Inequality-Averse Outcome-Based Matching

John Körtner* Ruben Bach†

June 1, 2026

Abstract

The increasing use of algorithmic decision-making in policy settings is accompanied by concerns about its potential to reinforce inequalities. In this article, we focus on algorithmic decision-making systems that allocate individuals based on predicted potential outcomes, also known as *outcome-based matching*. Outcome-based matching can amplify inequality in outcomes when treatment effects are heterogeneous and the reallocation's mean impact is driven by gains among individuals who were already relatively advantaged. We propose a modification to the objective function to incorporate inequality-averse preferences in the matching step. The modification applies a prioritarian welfare transformation to predicted potential outcomes, giving greater marginal weight to improvements for individuals who are worse off. We illustrate the modified allocation using data on programs for job seekers. By performing retrospective counterfactual impact evaluations, we show that gains can be shifted toward job seekers with worse predicted baseline outcomes, with moderate losses in mean impact. The loss in other contexts depend on the joint distribution of outcomes and heterogeneous treatment effects. Our approach can guide policymakers in regulating the distributional impact of outcome-based matching.

*Collegio Carlo Alberto. E-mail: john.kortner@carloalberto.org.

†MZES, University of Mannheim.

We would like to thank Giuliano Bonoli, Dominik Hangartner, Michael Knaus, Rafael Lalive, Cyrus Samii, Aleksey Tetenov, Sebastian Zezulka, and audiences at the MZES, EPSA (2023) and PolMeth (2024) for helpful comments and suggestions.

1 Introduction

As the use of algorithmic decision-making systems increases, so do concerns among researchers, policymakers, and the public alike that such systems may reproduce and reinforce inequalities (popularly formulated by Eubanks, 2018, in her book on "automating inequality"). In systems that automatically allocate individuals with the aim of maximizing the mean impact of a policy based on predicted conditional potential outcomes (introduced by Bansak et al. (2018) as "outcome-based matching") inequality in outcomes can indeed be amplified when gains from policy interventions are heterogeneous and when the largest improvements accrue to individuals who are already relatively well-off. In policy areas where goals extend beyond mean impact, such as for welfare policies that are judged not only by how much they help in aggregate, but also by whom they help (e.g. Danziger and Portney, 1982; Friedlander and Robins, 1997; Heckman, Smith and Clements, 1997), outcome-based matching can therefore run counter to policy goals.

In this paper, we propose a straightforward modification to the matching step of outcome-based matching, one that incorporates aversion to inequality in the distribution of outcomes. We formalize this using a prioritarian social welfare function, where a parameter ϵ governs the degree of inequality aversion, based on the social welfare function in Atkinson (1970). The function assigns more weight to improvements for individuals with worse predicted outcomes. Philosophically, the approach follows the position advanced by Parfit (2000), who argues in favor of helping those who are worse off. With $\epsilon \to \infty$, the allocation approaches a Rawlsian idea of justice (Rawls, 1971). The approach satisfies key axioms such as the Pigou-Dalton transfer principle (Pigou, 1912; Dalton, 1920) and Kolm's diminishing marginal utility (Kolm, 1976). When applied to program allocation, it allows policymakers to explicitly trade off average gains against distributional concerns, and to allocate resources in ways that reflect their normative priorities to make progress towards a distribution of outcomes they judge desirable.

We apply the approach to allocating job seekers to programs in public employment services. Using administrative data on active labor market policies in Switzerland (Lechner et al., 2020), we estimate conditional potential outcomes for each job seeker across five policy options with debiased machine learning. We compare standard outcome-based matching to inequality-averse matching through retrospective counterfactual policy evaluation (Samii, Paler and Daly, 2016). The results illustrate how distributional objectives change the allocation when treatment gains are heterogeneous and unevenly distributed across outcome levels. In our application, inequality aversion redirects gains toward job seekers with worse predicted baseline outcomes, while mean impact losses are moderate. Across contexts, losses depend on the joint distribution of outcomes and heterogeneous treatment effects: inequality aversion is less costly when larger gains are available among individuals with worse baseline outcomes, and more costly when larger gains accrue to individuals with better baseline outcomes. Since baseline outcomes may also differ across groups subject to non-discrimination protection, individual-level inequality aversion can have group-level implications even though group membership is not directly included in the objective function.

Our paper contributes to a growing literature on outcome-based matching (Bansak et al., 2018; Ferwerda et al., 2020; Bansak and Martén, 2021; Acharya, Bansak and Hainmueller, 2022; Bansak and Paulson, 2022) and on algorithmic fairness more broadly. To date, most work in algorithmic fairness has focused on predictive bias; that is, ensuring that predictions are equally accurate across groups (see Verma and Rubin, 2018; Mitchell et al., 2021). This has led to design efforts aimed at satisfying different predictive fairness constraints (e.g. Dwork et al., 2012; Zemel et al., 2013; Hardt, Price and Srebro, 2016), though tensions between competing criteria have also been highlighted (Chouldechova, 2017; Kleinberg, Mullainathan and Raghavan, 2017). For outcome-based matching, the focus on prediction alone offers limited guidance, where the key concern is how outcomes are ultimately distributed. Importantly, it also blurs the line between measurement and allocation, a distinction emphasized by recent work (Rambachan et al., 2020; Kuppler et al., 2022).

In our approach, we adopt a consequentialist perspective, in line with Bansak and Martén (2021) and others (e.g. Chohlas-Wood et al., 2023; Kasy, 2024), but extend the focus from impact parity to outcome equality. While our framework shares the general objective of regulating trade-offs between efficiency and distributional concerns with Bansak and Martén (2021), their approach adjusts penalties on group disparities, whereas we focus on regulating inequality across the distribution of outcomes.

Our approach connects to the literature on inequality considerations in treatment choice initiated by Manski (2004). Treatment choice typically focuses on low-dimensional assignment rules. While this leads to a different emphasis regarding the form of the model, the key insights on how to allocate programs according to different objectives extend to outcome-based matching. In particular, Kitagawa and Tetenov (2021) propose optimizing a rank-dependent social welfare function to capture "equality-minded" treatment choice. Our approach differs in that it tailors a prioritarian transformation directly into outcome-based matching. We operate on individualized predictions without constraining the treatment class, relying on flexible machine learning methods to estimate potential outcomes. This design enables straightforward integration with off-the-shelf predictive modeling and optimization tools.

The remainder of the paper proceeds as follows. Section 2 introduces inequality-averse outcome-based matching. Section 3 applies the approach to data from Switzerland, comparing allocations under different levels of inequality aversion. Section 4 discusses implications. Section 5 concludes.

2 Regulating the Distributional Impact of Outcome-Based Matching

We motivate the need for our perspective and introduce the inequality-averse objective function in three steps. We first introduce the building blocks of outcome-based matching. Then, we motivate the need to go beyond mean impact and make the case for distributional ob-

jectives. Finally, we introduce the inequality-averse objective function and discuss inherent trade-offs.

2.1 Preliminaries

At the core of algorithmic decision-making systems used by public bodies is supervised learning. Past data on individuals $j = 1, ..., n$, each described by a vector of features x_j, and an outcome y_j, are used to estimate a model to predict $\hat{y}_j$. As the predictive model, we define any function that predicts y_j from x_j. Once the model is trained, it can be used to make predictions of the outcome for individuals where the outcome is not (yet) known.[1]

While knowing future outcomes can have many applications on its own (see e.g. Kleinberg et al., 2015), in outcome-based matching it is only an intermediate step on the way to a decision. Let us formulate the decision as $D = \{0, 1, ..., k\}$, where k denotes policy options an individual can be allocated to. Taking into account the decision, we are interested in

$$\mu_k(x) = E[Y_j(k)|X_j = x], \tag{1}$$

where $\mu_k(x)$ represents conditional potential outcomes (Neyman, 1923; Rubin, 1974, see Keele, 2015 for an introduction). Having access to a prediction of $\mu_k(x)$ informs us about the expected potential outcomes of every intervention for all individuals. The idea of outcome-based matching is to assign individuals to options such that the mean impact of the policy is optimized. The matching step can be formalized as

$$\Phi_i = \arg\max_{\Phi \in \Pi} \sum_{j=1}^{n} \mu_{j\phi_{ij}}, \tag{2}$$

where $\Phi_i \in \Pi$ denotes a specific allocation out of all feasible ones, and ϕ_{ij} the individual assignment under allocation i (in line with Bansak and Martén, 2021).

[1] We can be agnostic to the way $\mu(x)$ is estimated. However, increasingly popular is the use of approaches from machine learning. The appeal of machine learning is that it allows for complex relationships in the historical data and that it estimates models that work well 'out-of-sample', i.e., on data that the model has not seen during the estimation, which is the aim of predictive algorithms for policy applications.

Outcome-based matching differs from predictive algorithms in the form of risk assessment (or profiling), where the quantity of interest is a baseline outcome $\mu_0(x) = E[Y_j(0) = 1|X_j = x]$ or $\Pr[Y_j(0) = 1|X_j = x]$ under no treatment i.e. $k = 0$ (Berger, Black and Smith, 2001). In practice, the model is often reduced to $\mu(x) = \Pr[Y_j = 1|X_j = x]$ ignoring the selected observability. Predicted outcomes are then used to prioritize individuals, e.g., those who have the highest risk to experience some undesired event. To go from the estimated risk to a decision, it is common to threshold the score, e.g., an option is provided to an individual if and only if the risk is higher than a given predefined cutoff value, $\Phi_i : \ k_j \geq 1 \ \leftrightarrow \ \mu_{0j} \geq c$, else $k = 0$, where the parameter c denotes the threshold to divide individuals into 'high-' and 'low-risk'. The selection of the specific intervention out of all $k \geq 1$ is not well defined in profiling, except when the decision contains only one option, $|D| = 2$. When $|D| > 2$, profiling only specifies whether an action should be taken or not. The choice of the policy option is left to the discretion of the deciding bureaucrat.

2.2 Criteria of Interest Besides the Mean

Outcome-based matching allocates individuals to policy options in order to maximize mean impact. However, outcome-based matching systems, as all algorithmic decision systems in the public sector, are deployed in the service of institutions that have to trade off political principles (Barry, 1965): besides optimal allocation, common principles are non-discrimination and distributional objectives such as service to particular groups and the reduction of inequality.

A large amount of literature on "algorithmic fairness" centers on the principle of equal treatment. It typically considers a protected group G (such as gender, ethnicity, or citizenship status) and evaluates whether predictive models or decisions exhibit disparities across groups. Common criteria include equal opportunity, which requires that the true positive rate is the same across groups, and predictive parity, which requires that predicted positives correspond to actual positives at the same rate across groups (see e.g. Verma and Rubin, 2018, for

an overview). While these criteria offer important diagnostic tools, they are not sufficient to assess outcome-based matching. First, the assignment is based on comparisons across multiple predicted outcomes rather than a binary threshold. Second, these criteria focus on prediction quality rather than the effect of interventions, and may conflate prediction error with real, heterogeneous treatment effects.

To assess fairness in a consequentialist sense, Bansak and Martén (2021) propose to take the average of the individual outcomes under the allocation to measure the potential outcome for allocation procedure Φ_i. To measure the average impact of a potential allocation, they propose to measure the change against the status quo. The status quo can either be the current policy allocation or the status without the policy.

$$\Psi_{\Phi_i} = \frac{1}{n} \sum_{j=1}^{n} \left(\mu_{j\phi_{ij}} - \mu_{j\phi_{0j}} \right). \tag{3}$$

The difference of Equation 3 to a treatment effect is that it takes into account the assignment, i.e., the effect only materializes for the individuals that are actually assigned to that intervention. For every individual that is not assigned to any program their measure is equal to 0. To then measure group impact, Bansak and Martén (2021) propose to compute the impact separately for each group with

$$\Psi_{\Phi_i}(g) = \frac{1}{\sum_{j=1}^{n} \mathbb{1}(G_j = g)} \sum_{j=1}^{n} \left(\mu_{j\phi_{ij}} - \mu_{j\phi_{0j}} \right) \cdot \mathbb{1}(G_j = g), \tag{4}$$

where G denotes membership in a protected group. As criteria for equal treatment, they propose to allocate individuals to policy options to achieve balanced impact across protected groups and allow an imbalance of ϵ at most, which the policymaker should decide. The policymaker could apply the "four-fifths rule," a guideline for determining adverse impact in hiring, and set the tolerance to $\epsilon = 1/5$ (Bansak and Martén, 2021, Equation 9 and 10).

The concern in this article is that outcome-based algorithms may inadvertently exacerbate inequality in outcomes. This critique has gained prominence with the rise of algorithmic

governance, especially in domains like welfare policy. Critics such as Eubanks (2018) have argued that algorithms risk reproducing existing social disadvantages (see also Allhutter et al., 2020; Desiere and Struyven, 2021). In social policy, the legitimacy of programs, particularly welfare interventions, is often tied not only to their overall effectiveness but also to their distributional impact. As Le Grand (2018) notes, a major rationale for the growth of the welfare state has been the pursuit of greater equality. Policymakers and the public frequently evaluate policies not just by how much good they do in aggregate, but by whom they benefit (Friedlander and Robins, 1997; Heckman, Smith and Clements, 1997; Bitler, Gelbach and Hoynes, 2006). Imagine a program that yields large benefits for already advantaged individuals and modest gains for disadvantaged ones. A matching algorithm that assigns individuals based on predicted gains would likely favor the former group, exacerbating existing inequalities, a dynamic also known as the "Matthew effect" (Merton, 1968), where those who have more receive more.[2]

This tension speaks directly to the philosophical distinction between utilitarianism and prioritarianism. While utilitarianism values total gains regardless of who receives them, prioritarianism gives greater moral weight to benefits accruing to the worse off. As Parfit (2000) argues: benefit "also depends on how well off the person is to whom this benefit comes. We should not give equal weight to equal benefits, whoever receives them. Benefits to the worse off should be given more weigh" (p. 101). In our context, this means evaluating not only the size of the impact from allocation algorithms, but its distribution.

To diagnose the distributional implications of a policy allocation, we propose measuring impact across the quantiles of a baseline outcome distribution. Extending the impact

[2]Matthew effects are a common theme of interest in the social policy literature on social investments (Bonoli, Cantillon and Van Lancker, 2017, provide an introduction and overview).

measure by Bansak and Martén (2021), we define the quantile-specific average impact as:

$$\Psi_{\Phi_i}(q^s) = \frac{1}{\sum_{j=1}^{n} \mathbb{1}(Q_j = q^s)} \sum_{j=1}^{n} \left(\mu_{j\phi_{ij}} - \mu_{j\phi_{0j}}\right) \cdot \mathbb{1}(Q_j = q^s), \tag{5}$$
$$\text{with } q^s = \inf\left\{q : \frac{1}{n}\sum_{l=1}^{n} \mathbb{1}(\mu_{l\phi_{0l}} \leq q) \geq s\right\}.$$

Here, $Q_j = q^s$ indicates that individual j belongs to quantile group s of the predicted outcome distribution under the reference allocation Φ_0. This formulation groups individuals based on their standing in the distribution of predicted outcomes under the status quo (e.g., no intervention), allowing us to examine who benefits most from the proposed allocation. By examining the impact of allocation decisions across quantiles, we can better understand whether algorithms direct benefits toward those who need them most or inadvertently reinforce inequality. The remaining question is how to diagnose inequality and achieve a desired impact in the allocation.

2.3 Modifying the Objective Function

A central question in the design of allocation algorithms is where and how normative priorities are expressed in a way that enforces them. In algorithmic fairness research, many proposals focus on modifying the prediction stage to ensure equal performance across protected groups. For instance, constraints are imposed so that error rates or predictive accuracy are balanced between groups defined by race, gender, or other protected attributes (Dwork et al., 2012; Zemel et al., 2013; Hardt, Price and Srebro, 2016). These approaches aim to ensure that predictions do not depend on group membership, either directly or through proxies.

However, this strategy has important limitations. First, modifying predictions can interfere with learning about true outcome heterogeneity, potentially harming those it aims to protect (e.g. Chohlas-Wood et al., 2023). Second, it neglects the second stage of algorithmic decision-making — the mapping of predictions to actual decisions. Even when predictions are unbiased, a decision rule may exacerbate inequalities depending on how those predictions

are used.

Instead of constraining prediction, we propose incorporating normative preferences directly into the objective function of the matching algorithm (in line with e.g. Rambachan et al., 2020). This allows policy-makers to transparently specify how much weight should be placed on efficiency versus distributional fairness. Following Berger, Black and Smith (2001), we write the general form of an allocation objective as

$$\Phi_i = \arg\max_{\Phi \in \Pi} \sum_{j=1}^{n} \mu_{j\phi_{ij}} \cdot \omega(\cdot), \tag{6}$$

where $\omega(\cdot)$ is a weighting function which allows the outcomes for different individuals to be weighted differently. In the canonical outcome-based matching formulation, given in Equation 2, it is implicitly assumed that weights are equal for all individuals, formalizing a strict utilitarian idea of social welfare. Introducing variation in $\omega(\cdot)$ allows us to consider alternative objectives to a strictly utilitarian one. The remaining question is how to derive $\omega(\cdot)$.

We derive $\omega(\cdot)$ from a normative framework rooted in prioritarianism, which holds that benefits to the worse off should count more (Parfit, 2000). We formulate the inequality-averse allocation problem as

$$\Phi_{\epsilon i} = \arg\max_{\Phi \in \Pi} \sum_{j=1}^{n} \mu_{j\phi_{ij}}^{1-\epsilon}/(1-\epsilon), \tag{7}$$

with $0 \leq \epsilon$ and $\epsilon \neq 1$. For $\epsilon = 1$, we write $\Phi_{1i} = \arg\max_{\Phi \in \Pi} \sum_{j=1}^{n} \log(\mu_{j\phi_{ij}})$. The function is strictly concave. With increasing ϵ, the welfare function becomes increasingly averse to inequality. As $\epsilon \to 0$, $W_\epsilon(\Phi_\epsilon)$ converges to the utilitarian welfare function. As $\epsilon \to \infty$, we approach Rawls (1971).[3]

The modified objective function is well grounded in distributive justice and social choice

[3]With $\epsilon > 1$ transformed values flip to negative. This is not a problem for the matching step. If transformed values should be compared however, adding a constant is helpful: $\Phi_{\epsilon i} = \arg\max_{\Phi \in \Pi} \sum_{j=1}^{n} (\mu_{j\phi_{ij}}^{1-\epsilon} - 1)/(1-\epsilon)$.

(e.g. Patty and Penn, 2019, provide an overview). $W_\epsilon(\cdot)$ satisfies the following axioms. First, it satisfies *symmetry*: welfare depends only on the distribution of predicted outcomes, not on who receives them. Swapping outcomes between individuals does not change overall welfare. Second, it satisfies *continuity*: small changes in outcomes lead to small changes in welfare, ensuring robustness to prediction variation. The function also adheres to the *Pareto principle*: if one individual's predicted outcome improves while all others remain the same, overall welfare must increase. This ensures that clearly beneficial allocations are preferred. More substantively, the function satisfies the *Pigou-Dalton principle* (Pigou, 1912; Dalton, 1920). This means that redistributing a small amount of predicted benefit from a better-off to a worse-off individual without changing the average increases social welfare. The intuition is that gains are worth more to those who begin with less. Finally, the function satisfies *Kolm's principle of diminishing transfers* (Kolm, 1976), which strengthens the previous axiom: for equal outcome improvements, welfare increases more when the gain goes to someone worse off. This expresses the core idea of prioritarianism (Parfit, 2000): marginal improvements matter more for those with lower starting values. The curvature of the welfare function, controlled by ϵ, formalizes this diminishing sensitivity and governs how strongly the planner prioritizes improvements among the disadvantaged. The axioms collectively provide a rigorous normative foundation for using the prioritarian welfare function in matching. The function is simple to implement and interpretable, while capturing rich trade-offs between efficiency and fairness.

Using the function as an approach to algorithmic equality and fairness has several important benefits. Disadvantaged individuals will receive greater weight in the allocations, proportional to their initial disadvantage in the sense that an equal predicted improvement at outcome level $\tilde{\mu}$ receives marginal welfare weight $\tilde{\mu}^{-\epsilon}$. At the same time, the allocation will not pursue equality if this would render the policy completely ineffective. The function would balance mean impact, or efficiency (as the resources would stay the same cf. Le Grand, 1990), and equality, and make the preferences for either side of the trade-off explicit through

the inequality-aversion parameter. This speaks to the widely studied policy trade-off between efficiency and inequality (e.g. Barry, 1965; Okun, 1975).

3 Empirical Illustration

We demonstrate our approach in the context of public employment services (PES). An important task of public employment services is to support job seekers in resuming work by assigning them to active labor market policies, that is, programs designed to increase job seekers' chances to resume work. In this context, algorithmic systems have been used since the 1990s to predict job seekers' re-employment chances (see e.g. Körtner and Bonoli, 2023, for an overview). Those predicted to struggle with resuming work and identified as high risk of becoming long-term unemployed are typically assigned to intensive counseling and active labor market policies. At the same time, algorithms in PES are increasingly subject to public criticism (e.g. Spiekermann, 2019; Allhutter and Mager, 2020; Allhutter et al., 2020; Desiere and Struyven, 2021).

3.1 Data

We use an observational dataset of active labour market policies in Switzerland (Lechner et al., 2020). In Switzerland, caseworkers decide on participation in programs based on subjective evaluations of a jobseeker's employment prospects. Job seekers are notified about their participation in a program one or two weeks in advance, and job seekers are not allowed to refuse participation once they are assigned to participate in a program (Lalive, Van Ours and Zweimüller, 2008). Our data includes individuals between 24 and 55 years old who registered as unemployed in 2003. Following Knaus (2022), we select individuals from the German-speaking part of Switzerland, participants and non-participants of vocational training, computer programs, language courses, and job search assistance, and we consider the first program participation within the first six months after the start of an unemployment spell. The outcome is months in employment (with zero meaning no employment) after

program start, or pseudo program start for non-participants.[4]

3.2 Evaluation Procedure

To evaluate our approach to regulating the distribution impact of outcome-based matching, we rely on the idea of a retrospective counterfactual policy evaluation (Samii, Paler and Daly, 2016): we train machine learning models and predict potential outcomes for all job seekers under all programs. We then re-allocate the job seekers in the original data to optimize according to the objective function we motivated in Section 2.3. Once we have all retrospective re-allocations, we evaluate the allocations based on their potential mean impact, potential quantile impact, and potential impact across protected groups.

The central object for assignment is the individual-level comparison of predicted potential outcomes across program options. A program may have a small or even zero average treatment effect while still being useful for assignment if its effects are heterogeneous. For example, a program that strongly benefits one subgroup and harms another can average to zero. Outcome-based matching can still improve outcomes by assigning the program to individuals with positive predicted gains and withholding it from those with negative predicted gains. The same logic motivates inequality-averse matching: the assignment rule does not only ask which option creates the largest predicted gain, but also whose outcome level receives priority in the social welfare objective.

Estimation. Assignment to programs in our data is non-random: caseworkers select participants based on characteristics that may also predict outcomes. To recover credible estimates of counterfactual potential outcomes, we apply a doubly robust (Robins, Rotnitzky and Zhao, 1994), cross-fitted machine learning estimator (Chernozhukov et al., 2018; Kennedy,

[4]Program non-participants comprise people that quickly came back into employment before they could be assigned to a training program. To account for this, we follow Knaus (2022) in estimating pseudo-program starts in the first six months. We draw pseudo-start values out of a Bernoulli distribution with the estimated pseudo-start as p. Individuals that are employed at the pseudo-starting point are dropped from the subsequent estimation (see also Lechner, 1999).

2023). For each program k, we estimate the conditional expectation:

$$\mu_k(x) = E[Y_j(k) \mid X_j = x] \tag{8}$$

where $Y_j(k)$ denotes the potential outcome under program k, and X_j is a vector of covariates. We model the propensity score $e_k(x) = \Pr[D_j = k \mid X_j = x]$ and the conditional mean outcome $y_k(x) = E[Y_j \mid D_j = k, X_j = x]$ using generalized random forests (Athey, Tibshirani and Wager, 2019), implementing 4-fold cross-fitting to avoid overfitting and maintain honest estimation. The doubly robust score for each individual-program pair is:

$$\hat{\Gamma}_k(x) = \hat{y}_k(x) + \frac{D_j(k)(Y_j - \hat{y}_k(x))}{\hat{e}_k(x)} \tag{9}$$

Final estimates of $\mu_k(x)$ are obtained by regressing $\hat{\Gamma}_k(x)$ again on covariates in a cross-fitted manner (Kennedy, 2023). Our identification relies on two key assumptions:

Unconfoundedness: $Y_j(k) \perp D_j \mid X_j$ for all k.

Overlap: $0 < e_k(x) < 1$ for all x.

Given rich covariate information and flexible machine learning models, these assumptions are plausible in our setting. Knaus (2022) and Knaus, Lechner and Strittmatter (2022) defend the assumptions for the same dataset to estimate treatment effects and benchmark causal learners, respectively. We show all variables used in the estimation and overlap in Appendix B.

Assignment. Having estimated individualized potential outcomes, we solve the following optimization problem:

$$\max_{a_{jk} \in \{0,1\}} \sum_{j=1}^{n} \sum_{k=0}^{K} u_\epsilon(\tilde{\mu}_{jk}) a_{jk} \tag{10}$$

subject to $\sum_{k=0}^{K} a_{jk} = 1, \forall j \in \{1, \ldots, n\}$ and $\sum_{j=1}^{n} a_{jk} \leq C_k, \forall k \in \{0, \ldots, K\}$. Here, $a_{jk} = 1$ indicates that individual j is assigned to option k, where $k = 0$ denotes no program. The no-program option is assigned a capacity equal to the sample size and is therefore effectively unconstrained. The other program capacities are bounded above by their observed levels, so that the counterfactual allocations do not expand program capacity relative to the status quo. In the empirical implementation, we set $\tilde{\mu}_{jk} = \max\{\hat{\mu}_{jk}, 10^{-6}\}$ before applying $u_\epsilon(\cdot)$, so that the logarithmic and constant-relative-inequality-aversion transformations are well-defined. Optimization is performed using a mixed-integer programming solver (Appendix A).

Evaluation. To evaluate the consequences of different allocation rules, we rely on measures for retrospective counterfactual impact evaluation (Samii, Paler and Daly, 2016; Bansak et al., 2018). For each allocation, we compute the mean change as specified in Equation 3 relative to baseline no-program outcomes, which corresponds to a "retrospective intervention effect" (RIE) in Samii, Paler and Daly (2016) and to causal impact in Bansak et al. (2018), assuming a population-level re-allocation. In addition, we compute the RIE within quartiles of the baseline distribution (Q-RIE, Equation 5) and impact by protected group status (G-RIE, Equation 4). We construct two-sided 95% confidence intervals by nonparametrically resampling individual-level retrospective impacts, conditional on the estimated potential outcome surfaces and the induced assignments.[5]

In addition to the assumptions underlying the debiased estimation, the evaluation relies on an *invariance* assumption as discussed by Heckman and Smith (1995) and Heckman (2020): we assume that the distribution of potential outcomes, conditional on covariates, remains stable across different assignment regimes. In other words, reallocating individuals based on predicted outcomes does not itself change the outcome production process.

Finally, our evaluation is model-based in the sense that the same estimated potential outcome surfaces enter both the assignment problem and the retrospective impact summaries.

[5]The intervals therefore summarize sampling uncertainty in the retrospective impact summaries given the fitted models. They should not be interpreted as incorporating all uncertainty from nuisance estimation, model selection, and optimization.

Recent work on outcome-based matching studies the robustness of such counterfactual impact evaluations to alternative off-policy evaluation methods (Bansak et al., 2026). Our empirical exercise should therefore be read as a model-based retrospective evaluation of how different social welfare objectives change allocations and implied outcome distributions, rather than as a full comparison of off-policy evaluation estimators.

3.3 Results

Estimates. We first report average treatment effects (ATE) and quartile-specific average treatment effects (Q-ATE) by program in Table 1.[6] Quartiles are defined by predicted outcomes under no program.

Program	N	ATE	Q-ATE			
			1st	2nd	3rd	4th
Computer Skills	905	3.40 [3.37, 3.43]	5.21 [5.15, 5.28]	3.97 [3.91, 4.04]	3.35 [3.29, 3.40]	1.05 [1.00, 1.09]
Language Skills	1'504	2.37 [2.34, 2.39]	1.66 [1.60, 1.71]	2.72 [2.66, 2.78]	2.87 [2.82, 2.93]	2.23 [2.18, 2.27]
Vocational Training	858	3.31 [3.28, 3.34]	6.44 [6.39, 6.49]	4.30 [4.25, 4.34]	2.63 [2.59, 2.66]	-0.13 [-0.17, -0.10]
Job Search Assistance	11'610	-1.01 [-1.02, -0.99]	0.67 [0.65, 0.69]	-0.87 [-0.89, -0.85]	-1.63 [-1.66, -1.60]	-2.19 [-2.22, -2.17]
No Program	47'598	—	—	—	—	—
Total sample	62'475					

Table 1: **Treatment Effects.** The table reports estimated average treatment effects (ATE) and quartile-specific average treatment effects (Q-ATE) on months in employment after program start. Quartiles are defined by predicted employment under no program, so that the first quartile contains individuals with the worst predicted baseline outcomes. Positive values indicate more months in employment relative to no program. Two-sided 95% confidence intervals are reported in brackets.

Computer skills, language skills, and vocational training increase cumulative months of employment on average, while job search assistance decreases employment on average.

[6]The definition is in line with conditional average treatment effects (CATE) and differs from quantile treatment effects (QTE) (as defined e.g. in Imbens and Rubin, 2015, 20.3.1). QTE are not necessarily informative because the particular individual located at a given quantile may differ across comparison groups (see also Manski, 2009, p. 157).

However, the average effects mask substantial heterogeneity across the baseline outcome distribution. Vocational training has large positive effects for individuals in the lowest baseline quartile, smaller gains in the middle of the distribution, and slightly negative effects in the top quartile. Computer skills show a similar gradient, with larger effects among individuals with worse predicted no-program outcomes. Language skills produce more even gains across quartiles.

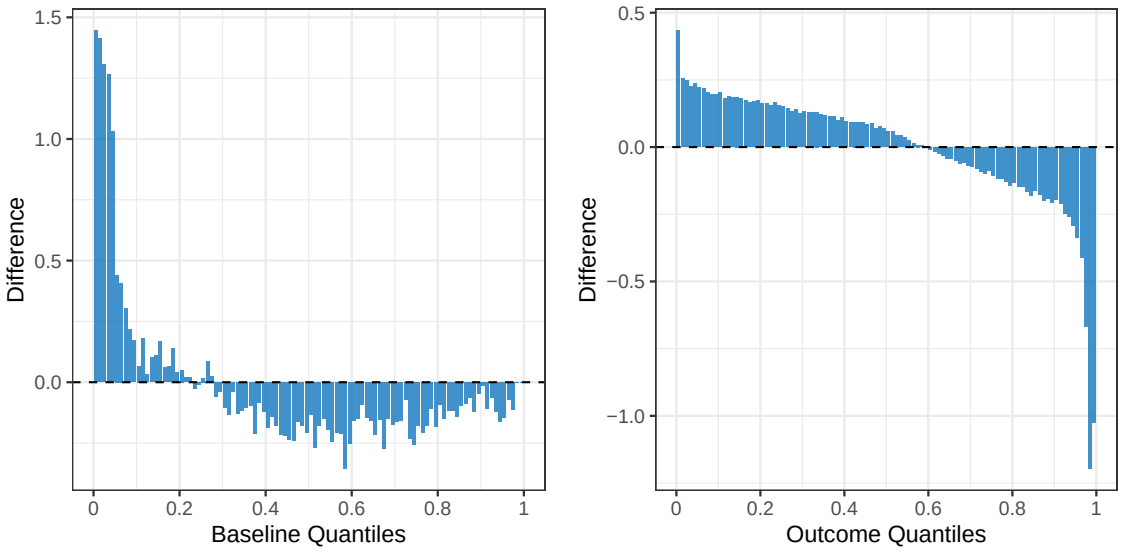
Impacts. Table 2 shows the results from implementing inequality-aversion into the assignment with different values of the inequality aversion parameter ϵ. Each row reports mean retrospective intervention effects (RIE) and quartile-specific retrospective intervention effects (Q-RIE) relative to assigning no program.

ϵ: Inequality Aversion	Q-RIE *over no intervention*				Mean RIE
	1st	2nd	3rd	4th	
$\epsilon = 0$	1.91	1.02	0.53	0.17	0.90
	[1.85, 1.95]	[0.97, 1.06]	[0.49, 0.56]	[0.15, 0.19]	[0.89, 0.92]
$\epsilon = 0.25$	2.10	0.95	0.44	0.11	0.90
	[2.05, 2.16]	[0.91, 1.00]	[0.41, 0.47]	[0.09, 0.12]	[0.88, 0.92]
$\epsilon = 0.5$	2.27	0.89	0.34	0.06	0.89
	[2.21, 2.32]	[0.85, 0.93]	[0.31, 0.36]	[0.05, 0.07]	[0.87, 0.91]
$\epsilon = 0.75$	2.40	0.80	0.26	0.04	0.87
	[2.34, 2.46]	[0.76, 0.83]	[0.24, 0.28]	[0.03, 0.04]	[0.85, 0.89]
$\epsilon = 1$	2.48	0.71	0.19	0.03	0.85
	[2.43, 2.54]	[0.67, 0.75]	[0.18, 0.21]	[0.02, 0.03]	[0.83, 0.87]

Table 2: **Retrospective Counterfactual Impact Evaluation.** The table reports mean retrospective intervention effects (RIE) and quartile-specific retrospective intervention effects (Q-RIE) relative to assigning no program. Quartiles are defined by predicted employment under no program. Larger values indicate larger predicted gains in months of employment under the counterfactual allocation. Two-sided 95% confidence intervals are reported in brackets.

As ϵ increases from 0 to 1, gains become increasingly concentrated among individuals in the lowest quartile of predicted no-program outcomes. For $\epsilon = 0$, individuals in the first quartile gain 1.91 months of employment on average; for $\epsilon = 1$, this rises to 2.48 months. This distributional shift comes at a moderate cost to mean impact: mean RIE declines from

0.90 to 0.85 months. Even a moderate value, $\epsilon = 0.5$, increases impact in the lowest quartile by 0.36 months relative to $\epsilon = 0$, while reducing mean impact by about 0.01 months. In this application, inequality aversion therefore produces a sizable shift in gains toward individuals with worse predicted baseline outcomes at limited cost to average impact.



(a) Impact Difference by Baseline Quantile (b) Outcome Difference by Realized Quantile

Figure 1: **Distributional Change.** Panel (a) plots the difference in retrospective impacts between inequality-averse matching with $\epsilon = 0.5$ and standard outcome-based matching with $\epsilon = 0$ across quantiles of predicted employment under no program. Positive values indicate that individuals at that baseline quantile gain more under the inequality-averse assignment. Panel (b) plots the corresponding difference across quantiles of the retrospectively realized outcome distribution.

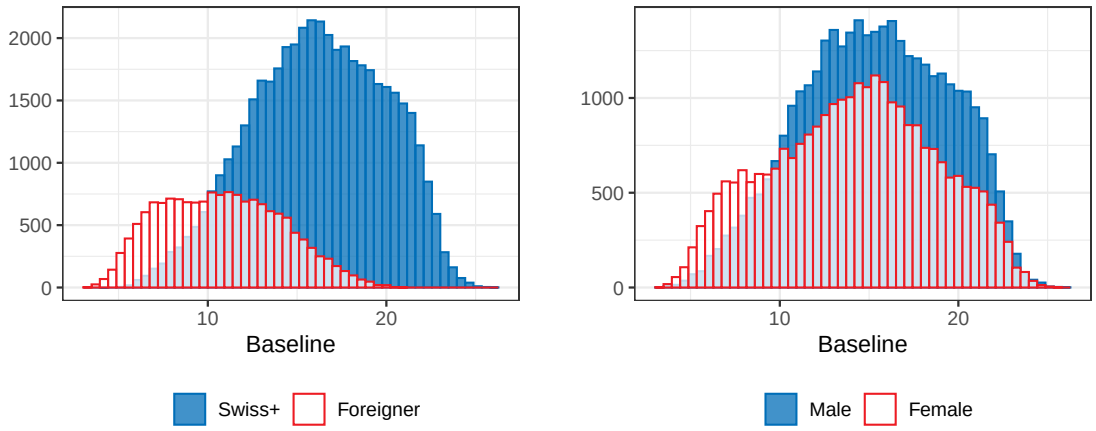
Figure 1 visualizes how inequality aversion changes the distribution of impacts. Panel (a) plots the difference in retrospective impacts, based on Equation 5, between inequality-averse matching with $\epsilon = 0.5$ and standard outcome-based matching with $\epsilon = 0$ across quantiles of predicted employment under no program. Positive values indicate quantiles that gain more under the inequality-averse assignment. The pattern shows that the modified objective shifts benefits toward the lower part of the baseline distribution. Panel (b) reports the corresponding difference across the retrospectively realized outcome distribution, using the same quantile-based comparison as in Kitagawa and Tetenov (2021, Figure 3), adapted to retrospective impacts. Specifically, we define $q^s_\Phi = \inf\{q : n^{-1} \sum_{j=1}^n \mathbf{1}(Y_j^\Phi \leq q) \geq s\}$, where

Y_j^{Φ} is the predicted realized outcome under allocation Φ, and plot $q^s_{\Phi_{\epsilon=0.5}} - q^s_{\Phi_{\epsilon=0}}$. Unlike Panel (a), this compares quantiles of realized outcome distributions rather than mean impacts within baseline quantiles.

To better understand where the changes stem from, Table A.2 in the Appendix reports assignment switches relative to standard outcome-based matching with $\epsilon = 0$. A switch occurs when an individual receives a different assignment under an inequality-averse objective than under the mean-maximizing objective. Switches are concentrated among individuals with the worst predicted no-program outcomes. At $\epsilon = 0.5$, 7.61% of individuals in the lowest baseline quartile switch assignments, compared with 2.54%, 2.05%, and 1.31% in the second, third, and fourth quartiles, respectively. At $\epsilon = 1$, the switching rate in the lowest quartile rises to 14.35%, while only 1.84% of individuals in the highest quartile switch. Inequality aversion therefore changes the allocation primarily at the bottom of the baseline outcome distribution rather than uniformly reshuffling assignments.

Group-Level Implications. We next consider group-level implications to show the connection to other work on algorithmic fairness, in particular to Bansak and Martén (2021). The inequality-averse assignment rule does not include group membership directly, but it can still have group-level consequences if protected groups differ in predicted baseline outcomes or treatment gains.

We focus on two attributes, shown in Figure 2. We define citizenship group as individuals who do not have Swiss citizenship and who did not report German, French, or Italian as a native language, in the sense of an extended citizenship category. This definition reflects the multilingual Swiss context: it distinguishes foreign citizens likely to face both citizenship- and language-related labor market barriers from foreign citizens from neighboring countries who share a Swiss official language. Gender provides a useful comparison because baseline differences are less pronounced. Considering both attributes helps assess when individual-level inequality aversion translates into group-level differences.



(a) Citizenship Status* (b) Gender

Figure 2: **Baseline Outcomes by Protected Group.** The figure shows the distribution of predicted months in employment under no program separately by citizenship status and gender. Panel (a) compares individuals with and without Swiss citizenship (*and who did not report German, French, or Italian as a native language). Panel (b) compares men and women.

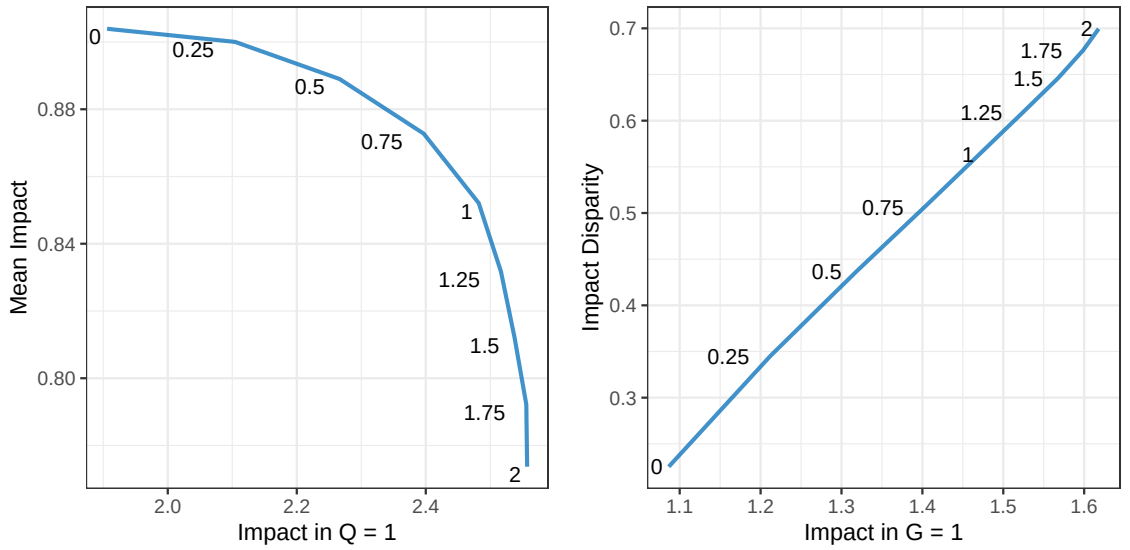
Figure 2 shows the predicted baseline outcomes. Individuals in the disadvantaged citizenship group are predicted to have fewer months in employment under no program. For gender, the difference is less pronounced.

Outcome-based matching with ϵ	G-RIE *over no intervention*				Mean RIE
	Swiss +	Foreigner	Male	Female	
$\epsilon = 0$	0.84	1.09	0.80	1.04	0.90
	[0.82, 0.87]	[1.05, 1.13]	[0.77, 0.83]	[1.00, 1.07]	[0.89, 0.92]
$\epsilon = 0.25$	0.79	1.21	0.79	1.04	0.90
	[0.77, 0.82]	[1.17, 1.25]	[0.76, 0.81]	[1.01, 1.08]	[0.88, 0.92]
$\epsilon = 0.5$	0.74	1.32	0.76	1.05	0.89
	[0.72, 0.77]	[1.27, 1.36]	[0.74, 0.79]	[1.02, 1.08]	[0.87, 0.91]
$\epsilon = 0.75$	0.69	1.40	0.74	1.04	0.87
	[0.67, 0.71]	[1.36, 1.45]	[0.72, 0.77]	[1.01, 1.07]	[0.85, 0.89]
$\epsilon = 1$	0.64	1.47	0.71	1.03	0.85
	[0.62, 0.66]	[1.42, 1.52]	[0.69, 0.74]	[1.00, 1.06]	[0.83, 0.87]

Table 3: **Group-Specific Retrospective Impact Evaluation.** The table reports group-specific retrospective intervention effects (G-RIE) by citizenship status and gender, together with the mean retrospective intervention effect (RIE). Two-sided 95% confidence intervals are reported in brackets.

Table 3 reports group-specific RIE (G-RIE) by citizenship status and gender. For each ϵ value, we report mean gains for Swiss and foreign citizens, and for males and females, respectively. At baseline ($\epsilon = 0$), foreign citizens and females already experience slightly larger gains than Swiss citizens and males, reflecting their relatively worse baseline employment prospects. As ϵ increases, the gap in gains between groups widens. For example, under $\epsilon = 1$, foreign citizens gain 1.47 months on average compared to 0.64 months for Swiss citizens. Thus, introducing inequality-aversion can also improve distributive fairness across protected groups, even though the optimization was conducted purely at the individual level without explicit group constraints.

Trade-Offs. Figure 3 summarizes two trade-offs induced by inequality-averse matching. Panel (a) plots mean retrospective impact against impact in the lowest baseline quartile across different values of ϵ.



(a) Mean Impact vs Priority (b) Mean Impact vs Group Impact Disparity

Figure 3: **Trade-Offs.** Panel (a) plots mean retrospective impact against impact in the lowest baseline quartile across values of ϵ (shown in the plot). Panel (b) plots the group impact disparity, defined following Bansak and Martén (2021), against group-specific impact.

Moving to larger values of ϵ increases gains for individuals with the worst predicted no-program outcomes, while reducing mean impact. In this application, the trade-off is relatively shallow for moderate levels of inequality aversion: sizeable gains for the lowest baseline quartile are achieved with only small losses in average impact.

Panel (b) shows the corresponding trade-off for group-specific impacts. Following Bansak and Martén (2021), we summarize group-impact disparity as $1 - \min\{\Psi_{\Phi_i}(G = 0), \Psi_{\Phi_i}(G = 1)\} / \max\{\Psi_{\Phi_i}(G = 0), \Psi_{\Phi_i}(G = 1)\}$. Larger values indicate greater imbalance in gains across protected groups. The figure shows that individual-level inequality aversion can increase group-impact disparities, especially by citizenship status. This occurs because the objective prioritizes individuals with worse predicted baseline outcomes, not parity between protected groups.

Together, the results show that modifying the objective function of outcome-based matching can shift gains toward individuals with worse baseline prospects. They also show that the same rule can have different implications for individual-level priority and group-level parity. When disadvantage is unevenly distributed across protected groups, prioritizing lower predicted outcomes can increase differences in average gains between groups.

4 Discussion

The empirical results show that inequality-averse outcome-based matching changes who benefits from algorithmic assignment. In the Swiss public employment service application, increasing ϵ shifts gains toward individuals with worse predicted no-program outcomes, while mean retrospective impact declines only moderately. Looking at assignment switches, we see that this is due to reallocation of marginal program capacity toward the bottom of the predicted baseline outcome distribution.

The cost of inequality aversion depends on the distribution of heterogeneous treatment effects. In Appendix C, we vary the relationship between baseline outcomes and treatment gains while holding the empirical baseline distribution and active-program capacity fixed.

When treatment gains favor the worse-off, inequality aversion is almost costless because efficiency and priority select similar individuals. When treatment gains are unrelated to baseline outcomes, the trade-off is moderate. When treatment gains favor the better-off, the efficiency cost becomes larger because the individuals who benefit most from treatment are not those with the worst baseline outcomes.

The emphasis on treatment effects is important for policy design. If available interventions are ineffective for disadvantaged individuals, an inequality-averse algorithm can assign more resources to them, but it cannot create large gains where the programs do not generate them. Inequality is therefore not only a property of the allocation rule; it is also a property of the policy: the available interventions and the heterogeneous effects they produce. Better algorithms can regulate how existing gains are distributed, but they cannot substitute for better programs.

The approach supports a consequentialist perspective on algorithmic fairness (Bansak and Martén, 2021; Chohlas-Wood et al., 2023; Kasy, 2024). The two closest connections are to Bansak and Martén (2021) and Kitagawa and Tetenov (2021). Bansak and Martén (2021) propose measuring and constraining disparities in impact across protected groups; we focus on inequality across the outcome distribution itself. This connects outcome-based matching to equality-minded treatment choice: Kitagawa and Tetenov (2021) show how rank-dependent welfare functions can trade off average welfare against distributional concerns in treatment assignment. We tailor this logic to outcome-based matching by incorporating a prioritarian transformation directly into the assignment problem over multiple policy options, predicted potential outcomes, and capacity constraints. The parameter ϵ makes explicit whether the planner seeks only to increase total predicted outcomes or also to prioritize improvements among those with worse prospects.

5 Conclusion

Outcome-based matching can improve the allocation of scarce policy resources, but its distributional consequences are not automatic. They depend on the joint distribution of baseline outcomes and heterogeneous treatment effects. When predicted gains are concentrated among already advantaged individuals, mean-maximizing allocation can increase inequality in outcomes.

We proposed an inequality-averse modification of the outcome-based matching objective. The approach incorporates prioritarian preferences directly into the assignment step, giving greater marginal welfare weight to improvements for individuals with worse predicted outcomes. In the empirical application to active labor market programs in Switzerland, the modified objective shifts gains toward the lowest baseline quartile with moderate losses in mean impact.

The broader implication is that algorithmic inequality is not only a problem of prediction quality or optimization design. It is also a problem of program effectiveness. Better algorithms can redirect existing treatment gains, but they cannot create gains where programs are ineffective. For policy areas such as employment services, child welfare, homelessness support, or social assistance, reducing inequality therefore requires both distributionally aware allocation rules and interventions that produce meaningful gains for disadvantaged individuals.

References

Acharya, Avidit, Kirk Bansak and Jens Hainmueller. 2022. "Combining Outcome-Based and Preference-Based Matching: A Constrained Priority Mechanism." *Political Analysis* 30(1):89–112.

Allhutter, Doris and Astrid Mager. 2020. "AMS-Algorithmus könnte zu struktureller und sozialer Ungleichheit beitragen." *A&W Blog* . December 14, 2020.

Allhutter, Doris, Florian Cech, Fabian Fischer, Gabriel Grill and Astrid Mager. 2020. "Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective." *Frontiers in Big Data* 3:5.

Athey, Susan, Julie Tibshirani and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics* 47(2):1148–1178.

Atkinson, Anthony B. 1970. "On the Measurement of Inequality." *Journal of Economic Theory* 2(3):244–263.

Bansak, Kirk and Elisabeth Paulson. 2022. Outcome-Driven Dynamic Refugee Assignment with Allocation Balancing. In *Proceedings of the 23rd ACM Conference on Economics and Computation*. pp. 1182–1183.

Bansak, Kirk, Elisabeth Paulson, Dominik Rothenhäusler, Jeremy Ferwerda, Jens Hainmueller and Michael Hotard. 2026. "Robustness of Refugee-Matching Gains to Off-Policy Evaluation Choices." *arXiv preprint arXiv:2605.06686* .

Bansak, Kirk, Jeremy Ferwerda, Jens Hainmueller, Andrea Dillon, Dominik Hangartner, Duncan Lawrence and Jeremy Weinstein. 2018. "Improving Refugee Integration Through Data-Driven Algorithmic Assignment." *Science* 359(6373):325–329.

Bansak, Kirk and Linna Martén. 2021. "Algorithmic Decision-Making, Fairness, and the Distribution of Impact: Application to Refugee Matching in Sweden." *PolMeth 2021* .

Barry, Brian. 1965. *Political Argument*. Routledge.

Berger, Mark C., Dan Black and Jeffrey A. Smith. 2001. Evaluating Profiling as a Means of Allocating Government Services. In *Econometric Evaluation of Labour Market Policies*, ed. Michael Lechner and Friedhelm Pfeiffer. Physica, Heidelberg pp. 59–84.

Bitler, Marianne P, Jonah B Gelbach and Hilary W Hoynes. 2006. "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments." *American Economic Review* 96(4):988–1012.

Bonoli, Giuliano, Bea Cantillon and Wim Van Lancker. 2017. Social Investment and the Matthew effect. In *The Uses of Social Investment*, ed. Anton Hemerijck. Oxford University Press, Oxford pp. 66–76.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters: Double/Debiased Machine Learning." *The Econometrics Journal* 21(1):C1–C68.

Chohlas-Wood, Alex, Madison Coots, Sharad Goel and Julian Nyarko. 2023. "Designing Equitable Algorithms." *Nature Computational Science* 3(7):601–610.

Chouldechova, Alexandra. 2017. "Fair Prediction With Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *Big Data* 5(2):153–163.

Dalton, Hugh. 1920. "The Measurement of the Inequality of Incomes." *The Economic Journal* 30(119):348–361.

Danziger, Sheldon H. and Kent E. Portney. 1982. "The Distributional Impacts of Public Policies." *Policy Studies Journal* 10(4):623–638.

Desiere, Sam and Ludo Struyven. 2021. "Using Artificial Intelligence to Classify Jobseekers: The Accuracy-Equity Trade-off." *Journal of Social Policy* 50(2):367–385.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. pp. 214–226.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Ferwerda, Jeremy, Nicholas Adams-Cohen, Kirk Bansak, Jennifer Fei, Duncan Lawrence, Jeremy M Weinstein and Jens Hainmueller. 2020. "Leveraging the Power of Place: A Data-Driven Decision Helper to Improve the Location Decisions of Economic Immigrants." *arXiv preprint arXiv:2007.13902* .

Friedlander, Daniel and Philip K. Robins. 1997. "The Distributional Impacts of Social Programs." *Evaluation Review* 21(5):531–553.

Hardt, Moritz, Eric Price and Nati Srebro. 2016. "Equality of Opportunity in Supervised Learning." *Advances in Neural Information Processing Systems* 29.

Heckman, James J. 2020. "Epilogue: Randomization and Social Policy Evaluation Revisited." *Annual Review of Economics* 12(1):735–749.

Heckman, James J. and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9(2):85–110.

Heckman, James J, Jeffrey Smith and Nancy Clements. 1997. "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *The Review of Economic Studies* 64(4):487–535.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Kasy, Maximilian. 2024. "Algorithmic Bias and Racial Inequality: A Critical Review." *Oxford Review of Economic Policy* 40(3):530–546.

Keele, Luke. 2015. "The Statistics of Causal Inference: A View From Political Methodology." *Political Analysis* 23(3):313–335.

Kennedy, Edward H. 2023. "Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects." *Electronic Journal of Statistics* 17(2):3008–3049.

Kitagawa, Toru and Aleksey Tetenov. 2021. "Equality-Minded Treatment Choice." *Journal of Business & Economic Statistics* 39(2):561–574.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105(5):491–95.

Kleinberg, Jon, Sendhil Mullainathan and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science (ITCS)*.

Knaus, Michael C. 2022. "Double Machine Learning Based Program Evaluation Under Unconfoundedness." *The Econometrics Journal* .

Knaus, Michael C., Michael Lechner and Anthony Strittmatter. 2022. "Heterogeneous Employment Effects of Job Search Programs A Machine Learning Approach." *Journal of Human Resources* 57(2):597–636.

Kolm, Serge-Christophe. 1976. "Unequal Inequalities. II." *Journal of Economic Theory* 13(1):82–111.

Körtner, John and Giuliano Bonoli. 2023. Predictive Algorithms in the Delivery of Public Employment Services. In *Handbook of Labour Market Policy in Advanced Democracies*. Edward Elgar Publishing pp. 387–398.

Kuppler, Matthias, Christoph Kern, Ruben L. Bach and Frauke Kreuter. 2022. "From Fair Predictions to Just Decisions? COnceptualizing Algorithmic Fairness and Distributive Justice in the Context of Data-Driven Decision-Making." *Frontiers in Sociology* 7:883999.

Lalive, Rafael, Jan C. Van Ours and Josef Zweimüller. 2008. "The Impact of Active Labour Market Programmes on the Duration of Unemployment in Switzerland." *The Economic Journal* 118(525):235–257.

Le Grand, Julian. 1990. "Equity Versus Efficiency: The Elusive Trade-off." *Ethics* 100(3):554–568.

Le Grand, Julian. 2018. *The Strategy of Equality: Redistribution and the Social Services*. Routledge.

Lechner, Michael. 1999. "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification." *Journal of Business & Economic Statistics* 17(1):74–90.

Lechner, Michael, Michael C. Knaus, Martin Huber, Stefanie Behncke, Giovanni Mellace and Anthony Strittmatter. 2020. Swiss Active Labor Market Policy Evaluation. Dataset Distributed by FORS, Lausanne. https://doi.org/10.23662/FORS-DS-1203-1.

Manski, Charles F. 2004. "Statistical Treatment Rules for Heterogeneous Populations." *Econometrica* 72(4):1221–1246.

Manski, Charles F. 2009. *Identification for Prediction and Decision*. Harvard University Press.

Merton, Robert K. 1968. "The Matthew Effect in Science: The Reward and Communication Systems of Science Are Considered." *Science* 159(3810):56–63.

Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour and Kristian Lum. 2021. "Algorithmic Fairness: Choices, Assumptions, and Definitions." *Annual Review of Statistics and Its Application* 8:141–163.

Neyman, Jerzey. 1923. "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes." *Roczniki Nauk Rolniczych* 10(1):1–51.

Okun, Arthur M. 1975. *Equality and Efficiency: The Big Tradeoff*. Brookings Institution Press.

Parfit, Derek. 2000. Equality or Priority? In *The Ideal of Equality*, ed. Matthew Clayton and Andrew Williams. Houndmills: Palgrave pp. 81–125.

Patty, John W. and Elizabeth Maggie Penn. 2019. "Measuring Fairness, Inequality, and Big Data: Social Choice Since Arrow." *Annual Review of Political Science* 22:435–460.

Pigou, Arthur Cecil. 1912. *Wealth and Welfare*. Macmillan and Company, limited.

Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig and Sendhil Mullainathan. 2020. An Economic Perspective on Algorithmic Fairness. In *AEA Papers and Proceedings*. Vol. 110 pp. 91–95.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Robins, James M., Andrea Rotnitzky and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89(427):846–866.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies." *Journal of Educational Psychology* 66(5):688.

Samii, Cyrus, Laura Paler and Sarah Zukerman Daly. 2016. "Retrospective Causal Inference With Machine Learning Ensembles: An Application to Anti-Recidivism Policies in Colombia." *Political Analysis* 24(4):434–456.

Spiekermann, Sarah. 2019. "Ist der Glaube an zeitsparende AMS-Algorithmen naiv?" *Der Standard* . September 27, 2019.

Verma, Sahil and Julia Rubin. 2018. Fairness Definitions Explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE pp. 1–7.

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi and Cynthia Dwork. 2013. Learning Fair Representations. In *International Conference on Machine Learning*. PMLR pp. 325–333.

Appendix

A Solver

We solve the assignment problem using the commercial-grade Gurobi solver (Gurobi Optimization, LLC, 2022). For each value of ϵ, the solver maximizes the transformed predicted outcomes subject to binary assignment and capacity constraints. The resulting problem is a mixed-integer linear program (MILP), which Gurobi solves using branch-and-bound over linear-programming relaxations. Gurobi offers flexibility for our use case, including capacity constraints, binary assignments, and integration with Python and R.

Bansak et al. (2018) use the RELAX-IV minimum-cost flow solver, which allows for very fast optimization (Bertsekas, 1998). RELAX-IV would also be suitable for our assignment structure. We use Gurobi because it provides a more accessible implementation for experimentation with alternative objective functions. The choice of the solver affects implementation, not the assignment problem itself.

B Application

Data. We rely on widely used observational data for active labor market policies in Switzerland (Lechner et al., 2020). Table A.1 provides descriptive statistics for the variables used in the estimation. Further information and more detailed descriptive statistics can be found in Knaus (2022) and Knaus, Lechner and Strittmatter (2022).

Variable	CS	LS	VT	JS	NP
Age	39.08	35.28	37.45	37.31	36.61
Mother tongue in canton's language	0.11	0.04	0.11	0.12	0.10
Lives in big city	0.11	0.23	0.21	0.19	0.19
Lives in medium city	0.15	0.15	0.12	0.13	0.12
Lives in no city	0.73	0.63	0.67	0.68	0.68
Caseworker age	44.59	44.61	44.81	44.10	44.10
Caseworker cooperative	0.42	0.45	0.41	0.50	0.48
Caseworker education: above vocational training	0.48	0.48	0.44	0.45	0.45
Caseworker education: tertiary track	0.16	0.21	0.17	0.21	0.20
Caseworker female	0.44	0.47	0.39	0.47	0.44

Missing caseworker characteristics	0.05	0.05	0.04	0.05	0.05
Caseworker has own unemployment experience	0.61	0.63	0.64	0.63	0.62
Caseworker tenure	5.83	5.61	5.73	5.44	5.48
Caseworker education: vocational degree	0.25	0.22	0.22	0.27	0.26
Fraction of months employed last 2 years	0.84	0.72	0.83	0.84	0.81
Number of employment spells last 5 years	0.86	0.78	0.93	0.97	1.21
Employability	1.97	1.85	1.93	1.98	1.93
Female	0.60	0.55	0.33	0.44	0.44
Foreigner with temporary permit	0.04	0.44	0.12	0.11	0.13
Foreigner with permanent permit	0.17	0.23	0.18	0.22	0.23
Cantonal GDP p.c.	0.53	0.54	0.51	0.53	0.52
Married	0.45	0.72	0.48	0.46	0.47
Mother tongue other than German, French, Italian	0.18	0.64	0.31	0.29	0.33
Past income	43213	37301	48654	46693	42541
Previous job: manager	0.09	0.07	0.10	0.08	0.08
Missing sector	0.16	0.29	0.15	0.15	0.18
Previous job in primary sector	0.05	0.05	0.09	0.06	0.09
Previous job in secondary sector	0.13	0.12	0.15	0.14	0.12
Previous job in tertiary sector	0.67	0.54	0.61	0.65	0.61
Previous job: self-employed	0.00	0.00	0.00	0.00	0.01
Previous job: skilled worker	0.75	0.43	0.65	0.65	0.60
Previous job: unskilled worker	0.15	0.48	0.22	0.24	0.29
Qualification: semiskilled	0.14	0.15	0.17	0.14	0.16
Qualification: some degree	0.72	0.38	0.63	0.62	0.58
Qualification: unskilled	0.12	0.40	0.17	0.20	0.23
Qualification: skilled without degree	0.02	0.07	0.02	0.03	0.03
Swiss citizen	0.79	0.34	0.70	0.67	0.63
Allocation of unemployed to caseworkers: by industry	0.51	0.64	0.58	0.67	0.60
Allocation of unemployed to caseworkers: by occupation	0.45	0.57	0.46	0.57	0.51
Allocation of unemployed to caseworkers: by age	0.06	0.05	0.04	0.04	0.04
Allocation of unemployed to caseworkers: by employability	0.08	0.06	0.10	0.07	0.09
Allocation of unemployed to caseworkers: by region	0.13	0.11	0.09	0.09	0.13
Allocation of unemployed to caseworkers: other	0.10	0.09	0.08	0.07	0.09
Number of unemployment spells last 2 years	0.37	0.43	0.52	0.39	0.57
Cantonal unemployment rate (in %)	3.36	3.63	3.41	3.59	3.52

Table A.1: **Descriptive Statistics.** The table reports means of all variables used in the estimation of $Y(k)$, separately by observed assignment option: computer skills training (CS), language skills training (LS), vocational training (VT), job search assistance (JS), and no program (NP).

Estimation. In Figure A.1, we show overlap. Propensity scores, overlap, and outcome-model diagnostics for the same Swiss active labor market policy data are also discussed in detail by Knaus (2022). We follow this setup but use the doubly robust scores $\hat{\Gamma}_k$ to estimate debiased individualized average potential outcomes rather than individualized treatment effects.

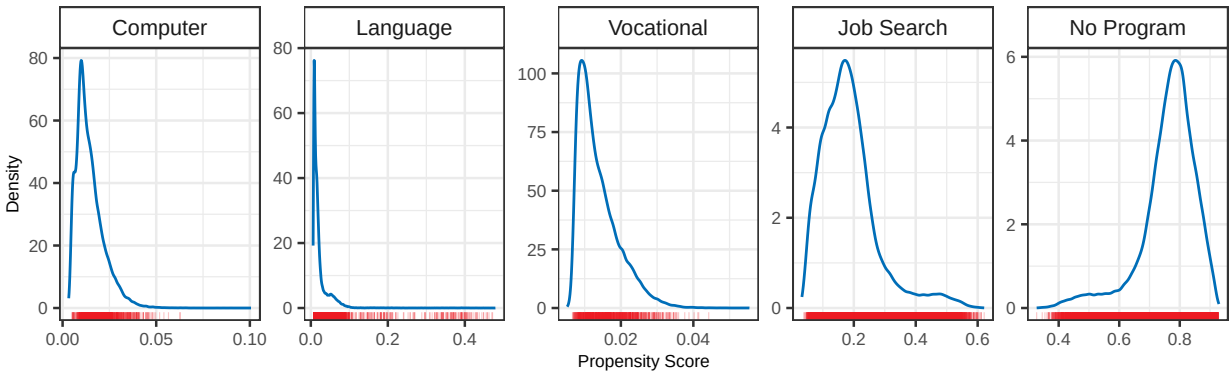


Figure A.1: **Overlap.** The figure shows estimated propensity scores for each assignment option. The blue density curve shows the distribution of $\hat{e}_k(X_i) = \Pr(D_i = k \mid X_i)$ across all individuals; the red rug marks show the estimated propensity scores for individuals actually observed in that option. Following Knaus (2022), small propensity scores are not interpreted as poor overlap by themselves because treatment-group sizes are highly imbalanced; the relevant question is whether observed participants have comparable individuals in the other assignment groups.

Assignment. Table A.2 reports assignment switches relative to standard outcome-based matching with $\epsilon = 0$. An assignment switch occurs when an individual is assigned to a different option under an inequality-averse objective than under the mean-maximizing objective. The table reports switches separately by quartiles of predicted no-program outcomes and overall.

ϵ	1st	2nd	3rd	4th	Overall
0.25	3.89	1.29	0.99	0.75	1.73
0.50	7.61	2.54	2.05	1.31	3.37
0.75	11.11	3.78	2.87	1.63	4.85
1.00	14.35	4.85	3.61	1.84	6.16

Table A.2: **Assignment Switches.** Entries report the percentage of individuals whose assigned option changes relative to standard outcome-based matching with $\epsilon = 0$. Columns report switches overall and by quartiles of predicted no-program outcomes. The first quartile contains individuals with the worst predicted baseline outcomes; the fourth quartile contains individuals with the best predicted baseline outcomes.

The results show that inequality aversion changes assignments primarily among individuals with worse baseline prospects. At $\epsilon = 0.25$, 3.89% of individuals in the lowest baseline quartile switch assignments, compared with 0.75% in the highest quartile. At $\epsilon = 1$, the

switching rate in the lowest quartile rises to 14.35%, while only 1.84% of individuals in the highest quartile switch. Overall switching remains moderate: even at $\epsilon = 1$, 6.16% of assignments differ from standard outcome-based matching.

C Simulation

We report a simulation that clarifies when inequality aversion is more or less costly in outcome-based matching. The simulation stays close to the empirical application while isolating one feature: the relationship between baseline outcomes and heterogeneous program gains. We keep the empirical distribution of predicted no-program outcomes and total active-program capacity fixed, but replace the four observed active labor market programs with a single simulated active program.

For each individual $j = 1, \ldots, n$, we set the no-program potential outcome equal to the estimated no-program potential outcome from the empirical application, $Y_j(0) = \hat{\mu}_{j0}$. The simulated program outcome is

$$Y_j(1) = \max\{Y_j(0) + \tau_j, 10^{-6}\},$$

where τ_j is the simulated individual-level gain. The lower bound ensures that all outcomes are strictly positive before applying the welfare transformation. To construct heterogeneous gains, let $p_j = (\text{rank}(Y_j(0)) - 1)/(n-1)$ denote the normalized rank of the empirical no-program outcome, with $p_j = 0$ corresponding to the lowest baseline outcome and $p_j = 1$ to the highest. Let $r_j = (\text{rank}(U_j) - 1)/(n-1)$, where $U_j \sim U(0,1)$, denote an independent random rank. This random rank introduces idiosyncratic treatment-effect heterogeneity

unrelated to baseline outcomes. We consider three scenarios:

$$s_j = \begin{cases} (1-\lambda)r_j + \lambda(1-p_j), & \text{HTE favor worse-off,} \\ r_j, & \text{HTE unrelated to baseline,} \\ (1-\lambda)r_j + \lambda p_j, & \text{HTE favor better-off.} \end{cases}$$

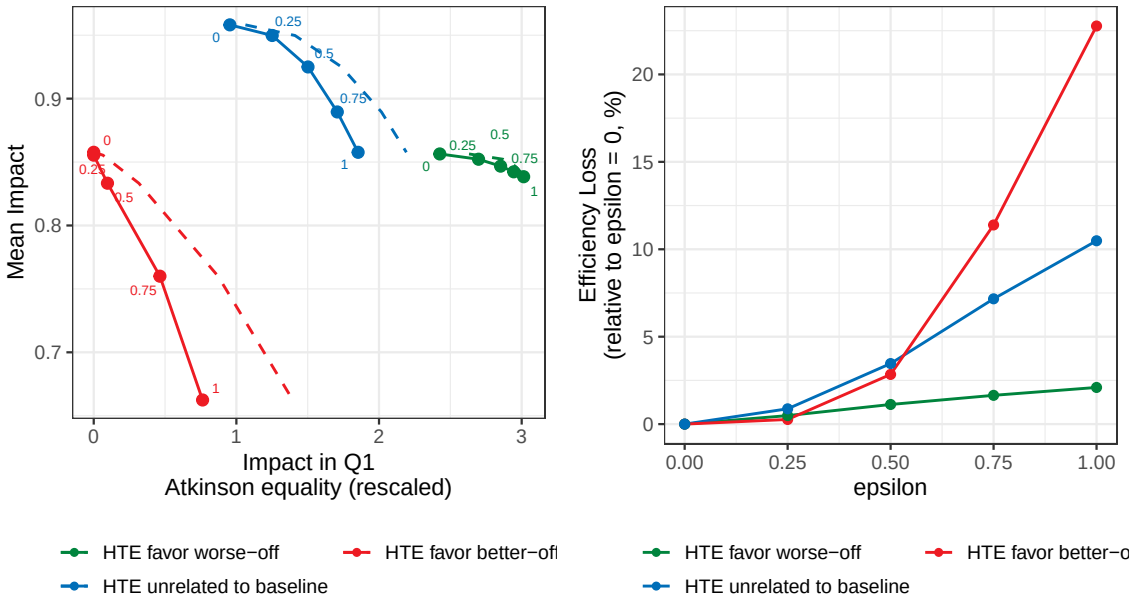
The simulated gain is $\tau_j = \tau_{\min} + \tau_{\text{range}} s_j$. The parameter $\lambda \in [0,1]$ controls how strongly gains align with baseline outcomes. When $\lambda = 0$, gains are unrelated to baseline outcomes. When $\lambda = 1$, gains are perfectly ordered by baseline rank in the worse-off and better-off scenarios. In the simulations reported here, we set $\lambda = 0.6$, so treatment effects are meaningfully but not perfectly aligned with baseline outcomes. For each scenario and each value of ϵ, we solve the same assignment problem as in the main text using two options: no program and the simulated active program. We then evaluate the resulting allocations using the same RIE-style impact measures as in the empirical analysis: mean impact, impact in the lowest baseline quartile, and efficiency loss relative to the $\epsilon = 0$ allocation.

Scenario	ϵ	Mean Impact	Q1 Impact	Loss	Loss (%)
HTE favor worse-off	0.00	0.86	2.43	0.00	0.0
	0.25	0.85	2.70	0.00	0.5
	0.50	0.85	2.85	0.01	1.1
	0.75	0.84	2.94	0.01	1.6
	1.00	0.84	3.01	0.02	2.1
HTE unrelated to baseline	0.00	0.96	0.95	0.00	0.0
	0.25	0.95	1.25	0.01	0.9
	0.50	0.92	1.50	0.03	3.5
	0.75	0.89	1.71	0.07	7.2
	1.00	0.86	1.85	0.10	10.5
HTE favor better-off	0.00	0.86	0.00	0.00	0.0
	0.25	0.86	0.00	0.00	0.3
	0.50	0.83	0.10	0.02	2.8
	0.75	0.76	0.46	0.10	11.4
	1.00	0.66	0.76	0.20	22.8

Table A.3: **Simulation Results.** The table reports impacts relative to no program under three simulated HTE structures. Mean Impact is the average gain; Q1 Impact is the gain for the lowest baseline quartile. Loss reports the reduction in Mean Impact relative to $\epsilon = 0$ within each scenario, in levels and percentages.

Table A.3 reports the simulation results. When heterogeneous treatment effects favor the worse-off, increasing ϵ raises impact in the lowest baseline quartile with little loss in mean impact. When treatment effects are unrelated to baseline outcomes, the trade-off is moderate. When treatment effects favor the better-off, the trade-off is steep: priority to the lowest quartile requires reallocating capacity away from individuals with larger simulated gains.

Figure A.2 visualizes the trade-off.



(a) Efficiency vs. Priority

(b) Efficiency Loss

Figure A.2: **Simulation Trade-Offs.** Panel (a) plots Mean Impact against Q1 Impact. Dashed lines show a reversed Atkinson index of assigned outcomes, computed with $\epsilon_A = 1$. Panel (b) plots efficiency loss relative to $\epsilon = 0$.

Panel (a) plots mean impact against priority to the worse-off. Solid lines use impact in the lowest baseline quartile; dashed lines show the reversed Atkinson index rescaled to the range of Q1 Impact, so that higher values correspond to lower inequality. Panel (b) reports efficiency loss relative to $\epsilon = 0$. The pattern mirrors Table A.3: inequality aversion is least costly when gains favor the worse-off and most costly when gains favor the better-off.

Figure A.3 shows the assignment mechanism. Under $\epsilon = 0$, the active program is assigned

to individuals with the largest simulated gains. When HTE favor the worse-off, this already concentrates assignment in the lowest baseline quartile. When HTE favor the better-off, the opposite occurs: standard outcome-based matching assigns the active program primarily to individuals in the highest baseline quartile. As ϵ increases, assignment shifts toward individuals with lower baseline outcomes in all scenarios. The shift is almost costless when gains favor the worse-off because the individuals prioritized by efficiency are also those prioritized by inequality aversion. In contrast, when HTE favor the better-off, the assignment pattern flips: inequality aversion moves capacity away from the high-baseline individuals with the largest gains and toward low-baseline individuals with smaller gains. This explains why the efficiency loss is largest in that scenario.

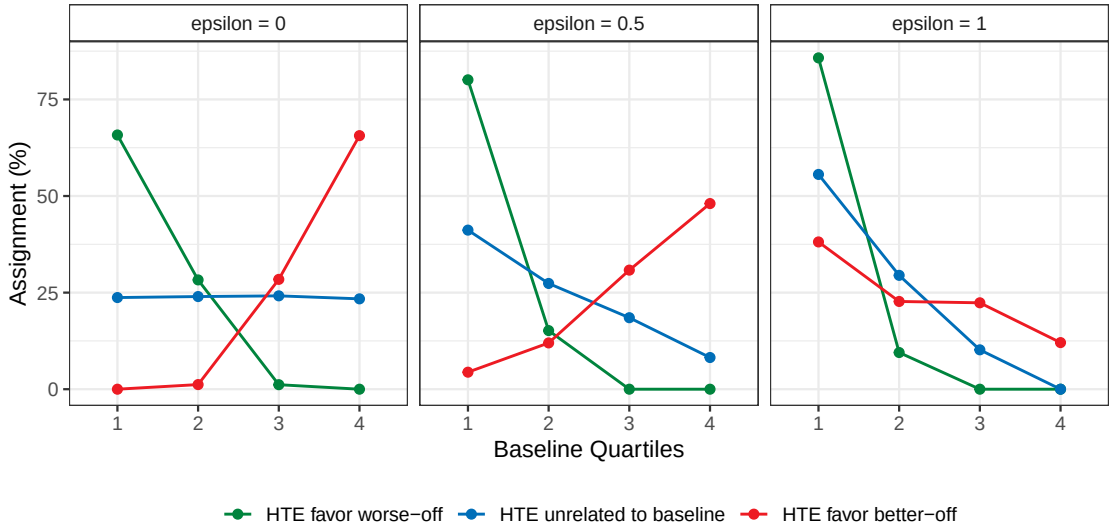


Figure A.3: **Assignment Distribution.** The figure shows assignment to the simulated active program by baseline outcome quartile. Under $\epsilon = 0$, the active program is assigned to individuals with the largest simulated gains. As ϵ increases, assignment shifts toward individuals with lower baseline outcomes.

References

Bansak, Kirk, Jeremy Ferwerda, Jens Hainmueller, Andrea Dillon, Dominik Hangartner, Duncan Lawrence and Jeremy Weinstein. 2018. "Improving Refugee Integration Through Data-Driven Algorithmic Assignment." *Science* 359(6373):325–329.

Bertsekas, Dimitri P. 1998. *Network Optimization: Continuous and Discrete Models*. Athena Scientific.

Gurobi Optimization, LLC. 2022. "*Gurobi Optimizer Reference Manual*.". https://www.gurobi.com.

Knaus, Michael C. 2022. "Double Machine Learning Based Program Evaluation Under Unconfoundedness." *The Econometrics Journal* .

Knaus, Michael C., Michael Lechner and Anthony Strittmatter. 2022. "Heterogeneous Employment Effects of Job Search Programs A Machine Learning Approach." *Journal of Human Resources* 57(2):597–636.

Lechner, Michael, Michael C. Knaus, Martin Huber, Stefanie Behncke, Giovanni Mellace and Anthony Strittmatter. 2020. Swiss Active Labor Market Policy Evaluation. Dataset Distributed by FORS, Lausanne. https://doi.org/10.23662/FORS-DS-1203-1.