Double Machine Learning based Program Evaluation under Unconfoundedness

Michael C. Knaus[†]

March 6, 2020

Abstract

This paper consolidates recent methodological developments based on Double Machine Learning (DML) with a focus on program evaluation under unconfoundedness. DML based methods leverage flexible prediction methods to control for confounding in the estimation of (i) standard average effects, (ii) different forms of heterogeneous effects, and (iii) optimal treatment assignment rules. We emphasize that these estimators build all on the same doubly robust score, which allows to utilize computational synergies. An evaluation of multiple programs of the Swiss Active Labor Market Policy shows how DML based methods enable a comprehensive policy analysis. However, we find evidence that estimates of individualized heterogeneous effects can become unstable.

Keywords: Causal machine learning, conditional average treatment effects, optimal policy learning, individualized treatment rules, multiple treatments

[*]Financial support from the Swiss National Science Foundation (SNSF) is gratefully acknowledged. The study is part of the project "Causal Analysis with Big Data" (grant number SNSF 407540_166999) of the Swiss National Research Program "Big Data" (NRP 75). I thank Martin Huber, Michael Lechner, Anthony Strittmatter, Stefan Wager, and Michael Zimmert for helpful comments and suggestions. The usual disclaimer applies.

[†]University of St. Gallen. Michael C. Knaus is also affiliated with IZA, Bonn, michael.knaus@unisg.ch.

1 Introduction

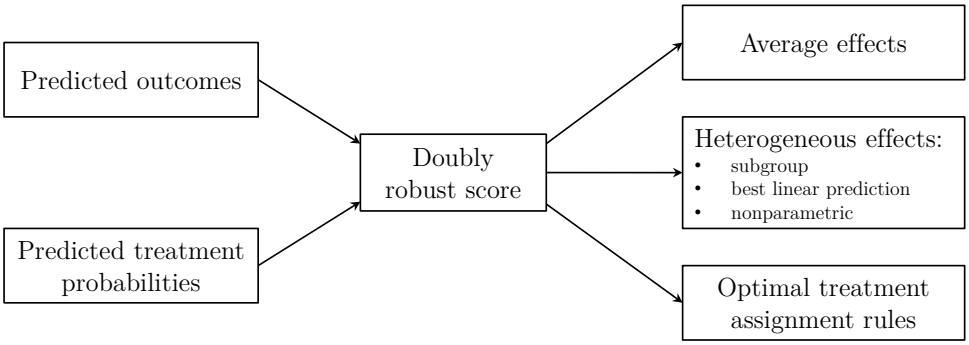
The adaptation of so-called machine learning to causal inference has been a productive area of methodological research in recent years. The resulting new methods complement the existing econometric toolbox for program evaluation along at least two dimensions (see for recent overviews Athey & Imbens, 2017, 2019; Abadie & Cattaneo, 2018). On the one hand, they provide flexible methods to estimate standard average effects. In particular, they provide a data-driven approach to variable and model selection in studies that rely on an unconfoundedness assumption[1] for identification. On the other hand, they enable a more comprehensive evaluation by providing new methods for the flexible estimation of heterogeneous effects and of optimal treatment assignment rules. However, the options for a concrete evaluation are steadily increasing and little guidance is available on how to choose suitable methods for a comprehensive and computationally feasible analysis.

This paper investigates the potential of Double Machine Learning (DML) as a framework for a thorough program evaluation. The DML framework seems attractive because (i) it can be combined with a variety of standard supervised machine learning methods (see for an overview Hastie, Tibshirani, & Friedman, 2009), (ii) it covers average effects for binary (e.g. Belloni, Chernozhukov, & Hansen, 2014; Chernozhukov, Chetverikov, et al., 2018), multiple (e.g. Farrell, 2015) as well as continuous treatments (e.g. Kennedy, Ma, McHugh, & Small, 2016; Semenova & Chernozhukov, 2017; Colangelo & Lee, 2019), (iii) it naturally extends to the estimation of heterogeneous treatment effects of different forms like canonical subgroup effects, the best linear prediction of the effect heterogeneity (Semenova & Chernozhukov, 2017), or nonparametric effect heterogeneity (Fan, Hsu, Lieli, & Zhang, 2019; Zimmert & Lechner, 2019), and (iv) it can be used to estimate optimal treatment assignment rules (e.g. Dudik, Langford, & Li, 2011; Athey & Wager, 2017; Zhou, Athey, & Wager, 2018). All these DML based methods have favorable statistical properties that allow the use of standard tools like t-tests, OLS or nonparametric regression for estimation and inference about the causal parameters of interest after flexibly controlling for confounding. Thus, the estimators can be implemented by combining standard statistical software.

This study contributes to the steadily growing literature of causal machine learning

[1] Also known as exogeneity, selection on observables, ignorability, or conditional independence assumption.

Figure 1: Double Machine Learning based program evaluation



for program evaluation in two ways. First, it consolidates the recent methodological contributions based on DML referenced in the previous paragraph. The consolidation highlights that the methods for different parameters all build on the same doubly robust score. The construction of this score might be computationally demanding because it requires the prediction of outcomes and treatment probabilities via machine learning methods. However, once constructed for the standard average treatment effect, the score might be reused for a variety of additional parameters of interest (see Figure 1 for a summary). This makes it particularly attractive for researchers who want to avoid using different frameworks for different parameters as the set of methods increases that integrate machine learning in the estimation of average treatment effects (e.g. van der Laan & Rubin, 2006; Athey, Imbens, & Wager, 2018; Avagyan & Vansteelandt, 2017; Tan, 2018; Ning, Peng, & Imai, 2018), heterogeneous treatment effects (e.g. Tian, Alizadeh, Gentles, & Tibshirani, 2014; Athey & Imbens, 2016; Wager & Athey, 2018; Athey, Tibshirani, & Wager, 2019; Künzel, Sekhon, Bickel, & Yu, 2017) and optimal treatment assignment (e.g. Bansak et al., 2018; Kallus, 2018). Second, we use DML based methods to provide a comprehensive and computationally convenient evaluation of four programs of the Swiss Active Labour Market Policy (ALMP) in a standard dataset (Huber, Lechner, & Mellace, 2017). The evaluation in this paper illustrates the potential of DML based methods for program evaluations under unconfoundedness and provides a potential blueprint for similar analyses. This adds to a small but steadily growing literature that applies causal machine learning to policy evaluation in general (e.g. Bertrand, Crépon, Marguerie, & Premand, 2017; Davis & Heller, 2017; Strittmatter, 2018; Farbmacher, Heinrich, & Spindler, 2019; Gulyas & Pytka, 2019; Knittel, 2019) and to evaluations based on unconfoundedness in

particular (e.g. Knaus, 2018; Jacob, Härdle, & Lessmann, 2019; Kreif & DiazOrdaz, 2019; Cockx, Lechner, & Bollens, 2020; Knaus, Lechner, & Strittmatter, 2020).

Overall, we find that DML based methods provide a promising set of methods for program evaluation. However, the estimation of individualized heterogeneous effects can become problematic by producing extreme effect estimates lying outside of the possible range. The estimated average program effects are mostly in line with the previous literature. We find that computer, vocational and language courses increase employment in the 31 months after programs start, while the effects are negative for job search training. The heterogeneity analysis shows substantial heterogeneities by gender and previous labor market success. These are picked up by the estimated optimal assignment rules.

The paper proceeds as follows. Section 2 defines the estimands of interest and shows how they are identified. Section 3 explains how DML based estimators can be used to estimate all estimands of interest. Section 4 introduces the application. Section 5 describes the implementation of the different estimators. Section 6 reports the results before Section 7 discusses the findings and concludes.

2 Estimands of interest

2.1 Definition

We define the estimands of interest in the multiple treatment version of the potential outcomes framework (Rubin, 1974; Imbens, 2000; Lechner, 2001). Let $\mathcal{W} = \{0, ..., T\}$ denote a set of programs.[2] We assume that each individual i $(i = 1, ..., N)$ has a potential outcome $Y_i(w)$ for all $w \in \mathcal{W}$. Without loss of generality, the discussion below assumes that higher outcome values are desirable.

The first estimand of interest is the average potential outcome (APO), $\gamma_w = E[Y_i(w)]$. It answers the question about the average outcome if the whole population was assigned to program w. However, the more interesting question is usually to compare different programs w and w'. To this end, we take the difference of the according individual potential

[2]For DML estimation with continuous treatments, see Colangelo and Lee (2019).

outcomes, $Y_i(w) - Y_i(w')$,[3] and aggregate them to different estimands: First, the average treatment effect (ATE), $\delta_{w,w'} = E[Y_i(w) - Y_i(w')]$. Second, the average treatment effect on the treated (ATET), $\theta_{w,w'} = E[Y_i(w) - Y_i(w') \mid W_i = w]$. Third, the conditional average treatment effect (CATE), $\tau_{w,w'}(z) = E[Y_i(w) - Y_i(w') \mid Z_i = z]$, where $Z_i \in \mathcal{Z}$ is a vector of observed pre-treatment variables.

The distinct aggregations accommodate the notion that treatment effects might be heterogeneous. ATE represents the average effect in the population, while ATET shows it for the subpopulation that is actually observed in program w. Thus, the comparison of ATE and ATET can be informative about the quality of the program assignment mechanism. For example, ATET being larger than ATE shows that the observed program assignment is better than random.

The ATET is defined by the observed program assignment and thus not subject to the choice of the researcher. In contrast, the conditioning variables Z_i of the CATE are specified by the researcher to investigate potentially heterogeneous effects across the groups of individuals that are defined by different values of Z_i. Such heterogeneous effects can be indicative for underlying mechanisms. Further, CATEs characterize which groups win and which lose by how much by receiving program w instead of w'.

The different average effects above provide a comprehensive evaluation of programs under the current program assignment policy. In many applications, however, we want to conclude the analysis with a recommendation how the assignment policy could be improved. This can either be done using the evidence on the different average effects defined above or by formally defining the objective of an optimal assignment rule. The latter is pursued by the literature on statistical treatment rules (e.g. Manski, 2004; Hirano & Porter, 2009; Stoye, 2009, 2012; Kitagawa & Tetenov, 2018; Athey & Wager, 2017, and references therein). Here we focus on the case with multiple treatment options as considered by Zhou et al. (2018). This requires more notation. Let $\pi(Z_i)$ be a policy that assigns individuals to programs according to their characteristics Z_i or, put more formally, the function $\pi(Z_i)$ maps observable characteristics to a program: $\pi : \mathcal{Z} \to \mathcal{W}$. In principle the policy function can be completely flexible and in the ideal world we would

[3] This would be $Y_i(1) - Y_i(0)$ in the canonical binary case that is a special case of what is discussed in the following.

assign each individual to the program with the highest conditional APO, $E[Y_i(w) \mid Z_i = z]$. However, in many cases we want to restrict the set of candidate policies denoted by Π to be interpretable for the communication with decision makers or to incorporate costs or fairness constraints. Each of these candidate policies has a policy value function, $Q(\pi) = E[Y_i(\pi(Z_i)] = E\left[\sum_w \mathbb{1}(\pi(Z_i) = w) Y_i(w)\right]$, where $\mathbb{1}(\cdot)$ is the indicator function. $Q(\pi)$ quantifies the average population outcome if policy π would be used to assign programs. The estimand of interest is then the optimal policy π^* with the highest value function for the set of candidate policies, or formally $\pi^* = \arg\max_{\pi \in \Pi} Q(\pi)$.

2.2 Identification

The previous section defined the estimands of interest in terms of potential outcomes. However, each individual is only observed in one program. Thus, only one potential outcome per individual is observable and the other potential outcomes remain latent. This is the fundamental problem of causal inference (Holland, 1986) and we need further assumptions to identify the estimands of interest. In this paper we consider the unconfoundedness assumption that assumes access to a vector of pre-treatment variables $X_i \in \mathcal{X}$ containing Z_i such that the following assumptions hold (see also, e.g. Imbens & Rubin, 2015):

Assumption 1

(a) Unconfoundedness: $Y_i(w) \perp\!\!\!\perp W_i \mid X_i = x, \forall\ w \in \mathcal{W}$, *and* $x \in \mathcal{X}$.

(b) Common support: $0 < P[W_i = w \mid X_i = x] \equiv e_w(x), \forall\ w \in \mathcal{W}$ *and* $x \in \mathcal{X}$.

(c) Stable Unit Treatment Value Assumption (SUTVA): $Y_i = Y_i(W_i)$.

The unconfoundedness assumption requires that X_i contains all confounding variables that jointly affect the program assignment and the outcome. Common support states that it must be possible to observe each individual in all programs. SUTVA rules out interference. These standard assumptions allow the identification of the average potential outcome (APO) conditional on X_i in three common ways:

$$E[Y_i(w) \mid X_i = x] = E\left[Y_i \mid W_i = w, X_i = x\right] \equiv \mu(w,x) \tag{1}$$

$$= E\left[\frac{\mathbb{1}(W_i = w)Y_i}{e_w(x)}\middle| X_i = x\right] \tag{2}$$

$$= E\left[\mu(w,x) + \frac{\mathbb{1}(W_i = w)(Y_i - \mu(w,x))}{e_w(x)}\middle| X_i = x\right] \equiv \Gamma(w,x) \tag{3}$$

The first line shows that the conditional APO is identified as a conditional expectation of the observed outcome. The second line shows that it is identified by reweighting the observed outcome with the inverse treatment probability. Finally, the third line adds the reweighted outcome residual to the conditional outcome representation of the first line. At first glance this seems redundant because the reweighted residual has expectation zero. However, it plays a crucial role for the estimation discussed in the next section. The estimators below exploit the double robustness property of representation 3 that allows identification if either $\mu(w,x)$ or $e_w(x)$ correspond to the correct function and the other one is allowed to be misspecified (see for a detailed discussion Glynn & Quinn, 2009).

For identification note that $\Gamma(w,x)$ defined in Equation 3 suffices to identify all estimands defined in the previous section:

- APO: $\gamma_w = E[Y_i(w)] = E[\Gamma(w,x)]$
- ATE: $\delta_{w,w'} = E[Y_i(w) - Y_i(w')] = E[\Gamma(w,x) - \Gamma(w',x)]$
- ATET: $\theta_{w,w'} = E[Y_i(w) - Y_i(w') \mid W_i = w] = E[\Gamma(w,x) - \Gamma(w',x) \mid W_i = w]$
- CATE: $\tau_{w,w'}(z) = E[Y_i(w) - Y_i(w') \mid Z_i = z] = E[\Gamma(w,x) - \Gamma(w',x) \mid Z_i = z]$
- Policy value: $Q(\pi) = E[Y_i(\pi(Z_i))] = E[\sum_w \mathbb{1}(\pi(Z_i) = w)\Gamma(w,x)]$
- Optimal policy: $\pi^* = \arg\max_{\pi \in \Pi} Q(\pi) = \arg\max_{\pi \in \Pi} E[\sum_w \mathbb{1}(\pi(Z_i) = w)\Gamma(w,x)]$

3 Estimation based on Double Machine Learning

3.1 The doubly robust scores

All Double Machine Learning (DML) based estimators for the estimands of interest discussed in the following build on the doubly robust scores of Robins, Rotnitzky, and Zhao (1994, 1995). This requires more notation where large Greek letters denote the scores corresponding to the small Greek letters used to define the estimands in Section 2.1.

The construction of the doubly robust scores requires the input of so-called nuisance parameters that are usually of secondary interest and seen as a tool to eventually obtain the parameters of interest. In our case the two nuisance parameters are $\mu(w, x) = E[Y_i \mid W_i = w, X_i = x]$ and $e_w(x) = P[W_i = w \mid X_i = x]$ for all w. $\mu(w, x)$ is the conditional outcome mean for the subgroup observed in program w. $e_w(x)$ is the conditional probability to be observed in program w, also known as the propensity score. Usually these functions are unknown and need to be estimated. Following Chernozhukov, Chetverikov, et al. (2018) they are estimated based on K-fold cross-fitting: (i) we randomly divide the sample in K folds of same size, (ii) we leave out fold k and estimate a prediction model for the nuisance parameters in the remaining $K - 1$ folds, (iii) we use this model to predict $\hat{\mu}^{-k}(w, x)$ and $\hat{e}_w^{-k}(x)$ in the left out fold k, and (iv) we repeat (i) to (iii) such that each fold is left out once. This procedure avoids overfitting in the sense that no observation is used to predict its own nuisance parameters. To avoid notational clutter, we ignore the dependence on the specific fold in the following notation and refer to the cross-fitted nuisance parameters as $\hat{\mu}(w, x)$ and $\hat{e}_w(x)$.

The major building block of the following estimators is the doubly robust score of the *APO*, which is the sample analogue of Equation 3:

$$\hat{\Gamma}_{i,w} = \hat{\mu}(w, X_i) + \frac{\mathbb{1}(W_i = w)(Y_i - \hat{\mu}(w, X_i))}{\hat{e}_w(X_i)}. \tag{4}$$

The *ATE* score for the comparison of treatment w and w' is then constructed as the difference of the respective APO scores:

$$\hat{\Delta}_{i,w,w'} = \hat{\Gamma}_{i,w} - \hat{\Gamma}_{i,w'} \tag{5}$$

The only estimator we consider that requires a different score is the *ATET* estimator. Although the identification result with the doubly robust APO score in the previous section is correct, it is not doubly robust. However, the doubly robust score for the ATET exists and is defined as

$$\hat{\Theta}_{i,w,w'} = \frac{\mathbb{1}(W_i = w)(Y_i - \hat{\mu}(w', X_i))}{\hat{e}_w} - \frac{\hat{e}_w(X_i)\mathbb{1}(W_i = w')(Y_i - \hat{\mu}(w', X_i))}{\hat{e}_w \hat{e}_{w'}(X_i)}, \tag{6}$$

where $\hat{e}_w = N_w/N$ is the unconditional treatment probability with N_w counting the number of individuals observed in program w (see also, e.g. Farrell, 2015).

3.2 Average potential outcomes and treatment effects

The estimation of the APOs, ATEs and ATETs boils down to taking the means of the previously defined doubly robust scores. For statistical inference we can rely on standard one-sample t-tests. Put more formally, the score's mean and the variance of this mean are the point and the variance estimate of the respective estimand of interest:

- APO: $\hat{\mu}_w = N^{-1} \sum_i \hat{\Gamma}_{i,w}$ and $\hat{\sigma}^2_{\mu_w} = N^{-1} \sum_i (\hat{\Gamma}_{i,w} - \hat{\mu}_w)^2$
- ATE: $\hat{\delta}_{w,w'} = N^{-1} \sum_i \hat{\Delta}_{i,w,w'}$ and $\hat{\sigma}^2_{\delta_{w,w'}} = N^{-1} \sum_i (\hat{\Delta}_{i,w,w'} - \hat{\delta}_{w,w'})^2$
- ATET: $\hat{\theta}_{w,w'} = N^{-1} \sum_i \hat{\Theta}_{i,w,w'}$ and $\hat{\sigma}^2_{\theta_{w,w'}} = N^{-1} \sum_i (\hat{\Theta}_{i,w,w'} - \hat{\theta}_{w,w'})^2$

Note that the estimated variances require no adjustment for the fact that we have estimated the nuisance parameters in a first step. The fundamental contribution of the methodological papers analyzing these estimators is to show that the scores contain all relevant information and inference that is solely based on them is valid. The resulting estimators are consistent, asymptotically normal and semiparametrically efficient under the main assumption that the estimators of the nuisance parameters are consistent and converge sufficiently fast (Belloni et al., 2014; Farrell, 2015; Belloni, Chernozhukov, Fernández-Val, & Hansen, 2017; Chernozhukov, Chetverikov, et al., 2018). In particular, the product of

the convergence rates of the outcome and propensity score estimators must be at least $n^{1/2}$. This allows to apply machine learning to estimate the nuisance parameters.[4] Flexible machine learning estimators converge usually slower than the parametric rate $n^{1/2}$ but several are known to be able to achieve $n^{1/4}$, which would be sufficiently fast if both nuisance parameter estimators achieve it.[5]

It is well known that the described estimators are doubly robust in the sense that they remain consistent if one of the parametric nuisance parameter models is misspecified. The innovation of the DML version is that it exploits what Smucler, Rotnitzky, and Robins (2019) call 'rate double robustness'. This robustness allows to estimate the parameters of interest with $n^{1/2}$ even if the nuisance parameters are estimated at slower rates using machine learning methods that do not require the specification of an actual parametric model. The same would not be possible for estimators based on only one nuisance parameter like outcome regressions or Inverse Probability Weighting (IPW), which could be motivated by the identification results in Equations 1 and 2, respectively.

3.3 Conditional average treatment effects

We can reuse the ATE score of Equation 5 to estimate conditional effects. In the following we discuss estimators that exploit the fact that the conditional expectation of the ATE score identifies CATE: $\tau_{w,w'}(z) = E[\Delta_{i,w,w'} \mid Z_i = z]$.[6] Thus, using the score $\hat{\Delta}_{i,w,w'}$ as a pseudo-outcome in a standard regression framework is a way to estimate CATEs.

First, consider to apply *ordinary least squares* (OLS) by plugging the pseudo-outcome in the standard OLS minimization problem,

$$\hat{\beta}_{w,w'} = \arg\min \sum_{i=1}^{N} \left(\hat{\Delta}_{i,w,w'} - \tilde{Z}_i \beta_{w,w'} \right)^2,$$

where $\tilde{Z}_i$ contains the original Z_i and a constant. The resulting coefficients $\hat{\beta}_{w,w'}$ have the

[4] Further results, regularity conditions and discussions can be found in section 5.1 of Chernozhukov, Chetverikov, et al. (2018).

[5] For example, versions of Lasso (Belloni & Chernozhukov, 2013), Boosting (Luo & Spindler, 2016), Random Forests (Wager & Walther, 2015), Neural Nets (Farrell, Liang, & Misra, 2018) or ensembles of those can be shown to achieve the required rates under certain conditions that can be found in the original papers.

[6] Note that this does not work for the ATET score in Equation 6 and suitable adaptations are beyond the scope of this paper.

same *ceteris paribus* interpretation as in a standard OLS model. The only difference is that instead of linearly modelling the level of an outcome, they model the level of a causal effect. Consequently, the fitted values[7] estimate CATEs if we specify a sufficiently flexible model or assume that we know the true functional form. If we allow for a misspecified model, the fitted values still provide the best linear predictor (BLP) of the CATE. Most importantly, Semenova and Chernozhukov (2017) show that standard heteroscedasticity robust standard errors are valid and that we can again ignore the fact that the nuisance parameters are estimated and potentially converge slower than $n^{1/2}$.

A complementary option for few continuous Z_i is proposed by Fan et al. (2019) and Zimmert and Lechner (2019). The pseudo-outcome can also be used in *nonparametric regressions* (NPR):

$$\hat{\tau}^{np}_{w,w'}(z) = \sum_{i=1}^{N} \frac{\mathcal{K}_h\left(Z_i - z\right) \hat{\Delta}_{i,w,w'}}{\sum_{i=1}^{N} \mathcal{K}_h\left(Z_i - z\right)}$$

where $\mathcal{K}_h(\cdot)$ is a suitable Kernel function with bandwidth h. Fan et al. (2019) and Zimmert and Lechner (2019) show that like in the OLS case the uncertainty of the nuisance parameter estimation can be neglected and standard statistical inference for nonparametric regression applies. However, there is a price to pay for this flexibility in terms of the required speed of convergence of the nuisance parameter estimators. Average effects or OLS CATE estimation can ignore the estimation of the nuisance parameters for statistical inference if their product achieves $n^{1/2}$ convergence. For nonparametric regressions this requirement depends on the dimension of Z_i. For example, the product of the convergence rates needs to achieve $n^{3/5}$ for a one dimensional continuous Z_i and further increases with more variables.

From a practical point of view the results for OLS and nonparamteric regression are very convenient because they allow to adopt standard statistical software for estimation and statistical inference in a modular way without adaptations.

In principle, we can also apply supervised machine learning to estimate CATEs as the conditional expectation of the pseudo-outcome. For example using random forest (RF) is one of the best performing CATE estimators in the simulation study of Knaus,

[7]Formally defined as $\hat{\tau}^{ols}_{w,w'}(z) = \langle \tilde{z}, \hat{\beta}_{w,w'} \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the inner product.

Lechner, and Strittmatter (2018). However, statistical inference for such DML based flexible CATE estimation is not yet well understood for low-dimensional Z_i and impossible for high-dimensional Z_i as discussed by Chernozhukov, Demirer, Duflo, and Fernandez-Val (2017).

3.4 Optimal treatment assignment

The APO score can further be reused as an input to estimate optimal treatment assignment. First, note that the value function of any policy $\pi(Z_i)$ can be estimated as

$$\hat{Q}(\pi) = N^{-1} \sum_{i=1}^{N} \sum_{w=0}^{T} \mathbb{1}(\pi(Z_i) = w) \hat{\Gamma}_{i,w}.$$

This means each individual contributes the score of the policy that she is assigned to under this policy. However, we are not necessarily interested in the value function of some policy, but want to estimate the optimal policy that maximizes this value function, $\hat{\pi}^* = \arg\max_{\pi \in \Pi} \hat{Q}(\pi)$. This is a non-convex optimization problem and requires to search over all candidate policies to find the optimum.

For example, consider the case of a binary covariate Z_i and a binary treatment W_i. We have four different policy options: treat nobody (π^1), treat only those with $Z_i = 1$ (π^2), treat only those with $Z_i = 0$ (π^3), or treat everybody (π^4). We illustrate this using two representative observations, $i = 1$ with $Z_1 = 0$, and $i = 2$ with $Z_2 = 1$:

i	Z_i	π^1	π^2	π^3	π^4	$\hat{Q}(\pi^1)$	$\hat{Q}(\pi^2)$	$\hat{Q}(\pi^3)$	$\hat{Q}(\pi^4)$
1	0	0	0	1	1	$\hat{\Gamma}_{1,0}$	$\hat{\Gamma}_{1,0}$	$\hat{\Gamma}_{1,1}$	$\hat{\Gamma}_{1,1}$
2	1	0	1	0	1	$\hat{\Gamma}_{2,0}$	$\hat{\Gamma}_{2,1}$	$\hat{\Gamma}_{2,0}$	$\hat{\Gamma}_{2,1}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The columns three to six show the assignments under the four potential assignment rules. For example, the first observations receive no treatment under policies π^1 and π^2, but is treated under policies π^3 and π^4. To find the optimal rule, we compare the means of the APO scores in the last four columns and pick the policy that corresponds to the largest mean. The number of policy values to compare increases exponentially in more general settings with multiple treatments and non-binary Z_i. Furthermore, we expect that in finite

samples and estimated nuisance parameters the identified optimum does not coincide with the true optimal policy. This is conceptualized as the 'regret' defined as the difference between the true and the estimated optimal value function, $R(\hat{\pi}^*) = Q(\pi^*) - Q(\hat{\pi}^*)$.

Zhou et al. (2018) show that the DML based procedure minimizes the maximum regret asymptotically under two main conditions: First, the same convergence conditions for the nuisance parameters that are required for ATE estimation (the product of the nuisance parameters achieves $n^{1/2}$). Second, the set of candidate policies Π is not too complex. In particular, they show that decision trees with fixed depth are a suitable policy class. Again the double robustness of the used scores results in statistical guarantees that are not achievable for methods based on outcome regressions or IPW alone.

4 Application: Swiss Active Labor Market Policy

We use a standard dataset of Swiss Active Labor Market Policy (ALMP) that is already basis of previous studies (Huber et al., 2017; Knaus et al., 2020; Lechner, 2018).[8] In particular, we start with the sample of 100,120 unemployed individuals of Huber et al. (2017) that consists of 24 to 55 year old individuals registered unemployed in 2003. We consider non-participants and participants of four different *program types*: job search, vocational training, computer programs and language courses.[9] As the assignment policies differ substantially across the three language regions, we focus only on individuals living in the German speaking part and remove those in the French and Italian speaking part to avoid common support problems.

This leaves us with 67,577 observations. We evaluate the first program participation within the first six months after the begin of the unemployment spell. One problem of this definition is that non-participants comprise people that quickly come back into employment before they would be assigned to a training program. This could result in an overly optimistic evaluation of non-participation. We follow Lechner (1999) and Lechner and Smith (2007) and assign pseudo program starting points to the non-participants and

[8] Gerfin and Lechner (2002), Lalive, van Ours, and Zweimüller (2008) and Knaus et al. (2020) among others provide a more detailed description of the surrounding institutional setting.

[9] The dataset contains also participants of an employment program and personality training. However, we leave them out to keep the number of obtained results manageable.

Table 1: Descriptive statistics of selected variables by program type

	No program	Job search	Vocational	Computer	Language
	(1)	(2)	(3)	(4)	(5)
No. of observations	47,544	11,610	858	905	1504
Outcome: months employed of 31	14.7	14.4	18.4	19.2	13.5
Female (binary)	0.44	0.44	0.33	0.60	0.55
Age	36.61	37.31	37.45	39.08	35.28
Foreigner (binary)	0.36	0.33	0.30	0.21	0.67
Employability	1.93	1.98	1.93	1.97	1.85
Past income in CHF 10,000	4.25	4.67	4.87	4.32	3.73

Note: Employability is an ordered variable with one indicating low employability, two medium employability and three high employability. The exchange rate USD/CHF was roughly 1.3 at that time. The full set of variables is reported in Table A.1.

keep only those who are still unemployed at this point.[10] This results in a final sample size of 62,421 observations.

The *outcome* of interest is the cumulated number of months in employment in the 31 months after program start, which is the maximum available time span. Row one of Table 1 provides the number of observations in each group. Roughly 75% participate in no program. By far the largest program is the job search program which is also called basic program. The more specific programs are much smaller with roughly 1000 observations each. Row two shows that the average outcomes substantially differ by different groups. However, it is not clear whether this is only due to selection effects because, as the remaining rows show, the observable characteristics are not comparable across groups. Especially the share of females, the share of foreigners and past income differ quite substantially across programs. The *control variables* comprise these five and 38 additional variables that are reported in Table A.1. They consist of socio-economic characteristics of the unemployed individuals, caseworker characteristics, information about the assignment process, information about the previous job and cantonal economic indicators.

[10]The assignment of the pseudo starting point is based on estimated probabilities to start a program at a specific time. The probability depends also on covariates and is estimated using the same random forest specification that is discussed later in Section 5.

5 Implementation

We estimate the nuisance parameters via Random Forest (Breiman, 2001) using the implementation with honest splitting in the `grf` R-package (Athey et al., 2019) and 5-fold cross-fitting. The tuning parameters in each regression are selected by out-of-bag validation. All regressions apply the full set of control variables listed in Table A.1. We run the outcome regressions for each treatment group separately to obtain $\hat{\mu}(w, x)$. Also the propensity scores are estimated for each treatment separately using a treatment indicator as outcome in the random forest. The propensity scores are then normalized to sum to one within an individual.

We estimate CATEs at different granularity across a set of five handpicked variables that are regularly used in the program evaluation literature. First, we investigate Group Average Treatment Effects (GATEs) for subgroups by gender, foreigners and three categories of employability. Usually these standard subgroup analyses require to re-estimate everything in the subgroups. However, after DML for average effects it can be performed at very low computational costs in a standard OLS regression with the pseudo-outcome described in Section 3.3 and dummy variables for all groups but the reference group. Second, we estimate nonparametric CATEs for the continuous variables age and past income. We use the R-package `np` package for the required nonparametric regressions (Hayfield & Racine, 2008). The regressions apply a second-order Gaussian kernel function and use 0.9 of the cross-validated bandwidth for undersmoothing as suggested by Zimmert and Lechner (2019). Third, we specify an OLS model in which all the five previously used variables enter linearly. Fourth, we use the same five characteristics in a Random Forest regression to avoid functional form assumptions.[11] Finally, we go beyond the handpicked variables and use all 43 control variables in a Random Forest to estimate what Knaus et al. (2018) call Individualized Average Treatment Effects (IATEs).

The optimal treatment assignment rule is estimated as decision trees of depth two, three and four. We follow Algorithm 2 of Zhou et al. (2018) and implement an exact tree-search. Usually classification and regression tress are build in a greedy fashion by splitting at each

[11] We use the same implementation and tuning as for the estimation of the nuisance parameters. In the absence of any theoretical guidance, we use the out-of-bag predictions to estimate the CATE for each individual.

Table 2: Steps of implementation

Step	Input	Operation	Output
1.	W_i, X_i	Predict treatment probabilities	$\hat{e}_w(x)$
2.	Y_i, W_i, X_i	Predict treatment specific outcomes	$\hat{\mu}(w,x)$
3.	Y_i, W_i, $\hat{e}_w(x)$, $\hat{\mu}(w,x)$	Plug in Equation 4	$\hat{\Gamma}_{i,w}$
4.	$\hat{\Gamma}_{i,w}$	mean, one-sample t-test	APOs
5.	$\hat{\Gamma}_{i,w}$	Take difference	$\hat{\Delta}_{i,w,w'}$
6.	$\hat{\Delta}_{i,w,w'}$	mean, one-sample t-test	ATEs
7.	$\hat{\Delta}_{i,w,w'}$, Z_i	Ordinary least squares	GATEs or BLP CATEs
8.	$\hat{\Delta}_{i,w,w'}$, Z_i	Nonparametric regression	NPR CATEs
9.	$\hat{\Delta}_{i,w,w'}$, Z_i	Random forest	RF CATEs
10.	$\hat{\Gamma}_{i,w}$, Z_i	Optimal decision tree	OTR

Notes: BLP means best linear predictor, NPR means nonparametric regression, and RF stands for random forest.

step to optimize the respective criterion. This is computationally convenient but usually results not in the optimal tree. To find the optimal tree, one needs to try every possible split, which is computationally more involved. As we report in the results below, the first split is different depending on the fixed depth of the tree. For a greedy tree algorithm, this split would be the same for all trees irrespective of their depth. We estimate the trees first with the five handpicked variables. However, these variables include gender and foreigner status that might be too sensitive to include. Thus, we investigate another set of 16 variables that includes only objective measures of individual education and labor market history of the unemployed persons that are immediately available to the caseworker from the administrative records.

Table 2 summarizes all required implementation steps. It highlights that a comprehensive DML based program evaluation can be run with few lines of code in any statistical software program that is capable of the operations in the third column. Thus, researchers can build their customized analyses in a modular fashion based on established code. Only the optimal tree in the final step might require to be implemented along the lines of Algorithm 2 in Zhou et al. (2018), which has been recently provided for R in the `policytree` package.

6 Results

6.1 Nuisance parameters

Before looking at the parameters of interest we discuss some features of the nuisance parameters. They are only a tool to remove confounding but it is still informative to understand which variables are the most important confounders. This is less straightforward for flexible tools like random forests than for the well-known regression outputs of parametric models.

We combine two sources of information to investigate important confounders. First, we use the variable importance measure of the `grf` package that essentially measures how often the random forest splits along a specific variable and sums up to one. In the following, we consider those variables as main confounders that receive on average over all treatment groups at least a value of 0.05 for either propensity score or outcome regressions.[12] Second, we enrich this measure with the direction of the correlation of the covariate with the predicted values. We multiply the variable importance measure by minus one if this correlation is negative. This is very condensed and ignores any possible non-linearities. However, it allows to compactly summarize the most important confounders in the estimation of the nuisance parameters in Figure A.1.

Figure A.1a shows that macroeconomic factors are important predictors of program selection. In particular, caseworkers in cantons with a higher GDP per capita and a high unemployment rate have a higher probability to assign job search programs. The other important predictors of assignment probabilities are previous labor market success and measures for being a foreigner. In line with intuition the latter seem to drive a large part of the selection into language courses. Also Figure A.1b depicting the outcome regressions shows intuitive patterns. Measures of qualification and previous labor market success are highly predictive for future employment with the correlations pointing in the expected directions. For example, holding a degree is positively and previous employment

[12] Completely random splitting of the 43 variables would result in 0.023 importance for every variable. We choose to consider those with twice this random benchmark as main confounders. The factor of two is ad-hoc and has no statistical basis. The variable importance measures of all variables are provided in Tables A.2 and A.3 of Appendix A.

as unskilled worker negatively associated with future employment.

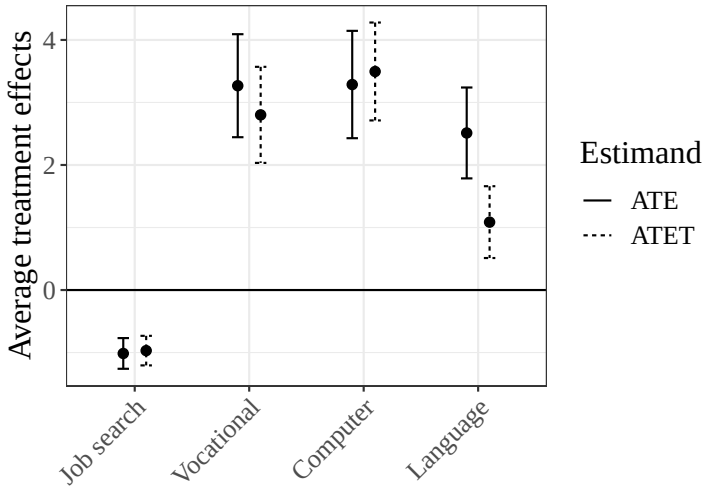
Finally, we investigate the propensity score distributions for all programs. Figure A.2 in Appendix A shows that propensity scores are quite variable. This indicates that selection into programs is not negligible. Further, we observe that some of the propensity scores get quite small with the smallest one being 0.003 for a computer training participant. This is not surprising given that the share of observations in computer and vocational training is only about one percent. However, the small propensity score *per se* is not an indicator of poor overlap. We could easily increase the smallest propensity score by randomly removing a large fraction of non-participants and participants of the job search program. This would discard valuable information and shows that the mere focus on the smallest propensity score can be misleading in cases with imbalanced treatment group sizes. More importantly, we observe overlap in the sense that all treatment groups contain individuals with similarly low propensity scores. Thus, overlap seems not to be a major issue in our application. As a robustness check, Appendix B reruns the whole analysis where we enforce overlap. This results in the trimming of 1,954 observations but keeps all main results unaltered. Such trimming shows best performance in the simulation study of Lechner and Strittmatter (2019) that focuses on propensity score methods. However, it is *a prior* not clear whether this holds for DML based methods as well because they additionally involve an outcome regression.

6.2 Average effects

In the following discussion we focus on the effects that compare the four programs to non-participation and do not discuss the effects comparing the programs with each other.[13] Recall that the outcome of interest is the cumulated number of months employed in the 31 months after program start. Figure 2 depicts ATE and ATET estimates and shows substantial differences in the effectiveness of programs. The job search program decreases the months in employment on average by about one month. In contrast, other programs that teach hard skills show substantial improvements with roughly three additional months in employment on average.

[13]The underlying APOs are shown in Figure A.3 of Appendix A.

Figure 2: Average treatment effects of participation vs. non-participation



Note: The figure shows the point estimates of the average treatment effects of participating in the program labeled on the x-axis vs. non-participation and their 95% confidence intervals.

For a better understanding of the underlying dynamics, Figure A.4 in Appendix A reports the effects of program participation on the employment probabilities over time. All program participants suffer from the well-known lock-in effect within the first months after program start (see, e.g. Wunsch, 2016). However, participants of the hard skill programs catch up and show a sustained increase in employment rates of up to 10 percentage points.

Comparing ATE and ATET shows no big differences for most programs. This suggests that there is either no effect heterogeneity correlated with observables or that the assignment does not take advantage of this heterogeneity. Usually we would expect to see ATETs being higher than ATEs if program assignment is well targeted. However, we find only evidence for the opposite as the actual participants of a language course show a 1.5 months lower treatment effect compared to the population.[14] The ATET is still positive but this difference suggests that there is substantial effect heterogeneity to uncover and the potential to improve treatment assignment.

6.3 Heterogeneous effects

This section studies effect heterogeneity at different granularity. We start by estimating *GATEs*. Panel A of Table 3 shows the result of an OLS regression with a female dummy as covariate, $\hat{\Delta}_{i,w,w'} = \beta_0 + \beta_1 female_i + error_i$. The constant ($\beta_0$) provides the GATE for the

[14]Figure A.5 in Appendix A shows the differences of ATET and ATE as well as their confidence intervals.

Table 3: Group average treatment effects

	Job search	Vocational	Computer	Language
	(1)	(2)	(3)	(4)
Panel A:				
Constant	-1.28***	3.82***	2.25***	3.31***
	(0.17)	(0.55)	(0.61)	(0.46)
Female	0.61**	-1.25	2.35***	-1.81**
	(0.25)	(0.85)	(0.87)	(0.76)
Panel B:				
Constant	-1.26***	2.56***	3.69***	3.54***
	(0.16)	(0.52)	(0.50)	(0.51)
Foreigner	0.66***	1.96**	-1.12	-2.84***
	(0.26)	(0.88)	(0.97)	(0.71)
Panel C:				
Constant	-0.18	5.57***	5.81***	2.65***
	(0.33)	(1.04)	(1.08)	(0.85)
Medium employability	-0.91**	-2.53**	-2.86**	-0.24
	(0.36)	(1.15)	(1.20)	(0.96)
High employability	-1.61***	-4.40***	-4.12**	0.47
	(0.50)	(1.49)	(1.68)	(1.48)
F-statistic	5.46***	3.67**	3.48**	0.17

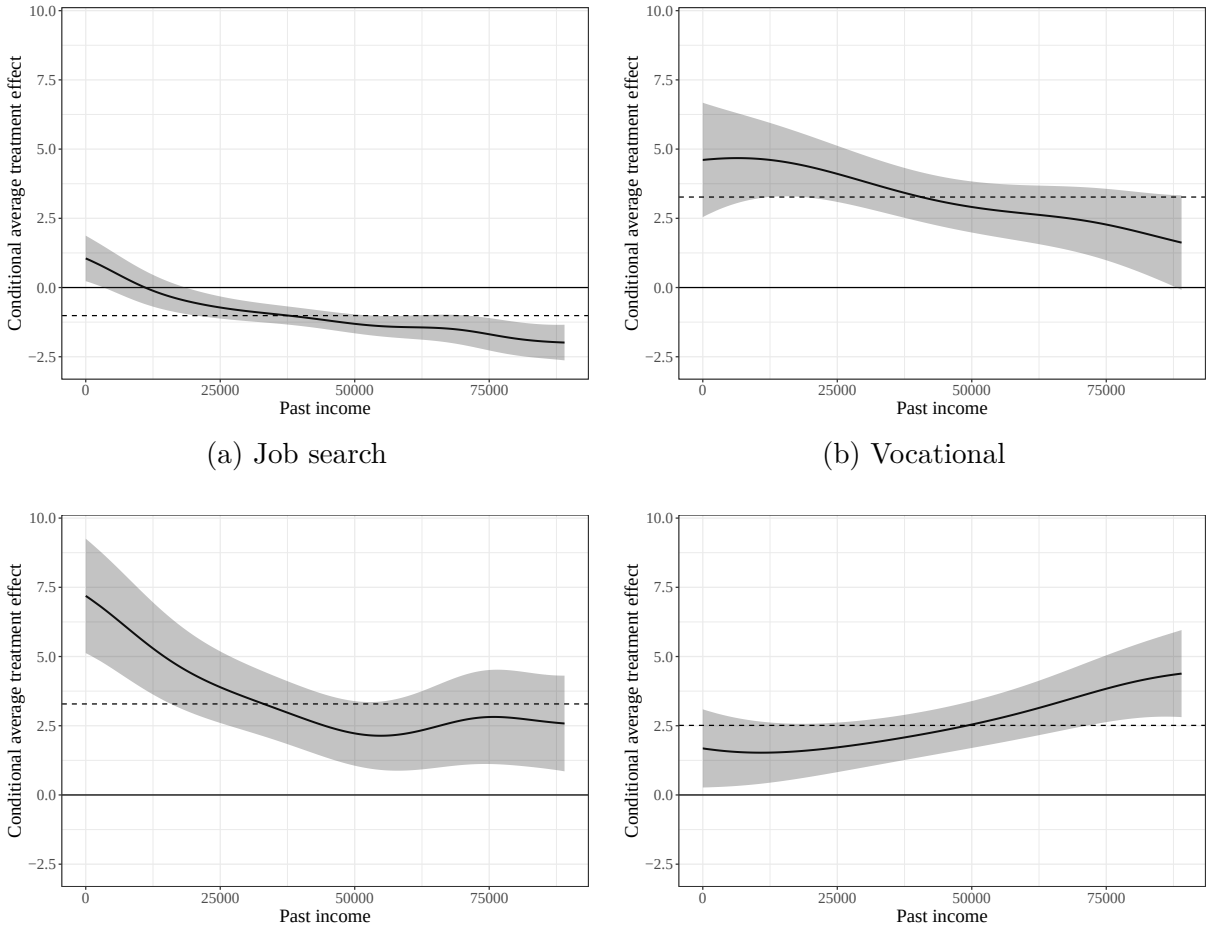
Note: This table shows OLS coefficients and their heteroscedasticity robust standard errors (in parentheses) of regressions run with the pseudo-outcome defined as described in Section 3.3. *p<0.1; **p<0.05; ***p<0.01

reference group men and the female coefficient (β_1) describes how much the GATE differs for women. We see substantial gender difference in the effectiveness of programs. Women significantly suffer less or profit more from job search and computer program participation. This gender gap in the effectiveness of ALMPs is also well-documented in the literature (Crépon & van den Berg, 2016; Card, Kluve, & Weber, 2018). In contrast to this, we find that women profit on average significantly less from language courses than men.

Panel B replaces the female dummy in the regression by a foreigner dummy. The striking result is here that Swiss as reference group show a big positive effect for participating in language courses but the effect disappears for foreigners. After adding the coefficient for foreigners to the constant the foreigners' CATE is only 0.7 ($3.54 + (-2.84)$, standard error: 0.62). A crucial information to better understand this finding would be to know which languages they learn, which is unfortunately not available in this dataset.

Panel C shows the results of a similar regression but now with two dummies indicating medium and high employability such that low employability becomes the reference group. The F-statistic in the last line tests the joint significance of the two dummies. It is

Figure 3: Effect heterogeneity regarding past income



(a) Job search

(b) Vocational

(c) Computer

(d) Language

Note: Dotted line indicates point estimate of the respective average treatment effect. Grey area shows 95%-confidence interval.

statistically significant at the 5%-level for the programs in the first three columns. They all show a common gradient that individuals with low employability benefit substantially more or at least suffer less from program participation.

While subgroup analyses are standard in policy evaluations, the estimation of *nonparametric regression CATEs* along continuous variables is rarely pursued. Here we estimate such CATEs along the continuous variables age and past income. We detect no significantly different CATEs for age.[15] However, effect sizes are clearly associated with past income. Figure 3 shows a decreasing effect size with higher past income for all but for language programs. The latter have only a negligible positive effect for individuals with low past income but it increases gradually with higher income. One potential explanation for these

[15]Figure A.6 shows the according results for completeness.

Table 4: Best linear prediction CATEs

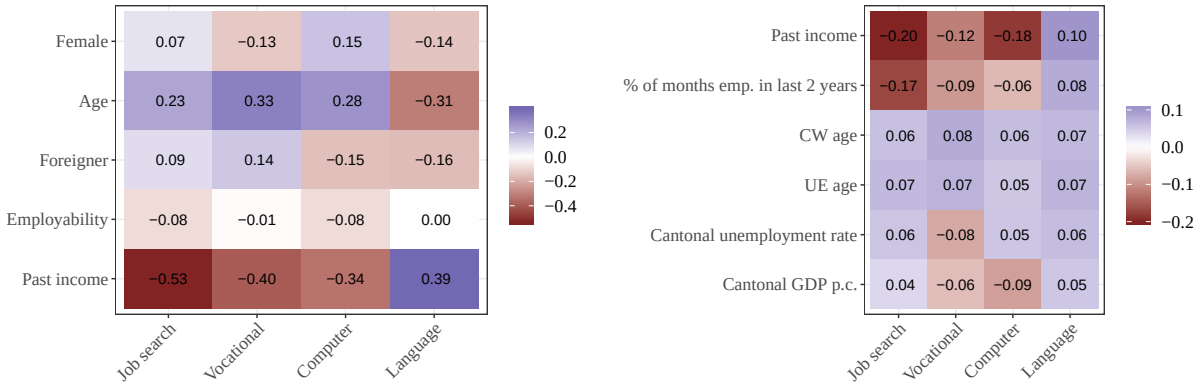
	Job search	Vocational	Computer	Language
	(1)	(2)	(3)	(4)
Constant	-0.50	4.45*	6.30***	5.26***
	(0.70)	(2.37)	(2.43)	(2.02)
Female	0.23	-2.11**	1.73*	-1.48*
	(0.27)	(0.91)	(0.92)	(0.80)
Age	0.03*	0.10**	0.02	-0.05
	(0.01)	(0.05)	(0.05)	(0.04)
Foreigner	0.45*	1.39	-1.50	-2.76***
	(0.26)	(0.89)	(0.99)	(0.73)
Medium employability	-0.61*	-1.75	-2.63**	-0.79
	(0.37)	(1.18)	(1.23)	(0.99)
High employability	-1.12**	-3.12**	-3.81**	-0.50
	(0.51)	(1.52)	(1.73)	(1.51)
Past income in CHF 10,000	-0.27***	-0.63***	-0.39**	0.29
	(0.06)	(0.23)	(0.19)	(0.18)
F-statistic	6.72***	3.95***	3.13***	3.86***

Note: This table shows OLS coefficients and their heteroscedasticity robust standard errors (in parentheses) of regressions run with the pseudo-outcome defined in Section 3.3. *p<0.1; **p<0.05; ***p<0.01

findings is that the value of language skills is larger for high-skilled workers in multilingual countries like Switzerland because they reduce information costs across language borders (Isphording, 2014).

The CATEs so far were nonparametric but only univariate. Now we model the CATE by specifying a multivariate OLS regression with the previously used covariates entering linearly. It is most likely misspecified and thus estimates the *best linear predictor* (BLP) of CATEs. However, it provides a compact and accessible summary of the effect heterogeneities. Additionally it holds the other variables constant allowing for a standard *ceteris paribus* interpretation of the coefficients that standard subgroup analyses are missing. Consider for example the coefficients for the female dummy in Table 4. Compared to Table 3 the female coefficients in the first three columns are smaller and the one for language courses is larger. The reason is that it represents a partial effect that holds other variables like past income fixed. The female coefficient in Table 3 partly picks up that women have lower past income and that lower income is associated with higher treatment effects for all but language courses. This discussion illustrates that the same strategies that are usually applied to interpret an outcome OLS model can now be used to interpret the effect OLS

Figure 4: Variable importance for CATE estimation



(a) Handpicked variables

(b) All variables

Notes: Variable importance measures are multiplied by minus one if the correlation between the covariate the the estimated CATE is negative.

model.

Running a random forest regression with the pseudo-outcome imposes no functional form restrictions and thus estimates CATEs. However, the interpretation of the detected heterogeneity is less straightforward than providing a standard OLS table. Following the procedure in Section 6.1, Figure 4a shows the variable importance measure with the sign of the correlation with the estimated CATEs for the Random Forest with five variables. We find that past income is by far the most important variable. The signs of the correlations are in line with the signs of the respective OLS coefficients in Table 4. To better understand the magnitudes of the correlations, we conduct a classification analysis as proposed by Chernozhukov, Fernandez-Val, and Luo (2018). This involves to define the groups of least and most affected individuals. Here we choose individuals in the first quintile of the CATE distribution as least affected and those in the fifth quintile as most affected. We compare then the covariate means of these two groups. Panel A of Table 5 shows the differences in these means. For comparability, we normalized all variables to have mean zero and variance one. For example, we observe that individuals that are most affected by a job search program have a 2.16 standard deviations lower past income compared to the least affected group. All these patterns are in line with the previous results.

Finally, we estimate the Individualized Average Treatment Effects (IATEs) based on all control variables. Figure 4b and Panel B of Table 4 show the variable importance and the

Table 5: Classification analysis of random forest CATEs

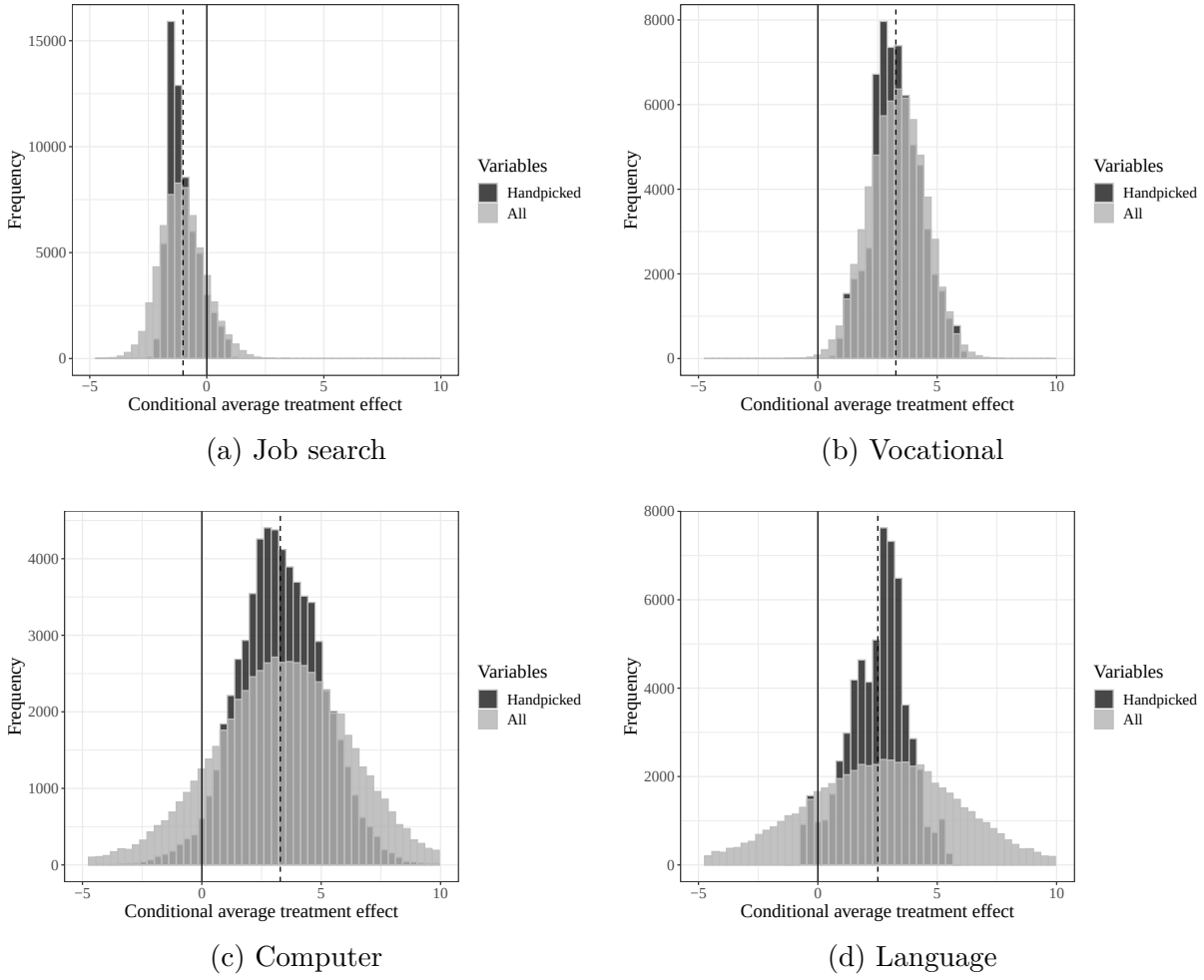
	Job search	Vocational	Computer	Language
	(1)	(2)	(3)	(4)
Panel A: 5 handpicked variables				
Female	1.06	-0.41	1.17	-0.88
Age	0.17	0.90	0.19	-0.16
Foreigner	0.61	1.15	-0.53	-1.75
Low employability	0.72	0.64	0.74	-0.32
Medium employability	-0.37	-0.33	-0.52	0.12
High employability	-0.35	-0.31	-0.15	0.23
Past income	-2.16	-1.55	-1.17	1.67
Panel B: All variables				
Past income	-1.32	-1.19	-0.86	0.59
% of months emp. in last 2 years	-1.16	-0.65	-0.37	0.18
Caseworker age	0.22	0.51	0.08	0.14
Unemployed age	0.07	0.49	0.16	-0.01
Cantonal unemployment rate	0.30	-1.08	-0.00	0.02
Cantonal GDP p.c.	0.20	-1.06	-0.03	0.17

Note: Table shows the difference in means of normalized covariates between the fifth and the first quintile of the respective estimated CATE distribution.

classification analysis of the six variables with the highest variable importance, respectively. The same pattern arises that past labor market success seems to be most predictive for the size of program effects. However, the distributions of the predicted IATEs suggest that these detailed effects can not be reliably estimated. Figure 5 plots the estimated distribution of the CATEs with the five handpicked variables and the IATEs estimated with all control variables. We observe that the low dimensional CATEs are relatively compact. However, the IATEs especially for the computer and language programs are very dispersed and show implausibly high effect sizes. The most extreme ones are censored in the Figure but some even exceed the maximum possible value of 31 with a maximum effect of 46. Such extreme behaviour of estimators based on the doubly robust score of Equation 4 is documented by Kang and Schafer (2007) for the estimation of average effects and called a violation of the boundedness property by Robins, Sued, Lei-gomez, and Rotnitzky (2007). This suggests that the estimation of too detailed heterogeneities is problematic in our application. Thus we do not investigate the IATEs further and leave a deeper investigation of this issue for future research.

In summary, we document substantial heterogeneities in the effectiveness of the different

Figure 5: Distribution of CATEs



(a) Job search

(b) Vocational

(c) Computer

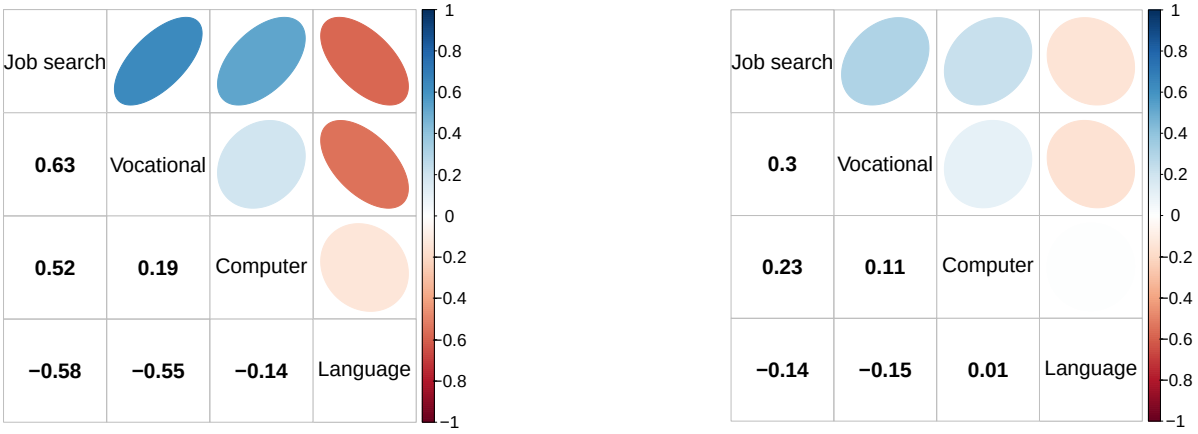
(d) Language

Note: The dark grey histogram shows the estimated CATE distribution with the five handpicked variables female, age, foreigner, employability and past income. The light grey histogram shows the estimated IATEs and is censored for computer and language programs.

programs. However, another interesting question is whether these CATEs are correlated over different treatments. This allows to assess whether the same group of people shows the highest effects for all programs or whether programs are complements in the sense that they work better for different groups. Figure 6 reports the correlations and shows that most programs are substitutes in the sense that they work better for the same groups of people. However, the CATEs of language programs are negatively correlated with other CATEs. This is not surprising given that we documented different signs of the correlation with past income in the heterogeneities of language programs compared to the other programs. Although the correlations do not take into account that the average effects are at different levels, these correlations suggest that different programs work better for

different groups and that this could be exploited for improved targeting.

Figure 6: Correlation of Random Forest CATEs across programs



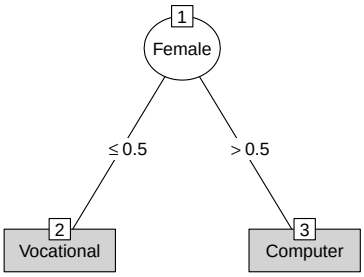
(a) Handpicked variables

(b) All variables

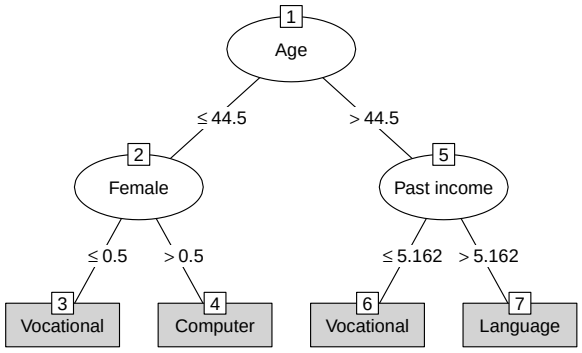
6.4 Optimal treatment assignment

The previous section documented substantial heterogeneities in the program effects. To leverage this heterogeneity for better targeting we apply the DML based optimal policy algorithm of Section 3.4. Figure 7a shows the simplest decision tree with only one split for the five handpicked covariates. It would allocate men to vocational training and women to computer courses. This split is probably similar to what we would have suggested given the evidence presented in Table 3. For a tree of depth three such an eyeballing approach has its limits and the algorithmic approach provides a systematic way to arrive at the optimal decision tree. The tree in Figure 7b splits first on age and then along gender or past income. The split on age might be surprising as it is not one of the main drivers of heterogeneity but the interaction with the female indicator for younger and with past income for the older are obviously important from a optimal assignment perspective. In the absence of the possibility to split on gender, the depth two tree in Figure 7c splits on past income roughly at the same value where the nonparametric CATEs of vocational and language training intersect in Figure 3. Appendix A.6 provides also the trees of depth four. Panel A of Table 6 summarizes the results of the different trees. It shows the percentage of individuals that are placed in the different programs. Not surprisingly all individuals are recommended to be placed into the positively evaluated hard skill enhancing programs.

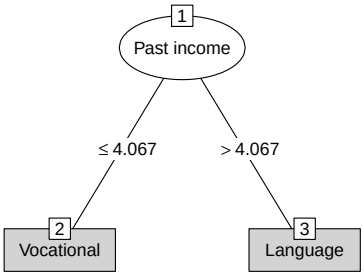
Figure 7: Optimal treatment assignment decision trees of depth two and three



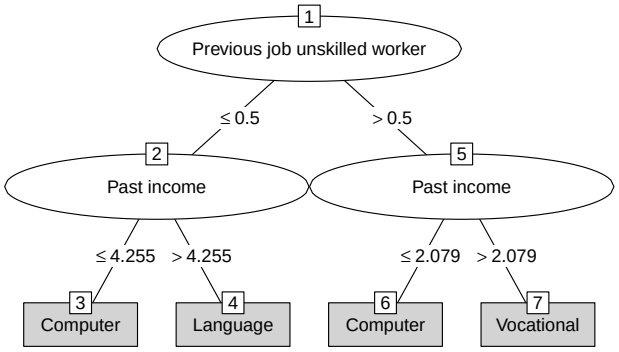
(a) Depth 2 & 5 covariates



(b) Depth 3 & 5 covariates



(c) Depth 2 & 16 covariates



(d) Depth 3 & 16 covariates

Notes: Past income is measured in CHF 10,000. Graphs are created with the R package `partykit` of Hothorn and Zeileis (2015).

One yet unsolved challenge is how to draw statistical inference about the quality and stability of the decision trees. Athey and Wager (2017) propose a form of cross-validation. To this end, we use the same folds that were used in the cross-fitting procedure to estimate the nuisance parameters. We build the decision tree in four folds and evaluate the value in the left out fold. First, we inspect how often the recommendations based on these trees coincide with the full sample policies. Figures A.9 and A.11 show that the cross-validated trees are not identical to the full sample ones. However, for the vast majority at least three cross-validated trees agree with the full sample. The exception is the tree of depth two with five variables that is more or less indifferent whether to split like in Figure 7a (two trees) or to split like in Figure 7c where nonparametric CATEs of past income of vocational training and language programs cross (three trees).

Table 6: Description of estimated optimal policies

	No program	Job search	Vocational	Computer	Language
	(1)	(2)	(3)	(4)	(5)
Panel A: Percent allocated to program					
Depth 2 & 5 variables	0	0	56	44	0
Depth 3 & 5 variables	0	0	56	35	9
Depth 4 & 5 variables	0	0	34	46	20
Depth 2 & 16 variables	0	0	48	0	52
Depth 3 & 16 variables	0	0	23	36	42
Depth 4 & 16 variables	0	0	49	26	25
Panel B: Cross-validated difference to APOs					
Depth 2 & 5 variables	3.30***	4.32***	0.03	0.01	0.79*
	(0.41)	(0.43)	(0.43)	(0.50)	(0.47)
Depth 3 & 5 variables	3.62***	4.64***	0.36	0.34	1.11**
	(0.40)	(0.41)	(0.46)	(0.47)	(0.46)
Depth 4 & 5 variables	3.30***	4.32***	0.04	0.02	0.79*
	(0.41)	(0.42)	(0.49)	(0.45)	(0.48)
Depth 2 & 16 variables	3.57***	4.59***	0.30	0.28	1.06**
	(0.43)	(0.44)	(0.49)	(0.51)	(0.42)
Depth 3 & 16 variables	3.77***	4.78***	0.50	0.48	1.25***
	(0.40)	(0.41)	(0.44)	(0.53)	(0.42)
Depth 4 & 16 variables	3.78***	4.79***	0.51	0.49	1.26***
	(0.41)	(0.42)	(0.48)	(0.47)	(0.45)

Note: Panel A show the percentage of individuals that are assigned to a specific program by the trees of different depth. Panel B show a t-test of the difference of the cross-validated policy (standard errors in parentheses) and the APOs of the programs. *p<0.1; **p<0.05; ***p<0.01

Zhou et al. (2018) propose another validation idea and test whether the optimal policies perform significantly better than sending all individuals to the same program. This is achieved by taking the difference of the APO score of the cross-validated policy and the APO score of the program w:

$$\hat{\Delta}^{cv}_{i,w}(\pi) = \sum_{t=0}^{T} \mathbb{1}(\hat{\pi}^{cv}(Z_i) = t)\hat{\Gamma}_{i,t} - \hat{\Gamma}_{i,w}$$

where $\hat{\pi}^{cv}(Z_i)$ is the policy that is estimated without individual i. A standard t-test on the mean of $\hat{\Delta}^{cv}_{i,w}(\pi)$ tests then whether the cross-validated policies are significantly better than sending everybody to the same program. Note that the cross-validated policies do not necessarily coincide with the trees in the full sample and the cross-validation estimates not the value function for that specific tree. This would require to hold out a test set

which would be viable for an application with bigger programs.

The results are provided in Panel B of Table 6. We can interpret the mean of $\hat{\Delta}^{cv}_{i,w}(\pi)$ as average treatment effect comparing a regime under the estimated assignment rule or a regime where everybody is sent to the same program. This effect is positive for all tree specifications indicating that the estimated rules can leverage the effect heterogeneities to improve the allocation. However, the cross-validated policies perform not significantly better than sending just everybody into vocational or computer programs. This would probably change if we could take costs or capacity constraints into account. However, we do not observe costs in this dataset and the optimal decision tree algorithm is currently not capable of incorporating capacity constraints in a systematic way. We leave both extensions for future research using a more detailed database on both costs and capacity constraints.

7 Discussion and conclusions

This paper considers recent methodological developments based on Double Machine Learning (DML) through the lens of a standard program evaluation under unconfoundedness. DML based methods provide a convenient toolbox for a comprehensive program evaluation as many different parameters of interest can be estimated using the same framework and a combination of standard statistical software. Building on the double robustness property of the underlying score, DML methods come further with attractive asymptotic statistical guarantees. The application to an Active Labor Market Policy evaluation shows that the methods also provide mostly plausible results in practice. However, several conceptual and implementational issues remain open for investigation and refinement.

In general, we know little about how to choose the estimator for the nuisance parameters. The pool of potential machine learning algorithms and their combinations is large and little is known, e.g., about the trade-off between high prediction performance and computation time in the causal setting. Also clear recommendations for the implementation of cross-fitting are missing. Another open question is how to deal with common support in general and for each estimand specifically. The literature on trimming rules is well developed

for propensity score based methods estimating average effects. However, we are not only interested in average effects and the propensity score is not the only nuisance parameter of DML. It remains an open question whether the established trimming methods are also sensible in these settings.

The estimation of flexible heterogeneous treatment effects provides an interesting new tool. However, it is currently not clear to what extent we can actually explore heterogeneity or to what extend we need to pre-define the heterogeneity of interest. Especially the instability of very individualized heterogeneous effects shows that the former should be used with caution and requires further investigation. For the latter, the possibility to summarize pre-defined heterogeneity of interest using OLS or nonparametric regressions provide valuable and easy to use options in applications.

The estimation of optimal treatment assignment rules is mostly unexplored in practice and many interesting issues in applications regarding inference, the implementation of different constraints, more flexible rules than decision trees, or the choice of variables that could or should enter the set of policy variables could be explored in future work.

The investigation of the DML specific questions but also the comparison with other probably more specialized causal machine learning methods for each estimand provides also an interesting direction of future research. Such evidence would help to understand and guide which choices are critical in applications similar to the one in this paper.

References

Abadie, A., & Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, *10*, 465–503.

Athey, S., & Imbens, G. (2019). Machine learning methods economists should know about. *Annual Review of Economics*, *11*.

Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Athey, S., Imbens, G. W., & Wager, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(4), 597–632.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics*, *47*(2), 1148 - 1178.

Athey, S., & Wager, S. (2017). *Efficient policy learning.* Retrieved from https://arxiv.org/abs/1606.02647

Avagyan, V., & Vansteelandt, S. (2017). *Honest data-adaptive inference for the average treatment effect under model misspecification using penalised bias-reduced double-robust estimation.* Retrieved from http://arxiv.org/abs/1708.03787

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., & Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. *Science*, *359*(6373), 325–329.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*(2), 521–547.

Belloni, A., Chernozhukov, V., Fernández-Val, I., & Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, *85*(1), 233–298.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, *81*(2), 608–650.

Bertrand, M., Crépon, B., Marguerie, A., & Premand, P. (2017). Contemporaneous and post-program impacts of a public works program: Evidence from Côte d'Ivoire. *World Bank Working Paper*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Card, D., Kluve, J., & Weber, A. (2018). What works? A meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association*, *16*(3), 894–931.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/Debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1-C68.

Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2017). Generic machine learning inference on heterogenous treatment effects in randomized experiments. Retrieved from http://arxiv.org/abs/1712.04802

Chernozhukov, V., Fernandez-Val, I., & Luo, Y. (2018). The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica*, *86*(6), 1911–1938.

Cockx, B., Lechner, M., & Bollens, J. (2020). *Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium.* CEPR Discussion Paper No. DP14270.

Colangelo, K., & Lee, Y.-Y. (2019). *Double debiased machine learning nonparametric inference with continuous treatments.* cemmap working paper CWP72/19.

Crépon, B., & van den Berg, G. J. (2016). Active labor market policies. *Annual Review of Economics*, *8*, 521–546.

Davis, J. M., & Heller, S. B. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, *107*(5), 546–550.

Dudik, M., Langford, J., & Li, L. (2011). Doubly robust policy evaluation and learning. Retrieved from http://arxiv.org/abs/1103.4601

Fan, Q., Hsu, Y.-C., Lieli, R. P., & Zhang, Y. (2019). Estimation of conditional average treatment effects with high-dimensional data. *arXiv:1908.02399*. Retrieved from http://arxiv.org/abs/1908.02399

Farbmacher, H., Heinrich, K., & Spindler, M. (2019). *Heterogeneous Effects of Poverty on Cognition.* MEA Discussion Paper No. 06-2019.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, *189*(1), 1–23.

Farrell, M. H., Liang, T., & Misra, S. (2018). *Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands.* Retrieved from http://arxiv.org/abs/1809.09953

Gerfin, M., & Lechner, M. (2002). A microeconometric evaluation of the active labour market policy in Switzerland. *Economic Journal*, *112*(482), 854–893.

Glynn, A. N., & Quinn, K. M. (2009). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, *18*(1), 36–56.

Gulyas, A., & Pytka, K. (2019). *Understanding the sources of earnings losses after job displacement: A machine-learning approach.* Discussion Paper Series – CRC TR 224 No. 131.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning - Data mining, inference, and prediction* (2nd ed.). Springer, New York.

Hayfield, T., & Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, *27*(5).

Hirano, K., & Porter, J. R. (2009). Asymptotics for Statistical Treatment Rules. *Econometrica*, *77*(5), 1683–1701.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Hothorn, T., & Zeileis, A. (2015). Partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, *16*, 3905–3909.

Huber, M., Lechner, M., & Mellace, G. (2017). Why do tougher caseworkers increase employment? The role of program assignment as a causal mechanism. *Review of Economics and Statistics*, *99*(1), 180–183.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical*

sciences. Cambridge University Press.

Isphording, I. E. (2014). *Language and labor market success* (No. 8572). IZA Discussion Papers.

Jacob, D., Härdle, W. K., & Lessmann, S. (2019). *Group Average Treatment Effects for Observational Studies.* Retrieved from http://arxiv.org/abs/1911.02688

Kallus, N. (2018). Balanced policy evaluation and learning. In *Advances in neural information processing systems* (pp. 8895–8906).

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523–539.

Kennedy, E. H., Ma, Z., McHugh, M. D., & Small, D. S. (2016). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*, 1229–1245.

Kitagawa, T., & Tetenov, A. (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica*, *86*(2), 591–616.

Knaus, M. C. (2018). *A double machine learning approach to estimate the effects of musical practice on student's skills.* Retrieved from https://arxiv.org/abs/1805.10300

Knaus, M. C., Lechner, M., & Strittmatter, A. (2018). *Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence.* Retrieved from http://arxiv.org/abs/1810.13237

Knaus, M. C., Lechner, M., & Strittmatter, A. (2020). Heterogeneous employment effects of job search programmes: A machine learning approach. *Journal of Human Resources*, forthcoming.

Knittel, C. R. (2019). *Using machine learning to target treatment: The case of household energy use.* NBER Working Paper No. 26531.

Kreif, N., & DiazOrdaz, K. (2019). *Machine learning in policy evaluation: new tools for causal inference.* Retrieved from http://arxiv.org/abs/1903.00402

Künzel, S., Sekhon, J., Bickel, P., & Yu, B. (2017). Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning. Retrieved from http://arxiv.org/abs/1706.03461

Lalive, R., van Ours, J., & Zweimüller, J. (2008). The impact of active labor market programs on the duration of unemployment. *Economic Journal*, *118*(525), 235–257.

Lechner, M. (1999). Earnings and employment effects of continuous gff-the-job training in east germany after unification. *Journal of Business & Economic Statistics*, *17*(1), 74–90.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In M. Lechner & E. Pfeiffer (Eds.), *Econometric evaluation of labour market policies* (pp. 43–58). Heidelberg: Physica.

Lechner, M. (2018). Modified causal forests for estimating heterogeneous causal effects. *arXiv:1812.09487*. Retrieved from https://arxiv.org/abs/1812.09487

Lechner, M., & Smith, J. (2007). What is the value added by caseworkers? *Labour Economics*, *14*(2), 135–151.

Lechner, M., & Strittmatter, A. (2019). Practical procedures to deal with common support problems in matching estimation. *Econometric Reviews*, *38*(2), 193–207. doi: 10.1080/07474938.2017.1318509

Luo, Y., & Spindler, M. (2016). *High-dimensional L2-boosting: Rate of Convergence.* Retrieved from http://arxiv.org/abs/1602.08927

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, *72*(4), 1221–1246.

Ning, Y., Peng, S., & Imai, K. (2018). *Robust estimation of causal effects via high-dimensional covariate balancing propensity score.* Retrieved from http://arxiv.org/abs/1812.08683

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, *90*(429), 106–121.

Robins, J. M., Sued, M., Lei-gomez, Q., & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable.

Statistical Science, *22*(4), 544–559.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Semenova, V., & Chernozhukov, V. (2017). Simultaneous inference for best linear predictor of the conditional average treatment effect and other structural functions. *arXiv:1702.06240*. Retrieved from http://arxiv.org/abs/1702.06240

Smucler, E., Rotnitzky, A., & Robins, J. M. (2019). *A unifying approach for doubly-robust L1 regularized estimation of causal contrasts.* Retrieved from http://arxiv.org/abs/1904.03737

Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics*, *151*(1), 70–81.

Stoye, J. (2012). Minimax regret treatment choice with covariates or with limited validity of experiments. *Journal of Econometrics*, *166*(1), 138–156.

Strittmatter, A. (2018). What is the value added by using causal machine learning methods in a welfare experiment evaluation? *arXiv:1812.06533*. Retrieved from http://arxiv.org/abs/1812.06533

Tan, Z. (2018). *Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data.* Retrieved from http://arxiv.org/abs/1801.09817

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, *109*(508), 1517–1532.

van der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *International Journal of Biostatistics*, *2*(1).

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

Wager, S., & Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv:1503.06388*. Retrieved from http://arxiv.org/abs/1503.06388

Wunsch, C. (2016). How to minimize lock-in effects of programs for unemployed workers. *IZA World of Labor*.

Zhou, Z., Athey, S., & Wager, S. (2018). *Offline multi-action policy learning: Generalization and optimization.* Retrieved from http://arxiv.org/abs/1810.04778

Zimmert, M., & Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv:1908.08779*. Retrieved from http://arxiv.org/abs/1908.08779

Appendices

A Results

A.1 Data

Table A.1: Means of control variables by program

	No	JS	Voc	Comp	Lang
	(1)	(2)	(3)	(4)	(5)
UE female	0.44	0.44	0.33	0.60	0.55
CW age	44.08	44.10	44.81	44.59	44.61
CW female	0.43	0.47	0.39	0.44	0.47
CW tenure	5.47	5.44	5.73	5.83	5.61
CW own unemp experience	0.62	0.63	0.64	0.61	0.63
CW above voc training	0.46	0.45	0.44	0.48	0.48
CW tertiary education	0.19	0.21	0.17	0.16	0.21
CW voc training	0.26	0.27	0.22	0.25	0.22
UE assignment by industry	0.60	0.67	0.58	0.51	0.64
UE assignment by occupation	0.51	0.57	0.46	0.45	0.57
UE assignment by age	0.04	0.04	0.04	0.06	0.05
UE assignment by employability	0.09	0.07	0.10	0.08	0.06
UE assignment by region	0.13	0.09	0.09	0.13	0.11
UE assignment by other	0.09	0.07	0.08	0.10	0.09
UE mother tongue non-Swiss	0.33	0.29	0.31	0.18	0.64
Cantonal unemployment rate	3.52	3.59	3.41	3.36	3.63
# of unemp. spells in last 2 years	0.57	0.39	0.52	0.37	0.43
% of months emp. in last 2 years	0.81	0.84	0.83	0.84	0.72
CW cooperative	0.48	0.50	0.41	0.42	0.45
Missing caseworker characteristics	0.05	0.05	0.04	0.05	0.05
Married	0.47	0.46	0.48	0.45	0.72
UE age	36.61	37.31	37.45	39.08	35.28
UE mother tongue in canton's language	0.10	0.12	0.11	0.11	0.04
Cantonal GDP p.c.	0.52	0.53	0.51	0.53	0.54
Past income	4.25	4.67	4.87	4.32	3.73
# employment spells in last 5 years	0.12	0.10	0.09	0.09	0.08
Employability	1.93	1.98	1.93	1.97	1.85
Lives in city	1.51	1.51	1.54	1.38	1.60
Qualification unskilled	0.23	0.20	0.17	0.12	0.40
Qualification semiskilled	0.16	0.14	0.17	0.14	0.15
Qualification no degree	0.03	0.03	0.02	0.02	0.07
Qualification degree	0.58	0.62	0.63	0.72	0.38
Swiss	0.63	0.67	0.70	0.79	0.34
Foreigner with B permit	0.13	0.11	0.12	0.04	0.44
Foreigner with C permit	0.23	0.22	0.18	0.17	0.23
Previous job self-employed	0.01	0.00	0.00	0.00	0.00
Previous job manager	0.08	0.08	0.10	0.09	0.07
Previous job skilled worker	0.60	0.65	0.65	0.75	0.43
Previous job unskilled worker	0.29	0.24	0.22	0.15	0.48
Sector missing	0.18	0.15	0.15	0.16	0.29
Previous job in primary sector	0.09	0.06	0.09	0.05	0.05
Previous job in secondary sector	0.12	0.14	0.15	0.13	0.12
Previous job in tertiary sector	0.61	0.65	0.61	0.67	0.54

Note: UE stand for unemployed individual, CW stand for caseworker.

A.2 Nuisance parameters

Figure A.1: Variable importance measures for nuisance parameters

(a) Propensity scores

(b) Outcomes

Notes: Each number represents the average over the five models used for cross-fitting. Variable importance measures are multiplied by minus one if the correlation between covariate the predicted values is negative. UE stands for unemployed individual, p.c. for per capita.

Table A.2: Variable importance for propensity score models

	No	JS	Voc	Comp	Lang	Mean
	(1)	(2)	(3)	(4)	(5)	(6)
Cantonal GDP p.c.	-0.29	0.34	-0.35	-0.09	0.02	0.22
Cantonal unemployment rate	-0.37	0.32	-0.05	-0.04	0.01	0.16
Past income	-0.05	0.11	0.29	0.02	-0.10	0.11
Foreigner with B permit	-0.00	-0.00	-0.00	-0.03	0.38	0.08
UE mother tongue non-Swiss	0.00	-0.01	-0.00	-0.14	0.16	0.06
% of months emp. in last 2 years	-0.04	0.03	0.06	0.01	-0.11	0.05
UE female	-0.00	-0.00	-0.06	0.16	0.01	0.05
Swiss	0.00	0.00	0.00	0.12	-0.09	0.04
# employment spells in last 5 years	0.12	-0.03	-0.02	-0.01	-0.01	0.04
UE age	-0.01	0.01	0.02	0.08	-0.01	0.03
Previous job unskilled worker	0.00	-0.01	-0.01	-0.08	0.01	0.02
# of unemp. spells in last 2 years	0.04	-0.04	-0.01	-0.00	-0.00	0.02
Previous job skilled worker	-0.00	0.00	0.00	0.07	-0.00	0.02
UE assignment by industry	-0.01	0.03	-0.00	-0.02	0.00	0.01
Qualification degree	-0.00	0.00	0.00	0.04	-0.01	0.01
CW age	-0.01	0.01	0.02	0.01	0.01	0.01
UE assignment by occupation	-0.01	0.02	-0.01	-0.01	0.00	0.01
CW tenure	0.01	-0.01	0.01	0.01	0.00	0.01
Married	-0.00	-0.00	-0.00	-0.00	0.03	0.01
Qualification unskilled	0.00	-0.00	-0.01	-0.02	0.01	0.01
UE mother tongue in canton's language	-0.00	0.00	-0.00	-0.01	-0.01	0.00
CW cooperative	-0.00	0.00	-0.01	-0.00	-0.00	0.00
Employability	-0.00	0.01	0.00	0.00	-0.00	0.00
Foreigner with C permit	0.00	-0.00	-0.00	-0.00	-0.01	0.00
UE assignment by region	0.00	-0.00	-0.00	0.00	-0.00	0.00
Lives in city	0.00	-0.00	0.01	-0.00	0.00	0.00
Previous job in primary sector	0.01	-0.00	0.00	-0.00	-0.00	0.00
CW voc training	-0.00	0.00	-0.00	0.00	-0.00	0.00
CW female	-0.00	0.00	-0.01	0.00	0.00	0.00
Sector missing	0.00	-0.00	-0.00	-0.00	0.00	0.00
CW tertiary education	-0.00	0.00	-0.00	-0.00	0.00	0.00
Previous job manager	-0.00	0.00	0.00	0.00	-0.00	0.00
Previous job in secondary sector	-0.00	0.00	0.00	-0.00	-0.00	0.00
UE assignment by employability	0.00	-0.00	0.00	-0.00	-0.00	0.00
Previous job in tertiary sector	-0.00	0.00	-0.00	0.00	-0.00	0.00
CW above voc training	0.00	-0.00	0.00	0.00	-0.00	0.00
CW own unemp experience	-0.00	0.00	0.00	-0.00	-0.00	0.00
Qualification semiskilled	0.00	-0.00	-0.00	-0.00	0.00	0.00
UE assignment by other	0.00	-0.00	-0.00	-0.00	-0.00	0.00
UE assignment by age	0.00	-0.00	-0.00	0.00	-0.00	0.00
Missing caseworker characteristics	0.00	-0.00	0.00	-0.00	0.00	0.00
Qualification no degree	-0.00	-0.00	-0.00	-0.00	0.00	0.00
Previous job self-employed	0.00	0.00	0.00	0.00	0.00	0.00

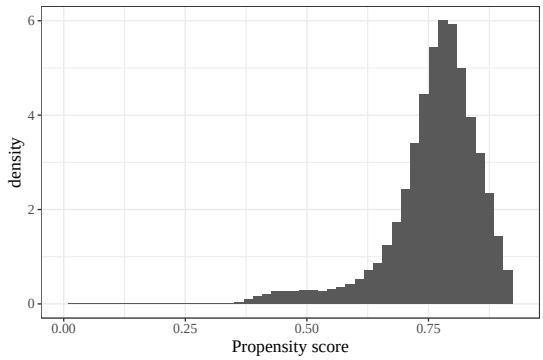
Note: UE stand for unemployed individual, CW stand for caseworker.

Table A.3: Variable importance for outcome models

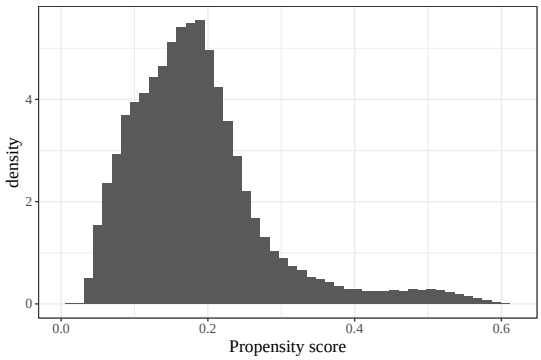
	No	JS	Voc	Comp	Lang	Mean
	(1)	(2)	(3)	(4)	(5)	(6)
Qualification degree	-0.29	0.33	0.04	0.04	-0.11	0.16
Previous job unskilled worker	0.22	-0.14	-0.05	-0.06	0.13	0.12
UE mother tongue non-Swiss	0.09	-0.12	-0.01	-0.15	0.18	0.11
Past income	-0.09	0.08	0.10	0.06	-0.20	0.11
Swiss	0.06	0.06	0.01	0.09	-0.14	0.07
# of unemp. spells in last 2 years	0.04	-0.03	-0.22	-0.04	-0.01	0.07
% of months emp. in last 2 years	-0.07	0.04	0.10	0.06	-0.03	0.06
UE age	-0.03	0.06	0.04	0.07	-0.02	0.04
Qualification unskilled	0.04	-0.06	-0.07	-0.03	0.02	0.04
Cantonal GDP p.c.	-0.01	0.00	-0.07	-0.06	0.01	0.03
Cantonal unemployment rate	-0.00	0.00	-0.06	-0.03	0.01	0.02
CW age	-0.00	0.01	0.03	0.04	0.01	0.02
Previous job skilled worker	-0.01	0.00	0.02	0.04	-0.02	0.02
Foreigner with C permit	0.01	-0.02	-0.01	-0.04	-0.00	0.02
# employment spells in last 5 years	0.00	-0.01	-0.02	-0.03	-0.00	0.01
Employability	-0.03	0.02	0.00	0.01	-0.00	0.01
Married	-0.00	-0.01	-0.00	-0.02	0.02	0.01
UE female	-0.00	-0.00	-0.01	0.02	0.02	0.01
Lives in city	0.00	-0.00	0.02	-0.02	0.01	0.01
CW tenure	0.00	-0.00	0.01	0.02	0.01	0.01
Sector missing	0.00	-0.00	-0.02	-0.00	0.01	0.01
Foreigner with B permit	-0.00	-0.00	-0.00	-0.00	0.01	0.00
UE assignment by occupation	-0.00	0.00	-0.01	-0.01	0.00	0.00
UE mother tongue in canton's language	-0.00	0.00	-0.00	-0.01	-0.00	0.00
CW female	-0.00	0.00	-0.01	0.01	0.00	0.00
CW cooperative	-0.00	0.00	-0.00	-0.01	-0.00	0.00
Previous job in secondary sector	-0.00	0.00	0.01	-0.00	-0.00	0.00
Previous job in tertiary sector	-0.00	0.00	-0.00	0.00	-0.00	0.00
CW voc training	-0.00	0.00	-0.00	0.00	-0.00	0.00
CW own unemp experience	-0.00	0.00	0.00	-0.00	-0.00	0.00
UE assignment by other	0.00	-0.00	-0.00	-0.01	-0.00	0.00
Qualification semiskilled	0.00	-0.00	-0.00	-0.00	0.00	0.00
CW above voc training	0.00	-0.00	0.00	0.00	-0.00	0.00
UE assignment by industry	-0.00	0.00	-0.00	-0.00	0.00	0.00
Previous job in primary sector	0.00	-0.00	0.01	-0.00	-0.00	0.00
CW tertiary education	-0.00	0.00	-0.00	-0.00	0.00	0.00
UE assignment by region	0.00	-0.00	-0.00	0.00	-0.00	0.00
Previous job manager	-0.00	0.00	0.00	0.00	-0.00	0.00
UE assignment by employability	0.00	-0.00	0.00	-0.00	-0.00	0.00
Qualification no degree	-0.00	-0.00	-0.00	-0.00	0.00	0.00
Missing caseworker characteristics	0.00	-0.00	0.00	-0.00	0.00	0.00
UE assignment by age	0.00	-0.00	-0.00	0.00	-0.00	0.00
Previous job self-employed	0.00	0.00	0.00	0.00	-0.00	0.00

Note: UE stand for unemployed individual, CW stand for caseworker.

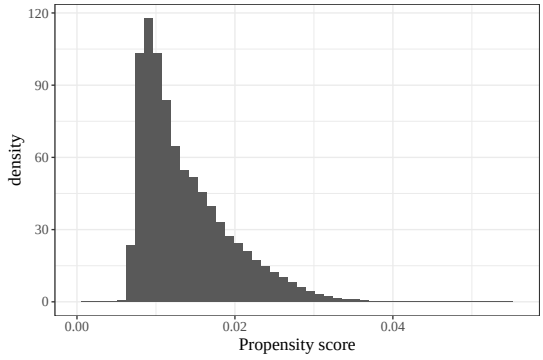
Figure A.2: Distribution of propensity scores



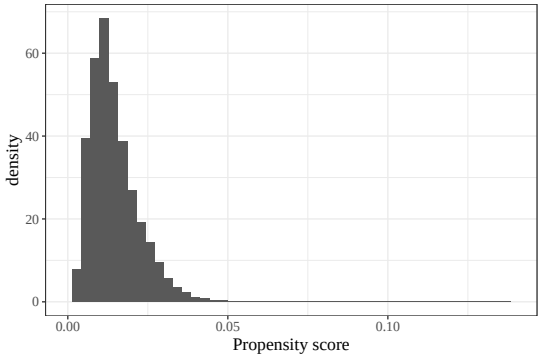
(a) No program



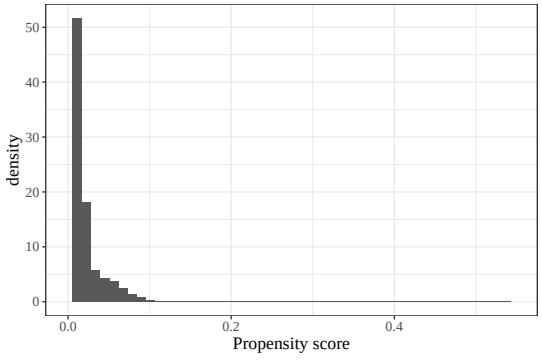
(b) Job search



(c) Vocational



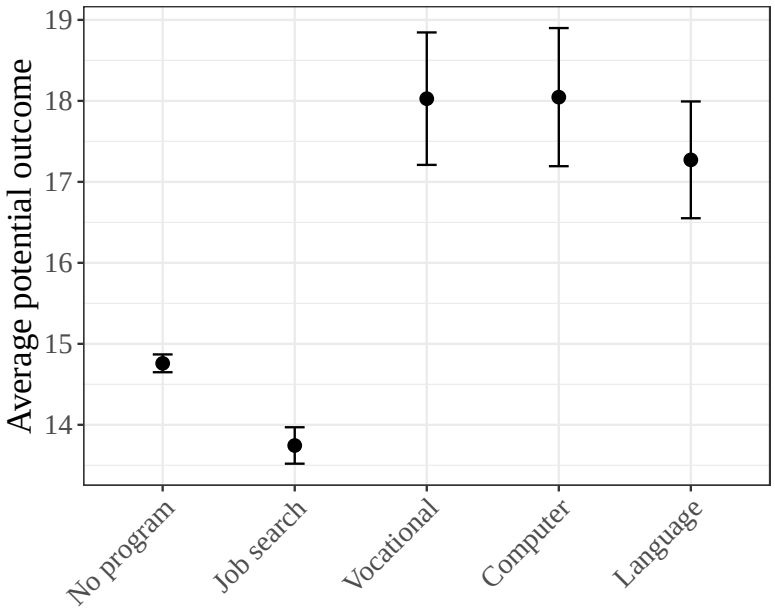
(d) Computer



(e) Language

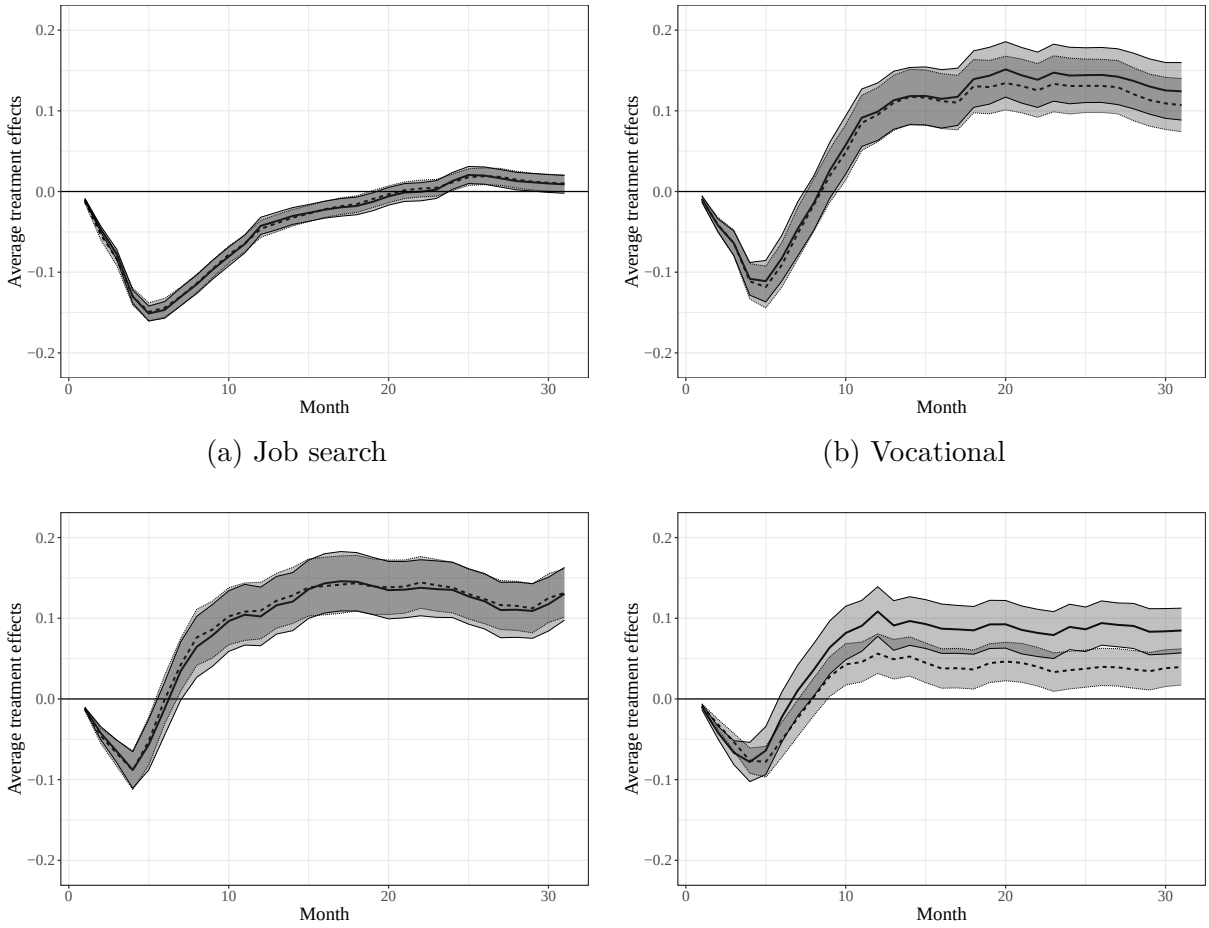
A.3 Average treatment effects

Figure A.3: Average potential outcomes



Notes: Average potential outcomes with 95% confidence intervals.

Figure A.4: Average treatment effects over time



(a) Job search

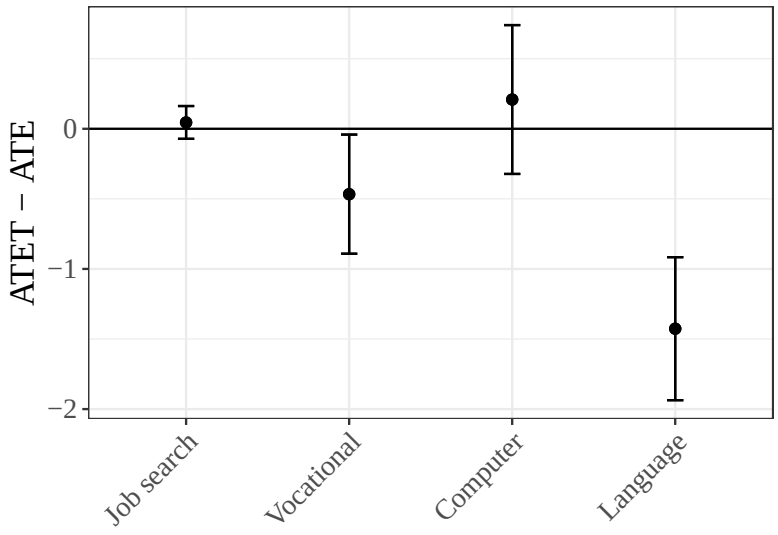
(b) Vocational

(c) Computer

(d) Language

Notes: Solid lines show ATE, dashed lines ATET. Grey area depict the 95% confidence intervals.

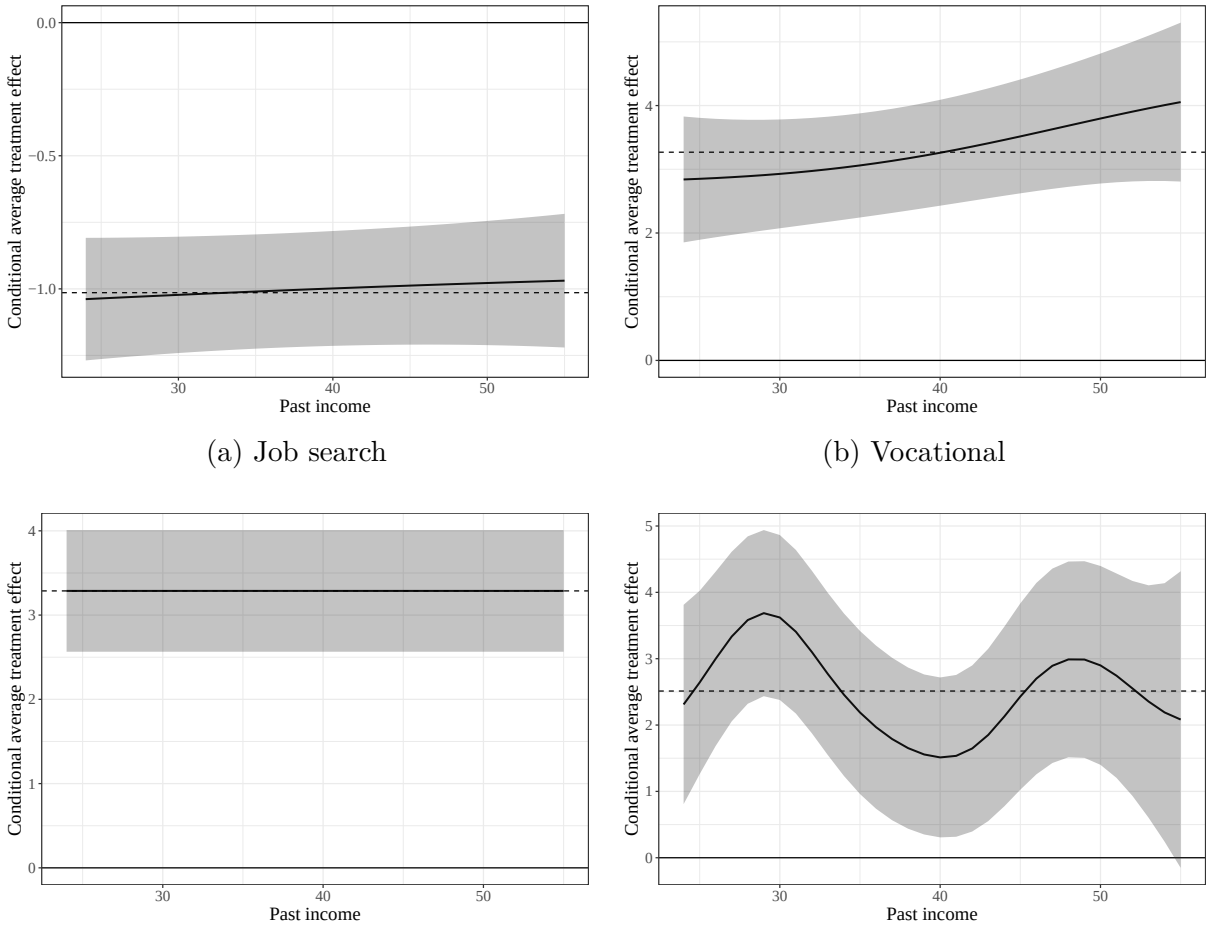
Figure A.5: Difference of ATET and ATE



Notes: Differences between ATET and ATE with 95% confidence intervals.

A.4 Nonparametric CATEs

Figure A.6: Effect heterogeneity regarding age



(a) Job search

(b) Vocational

(c) Computer

(d) Language

Dotted line indicates point estimate of the respective average treatment effect. Grey area shows 95%-confidence interval.

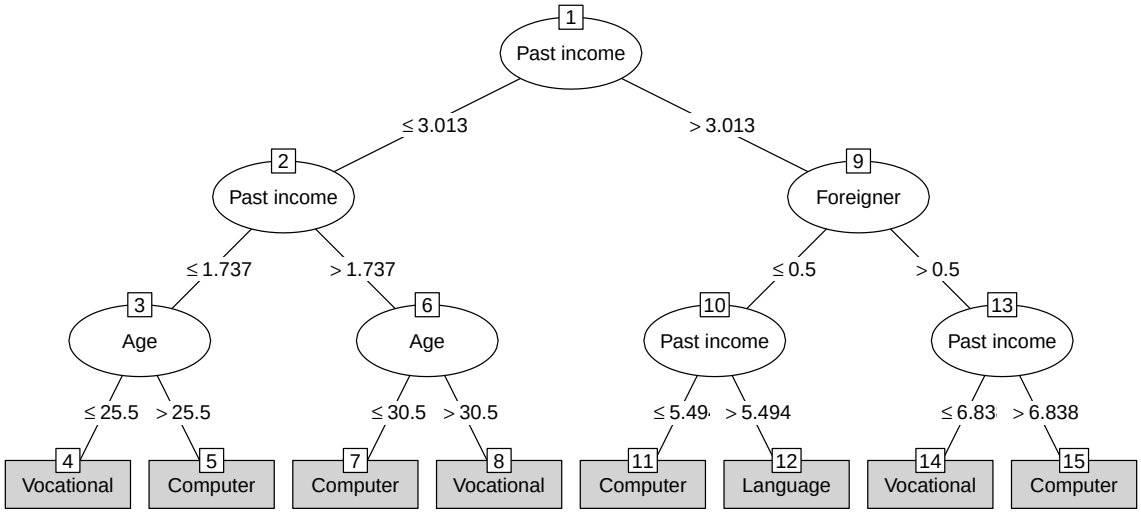
A.5 Random Forest CATEs

Figure A.7: Variable importance for Random Forest CATE estimation with all covariates

Notes: Variable importance measures are multiplied by minus one if the correlation between covariate the predicted values is negative. Only those variables shown with an average variable importance of at least 0.05.

A.6 Optimal treatment assignment

Figure A.8: Optimal decision tree of depth four with five covariates



Notes: Past income is measured in CHF 10,000, employability is an ordered variable with one indicating low employability, two medium employability and three high employability. Graph is created with the R package `partykit` of Hothorn and Zeileis (2015).

Figure A.9: Overlap of cross-validated policies with five covariates

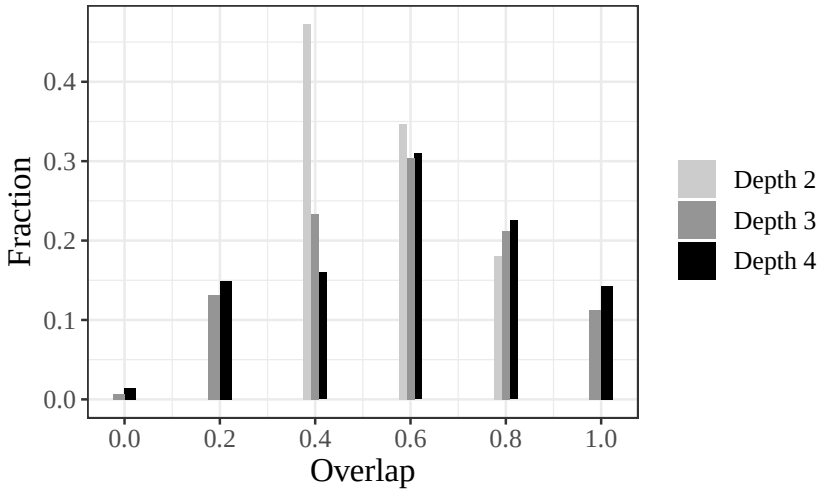
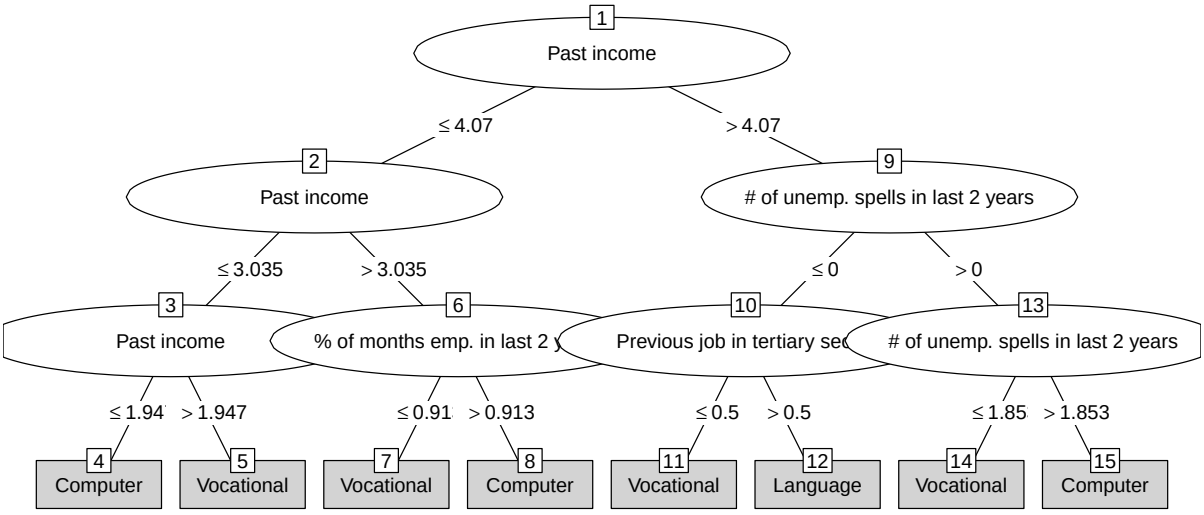
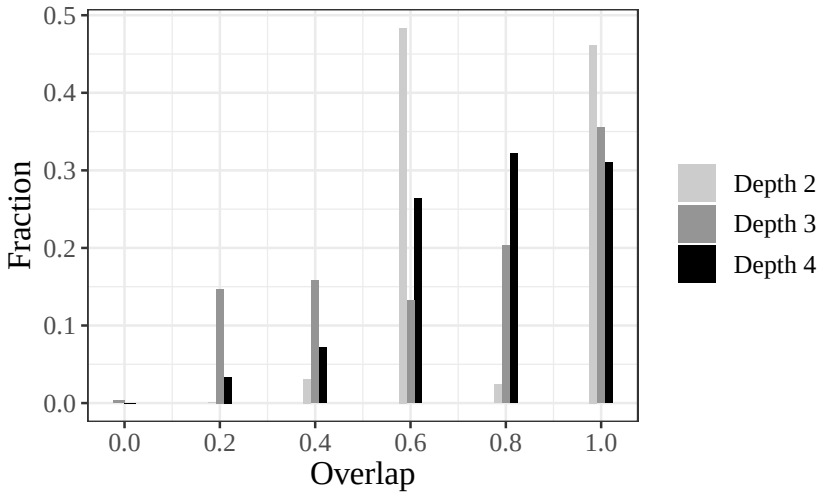


Figure A.10: Optimal decision tree of depth four with 16 covariates



Notes: Past income is measured in CHF 10,000, employability is an ordered variable with one indicating low employability, two medium employability and three high employability. Graph is created with the R package `partykit` of Hothorn and Zeileis (2015).

Figure A.11: Overlap of cross-validated policies with 16 covariates

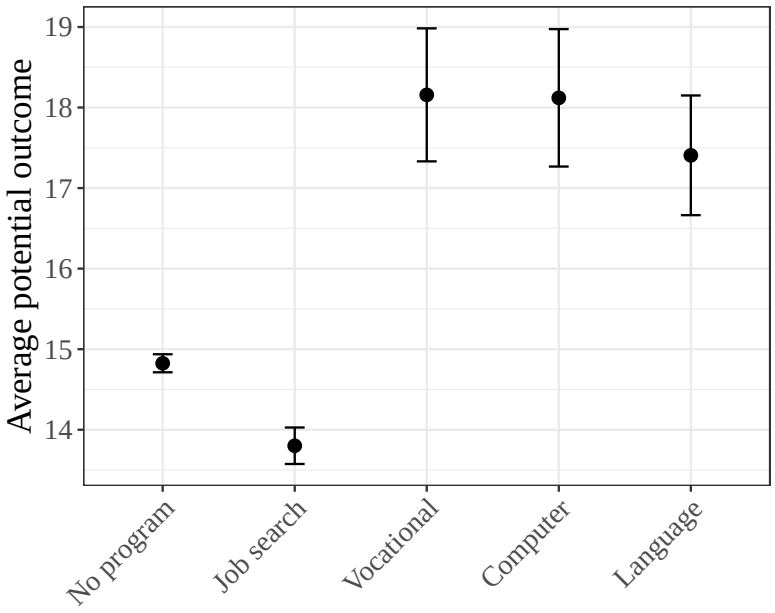


B Enforced common support

We enforce common support by trimming all observations with propensity scores below the largest minimum propensity score in the different treatment groups as well as propensity scores above the smallest maximum propensity score in the different treatment groups. This results in trimming of 1,954 observations. The results in the following document that this trimming changes the main results reported in the main text only marginally.

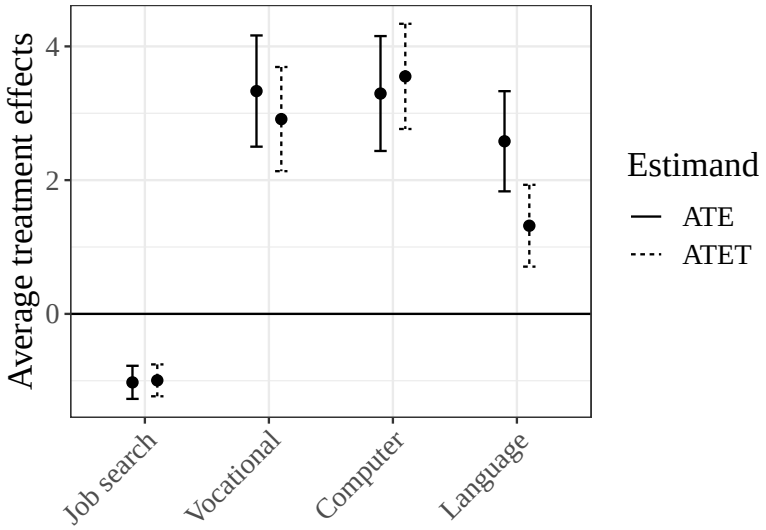
B.1 Average effects

Figure B.1: Average potential outcomes



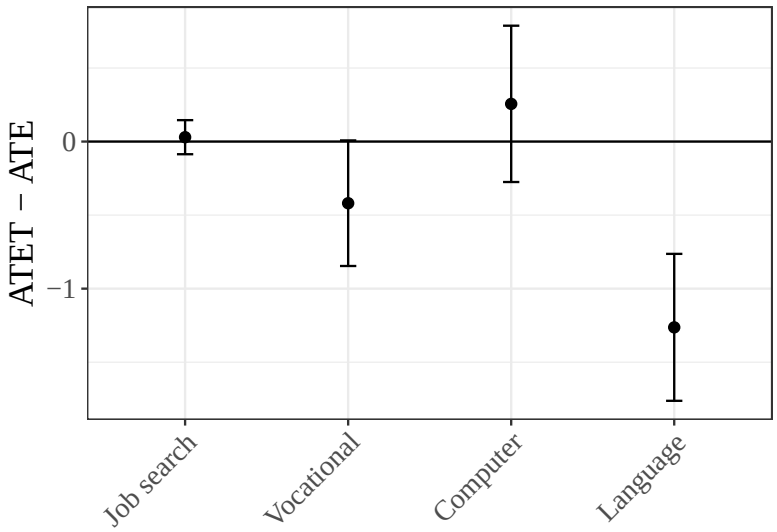
Notes: Average potential outcomes with 95% confidence intervals.

Figure B.2: Average treatment effects of participation vs. non-participation



Note: The figure shows the point estimates of the average treatment effects of participating in the program labeled on the x-axis vs. non-participation and their 95% confidence intervals.

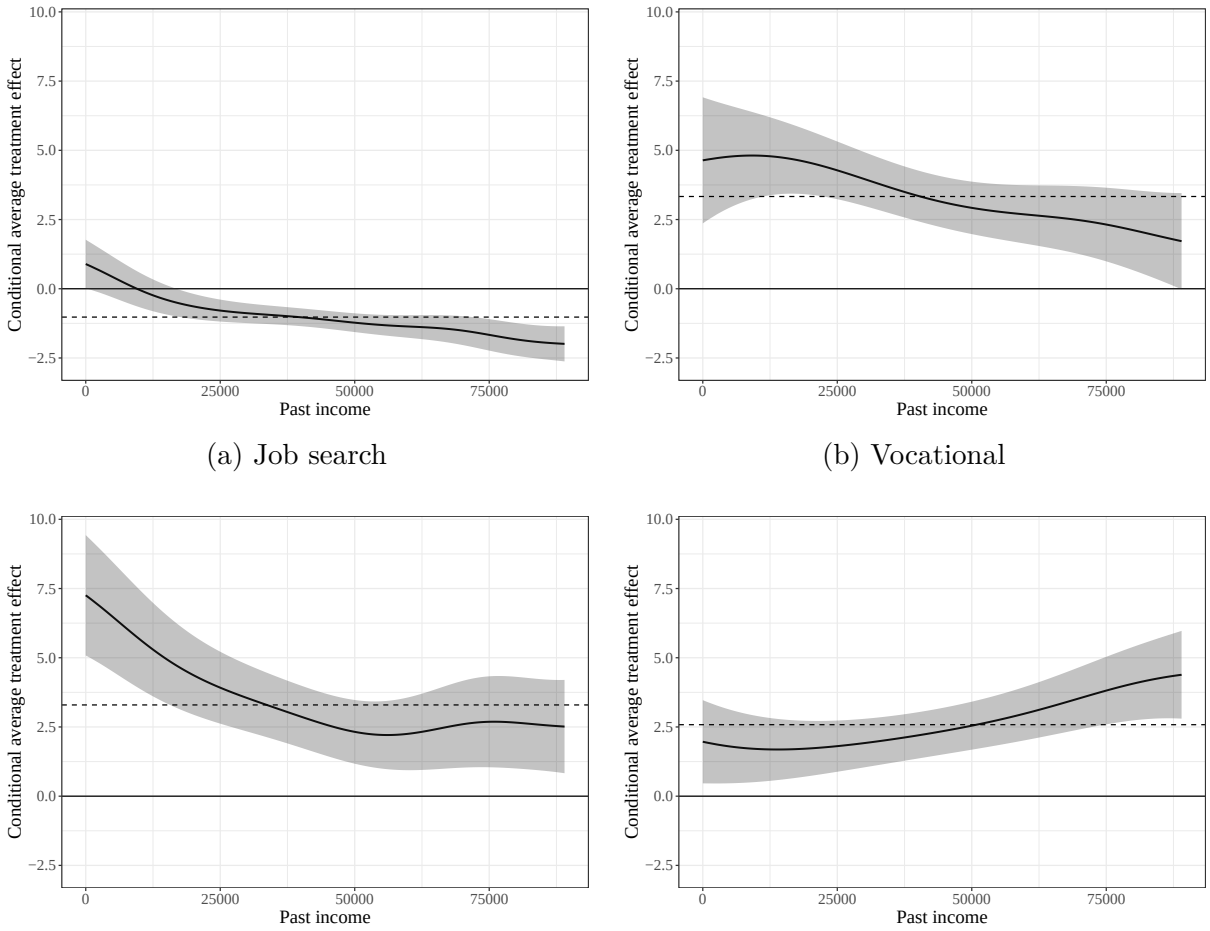
Figure B.3: Difference of ATET and ATE



Notes: Differences between ATET and ATE with 95% confidence intervals.

B.2 Heterogeneous effects

Figure B.4: Effect heterogeneity regarding past income



(a) Job search

(b) Vocational

(c) Computer

(d) Language

Note: Dotted line indicates point estimate of the respective average treatment effect. Grey area shows 95%-confidence interval.

Table B.1: Group average treatment effects

	Job search	Vocational	Computer	Language
	(1)	(2)	(3)	(4)
Panel A:				
Constant -1.27***	3.80***	2.27***	3.36***	
	(0.17)	(0.56)	(0.61)	(0.47)
Female	0.55**	-1.06	2.33***	-1.76**
	(0.26)	(0.86)	(0.87)	(0.78)
Panel B:				
Constant	-1.24***	2.64***	3.72***	3.49***
	(0.16)	(0.52)	(0.50)	(0.52)
Foreigner	0.60**	1.95**	-1.21	-2.56***
	(0.26)	(0.89)	(0.98)	(0.74)
Panel C:				
Constant	-0.25	5.40***	5.73***	2.73***
	(0.34)	(1.08)	(1.10)	(0.89)
Medium employability	-0.85**	-2.26*	-2.76**	-0.24
	(0.37)	(1.19)	(1.21)	(1.00)
High employability	-1.49***	-4.03***	-3.93**	0.31
	(0.50)	(1.51)	(1.69)	(1.52)
F-statistic	4.61***	2.95*	3.19**	0.10

Note: This table shows OLS coefficients and their heteroscedasticity robust standard errors (in parentheses) of regressions run with the pseudo-outcome defined as described in Section 3.3. *p<0.1; **p<0.05; ***p<0.01

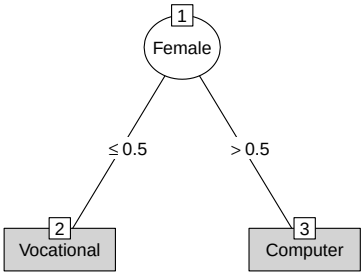
Table B.2: Best linear prediction CATEs

	Job search	Vocational	Computer	Language
	(1)	(2)	(3)	(4)
Constant	-0.67	4.14*	6.33***	5.18**
	(0.71)	(2.43)	(2.45)	(2.09)
Female	0.21	-1.89**	1.68*	-1.50*
	(0.27)	(0.91)	(0.93)	(0.82)
Age	0.03*	0.10**	0.02	-0.04
	(0.02)	(0.05)	(0.05)	(0.04)
Foreigner	0.43	1.43	-1.55	-2.53***
	(0.27)	(0.90)	(1.00)	(0.75)
Medium employability	-0.57	-1.48	-2.53**	-0.72
	(0.38)	(1.21)	(1.24)	(1.02)
High employability	-1.03**	-2.76*	-3.62**	-0.56
	(0.52)	(1.55)	(1.74)	(1.55)
Past income in CHF 10,000	-0.26***	-0.62***	-0.41**	0.26
	(0.07)	(0.23)	(0.19)	(0.18)
F-statistic	5.77***	3.54***	3.08***	3.11***

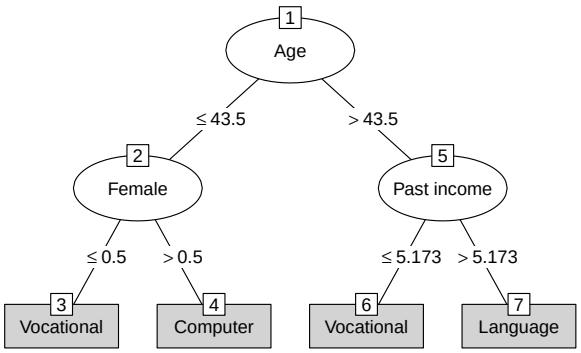
Note: This table shows OLS coefficients and their heteroscedasticity robust standard errors (in parentheses) of regressions run with the pseudo-outcome defined in Section 3.3. *p<0.1; **p<0.05; ***p<0.01

B.3 Optimal treatment assignment

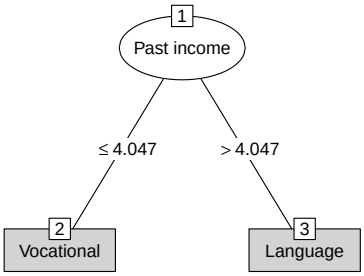
Figure B.5: Optimal treatment assignment decision trees of depth two and three



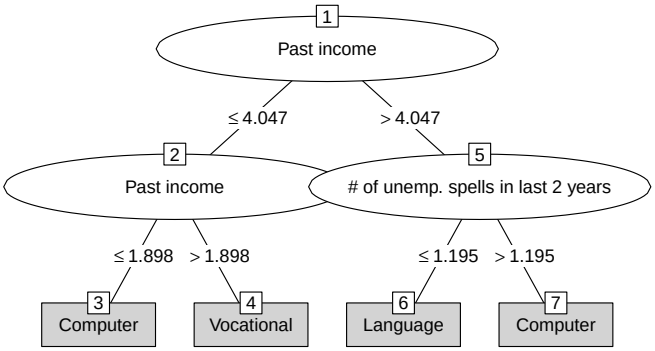
(a) Depth 2 & 5 covariates



(b) Depth 3 & 5 covariates



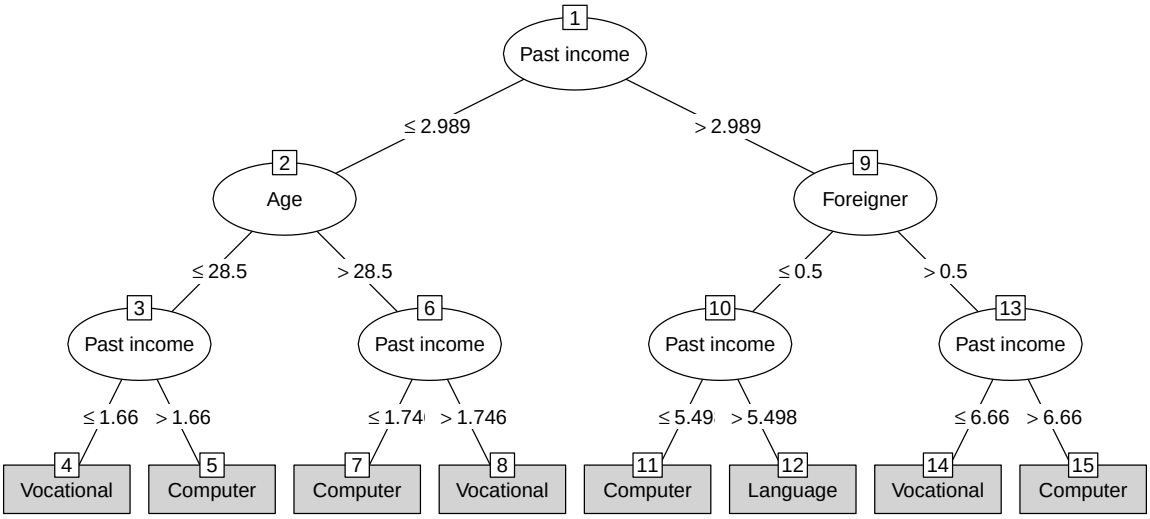
(c) Depth 2 & 16 covariates



(d) Depth 3 & 16 covariates

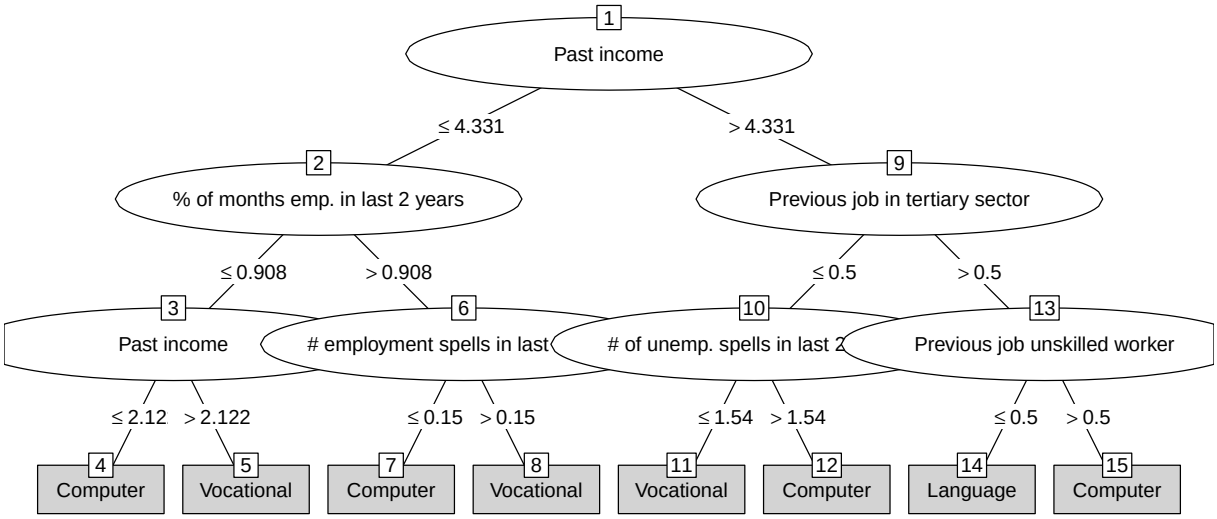
Notes: Past income is measured in CHF 10,000. Graphs are created with the R package `partykit` of Hothorn and Zeileis (2015).

Figure B.6: Optimal decision tree of depth four with five covariates



Notes: Past income is measured in CHF 10,000, employability is an ordered variable with one indicating low employability, two medium employability and three high employability. Graph is created with the R package `partykit` of Hothorn and Zeileis (2015).

Figure B.7: Optimal decision tree of depth four with 16 covariates



Notes: Past income is measured in CHF 10,000, employability is an ordered variable with one indicating low employability, two medium employability and three high employability. Graph is created with the R package `partykit` of Hothorn and Zeileis (2015).

Table B.3: Description of estimated optimal policies

	No program	Job search	Vocational	Computer	Language
	(1)	(2)	(3)	(4)	(5)
Panel A: Percent allocated to program					
Depth 2 & 5 variables	0	0	56	44	0
Depth 3 & 5 variables	0	0	57	33	10
Depth 4 & 5 variables	0	0	32	51	16
Depth 2 & 16 variables	0	0	46	0	54
Depth 3 & 16 variables	0	0	37	19	46
Depth 4 & 16 variables	0	0	42	30	28
Panel B: Cross-validated difference to APOs					
Depth 2 & 5 variables	3.46***	4.49***	0.13	0.17	0.88
	(0.43)	(0.44)	(0.60)	(0.62)	(0.57)
Depth 3 & 5 variables	2.92***	3.95***	-0.41	-0.36	0.34
	(0.42)	(0.43)	(0.59)	(0.61)	(0.56)
Depth 4 & 5 variables	2.97***	4.00***	-0.36	-0.32	0.39
	(0.42)	(0.43)	(0.50)	(0.47)	(0.47)
Depth 2 & 16 variables	3.60***	4.63***	0.27	0.32	1.02**
	(0.42)	(0.43)	(0.44)	(0.58)	(0.41)
Depth 3 & 16 variables	3.70***	4.73***	0.37	0.41	1.12***
	(0.41)	(0.43)	(0.48)	(0.52)	(0.43)
Depth 4 & 16 variables	3.58***	4.61***	0.25	0.30	1.01**
	(0.44)	(0.45)	(0.47)	(0.49)	(0.49)

Note: Panel A show the percentage of individuals that are assigned to a specific program by the trees of different depth. Panel B show a t-test of the difference of the cross-validated policy (standard errors in parentheses) and the APOs of the programs. *p<0.1; **p<0.05; ***p<0.01