Active Feature Acquisition for Personalised Treatment Assignment

Julianna Piskorz
University of Cambridge

Nicolás Astorga
University of Cambridge

Jeroen Berrevoets
University of Cambridge

Mihaela van der Schaar
University of Cambridge

Abstract

Making treatment effect estimation actionable for personalized decision-making requires overcoming the costs and delays of acquiring necessary features. While many machine learning models estimate Conditional Average Treatment Effects (CATE), they mostly assume that *all* relevant features are readily available at prediction time – a scenario that is rarely realistic. In practice, acquiring features, such as medical tests, can be both expensive and time-consuming, highlighting the need for strategies that select the most informative features for each individual, enhancing decision accuracy while controlling costs. Existing active feature acquisition (AFA) methods, developed for supervised learning, fail to address the unique challenges of CATE, such as confounding, overlap, and the structural similarities of potential outcomes under different treatments. To tackle these challenges, we propose specialised feature acquisition metrics and estimation strategies tailored to the CATE setting. We demonstrate the effectiveness of our methods through experiments on synthetic datasets designed to reflect common biases and data issues. In doing so, this work aims to bridge the gap between cutting-edge CATE estimation techniques and their practical, cost-efficient application in personalized treatment assignment.

1 INTRODUCTION

Personalised treatment assignment is crucial for optimising patient outcomes in high-stakes domains. Traditionally, randomised controlled trials (RCTs) have been the gold standard for developing treatment assignment guidelines. However, RCTs focus on estimating the *average treatment effect (ATE)*, overlooking the heterogeneity in individual responses. As a result, treatments deemed effective *on average* may not be beneficial for certain individuals.

To achieve truly personalised treatment assignment, it is necessary to shift the focus from ATE to the *conditional average treatment effect (CATE)*. Leveraging observational data is particularly beneficial in this context, as it reflects real-world patient variability and offers much more data than is typically available in RCTs. However, working with observational data presents its own challenges. To mitigate the confounding bias inherent in observational studies, models often incorporate a large number of potentially correlated covariates. While doing so can improve model validity, it also poses practical limitations.

Acquiring many covariates for every individual – e.g., through additional medical tests – is often unrealistic and costly. Each additional feature adds a model requirement, reducing the usability of the model in real-world settings where time and resources are limited. This leads us to pose the following question: Which variables are strictly *necessary* to accurately infer CATEs? A first realisation is that likely we do not require *every* variable to be present at inference time. A second realisation is that not every individual may require the *same* variables nor the same amount.

As such, we naturally arrive at the problem of *active feature acquisition* (AFA) for CATE inference. By selectively acquiring the most informative features for each individual, AFA can optimise treatment assignments with efficient use of available resources.

AFA in the context of CATE is challenging. For example, directly modelling causal quantities only on an arbitrary subset of the observed variables introduces confounding bias, contrasting the supervised learning setup. Moreover, since CATE is typically estimated as the difference between two potentially more complex functions, identifying the most relevant features is not straightforward.

In order to address these challenges, we present `AFA4CATE`: a unified framework for active feature

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

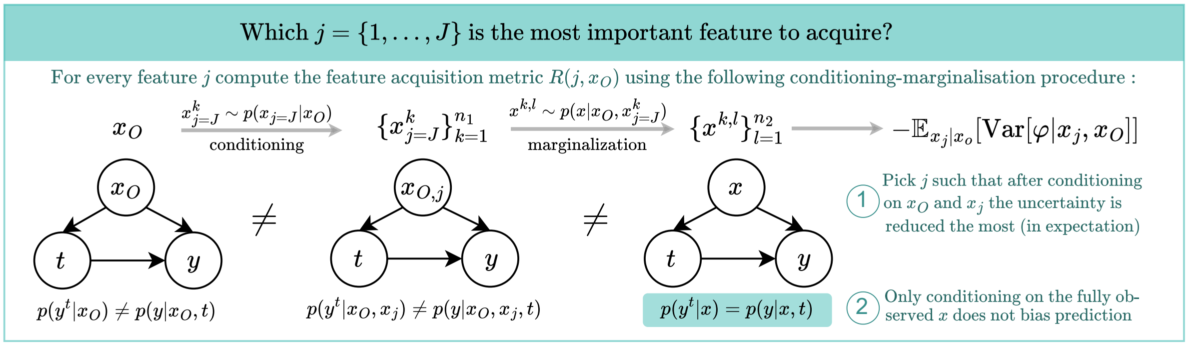

Figure 1: To avoid bias, estimation of the feature acquisition metrics in the context of CATE is possible only by *marginalising* the effect of the remaining unobserved variables.

acquisition in the context of CATE estimation. `AFA4CATE` comprises feature acquisition metrics as well as estimation strategies tailored to the CATE setting. To avoid confounding bias, our approach relies on principled marginalisation of the predicted causal quantities over the unobserved features (cf. Figure 1). We achieve this marginalisation efficiently by approximating the information gain, and relying on the law of total variance. With `AFA4CATE`, we also propose a set of metrics targeting the acquisition of different causal quantities and explain how the choice of the target variable can bias the acquisition towards predictive or prognostic variables, thus affecting the overall efficiency of the acquisition process.

We further evaluate the efficiency of `AFA4CATE` in the context of possible overlap violations, leading to high predictive uncertainty in CATE estimates. We note that in such cases, even acquiring all possible features does not yield reliable treatment recommendations. Thus, it is important to acquire features which allow the acquisition process to be stopped *early on*, minimising unnecessary costs. Such an approach advances the effectiveness and practicality of CATE-based decision-making in real-world scenarios.

Contributions. We introduce active feature acquisition for CATE estimation. Doing so results in the following contributions: 1 We identify a set of characteristics of the CATE estimation setting which make the existing AFA methods not directly applicable. 2 We propose a set of acquisition metrics and estimation strategies which respect the causal assumptions and allow for unbiased treatment effect estimation. 3 We empirically verify and compare the efficacy of the proposed metrics, providing guidelines for practical implementation of transparent AFA algorithms in the CATE setting. 4 We further evaluate `AFA4CATE` in overlap violating contexts, showing explicitly that careful choice of the acquisition metric can lead to cost savings.

1.1 Related Works

Active Feature Acquisition. The problem of active feature acquisition (sometimes referred to as *active sensing* (Yu et al., 2009; Alaa and van der Schaar, 2016; Yoon et al., 2018, 2019; Qin et al., 2023) or *active classification* (Greiner et al., 1996; Gao and Koller, 2011; Hollinger et al., 2017)) has garnered significant attention from the machine learning community, leading to various algorithms for supervised learning problems, adopting either a greedy acquisition strategy (Ma et al., 2019; Gong et al., 2019) or a sequential strategy using reinforcement learning (Shim et al., 2018; Li and Oliva, 2021; Li et al., 2021). However, to the best of our knowledge, no existing work addresses active feature acquisition in the context of CATE estimation. While active learning is already considered in CATE estimation (Jesson et al., 2021; Sundin et al., 2019; Qin et al., 2021), the challenges and objectives in active learning (aiming to improve model's generalisation performance by acquiring new labels) and active feature acquisition (aiming to improve model's predictive performance by acquiring new features) are fundamentally different, rendering active learning approaches not directly applicable to AFA in CATE.

Variable Selection for CATE. While standard variable selection has been considered before in the context of CATE estimation, its objectives and assumptions differ significantly from the active feature acquisition problem considered in this work. Standard variable selection methods (Heinze et al., 2018) aim to select a *global* subset of features $X' \subset X$ to fit the downstream outcome model: $\hat{Y} = f(X')$. When estimating CATE from observational data, this is challenging as one needs to ensure that the selected subset X' contains sufficient information to ensure that there is no hidden confounding between the treatment variable and the outcome (Greenewald et al., 2021; Shortreed and Ertefaie, 2017). In contrast, we propose fitting a single downstream CATE model using *all* available features, thereby avoiding hidden confounding. We then issue predictions for individuals with *different subsets of observed features* by sampling possible completions of all features before using the full model (cf. Section 3). Crucially, the aim of AFA4CATE is to *personalise* the feature acquisition, using each individual's already observed data to determine which additional features would maximally reduce the predictive uncertainty. This approach, which we consider complimentary to standard variable selection, should offer performance gains especially when the treatment effect depends on interaction terms between variables, as we further explain in Appendix A.

2 BACKGROUND AND PROBLEM SETTING

2.1 Conditional Average Treatment Effect (CATE) Estimation

Setup. In CATE estimation, we assume we have access to a fully observed training dataset $\mathcal{D} = \{(X^{(i)}, T^{(i)}, Y^{(i)})\}_{i=1}^{n}$, where $X \in \mathcal{X} \subseteq \mathbb{R}^d$ is the set of covariates, $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is the outcome, $T \in \{0,1\}$ is the binary treatment assignment. We will ignore the sample superscript unless explicitly needed. Adopting the potential outcome framework, we assume that for each individual there exist two *potential outcomes*, Y^1 and Y^0, corresponding to the outcome under treatment ($t = 1$) and under no treatment ($t = 0$), out of which we only observe the *factual* outcome, which (under the *consistency* assumption) corresponds to the assigned treatment: $Y = TY^1 + (1-T)Y^0$. Defining the individualised treatment effect (ITE) as $Y^1 - Y^0$, we note that this label of interest is never observed in the dataset and cannot be point identified. Instead, we can identify the *conditional average* treatment effect (CATE), defined as $\tau(x) = \mathbb{E}\left[Y^1 - Y^0 | X = x\right]$. This is possible under the standard causal assumptions of strong ignorability ($(Y^1, Y^0) \perp T|X$) and overlap ($0 < P(T = 1|X) < 1$), which allow to estimate the CATE as: $\tau(x) = \mu^1(x) - \mu^0(x)$, where $\mu^t(x) = \mathbb{E}\left[Y^t|X = x\right] = \mathbb{E}\left[Y|T = t, X = x\right]$.

Bayesian Inference for CATE. In this work, we consider parametric CATE estimators (Alaa and van der Schaar, 2017; Jesson et al., 2020, 2021), which assume that the outcomes are distributed according to a likelihood $p_{\omega_t}(y|x,t)$, specified using *all* the measured covariates $X = x$, the assigned treatment $T = t$ and a set of model parameters $\Omega_t = \omega_t$. For example, for binary outcomes we can assume that $Y^t|X = x \equiv Y|x,t \sim Ber(\mu^t_{\omega_t}(x))$ and for continuous outcomes we can assume that $Y^t|X = x \equiv Y|x,t \sim \mathcal{N}(\mu^t_{\omega_t}(x), \sigma^2_{\omega_t}(x))$. The mean functions $\mu^t_{\omega_t}(x)$ can be parametrised using, for example, neural networks (NNs), where we treat the weights as random variables and denote their prior distribution by $p(\omega_t)$, which is then updated to the posterior $p(\omega_t|\mathcal{D})$ after observing the training data $\mathcal{D}$. Importantly, this approach naturally incorporates a notion of model uncertainty around the predictions, and hence allows us to explore its role in the process of active feature acquisition.

With this model, for a new individual with $X = x$ and $T = t$ we can issue predictions by selecting a single value $\Omega_t = \omega^*_t$ of the parameters (e.g. the maximum a posteriori value) and then estimating the outcome using $\hat{\mu}^t(x) = \mu^t_{\omega^*_t}(x)$. Alternatively, one can obtain the estimate of the outcome by marginalising over the parameter values: $\hat{\mu}^t(x) = \mathbb{E}_{\omega_t \sim p(\omega_t|\mathcal{D})}\left[\mu^t_{\omega_t}(x)\right]$. In both cases a parametric estimate of the CATE is given by $\hat{\tau}(x) = \hat{\mu}^1(x) - \hat{\mu}^0(x)$.

We assume that the parametric CATE model is chosen and fitted *prior* to the implementation of the AFA procedure, using the fully-observed data. Then, with this downstream model fixed, we aim to develop an active feature acquisition algorithm which would allow to use the downstream model while acquiring *as few features as possible*. Thus, the questions of model selection in CATE estimation is orthogonal to the work at hand.

2.2 Active Feature Acquisition (AFA)

AFA algorithms aim to identify which unobserved features should be acquired to improve the accuracy of the model's prediction. For a new individual, instead of first collecting all the observations X, we would like to sequentially acquire features. That is, having observed $X_O = x_O$ (where $x_A := \{x_i : i \in A\}$) for some set of features[1] $O \subseteq [d]$, we would like to decide which additional feature X_j for $j \in [d] \backslash O$ we should measure to maximise our ability to accurately predict the outcome, Y. Importantly, we assume that the covariates X_i can be correlated with each other, indicating that what should be acquired next depends on the values of the features already acquired for each individual.

In this work, following (Ma et al., 2019; Gong et al., 2019), we approach the AFA problem using the principles of Bayesian Experimental Design (BED). Namely, we adopt the greedy approach to this problem, by proposing **acquisition metrics** indicating which feature should be acquired at the next time step. That is, our aim is to design functions $R : ([d]\backslash O) \times \mathcal{X}_O \to \mathbb{R}$, which can balance the utility of feature j with the cost of acquiring it C_j, and can then be used to select the

[1] We note that our framework also allows for $O = \varnothing$ initially.

'best' feature to acquire using:

$$j = \arg\max_j R(j, x_O) - C_j. \tag{1}$$

In what follows, we assume the cost of all features is constant ($X_j = C$), thus focusing only on maximising the acquisition metric $R(j, x_O)$.

Conditional Mutual Information Objective. The acquisition metric typically used in the supervised-learning setting is the *conditional mutual information* (CMI), I, between the proposed feature and some target, φ:

$$\begin{aligned} R(j, x_O) &= I(\varphi; X_j | x_O) && (2)\\ &= H(\varphi|x_O) - H(\varphi|X_j, x_O) && (3)\\ &= H(\varphi|x_O) - \mathbb{E}_{x_j|x_o}\left[H(\varphi|x_j, x_O)\right] && (4)\\ &\propto - \mathbb{E}_{x_j|x_o}\left[\mathbb{E}_{\varphi|x_j,x_O}\left[-\log(p(\varphi|x_j, x_O))\right]\right], && (5) \end{aligned}$$

where $H(A|b) = \mathbb{E}_{A|b}\left[-\log p(A|b)\right]$ is the Shannon's entropy and in supervised learning problems φ is the outcome variable, Y. CMI allows to balance the trade-off between *relevance* and *redundancy*. This can be seen by scrutinising the following decomposition:

$$\begin{aligned} &\arg\max_j I(\varphi; X_j|x_O)\\ &\qquad = \arg\max_j -H(\varphi, X_j|x_O) + H(X_j|x_O). \end{aligned}$$

Here, $H(\varphi, X_j|x_O)$ captures the joint uncertainty in the target variable φ and the candidate feature X_j, conditioned on x_O. Maximising the CMI prioritises the acquisition of features which, when observed, minimise this uncertainty, emphasising their *relevance*. In addition, $H(X_j|x_O)$ measures the uncertainty in X_j given the observed features x_O. If X_j is highly correlated with x_O, this term will be small, penalising *redundant* features.

Further, under an additional assumption of conditional independence of the features X, the CMI-based acquisition strategy can be shown to be approximately optimal, as we formalise in the following theorem, based on the work by Golovin and Krause (2017).

Theorem 2.1. *Assume that the features $X_1, \dots, X_d$ are conditionally independent given φ. Then, the greedy adaptive policy relying on maximising the conditional mutual information $I(\varphi; X_j|x_O)$ at each step starting from an empty set of features and ending after a fixed number of steps achieves $(1 - 1/e)$ approximation ratio with respect to the optimal policy.*

Proof. See Appendix B. □

While this result provides further theoretical justification for our approach, it is not a prerequisite – our method remains valid and well-motivated even in the absence of the conditional independence assumption.

Alternative Approaches to AFA. While reinforcement learning (RL) has also been proposed to address active feature acquisition (Li and Oliva, 2021; Shim et al., 2018; Li et al., 2021), we consider solutions relying on RL to be *secondary* to the question of design of feature acquisition metrics. That is, building RL solutions to the AFA problem relies on specifying acquisition metrics as one of the steps, making the design of acquisition metrics a more universal problem to study. Furthermore, as we are motivated by high-stakes domains, our goal is to design AFA algorithms which will be transparent and interpretable – a quality that many RL systems intrinsically lack.

What Makes AFA Difficult? Scrutinising the decomposition of the CMI objective, as shown in Equation (4), reveals that performing AFA requires estimates of $p(x_j|x_O)$ and $p(\varphi|x_j, x_O)$ *for every possible combination of the index j and set O*. Thus, the number of different probability functions that need to be estimated *grows exponentially* with the number of covariates. Finding efficient ways to model these probabilities has been the main focus of existing works, with the proposed solutions relying on partial variational autoencoders (Ma et al., 2019) or set encodings (Li and Oliva, 2021; Shim et al., 2018). Next, we explore how both the original acquisition metrics, as well as methods for modelling the distributions, need to be adjusted to be used in CATE.

3 DESIGNING ACQUISITION METRICS FOR CATE

3.1 How to Measure Information Gain?

The first step towards designing feature acquisition metrics for CATE estimation is to choose a suitable way of measuring information gain, satisfying the requirements imposed by CATE estimation.

Avoiding Confounding Bias. Existing feature acquisition metrics (Ma et al., 2019; Shim et al., 2018; Li and Oliva, 2021), designed for supervised learning problems, explicitly rely on *conditional mutual information* (Equation (2)). While a robust metric for measuring the level of dispersion (uncertainty) in a random variable, the challenge of using entropy in the context of CATE estimation lies in the need to estimate the probability $p(\varphi|x_j, x_O)$. In case of supervised learning problems, this probability can be learned directly from the training data using models which accept variable input size, by regressing φ only on a subset of the relevant variables $\{j\} \cup O$.

However, in case of CATE estimation this would no longer give an unbiased estimate of causal quantities such as Y^1, Y^0 or τ, due to the confounding effect of the remaining variables. As we have assumed that *all* covariates X are needed to respect the strong ignorability assumption, we have $p(y^t|x_j, x_O) \neq p(y|t, x_j, x_O)$ when $O \cup \{j\} \neq [d]$ (see Figure 1). To avoid bias, y^1 and y^0 have to be predicted using the *whole* vector of covariates x. For this reason, the existing AFA approaches do not directly apply to the CATE setting.

Hence, we propose to rely on **marginalisation**, by noticing that $p(\varphi|x_j, x_O)$ can be expressed as:

$$p(\varphi|x_j, x_O) = \mathbb{E}_{x|x_j,x_O}\left[p(\varphi|x)\right], \tag{6}$$

where x consists of x_j, x_O and the remaining covariates $x_{[d]\setminus(\{j\}\cup O)}$. This reformulation shows that obtaining the fully-observed unbiased model $p(\varphi|x)$, and then marginalising over the possible values of the unobserved covariates, allows AFA in the context of CATE.

Towards Efficient Estimation. While the marginalisation trick (Equation (6)) provides a principled way to use AFA in the context of CATE, computing the marginalisation can still be difficult and relies heavily on the parametric assumptions imposed on $p(\varphi|x_j, x_O)$. When φ is a categorical variable, the marginalisation can be computed exactly, by averaging class probabilities (Astorga et al., 2025). In case when φ is continuous, we propose to use the following *variance-based* approximation

$$I(\varphi; X_j|x_O) \approx -\mathbb{E}_{x_j|x_o}\left[\mathrm{Var}\left[\varphi|x_j, x_O\right]\right], \tag{7}$$

as a measure of information gain (this approximation is equivalent to the original formulation when $\varphi|x_j, x_O$ follows the normal distribution as we show in Appendix C; we also show empirically that this approach works well even when the distribution $p(\varphi|x_j, x_O)$ is not normal). The marginalisation is then easily achieved using the law of total variance:

$$\begin{aligned}&\mathrm{Var}\left[\varphi|x_j, x_O\right]\\ &\quad= \mathbb{E}_{x|x_j,x_O}\left[\mathrm{Var}\left[\varphi|x\right]\right] + \mathrm{Var}_{x|x_j,x_o}\left[\mathbb{E}\left[\varphi|x\right]\right],\end{aligned} \tag{8}$$

where x consists of x_j, x_O and the remaining covariates $x_{[d]\setminus(\{j\}\cup O)}$. Both of these terms are intuitively important when choosing the next feature to acquire: the first term ensures that upon acquiring the feature j, the variance in the prediction of φ will be small, while the second term ensures that acquiring feature j results in little variability in the prediction of φ (i.e., x_j successfully explained the heterogeneity of φ).

These insights lead to the following general form of the acquisition metric:

$$\begin{aligned}R(j, x_O) = &-\mathbb{E}_{x_j|x_O}\left[\mathbb{E}_{x|x_j,x_O}\left[\mathrm{Var}\left[\varphi|x\right]\right]\right]\\ &- \mathbb{E}_{x_j|x_O}\left[\mathrm{Var}_{x|x_j,x_o}\left[\mathbb{E}\left[\varphi|x\right]\right]\right].\end{aligned} \tag{9}$$

3.2 How to Estimate the Acquisition Metrics?

With the above decomposition, the acquisition metric $R(j, x_O)$ can be estimated using Monte Carlo sampling, as we describe in Algorithm 1. We use $\mu_l\{\cdot\}$ to denote the mean of the set calculated over index l (and analogically σ_l^2 denotes the variance). We note that this approach requires generating $n_1 \times n_2 \times n_3$ samples of the target variable φ for each unobserved feature j, which can be a computationally heavy process, particularly when the dimensionality of the covariate space is large, as then we require n_2 to be large to efficiently marginalise out the effect of the unobserved variables. See Appendix D for detailed time complexity analysis.

Algorithm 1 Estimating the acquisition metrics.

Input: observed variables x_O, candidate variable j, number of samples n_1, n_2, n_3

```
Condition: x_j^1, ..., x_j^{n_1} ~ p(x_j | x_O)
for k = 1 to n_1 do
    Marginalise: x^{k,1}, ..., x^{k,n_2} ~ p(x | x_j^k, x_O)
    for l = 1 to n_2 do
        Predict: φ^{k,l,1}, ..., φ^{k,l,n_3} ~ p(φ | x^{k,l})
        Calculate μ̂^{k,l} = μ_i{φ^{k,l,i}}
        Calculate (σ̂²)^{k,l} = σ_i²{φ^{k,l,i}}
    end for
    Calculate V̂_1^k = μ_l{(σ̂²)^{k,l}} and V̂_2^k = σ_l²{μ̂^{k,l}}
end for
```

Output: $\hat{R}(j, x_O) = -\mu_k\{\hat{V}_1^k\} - \mu_k\{\hat{V}_2^k\}$

We need only two models for the necessary samples:
① **Predictive model:** $\varphi \sim p(\varphi|x)$. This corresponds to the downstream CATE estimation model, as described in Section 2.1. Depending on the chosen acquisition metric, we might want to use $\hat{\mu}^t(x)$, $\hat{\tau}(x) = \hat{\mu}^1(x) - \hat{\mu}^0(x)$ or $\hat{S\tau}(x) = \mathbb{1}\{\hat{\tau}(x) > 0\}$ (cf. Section 3.3).

② **Generative surrogate model (GSM):** $x_j \sim p(x_j|x_O)$ and $x \sim p(x|x_j, x_O)$. This model is needed to generate samples from the conditional distributions of the covariates, allowing the marginalisation of the effect of x_j. As these distributions do not involve any causal quantities, standard AFA methods can be used to generate the necessary samples (Shim et al., 2018; Ivanov et al., 2018; Ma et al., 2019).

3.3 How to Choose the Target Variable?

Equipped with variance as an approximate measure of information gain, as well as an efficient estimation strategy which gives unbiased results in the CATE setting, we now turn to instantiating acquisition metrics $R(j, x_O)$. In particular, we propose metrics relying on different choices of the target variable φ. Since there is no prior research on AFA in the context of CATE, and existing studies on AFA are not directly applicable to the CATE setting, we conduct qualitative and quantitative research study to identify the most useful target variables.

As the CATE is defined as a difference of two potential outcome functions, $\tau(x) = \mu^1(x) - \mu^0(x)$, the most naive approach we can consider is to simultaneously maximise the information gain about *both* the potential outcome functions through the following metric:

$$R_{\mathrm{PO}}(j, x_O) := -\mathbb{E}_{x_j|x_O}\left[\mathrm{Var}\left[Y^1|x_j, x_O\right] + \mathrm{Var}\left[Y^0|x_j, x_O\right]\right].$$

This naive acquisition metric relies on the simple observation that if we reduce the uncertainty about both of the potential outcome functions, then we would also reduce the uncertainty about the treatment effect: $\mathrm{Var}\left[\tau\right] \leqslant \mathrm{Var}\left[Y^0\right] + \mathrm{Var}\left[Y^1\right]$.

However, this metric does not take into account the fundamental characteristic of CATE estimation: that $\tau(x)$ is a *difference* between the potential outcomes and as a result, it might be a *simpler* function than each of $\mu^t(x)$ separately (Curth and van der Schaar, 2021). In context of active feature acquisition, this implicit assumption about the reduced complexity of $\tau(x)$ in comparison to $\mu^t(x)$ gives rise to the distinction between *prognostic* and *predictive* variables.

Prognostic variables are predictive of the outcome regardless of the assigned treatment (and hence they express information which are shared between $\mu^1(x)$ and $\mu^0(x)$), while the predictive variables (effect modifiers) determine how a patient will react to different treatments (Hahn et al., 2020). Consider the following example:

$$Y^t = a_0 + a_1 X_1 + a_2 X_2 + t(a_3 X_3 - a_4 X_2). \tag{10}$$

Here, X_1 is a prognostic variable, X_3 is a predictive variable, while X_2 plays both a prognostic and predictive role. Assuming that for the purpose of decision making we are only interested in the treatment effect (and not the potential outcomes explicitly), significant performance gains in active feature acquisition for CATE can be obtained by designing acquisition metrics which prioritise the acquisition of *predictive* factors, as the prognostic factors are not explicitly necessary for the estimation of the treatment effects.

With these insights, we propose a metric explicitly optimising the acquisition of predictive factors:

$$\begin{aligned} R_\tau(j, x_O) &:= -\mathbb{E}_{x_j|x_O}\left[\mathrm{Var}\left[\tau|x_j, x_O\right]\right] \\ &= -\mathbb{E}_{x_j|x_O}\left[\mathrm{Var}\left[Y^1 - Y^0|x_j, x_O\right]\right] \\ &= -\mathbb{E}_{x_j|x_O}\Big[\mathrm{Var}\left[Y^1|x_j, x_O\right] + \mathrm{Var}\left[Y^0|x_j, x_O\right] \\ &\qquad - 2\mathrm{Cov}[Y^1, Y^0|x_j, x_o]\Big]. \end{aligned}$$

The last equality explains that the benefits of using R_τ stem from the fact that it directly adjusts for the possible correlation between the potential outcome functions. However, the extent to which choosing $\varphi = \tau$ can positively affect performance, by minimising acquisition costs, will depend on the downstream model's ability to correctly differentiate between predictive and prognostic factors. Thus, we expect that performance benefits can be obtained in particular when the downstream CATE model is equipped with inductive biases which encourage information sharing between $\mu^1(x)$ and $\mu^0(x)$ (Shi et al., 2019; Curth and van der Schaar, 2021), thus encouraging the distinction between the predictive and prognostic factors.

Following Sundin et al. (2019), we further note that if the treatment assignment decision is to be made solely based on the sign of the estimated treatment effect (e.g. assign $T = 1$ if and only if $\hat{\tau} > 0$) then it might be beneficial to change the target variable from continuous $\varphi = \tau$ to binary $\varphi = \mathbb{1}\{\tau > 0\}$. This could allow to more precisely select the features that need to be acquired, as well as obtain more targeted estimates of the uncertainty upon observing $X_O = x_O$. This leads to:

$$R_{S\tau}(j, x_O) := -\mathbb{E}_{x_j|x_O}\left[\mathrm{Var}\left[\mathbb{1}\{\tau > 0\}|x_j, x_O\right]\right].$$

As this acquisition metric focuses only on a subset of the relevant information involved in CATE estimation, its long-term performance in more complex problems might be sub-optimal.

4 NUMERICAL VERIFICATION

We verify the efficacy of the proposed active feature acquisition metrics and the proposed estimation approach using numerical experiments on a synthetically generated dataset.[2] The synthetic setup allows us to manipulate key properties of the data generating process so that we can efficiently compare and contrast the performance of the proposed acquisition metrics, as well as to access counterfactual outcomes for robust evaluation.

4.1 Setup

Datasets. *ACIC2016-based:* To evaluate the metrics' ability to identify the predictive rather than prognostic variables, we use the synthetic dataset proposed by Curth and van der Schaar (2021) (setup 'A'). In the original paper the dataset is constructed using the covariates from the ACIC2016 benchmark dataset (Dorie et al., 2017). In this work, to efficiently build a ground-truth GSM, we propose a modified approach: we calculate the mean and the covariance matrix of the continuous covariates from the ACIC2016 dataset, and

[2] Code to reproduce the experiments can be found at: https://github.com/j-piskorz/afa4cate.

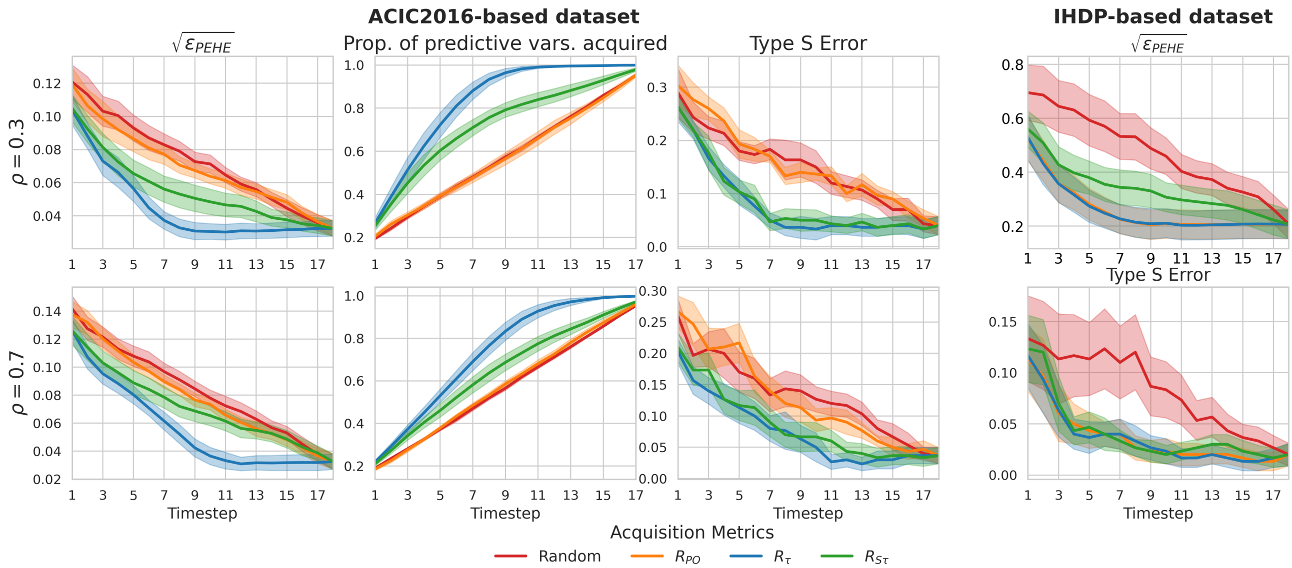

Figure 2: Comparing different acquisition metrics. Shaded areas mark 90% confidence intervals, over 10 seeds.

then sample our synthetic covariates from a multivariate normal distribution with this mean and covariance. With synthetic covariates obtained that way, we can then easily build a perfect GSM by using the conditional distribution of jointly normal variables. This procedure results in 21 covariates, which we simulate for 6300, 2700 and 1000 individuals in the train, validation and test sets respectively (we increase the sample size compared to the original dataset to obtain a good downstream CATE model). We obtain the potential outcomes using the formula described in Curth and van der Schaar (2021). Importantly, the parameter ρ allows to manipulate the proportion of predictive variables in the dataset: setting $\rho = 0.0$ corresponds to the lack of predictive variables (treatment effect equal to 0), while $\rho = 1.0$ corresponds to a setup where all variables are predictive. We further introduce confounding into the dataset, by designing a treatment assignment function based on a subset of the covariates. *IHDP-based:* In the ACIC2016-based dataset the coefficients take binary values $\{0, 1\}$, and hence the importance of features depends mostly on their relative magnitude. To evaluate the performance of the acquisition metrics in the setting where the importance of features is more heterogeneous, we also use a variant of the IHDP dataset (Hill and Su, 2013). We use the potential outcomes as proposed in the original dataset, but modify the covariates in the same way as in the ACIC2016-based dataset, to obtain ground-truth GSMs. We simulate the same number of samples as in the original dataset, to allow for the downstream model to have large predictive uncertainty in certain regions of the covariate space. In case of both datasets, we standardise the observed outcome (as well as the potential outcomes) before the subsequent model training and analysis. Details of both datasets can be found in Appendix F.3.

Models. *GSM:* As explained above, to efficiently compare the proposed acquisition metrics (without needing to take into account the potential bias of the GSM) we design a ground-truth GSM by relying on the properties of jointly normal variables. *CATE model:* Following Jesson et al. (2021), in our experiments we rely on the DUE model (van Amersfoort et al., 2022), which is an instance of Deep Kernel Gaussian Process. We describe the model parameters and training routine in the Appendix F.

4.2 Performance Improvement Over Time

Evaluation Routine. We evaluate the different acquisition metrics by comparing the performance of the downstream CATE model achieved by following the acquisition recommendations issued by each metric for a subgroup of 30 individuals from the test set. For each of the individuals, we start by observing m random covariates[3] (where $m = 3$ for the ACIC-based dataset and $m = 7$ for the IHDP-based dataset, resulting in 18 unobserved features in both cases). We then compute the value of the acquisition metric for each of the remaining unobserved features j following Algorithm 1 with $n_1 = 20$, $n_2 = 200$, $n_3 = 50$, and acquire the one with the highest score. After acquiring the variable, we evaluate the performance by sampling the remaining covariates from $p(x|x_j, x_O)$ and computing the average value of the CATE: $\mathbb{E}_{x|x_j,x_O}[\mathbb{E}[\hat{\tau}(x)]]$. We repeat this

[3]By randomly choosing m initially observed covariates we simulate real-world scenarios where baseline tests or pre-measured variables are available in the patient's records before further feature acquisition begins. However, we note that our framework also allows for $O = \varnothing$.

process until all the variables are acquired. For each metric, the final score is obtained by averaging the results obtained for all the 30 individuals. We repeat this process over 10 seeds, with each seed corresponding to a different realisation of the synthetic dataset.

Metrics. Across both datasets, at each acquisition timestep we measure both the *PEHE*, calculated as $\sqrt{\epsilon_{PEHE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}_i - \tau_i)^2}$, and the *type S error* (Sundin et al., 2019), which is the decision making error calculated as $\gamma = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\text{sign}(\hat{\tau}) \neq \text{sign}(\tau)\}$.

Results. The results of our evaluation are presented in Figure 2. For the ACIC2016-based dataset, we compare the performance of the three considered acquisition metrics (as well as the 'random acquisition' baseline) across two settings, with $\rho = 0.3$ and $\rho = 0.7$. Firstly, we note that the difference in performance across the metrics is much more pronounced when the number of predictive factors is small ($\rho = 0.3$). As predicted, R_τ consistently outperforms the other metrics, both in terms of PEHE and the type S error. As explained by the middle plots in Figure 2, this performance improvement can be attributed to the faster acquisition of the relevant predictive factors. Interestingly, while $R_{S\tau}$ still outperforms R_{PO}, we note that it does not allow to improve the decision making performance relative to R_τ (as it was aiming to do). Finally, the narrow gap between these targeted metrics and the random acquisition is related to the small variability in the relative importance of different features. In contrast, in the IHDP-based dataset, where this variability is much higher, all the proposed metrics clearly outperform the random baseline. Furthermore, we note that in the IHDP-based dataset, which does not differentiate between the predictive and prognostic factors, there is no significant difference in the performance of the R_{PO} and R_τ metrics.

4.3 Performance in the Presence of Possible Overlap Violations

Accurate CATE estimation relies on the overlap assumption, which ensures that for every individual, there is a non-zero probability of receiving each treatment. When overlap is violated, certain regions of the covariate space lack representation across treatment groups, leading to high predictive uncertainty and unreliable treatment effect estimates (Jesson et al., 2020; D'Amour et al., 2021). Traditional AFA methods typically focus on reducing uncertainty stemming from unobserved features, to maximise the predictive accuracy (Li and Oliva, 2021; Shim et al., 2018; Ma et al., 2019). However, they overlook the predictive uncertainty arising from overlap violations, which cannot be mitigated by additional feature acquisition.

In this work, we adopt a holistic view by considering both the uncertainty due to unobserved features and the predictive uncertainty stemming from overlap violations. We assume that decision-makers will only act on CATE estimates if the *total* uncertainty is below a threshold α. In regions where overlap is violated and predictive uncertainty is high, even acquiring all possible features does not yield reliable treatment recommendations. Therefore, it is crucial to acquire variables which determine whether overlap is violated or not *early on*, to avoid unnecessary acquisition costs. To achieve that, we consider an early-stopping criterion: if for at least 95% of the possible completions $x \sim p(x|x_j, x_O)$ the predictive uncertainty is above α, we stop the acquisition and decide to withhold treatment recommendation. In our instantiation of this approach, we quantify both the model and the total uncertainty using variance. The decision making process following these principles is described in Algorithm 2 (where $Q^{0.95}$ denotes the 95% quantile computed over the index l).

Algorithm 2 Decision making process.

Input: observed features x_O, threshold α

Sample $x^1, \ldots, x^{n_2} \sim p(x|x_O)$.
For each x^l, sample $\varphi^{l,1}, \ldots, \varphi^{l,n_3} \sim p(\varphi|x^l)$.
Calculate $\hat{\mu}^l = \mu_i\{\varphi^{l,i}\}$ and $(\hat{\sigma}^2)^l = \sigma_i^2\{\varphi^{l,i}\}$
Calculate $\hat{V}_1 = \mu_l\{(\hat{\sigma}^2)^l\}$ and $\hat{V}_2 = \sigma_l^2\{\hat{\mu}^l\}$
if $\hat{V}_1 + \hat{V}_2 < \alpha$ **then**
    Predict τ.
else if $Q_l^{0.95}\{(\hat{\sigma}^2)^l\} > \alpha$ **then**
    Withhold prediction.
else
    Acquire new feature.
end if

This formulation makes it clear that a good AFA metric should allow to acquire features that help detect overlap violations early in the decision-making process, to decrease costs. Here, we further evaluate the proposed metrics using this criterion.

Evaluation Routine. We perform our evaluation on the IHDP-based dataset, as it is designed to have regions where the overlap assumption does not hold (in Appendix G we consider also other settings, with more significant overlap violations). We adopt the same acquisition procedure as in the previous section, now introducing an early stopping criterion as described in Algorithm 2. Our early stopping criterion is instantiated using $\varphi = \tau$ across all the metrics (the results with other choices are included in Appendix G). Furthermore, we select the threshold value α by specifying the quantile of the model variance calculated for the observations in the training set (see Appendix E for detailed definition).

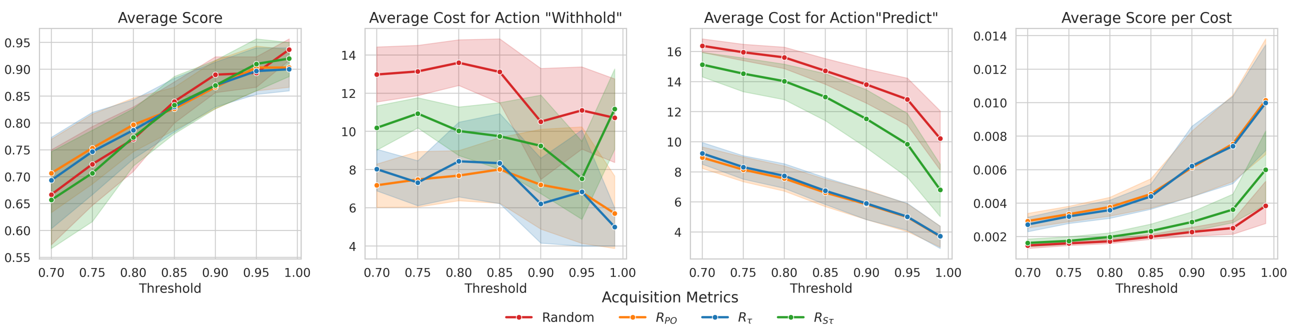

Figure 3: The performance of the different acquisition metrics when we explicitly allow to withhold treatment recommendations in case of large uncertainty. The error bars mark 90% confidence intervals, over 10 seeds.

Metrics. We evaluate the performance of the acquisition metrics using the measures of 'score' and 'cost'. For each individual, the score is assigned as: '0' when the recommendation is withheld, '1' when correct treatment is assigned, '-1' when incorrect treatment is assigned. This way, withholding recommendation does not explicitly decreases the score, but instead does not allow it to increase. We further compute the cost as the total number of features acquired before deciding on the final action (i.e. we assume that the cost is constant across the features).

Results. We present the results in Figure 3, averaged across 10 seeds. We see that as the accepted threshold α is kept fixed across the runs, all metrics allow to reach a similar final score, as expected. However, some metrics (R_τ and R_{PO}) allow to significantly decrease the overall acquisition costs needed to reach that total score, decreasing by approximately a half the number of acquisition steps needed to reach a decision, both for the individuals who in the end were and were not issued recommendations. As can be seen in the right-most plot in Figure 3, this allows to significantly improve the average score per cost, across all thresholds. Finally, we note that performance of the metric $R_{S\tau}$ starts to improve for larger values of the threshold α, when acquiring high-level information about the treatment effect, such as those optimised for by $R_{S\tau}$, is sufficient to reach the final action.

5 DISCUSSION AND FUTURE WORK

Limitations. Our method is computationally intensive (see Appendix D for time complexity analysis). While marginalization is necessary to condition the model on fully observed features, its computational burden could be reduced by employing more advanced approaches to selecting a potentially important subsets of features. For instance, employing large language models (LLMs) could allow effective sampling of the most relevant features by leveraging their extensive prior knowledge (Astorga et al., 2025). Additionally, advanced acquisition methods, such as those introduced by Kobalczyk et al. (2025), could proactively identify the most informative features or experimental queries, further improving efficiency. With such approaches in place, `AFA4CATE` could be scaled to higher-dimensional problems.

Future work. Although in this introductory work we have focused only on binary treatments, extending our framework to handle **multiple different treatments** is a natural next step. In such contexts, where there is no longer a unique treatment effect, the designed acquisition metrics need to have a multi-objective character. Coming up with solutions which can efficiently identify the predictive rather than prognostic factors might require more complex solutions.

Further, in this work we focused on acquiring only a single feature at each time step. However, allowing for the selection of multiple related features in batches (e.g. CT + MRI) could enhance the applicability of our work to real-world settings like healthcare. Considering **batch acquisition** methods designed for active learning, such as Batch-BALD (Kirsch et al., 2019), can allow to equip `AFA4CATE` with this added functionality.

Lastly, although in this work we base our feature acquisition procedure on the principles of Bayesian Experimental Design, by developing acquisition algorithms which greedily maximise the proposed acquisition metrics, alternative paradigms may be worth exploring in future work. **Reinforcement learning** (RL) approaches to active feature acquisition (e.g. Shim et al., 2018; Li and Oliva, 2021; Li et al., 2021) have shown promise in supervised learning, but adapting them to CATE estimation would require new reward functions or proxy losses to account for the fact that in the CATE estimation setting the true target, τ, is never directly observed. We believe the concepts and analysis presented here can guide those future developments.

Acknowledgements

We thank Cem Tekin, anonymous reviewers as well as members of the vanderschaar-lab for many insightful comments and suggestions. JP gratefully acknowledges funding from AstraZeneca. NA thanks W.D. Armstrong Trust for sponsorship and support. This work was supported by Azure sponsorship credits granted by Microsoft's AI for Good Research Lab.

References

Ahmed M. Alaa and Mihaela van der Schaar. Balancing Suspense and Surprise: Timely Decision Making with Endogenous Information Acquisition. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Nicolás Astorga, Tennison Liu, Nabeel Seedat, and Mihaela van der Schaar. Active Learning with LLMs for Partially Observed and Cost-Aware Scenarios. *Advances in Neural Information Processing Systems*, 37:20819–20857, January 2025.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, USA, 1991.

Alicia Curth and Mihaela van der Schaar. On Inductive Biases for Heterogeneous Treatment Effect Estimation. In *Advances in Neural Information Processing Systems*, volume 34, pages 15883–15894. Curran Associates, Inc., 2021.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, April 2021. arXiv: 1711.02582 Publisher: North-Holland.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34 (1):43–68, July 2017. arXiv: 1707.02641 Publisher: Institute of Mathematical Statistics.

Tianshi Gao and Daphne Koller. Active Classification based on Value of Classifier. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Daniel Golovin and Andreas Krause. Adaptive Submodularity: Theory and Applications in Active Learning and Stochastic Optimization, December 2017. arXiv:1003.3967.

Wenbo Gong, Sebastian Tschiatschek, Sebastian Nowozin, Richard E Turner, José Miguel Hernández-Lobato, and Cheng Zhang. Icebreaker: Element-wise Efficient Information Acquisition with a Bayesian Deep Latent Gaussian Model. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Kristjan Greenewald, Karthikeyan Shanmugam, and Dmitriy Katz. High-Dimensional Feature Selection for Sample Efficient Treatment Effect Estimation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 2224–2232. PMLR, March 2021. ISSN: 2640-3498.

Russell Greiner, Adam J. Grove, and Dan Roth. Learning active classifiers. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ICML'96, pages 207–215, San Francisco, CA, USA, July 1996. Morgan Kaufmann Publishers Inc.

P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, 15(3):965–1056, September 2020. Publisher: International Society for Bayesian Analysis.

Georg Heinze, Christine Wallisch, and Daniela Dunkler. Variable selection – A review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449, 2018.

Jennifer Hill and Yu-Sung Su. Assessing Lack of Common Support in Causal Inference Using Bayesian Nonparametrics: Implications for Evaluating the Effect of Breastfeeding on Children's Cognitive Outcomes. *The Annals of Applied Statistics*, 7(3):1386–1420, 2013. Publisher: Institute of Mathematical Statistics.

Geoffrey A. Hollinger, Urbashi Mitra, and Gaurav S. Sukhatme. Active Classification: Theory and Application to Underwater Inspection. In Henrik I. Christensen and Oussama Khatib, editors, *Robotics Research : The 15th International Symposium ISRR*, pages 95–110. Springer International Publishing, Cham, 2017.

Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational Autoencoder with Arbitrary Conditioning. September 2018.

Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models. *Advances in Neural Information Processing Systems*, 2020-December, July 2020. arXiv: 2007.00163 Publisher: Neural information processing systems foundation ISBN: 2007.00163v2.

Andrew Jesson, Panagiotis Tigas, Joost van Amersfoort, Andreas Kirsch, Uri Shalit, and Yarin Gal.

Causal-BALD: Deep Bayesian Active Learning of Outcomes to Infer Treatment-Effects from Observational Data. In *Advances in Neural Information Processing Systems*, volume 34, pages 30465–30478. Curran Associates, Inc., 2021.

Katarzyna Kobalczyk, Nicolas Astorga, Tennison Liu, and Mihaela van der Schaar. Active Task Disambiguation with LLMs, February 2025. arXiv:2502.04485 [cs].

Yang Li and Junier Oliva. Active Feature Acquisition with Generative Surrogate Models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6450–6459. PMLR, July 2021. ISSN: 2640-3498.

Yang Li, Siyuan Shan, Qin Liu, and Junier B. Oliva. Towards Robust Active Feature Acquisition, July 2021. arXiv:2107.04163 [cs].

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE, May 2019. arXiv:1809.11142 [cs, stat].

Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. Budgeted Heterogeneous Treatment Effect Estimation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8693–8702. PMLR, July 2021. ISSN: 2640-3498.

Yuchao Qin, Mihaela van der Schaar, and Changhee Lee. Risk-Averse Active Sensing for Timely Outcome Prediction under Cost Pressure. *Advances in Neural Information Processing Systems*, 36:6397–6411, December 2023.

Claudia Shi, David M. Blei, and Victor Veitch. Adapting Neural Networks for the Estimation of Treatment Effects, October 2019. arXiv:1906.02120 [cs, stat].

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint Active Feature Acquisition and Classification with Variable-Size Set Encoding. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Susan M. Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122, December 2017.

Iiris Sundin, Peter Schulam, Eero Siivola, Aki Vehtari, Suchi Saria, and Samuel Kaski. Active Learning for Decision-Making from Imbalanced Observational Data. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6046–6055. PMLR, May 2019. ISSN: 2640-3498.

Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. On Feature Collapse and Deep Kernel Learning for Single Forward Pass Uncertainty, March 2022. arXiv:2102.11409 [cs, stat].

Jinsung Yoon, William R. Zame, and Mihaela van der Schaar. Deep Sensing: Active Sensing using Multi-directional Recurrent Neural Networks. February 2018.

Jinsung Yoon, James Jordon, and Mihaela Schaar. ASAC: Active Sensing using Actor-Critic models. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, pages 451–473. PMLR, October 2019. ISSN: 2640-3498.

Shipeng Yu, Balaji Krishnapuram, Romer Rosales, and R. Bharat Rao. Active Sensing. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pages 639–646. PMLR, April 2009. ISSN: 1938-7228.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes:** see Section 2 for the description of the setting and assumptions, and section Section 3.2 for the algorithm used.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes:** see section Section 3.2 and Appendix D]
   (c) (Optional) Anonymised source code, with specification of all dependencies, including external libraries. [Code is available at: `https://github.com/j-piskorz/afa4cate`.]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes:** see Theorem 2.1 for the assumptions.]
   (b) Complete proofs of all theoretical results. [**Yes:** see Appendix B for the proof of Theorem 2.1.]
   (c) Clear explanations of any assumptions. [**Yes:** see Appendix B.]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**. The code to reproduce the experimental results can be found at: `https://github.com/j-piskorz/afa4cate`.]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**. See Appendix.]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**. See Appendix.]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).[**Yes**. See Appendix.]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**. See references in section Section 4.1]
   (b) The license information of the assets, if applicable. [**Yes**. See Appendix.]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A Comparison to Variable Selection Methods

As we explain in more details below, the advantages of using our AFA4CATE framework for personalised active feature acquisition are twofold. Firstly, our proposed approach ensures that all *potential confounder*s are included in the downstream CATE estimation model, addressing the risk of hidden confounding. Secondly, our method dynamically selects features to acquire for each individual based on their *already observed data*, thus making the acquisition process more efficient.

Ensuring that the 'no hidden confounding' assumption is satisfied. The aim of standard variable selection methods (Heinze et al., 2018) is to choose out of the available features X only a subset X' which should be used to fit the downstream outcome model. Namely, upon the selection of the subset X', the outcome Y will be predicted only using this subset X': $\hat{Y} = f(X')$. In the context of CATE estimated from observational data, however, the selection of such a subset X' is a particularly intricate problem, as one needs to ensure that the selected variables X' form an admissible set: namely, they contain sufficient information to ensure that there is no hidden confounding between the treatment variable T and the outcome Y (Greenewald et al., 2021; Shortreed and Ertefaie, 2017). Choosing too small of a set of variables X' might lead to bias in the CATE estimates.

In contrast, in our work we propose to fit a single downstream CATE model $\hat{\tau}(X)$ using *all the available features*, thus reducing the risk of biased CATE estimates. We then issue predictions for individuals with different subsets of observed features X_O by sampling possible completions of all features X (as we explained in detail in section Section 3.1).

Personalised Feature Acquisition. Another key advantage of our AFA4CATE framework is its ability to personalise feature acquisition based on the information already observed for each individual. Standard variable selection methods create a global non-personalised ranking of the importance of different features which is then used across *all* individuals. Meanwhile, AFA4CATE personalises this process, dynamically obtaining a ranking of features tailored to each individual and their already observed features. This personalization is particularly beneficial in scenarios where the treatment effect function depends on **interaction terms** between variables.

Example of Interaction Effects: Consider a treatment effect function that includes an interaction between X_1 and X_2: $\tau(X) = \mathbb{1}\{X_1 > 0\} \times X_2$. In this case the relevance of X_2 to the treatment effect is conditional on the value of X_1. If $X_1 \leqslant 0$, X_2 does not influence $\tau(X)$ and acquiring X_2 would not improve the estimation of the treatment effect for that individual.

Standard variable selection methods, which output a global non-personalised ranking of the importance of different features, would not be able to recognise that the importance of the variable X_2 is conditional on the value of the variable X_1. In contrast, our method uses the conditional mutual information $I(\varphi; X_j|x_O)$ to guide the acquisition process. This metric quantifies the expected reduction in uncertainty about φ upon acquiring X_j, *given the values of the already observed features* $X_O = x_O$. Thus, in the example above, the method would recognise that acquiring X_2 is only valuable when $x_1 > 0$. Thus, by personalising feature acquisition, we avoid unnecessary costs associated with acquiring features which are irrelevant for certain individuals, leading to a more cost-efficient estimation process.

Standard Variable Selection and AFA4CATE Offer Complementary Solutions. This aspect of personalisation of the feature acquisition process makes our AFA4CATE framework *complementary* to standard variable selection methods. Indeed, in the context of CATE, standard variable selection methods can be used to select an *admissible* subset of available features X' which can be used to fit the downstream CATE model. But even within this selective subset, not all variables might be equally important for all individuals (as we have demonstrated with our example of interaction effects). Thus, AFA4CATE could be used on top of a downstream CATE model $\tau(X')$, to make the acquisition process personalised and thus more cost-efficient.

B Submodular Analysis of the Greedy Acquisition Policy

As we have formalised in Theorem 2.1, under certain assumptions on the distribution of $(\varphi, X_1, \ldots, X_d)$, our adaptive policy relying on the greedy maximisation of the conditional mutual information $I(\varphi; X_j|x_O) = \mathbb{E}_{x_j|x_O}\left[H(\varphi|x_o) - H(\varphi|x_j, x_O)\right]$ can be shown to be $(1 - 1/e)$ optimal, by relying on the results derived by Golovin and Krause (2017).

Notation. We first introduce the necessary notation, bridging our work with the work of Golovin and Krause (2017). Firstly, let us assume that $\forall i \; X_i \in \mathcal{X}$. Let $\phi(i)$ be the *realisation* of the variable i, with $\Phi(i)$ denoting the random realisation. That is, we say that $X_i = \Phi(i)$ and $x_i = \phi(i)$. Further, let ψ denote a *partial realisation*, which we define as a set of item-observations pairs (i, x) for i in a subset of $[d]$ which we call the domain of ψ: $dom(\psi) = \{i \in [d] : \exists x \in \mathcal{X} : (i, x) \in \psi\}$. Additionally, we say that $\psi(i) = x$ if $(i, x) \in \psi$. Then, ψ is said to be consistent with ϕ (denoted $\phi \sim \psi$) if $\forall i \in dom(\psi), \; \psi(i) = \phi(i)$.

Proof of Theorem 2.1 With this notation in place, we now note that our objective of maximising the conditional mutual information $I(\varphi; X_j | x_O)$ is equivalent to maximising the conditional expected marginal benefit:

$$\Delta(i, \psi) = \mathbb{E}\left[f(dom(\psi) \cup \{i\}, \Phi) - f(dom(\psi), \Phi) | \Phi \sim \psi\right]$$

with the utility function $f(A, \phi)$ is defined as $-H(\varphi | \{X_i = \phi(i) : i \in A\})$.

As we demonstrate below, under certain conditions on the distribution of $(\varphi, X_1, \ldots, X_d)$ the function $f(A, \phi)$ is adaptive monotone and adaptive submodular. Let us then consider the greedy adaptive policy relying on maximising the conditional mutual information at each step, starting from an empty set of features and ending when we reach k features. The fact that the considered utility function is adaptive monotone and adaptive submodular allows us to apply the results derived by Golovin and Krause (2017) (Theorem 5.2) to conclude that *our proposed greedy policy is a* $(1 - 1/e)$*-approximation to the optimal policy in terms of expected utility.*

$f(A, \phi)$ is Adaptive Submodular and Adaptive Monotone. To see that $f(A, \phi)$ is adaptive monotone, we note that $\Delta(i, \psi) = I(\varphi; X_j | x_O) \geqslant 0$ by the properties of mutual information. To prove that $f(A, \phi)$ is adaptive submodular we need to show that $\Delta(i, \psi) \geqslant \Delta(i, \psi')$ for every $\psi \subseteq \psi$. This is equivalent to showing that $I(\varphi; X_j | x_O) \geqslant I(\varphi; X_j | x_{O'})$, where $O \subseteq O'$. To prove this, we assume that the features $X_1, \ldots, X_d$ are conditionally independent given φ. Then, we note that

$$I(\varphi; X_j | x_O) = H(X_j | x_O) - H(X_j | \varphi, x_O) = H(X_j | x_O) - H(X_j | \varphi)$$

by the conditional independence of X_j and x_O given φ, and hence

$$I(\varphi; X_j | x_O) - I(\varphi; X_j | x_{O'}) = H(X_j | x_O) - H(X_j | x_{O'}) \geqslant 0$$

by the "information never hurts" principle (Cover and Thomas, 1991). Thus, the function $f(A, \phi)$ is adaptive monotone and adaptive submodular, as needed.

Discussion. We note that in order to prove the adaptive submodularity of $f(A, \phi)$ we relied on the assumption that the features $X_1, \ldots, X_d$ are conditionally independent given φ. In real-world situations this assumption might not always be satisfied.

C Variance Approximation for the Information Gain

In here, we provide a justification for why using the variance-based approximation of the information gain is equivalent to the original formulation for normal variables (Section 3.1), in the context of active feature acquisition.

Assume that $\varphi | x_j, x_O \sim \mathcal{N}(\mathbb{E}[\varphi | x_j, x_O], \mathrm{Var}[\varphi | x_j, x_O])$. Then by the properties of the normal distribution we know that $H(\varphi | x_j, x_O) = \frac{1}{2} \log 2\pi \mathrm{Var}[\varphi | x_j, x_O] + \frac{1}{2}$. This gives the following result:

$$\begin{aligned}
\arg\max_j \; I(\varphi, X_j | x_O) &= \arg\max_j \; H(\varphi | x_O) - H(\varphi | X_j, x_O) \\
&= \arg\max_j \; -\mathbb{E}_{x_j | x_O}\left[H(\varphi | x_j, x_O)\right] \\
&= \arg\max_j \; -\mathbb{E}_{x_j | x_O}\left[\frac{1}{2} \log 2\pi \mathrm{Var}[\varphi | x_j, x_O] + \frac{1}{2}\right] \\
&= \arg\max_j \; -\mathbb{E}_{x_j | x_O}\left[\mathrm{Var}[\varphi | x_j, x_O]\right]
\end{aligned}$$

where the last line follows as $f(x) = \frac{1}{2} \log 2\pi x + \frac{1}{2}$ is a monotone transformation.

D Time Complexity Analysis of Algorithm 1

Below, we present the complexity analysis of Algorithm 1 presented in our work.

Sampling Complexity The time complexity of our method depends on the time complexity of the generative surrogate model used. In our experiments we sample the covariates X from a jointly normal distribution, and then use the oracle conditional normal distributions to sample the features. This leads to the following complexity:

1. **Sampling from $p(x_j|x_o)$.** Computing the conditional mean and covariance for $x_j \sim p(x_j|x_o)$ in a multivariate normal distribution involves matrix operations scaling as $O(|x_O|^3)$. Then, generating n_1 samples of x_j scales as $O(n_1 \cdot |x_O|)$. Thus, the total cost for all $j \in [d]\backslash O$ is $O(|x_O|^3 + n_1 \cdot |x_O| \cdot (d - |x_O|))$.

2. **Sampling from $p(x|x_o, x_j)$.** Sampling n_2 samples from $p(x|x_o, x_j)$ involves similar matrix operations and hence scales as $O((d - |x_O|) \cdot n_1 \cdot ((|x_O| + 1)^3 + n_2 \cdot |x_O|)$.

Prediction Complexity The time complexity of our method is further dependent on the time complexity of the downstream CATE model used to obtain the treatment effect estimate. In our experiments we use a sparse Deep Kernel Gaussian Process with M inducing points. The cost of predicting n_3 samples scales as $O(n_3(M^2 + MN))$, where N is size of the training dataset. Thus the total cost of prediction is $O((d - |x_O|) \cdot n_1 \cdot n_2 \cdot n_3(M^2 + MN))$

Repeating the Process for Multiple Acquisition Steps Since the algorithm iterates over multiple acquisition steps, where $|x_O|$ starts small initially (e.g., $|x_O| = s$ where in our experiments $s = 3$), and grows incrementally to $d - 1$ (if no stopping criterion is used), the total time complexity is the sum of the costs across all steps:

$$O\left(\sum_{|x_O|=s}^{d-1} (d - |x_O|)\left(|x_O|^3 + n_1 \cdot |x_O| + n_1(|x_O| + 1)^3 + n_1 \cdot n_2 \cdot |x_O| + n_1 \cdot n_2 \cdot n_3(M^2 + MN)\right)\right).$$

After simplifying, the total time complexity is:

$$O\left(n_1(d - s)s^3 + n_1 n_2(d - s)s + n_1 n_2 n_3(d - s)^2(M^2 + MN)\right)$$

The Effect of s Increasing s (i.e. starting with a large number of features already observed) significantly reduces the time complexity because:

1. The number of acquisition steps $d - s$ decreases, which has a multiplicative effect on all terms.

2. The significant prediction cost $(O\left((d - s)^2 n_1 n_2 n_3(M^2 + MN)\right))$, which dominates the time complexity particularly when n_1, n_2 and n_3 are large, is reduced as we increase s.

Thus, while the per-acquisition-step cost increases as we increase s (because s^3 is larger), this increase is outweighed by the reduction in the number of steps.

D.1 Empirical Results

Below we validate the conclusions of our theoretical time complexity analysis by presenting the average time needed to complete the full acquisition process for a single individual within our implementation, when using $n_1 = 20$, $n_2 = 200$ and $n_3 = 50$ on the ACIC2016-based dataset with $\rho = 0.3$ which has $d = 21$. We validate how changing the number of initially observed variables, s, affects the time taken to complete the process. We run the experiment over 10 seeds, each time running the acquisition process for 10 randomly chosen individuals. We use the R_τ acquisition metric.

As predicted by our analysis, increasing s (which is equivalent to decreasing the number of features that need to be acquired) allows to significantly decrease the time needed to run the acquisition for a single individual.

E Details of the Early Stopping Criterion

In the numerical experiments in Section 4.3 we proposed introducing an early stopping criterion into the decision making process (see Algorithm 2). This early-stopping criterion allows to stop acquisition when a sufficient level of confidence in the predictions has been reached, to save costs. It depends on a threshold value α, where larger values of α indicate that more uncertainty (variance) is allowed in the final estimates provided by the CATE model.

s	time
0	34.73
2	28.73
4	23.06
6	18.1
8	13.64
10	9.99
12	6.84
14	4.31
16	2.35
18	0.99

Table 1: Results of the time complexity analysis indicate that increasing the number of features observed initially for each individual, s, significantly decreases the computational time.

In our numerical experiments we chose the value of α based on the value of predictive variance calculated for the individuals in the training set, when issuing CATE predictions using the fully-observed data. That is, assuming that the training set consists of individuals $x^1, \dots, x^n$, we calculate α as follows:

1. For each $l = 1, \dots, n$ sample $\varphi^{l,1}, \dots, \varphi^{l,m} \sim p(\varphi|x^l)$.
2. For each $l = 1, \dots, n$ calculate $(\hat{\sigma}^2)^l = \sigma_i^2\{\varphi^{l,i}\}$.
3. Set $\alpha = Q_l^q\{(\hat{\sigma}^2)^l\}$.

Here, by Q_l^q we denote the q^{th} quantile of the given set of values, indexed by l. In our experiments, we vary the value of q from 0.7 to 0.99 (we avoid setting $q = 1.0$ to avoid including potential outlier-values of the predictive variance). Changing the value of q determines what is the proportion of the individuals in the training set for which the model is informed enough to issue a recommendation.

In practical applications, we suggest tuning the value of α with the help of domain experts, to quantify the allowed level of uncertainty. Further guidance can be provided by exploring the trade-off between costs and potential errors in treatment assignment induced by different values of α (e.g. by calculating the value of average score per cost).

F Details of the Numerical Verification

F.1 Code

Code to reproduce the experiments can be found at `https://github.com/j-piskorz/afa4cate`.

F.2 Assets and Licensing Information

The following existing assets were used to produce the experimental results:

- *causal-bald* python library (Jesson et al., 2021): Apache 2.0 License (which includes an implementation of the IHDP dataset)
- *CATENets* python library (Curth and van der Schaar, 2021): BSD 3-Clause License (which includes an implementation of the ACIC2016 dataset)

F.3 ACIC2016-based Dataset.

Covariates. We simulate the covariates based on the distribution of the continuous covariates in the ACIC2016 datasets Dorie et al. (2017). Namely, we calculate the mean μ_{ACIC} and the variance σ^2_{ACIC} of the whole ACIC2016 dataset, and then simulate 6300, 2700 and 1000 samples of the covariates for the training, validation and test set respectively as samples from the distribution $\mathcal{N}(\mu_{ACIC}, \Sigma_{ACIC})$. This results in 21 continuous covariates.

Treatment Assignment. To introduce confounding into the dataset, we compute the propensity score for each individual in the following way:

$$q^l = \sum_{i=1}^{d} \beta_i X_i^l$$
$$p^l = \text{sigmoid}(\frac{q^l}{r/2} \times 2.2)$$

where $\beta_i \sim \text{Ber}(0.4)$, r is the range q^l ($r = \max_l q^l - \min_l q^l$) and $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$. We scale by the factor of 2.2 to make sure that the probabilities fall between 0.1 and 0.9, to avoid the violation of the overlap assumption.

Potential Outcomes. We simulate the potential outcomes following the setup introduced by Curth and van der Schaar (2021) (setup A). Namely, we simulate the potential outcome functions as follows:

$$Y^t = c + \sum_{j=1}^{d} \beta_j X_j + \sum_{j=1}^{d} \sum_{l=1}^{d} \beta_{j,l} X_j X_l + \sum_{j=1}^{d} \gamma_j X_j,$$

where $c = 1.0$, $\beta_j \sim Ber(0.6)$, $\beta_{j,l} \sim Ber(0.3)$ and $\gamma_j \sim Ber(\rho)$. We then set $Y = TY^1 + (1-T)Y^0 + \epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$. As we explain Section 4, the value of the parameter ρ quantifies the strength of the *predictive factors* (and hence also the heterogeneity of the treatment effect).

Train-Test Split and Standardisation. We generate 6300, 2700 and 1000 points for the train, validation and test set respectively. This is larger than in the original ACIC2016 dataset as we want to ensure that the downstream CATE model is well-trained. After generating the potential outcomes, the observed outcomes, the potential outcomes and the treatment effects are all standardised using the mean and standard deviation calculated using the values of the observed outcomes in the training set. The values remain standardised throughout the remaining analysis.

Test Subset Selection. For both datasets, we randomly select a group of 30 individuals from the generated test set to conduct the feature acquisition procedure for. For a given seed, the choice of individuals which were added to the test subset remains fixed across acquisition runs performed with different acquisition metrics.

Observed Covariates Selection. For each individual in the test subset, we randomly choose $m = 3$ features which are initially observed, resulting in 18 unobserved covariates. For a given seed, the choice of features which were observed for each individual remains fixed across acquisition runs performed with different acquisition metrics.

F.4 IHDP-based Dataset.

We note that the IHDP benchmark dataset for causal inference (Hill and Su, 2013) is created by subsampling the results of a randomised clinical trial targeting premature infants with low birth weight with an intervention. Namely, the dataset is made 'observational' by excluding a non-random proportion of the treated individuals (those with nonwhite mothers). We consider two alternative formulations of the IHDP-based dataset, differing in how the treatments and covariates are generated.

Jointly Regenerating Covariates and Treatments. We calculate the mean μ_{IHDP2} and the covariance matrix Σ_{IHDP2} of all the continuous and binary covariates, the indicator of whether or not the mother is white ('momwhite' variable) as well as the treatment variable (27 variables in total). We then simulate the substitute variables by generating samples from the multivariate normal distribution $\mathcal{N}(\mu_{IHDP2}, \Sigma_{IHDP2})$. We draw as many samples as there were in the original IHDP dataset (before it was subsampled to exclude treated non-white mothers). We then binarise the treatment variable by assigning treatment ($t = 1.0$) to those individuals whose value of the generated (continuous) treatment variable is above 0.5 and assign $t = 0.0$ otherwise. We do not binarise any of the other originally binary variables, as this resulted in poor performance in exploratory analysis. After doing that, we obtain the final dataset by following the same procedure as in the original dataset: we exclude from the dataset all those *treated* individuals who had non-white mothers. The final dataset results in 25 covariates (because we do not include the 'momwhite' variable in the dataset).

Hyperparameter	Search Space
kernel	[RBF, Matern]
ν (Matern kernel)	[0.5, 1.5, 2.5]
no. inducing points	[100, 200, 500]
hidden units	[128, 256, 512]
network depth	[2, 3, 4]
negative slope	[-1.0, 0.0, 0.1, 0.2]
dropout rate	[0.05, 0.1, 0.2, 0.5]
spectral norm	[0.0, 0.95, 1.5, 3.0]
learning rate	[1e-4, 1e-3, 1e-2, 1e-1]
batch size	[64, 128, 256]

Table 2: Search space for the hyperparameter tuning of DUE.

Potential Outcomes. In both cases, we obtain the potential outcomes in the following way.

$$Y^0 = \exp\left(\sum_{j=1}^{d} \beta_j (X_j + 0.5)\right) + \epsilon$$

$$Y^1 = \sum_{j=1}^{d} \beta_j (X_j + 0.5) + \epsilon$$

where β_j are randomly chosen to have values $[0.0, 0.1, 0.2, 0.3, 0.4]$ with probabilities $[0.6, 0.1, 0.1, 0.1, 0.1]$ and $\epsilon \sim \mathcal{N}(0, 1)$.

Train-Test Split and Standardisation. We generate the dataset of the size which is equal to the original IHDP dataset. This is to ensure that there are areas of the covariate space (related to low overlap) where the model has high predictive uncertainty. The dataset is then split randomly into train, validation and test sets which relatively take 63%, 27% and 10% of the original dataset. After generating the potential outcomes, the observed outcomes, the potential outcomes and the treatment effects are all standardised using the mean and standard deviation calculated using the values of the observed outcomes in the training set. The values remain standardised throughout the remaining analysis.

Test Subset Selection. For both datasets, we randomly select a group of 30 individuals from the generated test set to conduct the feature acquisition procedure for. For a given seed, the choice of individuals which were added to the test subset remains fixed across acquisition runs performed with different acquisition metrics.

Observed Covariates Selection. For each individual in the test subset, we randomly choose $m = 7$ features which are initially observed, resulting in 18 unobserved covariates. For a given seed, the choice of features which were observed for each individual remains fixed across acquisition runs performed with different acquisition metrics.

F.5 CATE Model

In our numerical experiments we have relied on the DUE model, as implemented by Jesson et al. (2021) in their open-source python library. For each dataset (and for each combination of the hyperparameters used to instantiate the ACIC2016-based dataset) we have run hyperparameter tuning using `optuna` search algorithm. We run the hyperparameter tuning over the following search space:

As the evaluation of the performance of the downstream CATE model was not the main objective of this work, we have limited the hyperparameter tuning time to 10 minutes. We run the hyperparameter tuning for each seed separately. After tuning the hyperparameters, we train the model for 300 epochs, with patience of 50 epochs. We train the model once for each seed, then save the model and reload it when running the acquisition for each of the metrics.

F.6 GSM

As we simulate our covariates as samples from the jointly normal distribution, we use the properties of the normal distribution to instantiate the generative surrogate model. Namely, we have that $X \sim \mathcal{N}(\mu, \Sigma)$ where

$\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$. After observing $X_O = x_O$, we then obtain the samples from $p(x_j|x_O)$ as samples from the multivariate normal distribution with the following mean and variance:

$$\mathbb{E}\left[X_j|X_O = x_O\right] = \mu_j + \Sigma_{jO}\Sigma_{OO}^{-1}(x_O - \mu_O)$$
$$\mathrm{Var}\left[X_j|X_O = x_O\right] = \Sigma_{jj} - \Sigma_{jO}\Sigma_{OO}^{-1}\Sigma_{Oj}$$

F.7 Computing Infrastructure Used

Hardware. All experiments were conducted using an NC24rs_v3 instance on the Microsoft Azure cloud platform. This instance is part of Azure's GPU-optimised virtual machine series and has the following hardware specifications:

1. CPU: Intel Xeon E5-2690 v4 (Broadwell) with 24 cores
2. GPUs: 4 NVIDIA Tesla V100 GPUs, each with 16 GB VRAM
3. Memory: 448 GB RAM
4. CUDA Version: 11.3
5. Operating System: Ubuntu 20.04 LTS

Once for each dataset and each seed the CATE model was tuned and trained, conducting the complete feature acquisition procedure for 30 individuals using a single acquisition metric took no longer than 30 minutes.

G Additional Numerical Results

G.1 Performance in the Presence of Possible Overlap Violations: Different Early Stopping Criteria

Here we present the results of the numerical experiments from Section 4.3, where we instantiate the decision making process presented in Algorithm 2 with alternative threshold variables φ. While the results in the main manuscript were conducted using τ as the threshold variable, in here we also compare the performance of different metrics when the signed treatment effect $S\tau$ and the potential outcomes Y^1 and Y^0 are used for thresholding.

The procedure for using $S\tau$ as a threshold does not differ from using τ and is described in the main text. For using the potential outcomes for thresholding, we note that we calculate the threshold α as the minimum of the thresholds for each of the potential outcomes separately (cf. Appendix E). Then, we say that prediction should be issued if the variance in the predictions of *both* of the potential outcomes falls below the threshold and the prediction should be withheld if 95% of the variance in *either* of the potential outcomes falls above α.

We present the results calculated for the IHDP-based dataset in Figure 4 and Figure 5 below. We note that the middle two plots of Figure 5 have incomplete error bars, as for certain seeds and certain thresholds there were no individuals assigned action 'predict' or no individuals assigned action 'withhold'.

We note that compared to using $\varphi = \tau$ for thresholding, using the potential outcomes (Figure 4) generally achieves lower cost for the action 'withhold', particularly for lower thresholds, but higher cost for action 'predict'. We further note that when using the potential outcomes for thresholding, the performance of the R_{PO} metric surpasses the performance of the R_τ metric. Thus we note that choosing the optimal acquisition strategy might involve choosing a good combination of both the threshold and the acquisition metric, as they are interdependent.

Scrutinising Figure 5, we further observe using $\varphi = S_\tau$ for thresholding, while leading to very poor results for low threshold values (see the left-most plot in Figure 5), can achieve very good average score per cost ratio compared to the other thresholding variables, across all the metrics, for large values of the threshold. This can be explained by very small proportion of the individuals being assigned action 'withhold' (since S_τ contains relatively high-level information, very few individuals have large enough levels of uncertainty that would doom them to being unsuitable for a treatment recommendation). Thus, the choice for the most suitable acquisition metric and threshold will further depend on the desired level of "safety" and confidence in predictions.

G.2 Performance in the Presence of Possible Overlap Violations: Other datasets

To further evaluate the performance of the different metrics in the presence of possible overlap violations, by considering the explicit early stopping criteria and the decision process described in Algorithm 2, we run the numerical experiments also on the datasets obtained as a modified version of the ACIC2016-based dataset, which more explicitly creates overlap violations. Namely, we modify the data generating process as follows:

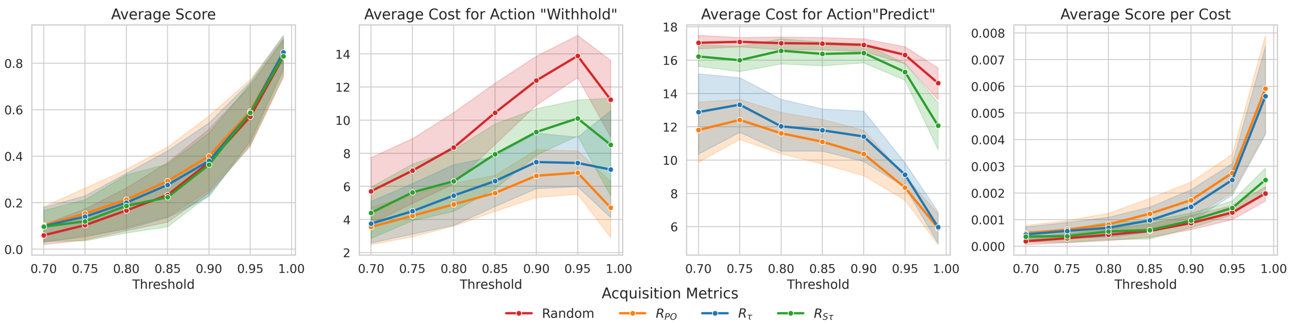

Figure 4: The performance of the different acquisition metrics, when we explicitly allow to withhold treatment recommendations in the presence of large uncertainty. The error bars mark 90% confidence intervals. The results were obtained using the IHDP-based dataset, when the potential outcomes were used for thresholding.

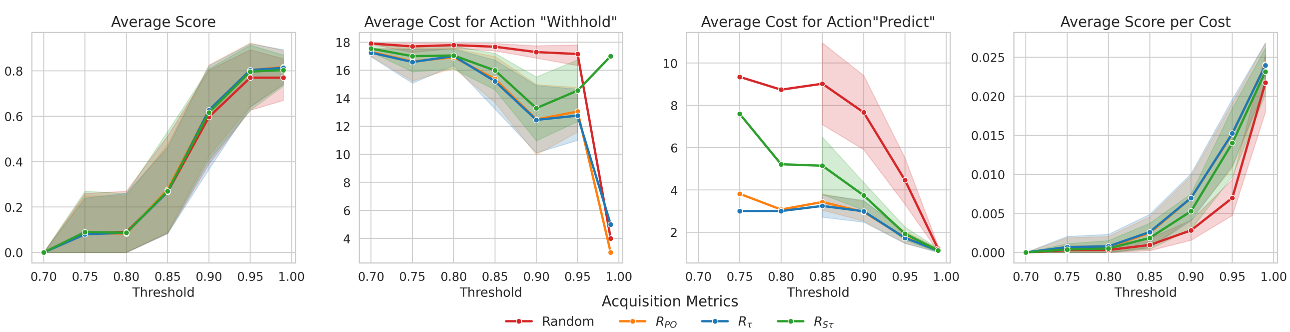

Figure 5: The performance of the different acquisition metrics, when we explicitly allow to withhold treatment recommendations in the presence of large uncertainty. The error bars mark 90% confidence intervals. The results were obtained using the IHDP-based dataset, when $S\tau$ was used for thresholding.

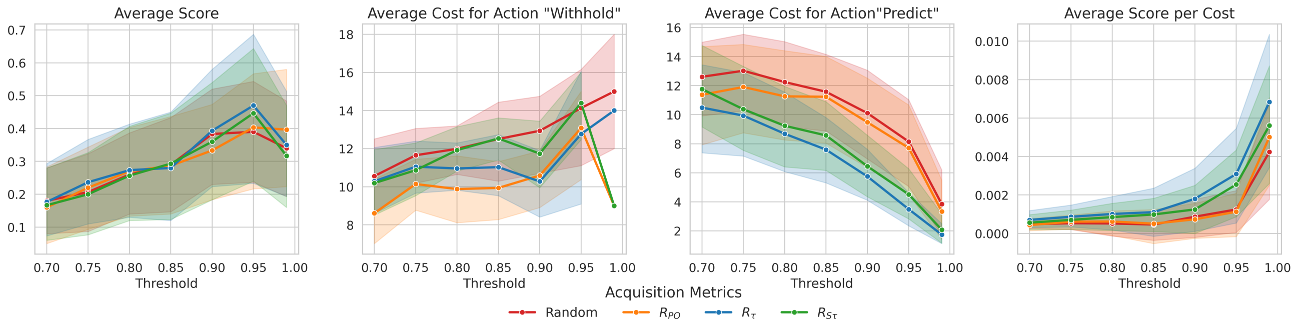

Figure 6: The performance of the different acquisition metrics, when we explicitly allow to withhold treatment recommendations in the presence of large uncertainty. The error bars mark 90% confidence intervals, computed over 10 seeds. The results were obtained on the modified ACIC-2016 dataset, with $\rho = 0.3$.

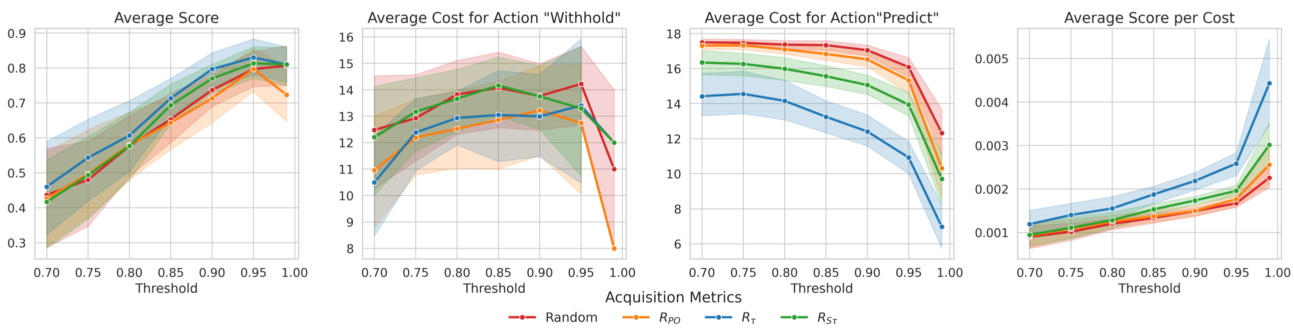

Figure 7: The performance of the different acquisition metrics, when we explicitly allow to withhold treatment recommendations in the presence of large uncertainty. The error bars mark 90% confidence intervals, computed over 10 seeds. The results were obtained on the modified ACIC-2016 dataset, with $\rho = 0.7$.

Treatment Assignment. We begin by creating a propensity score as described in Appendix F.3. We then randomly sample 4 variables X_j and do the following: for two of these variables, if the value of that variable for some individual is larger than the 90^{th} percentile of values in the entire dataset, we set the treatment probability to 1.0. For the remaining two variables, if the value of that variable for some individual is smaller than the 10^{th} percentile of values in the entire dataset, we set the treatment probability to 0.0. This results in a dataset where $\sim 30\%$ of individuals violate the overlap assumption.

Potential Outcomes. To ensure that the four variables which were chosen to influence overlap violations are not also important predictive or prognostic factors, we explicitly set their coefficients β_j and γ_j to zero.

The results are presented in Figures 6-7. The conclusions drawn from the results largely agree with the results for the IHDP-based dataset. We note, however, that in the case of the more explicit overlap violations, the average score decreases, particularly for the case when $\rho = 0.3$. This signifies the importance of designing solutions which will more explicitly target overlap violation settings.