ability to let the data discover the form of the relations between the estimated treatment effects and baseline characteristics.

The second component of machine learning models is the ability to model large numbers of variables in a single model. The key to this is incorporating regularisation in order to prevent the model overfitting to the data that are used to estimate the model. Traditional machine learning methods use sample-splitting and resampling methods to control overfitting. Statistical machine learning models can impose smoothing and shrinkage on parameter estimates through penalty terms or equivalently Bayesian priors. The benefit of including large numbers of variables in impact evaluation is the ability to explore potential causal factors and mediators in a data-driven manner.

Causal inference is used in impact evaluation to establish the causal effects of the exposure of interest, and further to identify plausible explanations for these effects in terms of baseline or intermediary variables. For example, as well as characterising the causal effects of interventions on outcomes, we may be interested in identifying the most important variables that modify the causal effect of the exposure or intervention. The benefit that machine learning approaches can bring here is to enable the inclusion of large numbers of potential confounders and effect modifiers in a single analysis. Further, the relations between outcomes, confounders and effect modifiers can be modelled in a flexible manner, without imposing restrictive parametric assumptions.

The main theme which emerges across all the CEDIL-funded studies is that of estimating heterogeneous treatment effects. Going beyond the average treatment effect, investigators want to find out which individuals can best benefit from a given intervention. Thus the aim is to characterise individual treatment effects through measured baseline variables. In a potential outcomes framework, the individual treatment effects can be conceptualised as the difference in potential outcomes for different treatment regimes. However, these can never be directly observed: they can only be estimated by combining observations from multiple individuals who are assumed to be in some sense equivalent. This equivalence is defined by having similar characteristics, and conditional treatment effects can be estimated within groups of people with similar covariate values (after making appropriate causality assumptions).

Hence, the problem becomes that of estimating conditional average treatment effects (CATEs) as functions of observed covariates. Since the CATEs themselves are not directly observed, the statistical modelling problem involves simultaneous estimation of the CATEs and their functional relationship on the covariates. Machine learning models and other statistical models with latent structure can provide a way to do this.

Section 2 of this paper gives a brief description of each of the studies, outlines the structure and scope of the studies' designs, and discusses common research aims and methods across the studies. Sections 3 gives an overview of the estimation of CATEs using machine learning, and reflects on questions of interpretation. Section 4 discusses possible ways to gain insight into considerations of context and mechanisms. Section 5 discusses further methodological challenges and future directions, and Section 6 summarises our main conclusions.

2. CEDIL studies using machine learning

CEDIL is funding three studies that aim to use machine learning methods in impact evaluation. At the time of writing, these studies are ongoing and have not produced their final results. In this paper we discuss the aims and methodology proposed by the studies in the context of machine learning.

All three of the studies are evaluating complex real-world interventions, which typically exhibit substantial heterogeneity across subjects, due to large differences in individual and environmental factors affecting study participants. It is therefore of general interest to investigate whether there are groups of individuals who respond better than others to a given intervention. Large numbers of variables are available to the studies, which they can use to characterise better or worse responders. These can be split into baseline variables and variables collected during follow-up. Therefore, all three CEDIL-funded studies aim (1) to estimate and interpret heterogeneous treatment effects using baseline covariates and (2) to investigate possible mediation mechanisms using follow-up variables. The focus of interest is to investigate how machine learning methods can contribute to pursuing these research questions. Three of the studies providing data for the machine learning research have already been completed or are ongoing. Here we briefly describe the designs of these studies in order to highlight common statistical issues. In later sections we put into context the analysis methods proposed by the three studies

Chioda and Gertler are investigating the impact of the two youth skills training programmes in Uganda: Educate! (sample size: n = 1,942) and SEED (sample size: n = 4,400). The Educate! programme is a cluster randomised trial that is being delivered through 48 schools in the most populous districts of Uganda, and lasts for two years. The SEED intervention is a three-week-long intensive course, with participants and controls randomised from a large pool of applicants across the country. Participants have been followed up at four years and will be followed up again at eight years. The outcome variables in these studies are educational investments, economic outcomes and social spillovers. Potential mediators are the primary targets of the skills programmes (a mix of soft and hard skills). Demographics and soft skills measured at baseline are used as confounders and variables to characterise effect heterogeneity.

Handa *et al*. are studying four interventions involving unconditional cash transfers: the Malawi Social Cash Transfer Programme SCTP (n = 3,530), the Zambia Child Grant Programme (n = 2,512), the Ghana Livelihood Empowerment Against Poverty (n = 2,497) and the Zimbabwe Harmonised Social Cash Transfer Programme (n = 2,063). The first two studies of these interventions are randomised controlled trials, and the last two use matched designs. All are household-level studies: the first two focus on labour-constrained households and the last two on households with young children. The studies had one, two or four follow-ups in the first few years following baseline. The outcome variables are indicators of household consumption. Potential mediators of the effect of the interventions on household consumption include a large set of secondary variables relating to the macroenvironment (market access, climate, topology and land cover), as well as financial and other assets, and agricultural activities. Baseline confounders and effect modifiers include demographics and psychological indicators, as well as financial and other assets.

Baliki *et al*. are focusing on two intervention packages delivered by the Food and Agriculture Organization of the United Nations (FAO) in Syria, which is running over two years. The unit of analysis is households (n = 3,907), in 312 villages. Interventions are at the

village level (the rehabilitation of local agricultural infrastructure, in the form of irrigation and drainage systems) and at the household level (providing agricultural assets in the form of crop and vegetable seeds and livestock kits). The study follows a matched design at the village level. Households will be followed up yearly, resulting in two pre-intervention and three post-intervention time-points. Outcome variables are welfare indicators (food security, diet quality, biomarkers of maternal and child nutritional status, and economic well-being), behavioural indicators (such as agricultural and livelihood decisions) and stability indicators, including exposure to violence. Potential mediators to be collected in post-intervention follow-ups include measures of agricultural productivity derived from satellite data. Baseline variables used as confounders and explanatory variables include household and farmland characteristics, district demographics and measures of exogenous shocks during the year before the intervention.

The data structure of the studies described above can be expressed in a common framework, which naturally allows for similar research questions and for raising similar statistical issues.

Interventions: All three studies considered here involve well-defined interventions, which can be treated as exogenous. Most evaluation designs involve some level of randomisation, though there are different specific designs, including randomised control trials, matched design and discontinuity design. This allows the formulation of causal questions, in particular estimating treatment effects for the intervention in question.

Outcome variables: The threestudies consider large numbers of outcome variables, across multiple domains. Many of these variables will reflect different aspects of the same underlying traits, and hence possess an underlying structure. The CEDIL studies, and other studies in the literature, currently tend to treat outcome variables separately, even if they are known to be related. In addition to the latent correlation structure between different outcome measures, there is also the time-dependent structure. All studies have at least one pre- and one post-intervention data collection, with most having multiple post-intervention follow-ups, with outcome variables collected at more than one time-point. Hence the evolution of treatment effects can be assessed over time.

Baseline covariates: All of the studies have large numbers of variables measured at baseline. This allows for the investigation of which individuals respond best or worst to the interventions, using baseline covariates to characterise the responders. The primary use of machine learning methods in these studies is to accommodate baseline covariates in the estimates of CATEs, which are interpreted as heterogeneous treatment effects.

Time-varying covariates: The studies also measure many covariates at later follow-ups. The investigators aim to use these intermediate variables to further explore possible treatment mechanisms, either through *post hoc* exploratory analysis comparing estimated CATEs with intermediate variables, or by causal mediation analysis treating time-varying covariates as potential mediators.

Multi-level structure: Several studies involve grouped data – for example, individuals in villages or in schools – with different balances between group size and sample size within groups. This is an important aspect of the design, which should be accounted for in the analysis, with respect to exposures and outcomes at different levels, and for estimating uncertainty.

3. Estimating heterogeneous treatment effects

Machine learning has become popular in causal inference as a way to estimate so-called heterogeneous treatment effects (HTEs), in order to gain further insight into who may or may not benefit from a particular intervention. It is important to understand that these HTEs are in fact CATEs, defined in relation to a specific set of explanatory variables. Heterogeneity of treatment effects cannot be estimated independently of a choice of potential explanatory variables.

Average treatment effect (ATE) is defined as the comparison between potential outcomes across the different treatment groups. For a binary treatment and continuous outcomes, the average treatment effect is usually defined as the difference in expected values:

$$ATE \equiv E(Y(1) - Y(0)) \quad (1)$$

where $Y(a)$ is the potential outcome for treatment group a. The potential outcome for an individual under treatment a is the (possibly counterfactual) outcome that individual would experience if they received treatment a. Since only one potential outcome can be observed for each individual, causal inference proceeds by making certain assumptions (see for example Kreif and DiazOrdaz, 2019), under which the ATE is then given by the expected difference in outcomes between people in different treatment groups:

$$ATE = E(Y|A = 1) - E(Y|A = 0) \quad (2)$$

where A and Y are the treatment and outcome variables, respectively.

CATEs (here denoted by τ) are defined as the treatment effect for people with a given set of covariate values:

$$\tau(x) \equiv E(Y(1) - Y(0)|X = x) \quad (3)$$

where X may stand for a set of several covariates. Within a group of people with identical covariate values, estimation of the CATE can proceed exactly as for the ATE. The same causal assumptions are made, and the CATE is estimated by comparing outcomes between individuals in different treatment groups:

$$\tau(x) = E(Y|A = 1, X = x) - E(Y|A = 0, X = x). \quad (4)$$

The CATE is an expectation, conditional on covariate values. By definition, individuals with identical covariate values must have identical CATEs. In a study with pre-specified groups of people with identical covariate values, which could happen for example if the covariate set is a small number of categorical covariates, the CATEs can be directly estimated by the differenced averages with groups. However, in large-scale intervention studies with many variables and mixed data types, there will not usually be any obvious pre-specified groups. Here the use of flexible modelling strategies can help to identify relevant groupings of study participants, or, as is usually done, to model the CATEs as smooth functions of the available covariates.

One way to identify groups of participants in a data-driven way is to use decision trees. A decision tree is a statistical model in which a study sample is partitioned into groups which share similar covariate values and similar outcome values. Decision trees are sometimes used to provide simple, interpretable models of outcome/response variables in terms of a number of covariate/input variables. Classification trees model the probabilities of categorical outcomes, while regression trees model the expected values of continuous

outcomes. The leaves of the tree correspond to groups of individuals with similar covariates and similar outcomes. Hence a decision tree model provides a data-driven partition of the study sample into groups of individuals based on covariate values, who also share similar outcomes. The estimation procedure for standard decision trees maximises the difference in average outcomes between leaves. Random forests are ensembles of decision trees where the predicted outcome is modelled as an average over many trees. This provides a smooth, flexible model of predicted outcomes.

Causal decision trees have been developed to identify relevant groupings of people for estimating CATEs (Athey and Imbens, 2016). In the causal inference framework, the outcome in the tree model is not a single observed variable for each individual, but the CATE $\tau(x)$, which is an unknown parameter of the model. In contrast to traditional decision trees, the estimation procedure for causal trees maximises the predicted average treatment effect difference between leaves. The machine learning aspect of trees comes in the exploration of the space of data partitions. The data partition is a parameter of the model and must be estimated, along with the CATEs themselves. A partition of the data sample can be completely specified by a vector of group labels $\boldsymbol{z}$, where z_i is the label of the group individual i belongs to. Thus, the learning algorithm must iterate between estimating the partition $\boldsymbol{z}$ and the CATEs $\boldsymbol{\tau}(\boldsymbol{z})$. The partition $\boldsymbol{z}$ is a structural parameter, and as such needs careful modelling to avoid overfitting. This can be done by using sample-splitting methods or Bayesian approaches. Athey and Imbens (2016) use a sample-splitting method where the structure of the tree is estimated on one subsample of the data and the CATEs are estimated on a different subsample. This is referred to as honest estimation.

Another popular approach is to use causal forests, which are ensemble averages of many tree estimators. These, again, can be estimated in different ways, including resampling methods (Wager and Athey, 2018) and Bayesian model averaging (Hill, 2011). Causal forests average over multiple trees, and therefore can produce a different CATE for each individual. However, these constructed CATEs should be seen as smooth functions of the covariates. By extending the tree-based models to forest-based models, the partition of the study sample into groups is lost. The forest approach provides a smoothed flexible functional form for the CATE in terms of covariate values. Note that other flexible modelling approaches can be used for this purpose, such as Gaussian processes or neural networks.

More recently, however, the focus in the causal inference literature has turned to methods which model so-called transformed outcomes: combinations of variables which allow the use of estimated propensity score functions to balance the covariates between treated and control participants, and which often possess some type of double robustness property. One example of a transformed outcome is given by:

$$\tau(X) = E\left[\frac{A(Y-\mu_1(X))}{e(X)} - \frac{(1-A)(Y-\mu_0(X))}{1-e(X)}\right] + \mu_1(X) - \mu_0(X) \qquad (5)$$

where $e(X)$ is the propensity score $P(A = 1 | X)$ and $\mu_a(X)$ is the conditional expectation of the outcome in treatment group a, i.e. $E(Y | A = a, X)$. The estimator based on this expression combines the standard inverse probability weighted estimator (IPWE) and the direct estimator using the outcome model ($\mu_1(X) - \mu_0(X)$), and is robust to mis-specification of either the propensity score model or the outcome model.

This expression for $\tau(X)$ can be estimated using plug-in estimators of $\mu_0(X)$, $\mu_1(X)$ and $e(X)$. In order to obtain HTEs, flexible machine learning models can be used for these plug-in functions. Using this approach (rather than modelling τ directly as a function of X) has the added advantage that existing machine learning software can be used to obtain these functions, since these are obtainable using observable data. A recent approach using different forms of transformed outcomes is the R-learner (Chernozhukov *et al*., 2018; Nie and Wager, 2021), which first constructs transformed outcomes using plug-in estimators found using machine learning prediction algorithms, and then regresses those transformed outcomes on the covariates to obtain estimates of τ as a non-linear function of the covariates. The regression in the second step may also be carried out using flexible models. The DR-learner (Kennedy, 2022) uses this approach with the doubly robust transformed outcome referred to above.

Bayesian models have also been developed using the ideas behind transformed outcomes to reparametrise the likelihood in order to put flexible data-driven priors on both τ and the propensity score. Putting flexible priors on τ is equivalent to modelling the unknown CATEs using causal forests, but modelling the propensity scores at the same time aims to improve the robustness to mis-specification, as in classical, non-probabilistic machine learning approaches. Hahn *et al*. (2020) use Bayesian additive regression trees (BART) to model τ, μ_0 and e. Zaidi and Mukherjee (2018) use the doubly robust transformed outcome referred to above, with Gaussian process priors for flexible modelling of both the direct estimator $\mu_1(X) - \mu_0(X)$ and the IPWE.

All three of the CEDIL-funded studies are using or have proposed to use causal trees or causal forests to estimate CATEs using baseline variables. Some studies involve multiple outcomes, possibly measured at more than one follow-up. To date, most research estimating CATEs treats each outcome variable separately.

4. Investigating contextual factors and mechanisms

The second major area of interest for all three of the CEDIL-funded studies discussed here is to further elucidate the mechanisms by which the affected population benefit from the intervention. This aim can be separated into (1) characterising people who respond well or less well to the intervention in terms of pre-intervention variables, and (2) identifying which post-intervention variables might be mediating the effect of the intervention. Note that "responders" here refers to individuals whose outcomes improved due to the intervention, even if they received only a group-level intervention.

The first of these aims is stated in the project proposals as identifying better-responding and worse-responding individuals. Since the HTEs or CATEs described above are constructed using the pre-intervention (baseline) variables, this aim is equivalent to identifying the baseline variables which contribute to variation in HTEs. These can be seen as effect modifiers of the intervention effect. Thus, characterising individuals as better or worse responders in terms of their estimated HTEs is equivalent to identifying people with particular values/ranges of the effect modifiers.

Therefore, in order to identify and characterise better and worse responders, we need to be able to identify which of the pre-intervention variables are the most important for explaining the variability in the HTEs. The issues involved in identifying the most important effect modifiers are the same as those involved in variable importance/feature selection in any statistical or machine learning inference scenario. Modern machine learning algorithms usually provide variable importance measures, which characterise the effect of leaving a particular variable out of the model. These would be the most direct measures for deciding which are the most important pre-intervention effect modifiers. However, it may also be useful to perform some simple analyses comparing the estimated CATEs with the baseline variables directly, using, for example, linear models or hypothesis tests. Machine learning methods can provide complex non-linear functions which are not necessarily straightforward to summarise. Post-analysis summary statistics and hypothesis tests may be useful to give an overview of which baseline variables are the most relevant.

The post-intervention variables may be contextual factors or individual time-varying covariates, measured at intermediate time-points (after the conclusion of the intervention). Methodologically, two broad types of approach are used to investigate potential mediating variables. Firstly, there is the data-driven approach, which proposes to use flexible machine learning models to explore the relations between estimated HTEs (CATEs) and the post-treatment variables of interest. These can be seen as exploratory analyses which can be used to suggest further directions of research and risk factors to focus on. The advantage of taking this approach is that large numbers of potential mediators can be explored simultaneously. The drawback is that these results cannot be interpreted in a causal manner.

With regard to the CEDIL studies, Handa *et al*. and Baliki *et al*. are using a combination of exploratory and testing approaches, both within-model and post-analysis, to explore which variables are the most relevant to characterising the better and worse responders.

The second methodological approach is to use and adapt causal mediation methods. Machine learning methods have been proposed for estimating causal mediation effects (Farbmacher *et al*., 2021; Huber *et al*., 2018; Tchetgen and Shpitser, 2012) in a robust way, while adjusting for

large numbers of confounders. This approach can be used to more robustly suggest which intermediate factors are responsible for the response to an intervention.

In mediation analysis, the natural direct and indirect effects are defined as:

$$NDE(x,a) \equiv E(Y(1,M(a)) - Y(0,M(a))|X=x)$$
$$NIE(x,a) \equiv E(Y(a,M(1)) - Y(a,M(0))|X=x) \quad (6)$$

where M is a potential mediator variable, and $Y(a, M(a'))$ is the nested counterfactual: the potential outcome for treatment a when the mediator variable is set to its natural value under treatment at. Under similar causal assumptions to those used for estimating CATEs, these can be estimated in terms of observable quantities. Tchetgen and Shpitser (2012) proposed a triply robust estimator of the nested counterfactual, an analogue of the doubly robust estimator of the CATE, involving estimates of the outcome model $E(Y|M=m, A=a, X=x)$, the propensity score $P(A|X=x)$ and the mediator model $E(M|X=x, A=a)$. Farbmacher *et al*. (2021) use a similar triply robust estimator, using flexible machine learning methods to obtain the plug-in estimates of the three component models.

Chioda and Gertler, in their CEDIL-funded project, propose to extend the causal mediation approach even further, investigating some approaches to estimating HTEs as a function of mediators as well as baseline variables – a form of moderated mediation analysis. This has the potential to provide further understanding of which post-intervention and baseline variables are responsible for heterogeneity in treatment response between individuals.

5. Methodological challenges and future directions

The machine learning models which have been adapted to work with causal inference methods are very recent methodological developments in the field of causal machine learning. As is usual in methodological work, these approaches have first been demonstrated in relation to simple scenarios with clean data. In the CEDIL studies, where the emphasis is on obtaining robust results in new intervention studies, the researchers additionally have to deal with extra issues which typically arise in real data analysis. A major issue, as always, is missing data, and the assumptions that must be made about missingness mechanisms. Another important aspect is multi-level correlation structure, due to grouped individuals (e.g. in villages or schools). This affects inference through the assessment of uncertainty. Obtaining valid confidence intervals in machine learning models, even in the absence of grouped data, is an active area of research. Athey and Wager (2019) demonstrate a sampling method for causal forests which accounts for the structure of grouped data.

As the field develops, it will be interesting to see the use of models which use larger data structures in a single model, such as incorporating all follow-up times together in trajectory models. This approach gives higher statistical power to detect patterns and associations between outcomes and baseline or mediating variables, though often at the cost of making stronger assumptions about time-dependent trends in the model. A related point is accounting for correlated outcomes. These are very common in observational studies monitoring multiple related outcomes, such as developmental, educational or economic variables. Again, modelling multiple variables with some known structure results in higher power for statistical inference.

All three of the studies in the CEDIL programme have large numbers of variables relating to potential mediating mechanisms, collected at multiple follow-ups after the intervention. Existing methods for investigating mediation deal with one variable at a time, which can give some initial insight into which intermediate factors are important in the causal pathway from intervention to outcome. In reality, there will be a network of causal relations between different mediating factors. Untangling the network of effects is a difficult task. Exploring methodologies for using multiple mediators in a single model is an active area of research in causal inference. Benkeser and Ran (2021), for example, propose a non-parametric method for estimating interventional mediation effects using machine learning.

Very often, machine learning methods are proposed as purely data-driven approaches where new hypotheses can be generated from large data sets. Most of the analyses discussed so far in this paper follow this path. An important direction of future research will be to unite large-scale models with existing theory. This could be done through the use of Bayesian models which can explicitly incorporate prior information and beliefs, resulting in inferences which combine existing theory and new data together. Another approach is to use causal machine learning approaches that focus on particular hypotheses of interest. This would be a hypothesis-testing approach, restricted to the testing of existing theory but in a manner that can flexibly model the large numbers of variables available.

6. Conclusions

There is a recent and increasing field of research that is using machine learning methods to aid in the estimation of causal estimands. The three CEDIL-funded studies discussed in this paper have proposed to use causal trees and forests for the outcome model. However, when using robust estimation methods, in principle any flexible modelling approach can be used for the plug-in estimates of the outcome, mediator and propensity score models. It should be noted that care is needed in interpreting estimates of HTEs: these are CATEs, and cannot be defined without specifying a set of covariates to condition on. Estimated HTEs will differ depending on the set of covariates available to be modelled.

Machine learning models (models which adapt to non-linear functions of covariates and accommodate interactions between covariates) are well-suited to obtaining flexible models of treatment effects, since the required functions are predictive models. If we are only interested in estimates of the treatment effects then we may not care about the specifics of how those functions are composed in terms of the covariates. However, if, as in the three CEDIL studies, we are also interested in the variables which are responsible for the differences in response to treatment, or in the mediating variables, then there are further issues. For example, if covariates are co-linear there will be many different possible functional forms that would give the same estimated treatment effects. Hence, deciding which are the most important effect modifiers and mediators is not so straightforward.

An exciting future direction of research is to use and adapt mediation analyses for impact evaluation studies. Outcome variables measured at intermediate time-points may be treated as outcomes or mediators in different analyses. There are already some robust estimation methods available in the literature that use machine learning for flexible estimation. Multiple mediators are an ongoing research area. One of the CEDIL-funded studies is hoping to develop its own extension to causal mediation analysis, aiming to relax some of the causal assumptions that are usually made. This will be a very interesting result of the programme.

References

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* **113**, 7353–7360.

Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies* **5**, 37–51.

Benkeser, D. and Ran, J. (2021). Nonparametric inference for interventional effects with multiple mediators. *Journal of Causal Inference* **9**, 172–189.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21**, C1–C68.

Farbmacher, H., Huber, M., Laff´ers, L., Langen, H., and Spindler, M. (2021). Causal mediation analysis with double machine learning. FSES Working Papers 515, Faculty of Economics and Social Sciences, University of Freiburg/Fribourg Switzerland.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis* **15**, 965–1056.

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

Huber, M., Lechner, M., and Strittmatter, A. (2018). Direct and indirect effects of training vouchers for the unemployed. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **181**, 441–463.

Kennedy, E. H. (2022). Towards optimal doubly robust estimation of heterogeneous causal effects. ArXiv:2004.14497 [math, stat].

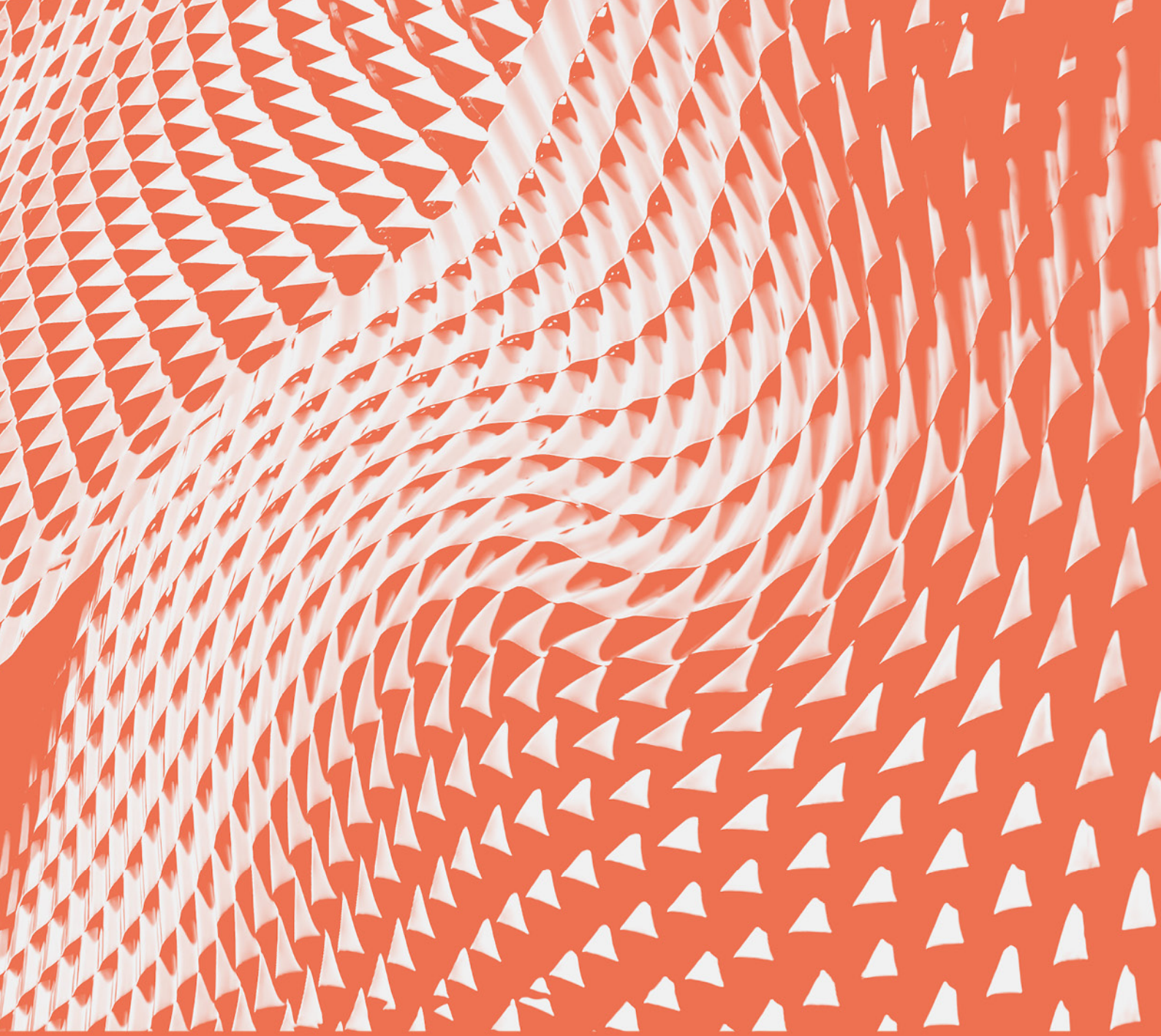
Kreif, N. and DiazOrdaz, K. (2019). *Machine Learning in Policy Evaluation: New Tools for Causal Inference.* arXiv preprint arXiv:1903.00402, 2019 ISBN: 9780190625979.

Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108**, 299–319.

Tchetgen, E. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *Annals of Statistics* **40**, 1816–1845.

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

Zaidi, A. and Mukherjee, S. (2018). Gaussian process mixtures for estimating heterogeneous treatment effects. *arXiv: Methodology*.



@CEDILProgramme

Centre of Excellence for Development Impact

Centre of Excellence for Development Impact and Learning

UKaid from the British people

About CEDIL

The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by the UK Government through UKaid. The mission of the centre is to test innovative methodologies in evaluation and evidence synthesis and to promote evidence-informed development. CEDIL-supported projects fall into three programmes of work: evaluating complex interventions, enhancing evidence transferability, and increasing evidence use.

For more information on CEDIL, contact us at **cedil@opml.co.uk** or visit our website **www.cedilprogramme.org**