WHEN IS OFF-POLICY EVALUATION USEFUL? A DATA-CENTRIC PERSPECTIVE

Hao Sun*, Alex Chan*, Nabeel Seedat, Alihan Hüyük, Mihaela van der Schaar
DAMTP, University of Cambridge
hs789@cam.ac.uk

ABSTRACT

Evaluating the value of a hypothetical target policy with only a logged dataset is important but challenging. On the one hand, it brings opportunities for safe policy improvement under high-stakes scenarios like clinical guidelines. On the other hand, such opportunities raise a need for precise off-policy evaluation (OPE). While previous work on OPE focused on **improving the algorithm** in value estimation, in this work, we emphasize the importance of the **offline dataset**, hence putting forward a data-centric framework for *evaluating OPE problems*. We propose DataCOPE, a data-centric framework for ***evaluating* OPE**, that answers the questions of whether and to what extent we can evaluate a target policy given a dataset. DataCOPE (1) forecasts the overall performance of OPE algorithms without access to the environment, which is especially useful before real-world deployment where *evaluating OPE is impossible*; (2) identifies the sub-group in the dataset where OPE can be inaccurate; (3) permits evaluations of datasets or data-collection strategies for OPE problems. Our empirical analysis of DataCOPE in the logged contextual bandit settings using healthcare datasets confirms its ability to evaluate both machine-learning and human expert policies like clinical guidelines.

1 INTRODUCTION

Introducing novel policies and guidelines in high-stakes settings such as healthcare and criminal justice comes with great potential harm, and should be backed by appropriate data and evidence before enacting the policy or guideline (Woolf et al., 1999; Suresh and Guttag, 2021). The challenge is that it can be hard to actually know when your data is sufficient, and without the ability to run a policy to see what *actually* happens, we must resort to estimating its effect given a logged *offline dataset* of previously seen observations and actions. Generally, this is known as **off-policy evaluation** (OPE) (Precup, 2000; Beygelzimer and Langford, 2009; Dudík et al., 2011), using data collected previously under one policy to predict the performance of *another* policy. While this terminology emerged in the reinforcement learning (RL) literature, it's rooted in a causal problem and is often considered under the guise of treatment effect estimation when intervention occurs only in one time step (Powers et al., 2018). The causal nature highlights why the offline RL problem is hard - there is a limit to what can be said about a policy from *observational* data alone (Pearl, 2009).

OPE has been extensively studied both methodologically and empirically (Strehl et al., 2010; Jiang and Li, 2016) - however, this has mostly been from the perspective of improving estimators. As such, we challenge the underlying assumption that the dataset is suitable for all potential OPE methods, a neglected focus so far. In this work, we take a *data-centric* approach, looking at the problem:

Given a *dataset*, how accurately can OPE algorithms evaluate *a specific target policy*?

Throughout the paper, we will return to the concrete example of introducing a new clinical guideline (*cf. policy*), such as the Model for End-stage Liver Disease (MELD) scoring in liver transplant allocation (Habib et al., 2006). These are highly impactful, committee-made decisions, for which we really must know the answers to the question of whether it will be effective and if so, who will it actually benefit?

What is needed from the community? We require a method that can *evaluate whether an OPE method will be reliable* and *identify for what contexts/policies an evaluation is uncertain*. For clarity,

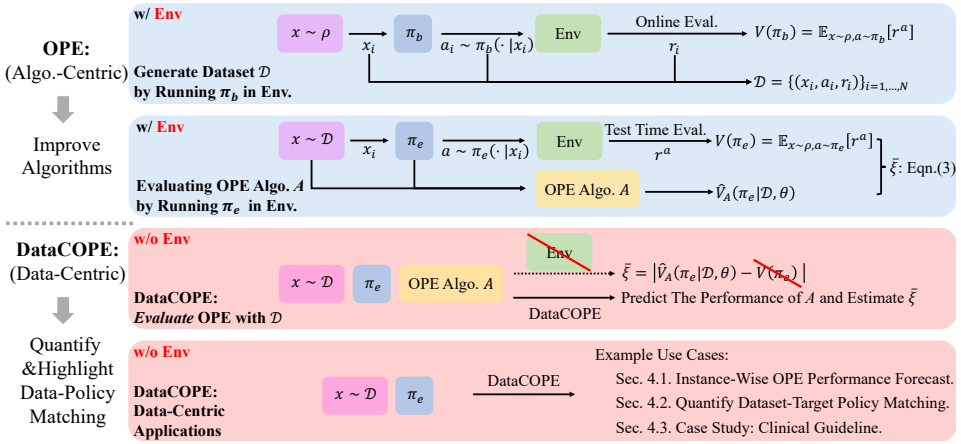

Figure 1: *Road map of DataCOPE.* We first highlight the difference between DataCOPE and classical OPE literature that the objective of DataCOPE is to evaluate whether OPE problems are well-defined, while OPE focus on finding better estimators. **1-st row:** illustrates the offline dataset collection process. In the context of healthcare, it corresponds to treatment records abiding by an existing guideline. **2-nd row: (Sec.2)** The collected dataset $\mathcal{D}$ is then used for OPE. For an OPE algorithm, Equation (3) calculates the test-time residual (error) between the value estimation result from an algorithm and the true value. **3-rd row: (Sec.3)** DataCOPE is proposed as a proxy for such an evaluation residual, that can work in test-time as an OPE performance indicator without access to the true value $V(\pi_e)$ and the environment. **4-th row: (Sec.4)** DataCOPE can be applied to various use cases, which is demonstrated with extensive empirical studies. Notions are explained in Section 2.

consider the following desiderata: **(D1)** Method-Agnostic (Data-Centric): it should be robust to change of underlying OPE methods, being able to capture the intrinsic difficulty of OPE problems regardless of the methods being selected. **(D2)** Individualized Evaluation: for hard-to-evaluate policies, it should be able to identify the reason (examples) that causes the difficulty in evaluation. Specifically, it will be useful if hard-to-evaluate examples can be discovered. **(D3)** Assess Data-Policy Matching: It should be able to identify which dataset, or data collection strategy, is more appropriate in evaluating a certain target policy.

Fulfilling D1-D3, in this work we propose Data-Centric Off-Policy Evaluation (DataCOPE), a framework that evaluates the inherent difficulty of OPE problems and is able to predict the general performance of OPE algorithms by decomposing the estimation uncertainty in the problem. DataCOPE detects dataset-target policy mismatch and thus compares data collection strategies for more accurate OPE without an environment. Figure 1 illustrates how this paper develops. Contributions of our work are threefold:

1. Methodologically, our research diverges from previous algorithm-centric studies on OPE, which have primarily concentrated on developing algorithms. Instead, our investigation places a significant emphasis on the crucial role of data, particularly with regard to the target policy, in OPE problems. Thus, this work is the first to bring ideas of data-centric AI from supervised learning to address the unique issues pertinent to OPE.
2. Practically, we introduce DataCOPE as an evaluation proxy for OPE problems. Traditional evaluation of OPE algorithms requires access to the true target policy value or live environment, but DataCOPE can serve as a proxy to predict whether OPE algorithms perform well in the absence of an environment.
3. Empirically, we have demonstrated that DataCOPE (1) serves as an effective evaluation proxy for OPE, (2) provides a detailed performance prediction on instance-wise value estimation, and (3) can be applied to real-world datasets, such as evaluating clinical guidelines.

2 Preliminaries

We focus on a **logged contextual bandit** setting (Joachims et al., 2018). In particular, we consider **contexts** $x \in \mathcal{X}$, **actions** $a \in \mathcal{A} := \{1, 2, ..., k\}$, and **rewards** $r^a \in \mathbb{R}^+$ generated by the stochastic reward generation process taking action a given context x. In this environment, one can act according

to a **policy** $\pi \in \Pi := \Delta(\mathcal{A})^{\mathcal{X}}$, the **value** of which is given by:

$$V(\pi) = \mathbb{E}_{x\sim p(X), a\sim\pi(x)}\left[r^a\right], \tag{1}$$

the expected reward obtained by executing the given policy. A **contextual bandit** problem then involves finding the solution to:

$$\pi^* = \arg\max_{\pi\in\Pi} V(\pi), \tag{2}$$

considered the **optimal policy** π^*. Typically this is solved via repeated interaction with the environment, which is not available when we move to the **logged** setting. In this case, we are unable to interact with the environment but alternatively have access to a **dataset** $\mathcal{D} = \{(x_i, a_i, r_i)\}_{i=1}^N$ of context, action, reward tuples. These have been generated via a **behavior policy** $\pi_b \in \Pi$ that we *do not* observe but has previously interacted with the environment as follows:

1. An examples $x_i \sim p(X)$ is drawn.
2. The behavior policy π_b selects action $a_i \sim \pi_b(x_i)$.
3. A stochastic reward r^a is observed.
4. (x_i, a_i, r_i) is added to $\mathcal{D}$.

This makes the optimization task in (2) difficult as in the dataset $a_i \sim \pi_b$, which will introduce significant bias if we attempt to Monte Carlo estimate the expectation directly using our given samples. Thus, a large part of the logged contextual bandit problem revolves around the accurate estimation of (1), a task referred to as **Off-Policy Evaluation** (OPE), as we wish to *evaluate* a policy using data collected *off-policy* (i.e., with a behavioral policy).

Many algorithms have been proposed for such estimation, the learning objective of which is normally to minimize the mean-square error $\bar{\xi}$ (**OPE residual**) between real and estimated values. Considering a value estimator $\hat{V}(\pi|\mathcal{D}, \theta)$, built with dataset $\mathcal{D}$ and parameterised by θ, the objective is to minimise:

$$\bar{\xi} := \mathrm{MSE}(\hat{V}) = \mathbb{E}\left[\left(V(\pi) - \hat{V}(\pi|\mathcal{D}, \theta)\right)^2\right]. \tag{3}$$

Plug-in Estimation of the Value Function. An important sub-class of OPE algorithms, based on the **Direct Method** (DM) (Beygelzimer and Langford, 2009) revolves around constructing a directly parameterized estimator of the reward, a function $\hat{q}_\theta$ taking contexts and actions and returning the predicted reward. This is typically learned using supervised learning based on the dataset $\mathcal{D}$:

$$\hat{q}_\theta = \arg\min_{\hat{q}_\theta} \mathbb{E}_{(x,a,r)\sim\mathcal{D}}\left[\,(r - \hat{q}_\theta(x, a))^2\,\right]. \tag{4}$$

Armed with $\hat{q}_\theta$, DM estimates the predicted reward of taking actions according to the policy and plugs them into the value function calculation, producing an estimator:

$$\hat{V}_{\mathrm{DM}}(\pi|\mathcal{D}, \theta) = \mathbb{E}_{x\sim\mathcal{D}, a\sim\pi}\left[\hat{q}_\theta(x, a)\right]. \tag{5}$$

This class of methods is by far one of the most popular (Saito et al., 2020; Fu et al., 2021) (and the one we shall build our method around).

Data-Centric AI Characterization. Previous data-centric works like Data Maps (Swayamdipta et al., 2020) and Data-IQ (Seedat et al., 2022) evaluate the data in classification settings with the goal of characterizing examples in a dataset into easy, hard and ambiguous based on analyzing the training dynamics of individual examples. Such methods assume access to the prediction probability for the ground-truth class to compute uncertainty measures and are largely focused on curating a high-quality training dataset. However, the problem of OPE has three distinct differences: (1) we are focused on test time evaluation, where (2) the ground-truth label in value prediction is not available, therefore, the prediction confidence used in prior works is unavailable and (3) these methods are unable to tackle regression tasks without a softmax probability, which are always needed in OPE. Extended discussions on related work are elaborated in Appendix A.

3 When is OPE Useful?

Given a dataset, target policies are not created equal for OPE. Consider an offline dataset generated by some behavior policy π_b, then intuitively the more similar a target policy π_e is to the behavior policy the easier it should be to evaluate the performance of π_e using the dataset generated by π_b. When the decisions made by π_e and π_b are similar, outcomes of the actions from π_e should be well represented in the dataset, while the unsupported actions' values will be challenging to predict. However, without explicit access to the behavior policy π_b, measuring its distance to the target policy is a highly non-trivial task.

3.1 INHERENT DIFFICULTY OF OPE: A DATA-CENTRIC PERSPECTIVE

A critical objective of this work is to provide a forecast of how accurately an OPE problem can be solved from a data-centric perspective: we emphasize the importance of *data*, rather than the algorithms. Such a perspective permits a hierarchical analysis of the problem:

First and foremost, we want to provide an overall description of the *inherent difficulty* of the OPE problem. Given the offline dataset $\mathcal{D}$, and target policy π_e, we formally write the difficulty of the OPE problem as

$$h(\mathcal{D}, \pi_e) \propto \mathbb{E}_\theta \mathrm{MSE}(\hat{V}_A(\pi_e|\mathcal{D}, \theta)), \forall \hat{V}_A \tag{6}$$

where A is an arbitrary OPE algorithm and θ denotes its instantiation. Then, an OPE problem evaluation method should permit a case-by-case analysis of the OPE problem, and give fine-grained explanations of the difficulties identified above. In doing so, it should identify the sources of OPE difficulty. While in some cases the difficulty originates from inherent stochasticity, in some other cases it requires more diverse samples to match the target policy for an accurate OPE.

3.2 DISTRIBUTIONAL DIRECT METHOD FOR UNCERTAINTY DECOMPOSITION

A central part of our idea for quantifying the difficulty of OPE is to quantify decomposed uncertainty and use that to build a model that is capable of predicting the OPE residual. As is common in data-centric approaches, to satisfy the previous desiderata we propose to separate the aleatoric (data) uncertainty and epistemic (model) uncertainty parts of the prediction (Seedat et al., 2022), which in our case is the value estimator. This has many benefits, especially for comparing datasets for OPE, since the only relevant part here is the epistemic (model) uncertainty.

This section is organized as follows: first, we propose our tailored DM that has a distributional reward estimator; we then introduce the uncertainty decomposition used for breaking down this estimator; finally, we present a practical method for predicting OPE difficulty using these components.

DM with Distributional Reward Estimators To achieve such a separation, instead of using a regular regression model that only learns an expected value, we leverage a probabilistic network to *capture the distributional information* in value estimation. When the reward model is binary (i.e., only success with $+1$ or failure with 0), a network with a standard softmax output can be considered to output the logits of a Bernoulli distribution. For continuous reward models, normal regression models are insufficient and so we adopt the DM with a mixture density network (MDN) (Bishop, 1994) as a reward estimator. Specifically, an MDN predicts a set of parameters used in a mixture of Gaussian predictive distribution so as to maximize the log-likelihood of observed data. The likelihood with K Gaussian is

$$\mathcal{L} = \sum_{i=1}^{i=\mathbb{N}} \mathcal{L}(r_i|x_i, a_i, \theta) = \sum_{i=1}^{i=\mathbb{N}} \sum_{k=1}^{K} w_k(x_i, a_i, \theta) \times \phi(r_i|\mu_k(x_i, a_i, \theta), \sigma_k(x_i, a_i, \theta)), \tag{7}$$

where θ denotes the parameters of networks with three branches of outputs $\{w_k, \mu_k, \sigma_k\}_{k=1,\dots,K}$, ϕ is the probability density function of normal distribution , and $\sum_{i=1}^{K} w_i = 1$ is the normalized weight. In this case, the reward prediction, given context x and action a, is a random variable, denoted as $\hat{R}_x^a$.

Uncertainty Decomposition As we are interested in the predictive random variable $\hat{R}_x^a$, we model the uncertainty with value predictive variance: $v(x, a) = \mathbb{V}_{\hat{R}^a|X=x,A=a}(\hat{R}^a|X = x, A = a)$. According to the law of total variance, we have:

$$v(x, a) = \mathbb{V}_\Theta \left[\mathbb{E}_{\hat{R}^a|X=x,A=a}\left(\hat{R}^a|X = x, A = a, \Theta\right)\right] + \mathbb{E}_\Theta \left[\mathbb{V}_{\hat{R}^a|X=x,A=a}\left(\hat{R}^a|X = x, A = a, \Theta\right)\right], \tag{8}$$

where Θ is a random variable that has an empirical distribution over the set of parameters with different instantiations, i.e., ensemble.

In this way, the overall uncertainty is split into two components: the *epistemic uncertainty* component $v_{\mathrm{ep}} = \mathbb{V}_\Theta \left[\mathbb{E}_{\hat{R}^a|X=x,A=a}\left(\hat{R}^a|X = x, A = a, \Theta\right)\right]$, and the *aleatoric uncertainty* component $v_{\mathrm{al}} = \mathbb{E}_\Theta \left[\mathbb{V}_{\hat{R}^a|X=x,A=a}\left(\hat{R}^a|X = x, A = a, \Theta\right)\right]$. Regarding the epistemic component v_{ep}, the variance originates from the model's oscillation due to a lack of training data. In this case, collecting more data

could help to reduce such variance. On the other hand, in the aleatoric component v_{al}, the variance originates from the inherent uncertainty in the data, which results in the inability to make specific predictions in some cases. In this case, collecting more informative features rather than more data could help to reduce such variance.

Residual Prediction through Uncertainty Now that we are equipped with a decomposition of the uncertainty, we are able to compare the evaluation confidence of target policies: naturally, policies that induce (x, a) pairs with higher epistemic and aleatoric uncertainties are harder to evaluate. While the former can be alleviated by collecting more examples in the training data, the latter is an inherent problem of the task. Moreover, we introduce a practical method to quantitatively estimate the accuracy of OPE algorithms with the help of training data: Given the uncertainty decomposition of (x, a) pairs, we are able to build a linear regression model h that takes v_{al}, v_{ep} as the inputs and outputs the OPE residual. We call this model h the hardness predictor, and fit it with a group of held-out training data from $\mathcal{D}$ where true residual can be obtained without the environment:

$$h = \arg\min_{h} \mathbb{E}_{(x,a)\sim\mathcal{D}}(\bar{\xi} - h(v_{\mathrm{ep}}(x,a), v_{\mathrm{al}}(x,a)))^2, \tag{9}$$

where $\bar{\xi}$ is defined in Eqn.(3). In this way, such a model h is able to work as a hardness indicator function of the OPE problem. Pseudo-code of DataCOPE and more implementation details can be found in Appendix D and Appendix E.

4 EXPERIMENTS

Recall the three desiderata (and hence goals of our method) that we established in the Introduction: we want to be able to 1) identify an instance-wise difficulty in evaluation, 2) be able to compare datasets for evaluating a certain policy by comparing coverage, and 3) robustly evaluate arbitrary OPE algorithms. In experiments, we evaluate DataCOPE's ability to fulfill them point-by-point in Section 4.1, Appendix B and C limited by the space, alongside ablations that consider how our ability is affected by factors such as policy complexity and bias, as well as data coverage. Section 4.2 further verifies those abilities on a real-world clinical dataset — We demonstrate how DataCOPE can be applied to evaluate clinical guidelines.

Synthetic Dataset Generation Following standard procedures in the OPE literature (Chu et al., 2011; Li et al., 2012; Agrawal and Goyal, 2013), we adapt supervised learning datasets into example logged bandit datasets where we know the true underlying generative process. In particular, we use two medical tabular datasets, namely `Breast Cancer` and `Diabetes` (Dua and Graff, 2017) — representing both classification and regression tasks respectively — in order to validate that DataCOPE is effective and powerful. Further validation on other common UCI datasets is provided in Appendix F to verify the generalization of DataCOPE.

We use linear models for the behavior policy π_b, which is trained on the dataset examples $\{(x_i, y_i)\}_{i=1}^N$. Given context x_i with corresponding labels y_i we learn a policy to generate actions by minimizing the expected negative log-likelihood $\mathbb{E}_{x,y}\big[\mathrm{NLL}(\pi_b(x_i), y_i)\big]$, under a predictive Gaussian/Bernoulli distribution for regression/classification respectively.

We then evaluate (x_i, a_i) with the help of ground-truth labels y_i. For classification tasks, the reward of action a is given as $r_i^a = \mathbf{1}(a_i = y_i)$, where $\mathbf{1}$ is the indicator function; while for regression tasks, the reward of action a_i is given by the coefficient of determination of the prediction. i.e., $r_i^a = 1 - \frac{u}{v}$, where u is the residual sum of squares $u = ||y - a||_2^2$ and v is the total sum of squares $v = ||y - \bar{y}||_2^2$, $\bar{y}$ denotes the mean of y. In this way, we can generate $\{(x_i, a_i, r_i^a)\}_{i=1}^N$ tuples as our dataset containing N examples.

Target Policy We also generate target policies using neural network models trained on the *same* training data $\{(x_i, y_i)\}_{i=1}^N$, but with injected noise, that will depend on the particular experiment, as π_e. The ground-truth performance of policy π_e is then given by $r_j^b = \mathbf{1}(b_j = y_j)$, where $b_j = \pi_e(x_j)$ - importantly this is available to us and thus allows for the evaluation of the quality of the OPE. In particular, the off-policy evaluation problem is by definition estimating $\mathbb{E}[r_j^b]$, for $j = 1, ..., N$. As we have access to $y_j, j = 1, ..., N$, we can quantitatively evaluate our method and compare results with the ground-true policy values.

Table 1: DataCOPE is able to predict the instance-wise evaluation error of various OPE algorithms. (Reported numbers: correlation for DataCOPE rows, and for the ablation studies we report the performance difference compared with DataCOPE, **higher is better**)

Dataset	Ablations	DM	DR	RM	SNIPW	IPWsk	SNDR
Breast Cancer	DataCOPE	0.708	0.899	0.936	0.936	0.936	0.901
	w/o v_{al}	−0.136	−0.360	−0.450	−0.450	−0.450	−0.364
	w/o v_{ep}	−0.129	−0.044	−0.022	−0.022	−0.022	−0.043
	w/o Decomposition	−0.129	−0.044	−0.022	−0.022	−0.022	−0.043
Diabetes	DataCOPE	0.743	0.743	0.745	0.745	0.744	0.744
	w/o v_{al}	−0.015	−0.014	−0.013	−0.013	−0.013	−0.013
	w/o v_{ep}	−0.357	−0.365	−0.371	−0.371	−0.370	−0.370
	w/o Decomposition	−0.357	−0.364	−0.370	−0.370	−0.370	−0.369

General Experiment Settings We conducted all our experiments using 8 random seeds and reported the mean and standard deviation of the results. The observed performance differences with 8 trials were found to be statistically significant. For the ensemble-based uncertainty quantification, we determined that using 100 models was computationally feasible for tabular datasets in the healthcare domain. However, we also found that using 10 models produced satisfactory results. We note that DataCOPE is the first to tackle OPE problems by evaluating the problem themselves.

To maintain consistency with prior literature, we differentiated between the behavior policy and target policy in two distinct ways. The first approach involved injecting random noise into the labels or regression target during the generation of behavior policies, thereby ensuring that the behavior trajectories (i.e., dataset) did not completely conform to the target policy behavior. The second, slightly more advanced method involved biasing the policy through dataset sculpture, as described in Section 4.2. The aim of both approaches was to evaluate the performance of DataCOPE on various datasets with differing degrees of mismatch between the behavior and target policies.

4.1 Instance-Wise Difficulty Indication

In this section, we zoom in on the individuals who will be acted on by the policies and use DataCOPE to predict instance-wise OPE difficulty. We provide an analysis at the algorithm level, asking how well will an OPE method do *on average* in Appendix B.

Experimental Setup For each dataset, we run different OPE algorithms and are able to calculate the value estimation residual (error) over every datum, individually comparing this to the output of DataCOPE, which has decomposed the uncertainties into aleatoric and epistemic components. The correlation between the OPE residuals and the uncertainties is then calculated and reported, noting that a *high* correlation implies that the DataCOPE uncertainties have *predictive power* when it comes to determining if the OPE is accurate.

Results As shown in Table 1, DataCOPE effectively predicts the OPE residual, but we can see the effect on an individual level more clearly in the scatter plot of Figure 2 which qualitatively shows the relation between two uncertainties and the averaged OPE performance in terms of residual over 6 algorithms. Normally, the examples whose value can not be precisely estimated are located at the upper right in the scatter plots, meaning they either have high aleatoric or epistemic uncertainty.

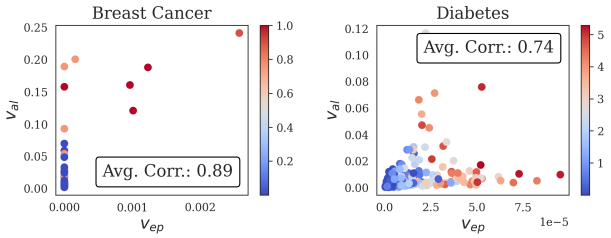

Figure 2: DataCOPE decomposes the uncertainty and provides an instance-wise prediction of the estimation error (Shallow color indicates larger residual value). We find in our experiments that such a prediction is highly correlated with the instance-wise OPE residual: examples with high uncertainty always have high-value estimation errors. The result holds with all datasets and various OPE algorithms.

Our ablation studies also serve to demonstrate how the decomposition is important: In `Diabetes`, the epistemic uncertainty is more important than the aleatoric component in predicting the OPE residual, indicating that such difficulty can potentially be alleviated by collecting more data. While in `Breast Cancer`, we find aleatoric uncertainty dominates the performance prediction. This is important to be aware of since the aleatoric dominance in `Breast Cancer` suggests there is limited opportunity in the future to be able to reduce this uncertainty by collecting additional data -

which is not the case at all for `Diabetes`. In both cases, the decomposition is vitally important as it manifests the uncertainty components even at a different magnitude.

Take-away: *DataCOPE is able to predict instance-wise difficulty in OPE for both discrete and continuous tasks. The uncertainty decomposition step in DataCOPE isolates the uncertainty components' effect even when they are of different magnitudes.*

4.2 Case Study: The Introduction of MELD

We apply DataCOPE to a real-world healthcare guideline and examine its potential to benefit high-stakes OPE problems, with a specific focus on evaluating organ transplant allocation policy.

Background In the medical field, the official guidance on organ transplantation and which potential recipients are offered organs when they become available has evolved multiple times over the past few decades (Starzl et al., 1982; Adam et al., 2012; Hüyük et al., 2022). This setting can be seen as a contextual bandit problem where every arriving patient and organ feature instance is considered the context, while the policy makes an allocation decision as the action, with the corresponding survival time for the patient perceived as an internal reward. We examine data from the Organ Procurement & Transplant Network (OPTN) (Leppke et al., 2013), which includes information on patients registered for liver transplants from 1995 to 2020. More details on data processing can be found in Appendix E.

We specifically focus on the deployment of the prevailing allocation policy — MELD (Bernardi et al., 2011), and study the OPE problem of MELD before its deployment time, 2002. We find DataCOPE:

1. has a strong correlation with the OPE residual;
2. predicts and explains when and why additional collected data improves OPE;
3. identifies vulnerable sub-groups of patients that are more likely to suffer from inaccurate OPE estimation, hence potentially having high risk in medical practice.

This is important to consider whether there actually was sufficient evidence at point of deployment to justify if the guidance would be - or if there were some (sub-)groups who might be left worse off.

Experiment Setting We consider the behavior policy π_b to be that which was deployed before 2002. In order to approximate the process of selecting the most in-need patient, for each patient who receives an organ at some time point, we add 9 other contemporary patients not being allocated to the organ as reference examples. π_b then identifies the selected patient out of the 10 patients — with a discrete action, and the corresponding reward is given by the survival time of the selected patient after receiving the donated liver.

This is then used to estimate the value of the MELD policy, the π_e in the context of this OPE problem, that is deployed after 2002. Similar to the training dataset, we collect test-time data using the transplant records between 2003 and 2005. As MELD was applied as the allocation policy during this period, we have the real value of π_e — the average survival time of the patients who receive organs during this period. We apply the direct method as a demonstrative OPE solver in experiments considering it does not require a parameterized policy.

Results: Correlation. We calculate the Pearson R between uncertainty components calculated by DataCOPE and the OPE residual. Table 2 shows that DataCOPE is highly correlated with OPE performance. Our ablation studies highlight the importance of uncertainty decomposition.

Table 2: DataCOPE correlates with the OPE residual strongly on the Organ Transplant dataset.

DataCOPE	0.712 ± 0.069
w/o v_{ep}	0.654 ± 0.044
w/o v_{al}	0.309 ± 0.198
w/o Decomposition	0.654 ± 0.044

Results: Evaluating MELD Evaluation over time. By utilizing DataCOPE, we are able to examine snapshots of off-policy datasets taken at different points in time, analyze the uncertainties, and compare them with the residuals resulting from OPE, as illustrated in Figure 3.

1. Before the year 2000, the Institute of Medicine allocation recommendations (*IoM Alloc. Rec.*) were used as the behavior policy. Collecting more data during this period can improve the evaluation of the MELD policy. By comparing the results provided by DataCOPE before and at 2000, we can see that the epistemic uncertainty does not change significantly, while the aleatoric uncertainty decreases. This suggests that the OPE task should become easier, which is verified by the reduction in OPE residuals at 2000.

2. After the year 2000, the *OPTN Final Rule* was implemented as the allocation policy. This change in policy affected the pattern of collected data and subsequently, the performance of OPE for MELD. During the period when OPTN was in operation (2000-2002), the aleatoric component of uncertainty decreased while the epistemic component increased. This reminds us to be cautious when using data from this period to evaluate MELD. In fact, the OPE residual using data from this period increases.

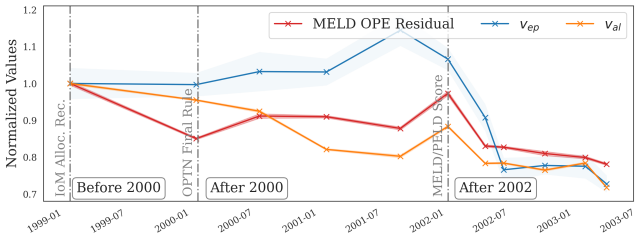

Figure 3: Evaluating MELD over time. DataCOPE can be applied to monitor the OPE performance without the real policy value.

3. After 2002, the *MELD* guideline was implemented as the allocation policy, replacing OPTN. As a result, the data collected since 2002 is unbiased. Both uncertainty components decreased significantly and so did the OPE residual when looking back in hindsight.

To summarize, our study on the Organ Transplant dataset shows that in order to achieve more accurate OPE results, efforts should be made to reduce the epistemic uncertainty as identified by DataCOPE.

Results: Vulnerable Sub-Groups Identification with v_{ep}. In this section, we use DataCOPE to investigate the sub-group with high epistemic uncertainty in OPE, as improving the performance of this sub-group is possible through collecting more data. We examine the examples with the highest epistemic uncertainty in OPE, and found those selected patients, in general, have a larger *Weight Difference*.

To manifest how the performance of OPE evolves over such a sub-group, we compute the averaged OPE residual on the sub-group and compare it with the overall performance. To be specific, we choose to use the data snapshots from the years 2000 (*IoM*), 2002 (*OPTN*), and 2003 (*MELD*) respectively to contrast the effect induced by data collected by different behavior policies.

Table 3: Vulnerable sub-group identification. DataCOPE discovers the vulnerable group in OPE problem of MELD, and can be used to forecast the performance. (Ratio: **Lower is Better**)

	Data	Sub-Group	Overall	Ratio
$v_{\text{ep}}(10^{-5})$	2000 *(IoM)*	2.261 ± 0.124	1.252 ± 0.050	$\times 1.808$
	2002 *(OPTN)*	2.238 ± 0.146	1.358 ± 0.054	$\times 1.648$
	2003 *(MELD)*	1.443 ± 0.081	1.151 ± 0.048	$\times 1.254$
$\bar{\xi}(10^{-2})$	2000 *(IoM)*	8.869 ± 0.116	6.625 ± 0.022	$\times 1.339$
	2002 *(OPTN)*	7.885 ± 0.125	6.838 ± 0.023	$\times 1.153$
	2003 *(MELD)*	6.351 ± 0.043	6.098 ± 0.022	$\times 1.041$

Results are presented in Table 3. In addition to the epistemic uncertainty and OPE residuals of both sub-group and in population, we additionally calculate the ratio to highlight the vulnerability of the sub-group. We find the inaccuracy of OPE in this sub-group remains clearly higher than average until at least 2003, a year or so after MELD was introduced. DataCOPE successfully identifies and monitors such vulnerability with the epistemic uncertainty component, and would have been able to highlight this uncertainty to clinicians implementing the policy, allowing them to take more care in decisions for this sub-group.

Take-Away: *On a real-clinical dataset, we demonstrate DataCOPE accurately mirrors the behavior of the OPE residual and empowers decision-makers with deeper insights into their policy choices and evaluation. Furthermore, our examination of a particular subgroup where the OPE proves inaccurate underscores the practical value that DataCOPE can guide real-world policy deployment.*

5 Conclusion

In this work, we propose DataCOPE to evaluate problems in off-policy evaluation (OPE) which is widely applicable in high-stakes real-world tasks like healthcare. DataCOPE serves as an effective proxy for the true OPE residual without access to the environment. While previous work focused on improving estimators based on offline datasets without considering the inherent difficulty introduced by the mismatch between such datasets and the target policy, we propose to re-think the problem of OPE from a data-centric perspective. The benefits of DataCOPE is demonstrated through empirical studies in both synthetic and real-world healthcare datasets.

REPRODUCIBILITY STATEMENT

The link to our code is attached in our appendix. All implementation details are also elaborated in Appendix D. Our code will be made publicly available.

BROADER IMPACT STATEMENT

Our research answers the question of *when Off-Policy Evaluation (OPE) can be effective* from a data-centric perspective. The problem of OPE is generally important and has been widely applied to real-world applications like healthcare and recommendation systems. However, previous work focused on algorithm-centric development of the methods and neglected the importance of *data*. In our work, we highlight the critical role of *data* in OPE: to evaluate a certain target policy, having appropriate data that matches such a target policy evaluation is a must. Our work calls on the community of OPE research to not only focus on developing estimation algorithms but also rethink the problems from a data-centric perspective and be skeptical of using arbitrary data to evaluate arbitrary target policies.

REFERENCES

Steven H Woolf, Richard Grol, Allen Hutchinson, Martin Eccles, and Jeremy Grimshaw. Potential benefits, limitations, and harms of clinical guidelines. *Bmj*, 318(7182):527–530, 1999.

Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, pages 1–9. 2021.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 129–138, 2009.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie, and Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, 37(11):1767–1787, 2018.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. Learning from logged implicit exploration data. *Advances in neural information processing systems*, 23, 2010.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

Shahid Habib, Brian Berk, Chung-Chou H Chang, Anthony J Demetris, Paulo Fontes, Igor Dvorchik, Bijan Eghtesad, Amadeo Marcos, and A Obaid Shakil. Meld and prediction of post–liver transplantation survival. *Liver transplantation*, 12(3):440–447, 2006.

Thorsten Joachims, Adith Swaminathan, and Maarten De Rijke. Deep learning with logged bandit feedback. In *International Conference on Learning Representations*, 2018.

Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. *arXiv preprint arXiv:2008.07146*, 2020.

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R Zhang, Yutian Chen, Aviral Kumar, et al. Benchmarks for deep off-policy evaluation. *arXiv preprint arXiv:2103.16596*, 2021.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv preprint arXiv:2009.10795*, 2020.

Nabeel Seedat, Jonathan Crabbé, Ioana Bica, and Mihaela van der Schaar. Data-iq: Characterizing subgroups with heterogeneous outcomes in tabular data. *arXiv preprint arXiv:2210.13043*, 2022.

Christopher M Bishop. Mixture density networks. 1994.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Lihong Li, Wei Chu, John Langford, and Robert Schapire. Contextual-bandit approach to personalized news article recommendation, January 19 2012. US Patent App. 12/836,188.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Thomas E Starzl, Shunzaburo Iwatsuki, David H Van Thiel, J Carlton Gartner, Basil J Zitelli, J Jeffrey Malatack, Robert R Schade, Byers W Shaw Jr, Thomas R Hakala, J Thomas Rosenthal, et al. Evolution of liver transplantation. *Hepatology (Baltimore, Md.)*, 2(5):614, 1982.

René Adam, Vincent Karam, Valérie Delvart, John O'Grady, Darius Mirza, Jurgen Klempnauer, Denis Castaing, Peter Neuhaus, Neville Jamieson, Mauro Salizzoni, et al. Evolution of indications and results of liver transplantation in europe. a report from the european liver transplant registry (eltr). *Journal of hepatology*, 57(3):675–688, 2012.

Alihan Hüyük, Daniel Jarrett, and Mihaela van der Schaar. Inverse contextual bandits: Learning how behavior evolves over time. In *International Conference on Machine Learning*, pages 9506–9524. PMLR, 2022.

Susan Leppke, Tabitha Leighton, David Zaun, Shu-Cheng Chen, Melissa Skeans, Ajay K Israni, Jon J Snyder, and Bertram L Kasiske. Scientific registry of transplant recipients: collecting, analyzing, and reporting data on transplantation in the united states. *Transplantation Reviews*, 27(2):50–56, 2013.

Mauro Bernardi, Stefano Gitto, and Maurizio Biselli. The meld score in patients awaiting liver transplant: strengths and weaknesses. *Journal of hepatology*, 54(6):1297–1306, 2011.

Aaron David Tucker and Thorsten Joachims. Variance-optimal augmentation logging for counterfactual evaluation in contextual bandits. *arXiv preprint arXiv:2202.01721*, 2022.

Runzhe Wan, Branislav Kveton, and Rui Song. Safe exploration for efficient policy evaluation and comparison. In *International Conference on Machine Learning*, pages 22491–22511. PMLR, 2022.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Muhammad Faaiz Taufiq, Jean-Francois Ton, Rob Cornish, Yee Whye Teh, and Arnaud Doucet. Conformal off-policy prediction in contextual bandits. *arXiv preprint arXiv:2206.04405*, 2022.

Yingying Zhang, Chengchun Shi, and Shikai Luo. Conformal off-policy prediction. *arXiv preprint arXiv:2206.06711*, 2022.

Takuma Udagawa, Haruka Kiyohara, Yusuke Narita, Yuta Saito, and Kei Tateno. Policy-adaptive estimator selection for off-policy evaluation. *arXiv preprint arXiv:2211.13904*, 2022.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pages 9167–9176. PMLR, 2020.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28, 2015.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32, 2019.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020a.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. Gradientdice: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, pages 11194–11203. PMLR, 2020b.

Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *Advances in Neural Information Processing Systems*, 33:6551–6561, 2020.

Cameron Voloshin, Hoang M Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.

David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19. ACM New York, NY, USA, 1995.

Paola Sebastiani and Henry P Wynn. Maximum entropy sampling and optimal bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):145–157, 2000.

Stephen Mussmann and Percy Liang. On the relationship between data efficiency and error for uncertainty sampling. In *International Conference on Machine Learning*, pages 3674–3682. PMLR, 2018.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122, 2022.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232, 2002.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

A Extended Related Works

A.1 High-Level Difference: DataCOPE is a Framework for Evaluating OPE, rather than an OPE Estimator.

Although DataCOPE is situated within the general field of OPE, it differs from typical OPE papers that propose methods for solving existing OPE problems. Instead, DataCOPE functions as a type of Meta-OPE that "COPE with the Data" — it evaluates general OPE algorithms by forecasting whether the conditional expected return can be estimated accurately.

It is important to note that not all OPE problems are equal, as some can be intrinsically difficult, while others may be much easier. In our work, we focus on evaluating OPE problems (i.e., a dataset-target policy pair) rather than proposing or evaluating OPE algorithms. We present DataCOPE as a practical method and a first step toward quantifying the difficulty of OPE problems. This difficulty is an inherent quantity that relates to the mismatch between the behavior dataset and the target policy. We are motivated to introduce DataCOPE as a data-centric solution that captures the properties of the dataset and target policy.

A.2 OPE Estimators

We provide a further discussion of the OPE literature in this section. Specifically, we connect and differentiate DataCOPE with (Tucker and Joachims, 2022; Wan et al., 2022; Jiang and Li, 2016; Thomas et al., 2015; Taufiq et al., 2022; Zhang et al., 2022; Udagawa et al., 2022).

Inverse Probability Weighting (IPW) (Precup, 2000; Strehl et al., 2010), a weighted sum of behavior rewards is used in estimating the value of a target policy. The Doubly Robust (DR) method (Dudík et al., 2011; Jiang and Li, 2016; Su et al., 2020) improves the DM and IPW by leveraging the strength of both. Shrinkage techniques (Su et al., 2020), self-normalization (Swaminathan and Joachims, 2015), and switch method (Wang et al., 2017) further address the variance issue of those estimators. The recent advance of distribution correction estimation (DICE) family (Nachum et al., 2019; Zhang et al., 2020a;b) achieve promising performance and are unified as regularized Lagrangians of the same linear program (Yang et al., 2020). Differently, in our work, we focus on the *importance of the dataset* used for those OPE estimators, which is why we emphasize the specific dependence on $\mathcal{D}$ in Equation (5).

One of the fundamental differences between our work and the literature above is that DataCOPE challenges the OPE problem itself, rather than aiming to optimize for a better OPE estimator. Our perspective is to question whether evaluating a specific target policy is reasonable, given the current dataset. This data-centric perspective is a significant contribution to the field. Instead of developing OPE algorithms without considering the data quality, we propose to evaluate whether a particular target policy can be appropriately evaluated using the available dataset. Not all target policies can be evaluated with the same level of confidence, and some may be harder to evaluate than others. DataCOPE aims to identify the easy problems and subsets from the hard ones, rather than focusing on developing OPE algorithms that perform well on some problems but not others.

To elaborate on the connections and differences between our work and the above literature in more detail:

We emphasize the importance of the data-centric perspective of OPE throughout the paper, and our contribution is not a method for data collection that treats other approaches as baselines under specific settings (budget, safety, etc.). Instead, we provide insight into the importance of a match between the dataset and the target policy. This perspective is supported by the results of (Tucker and Joachims, 2022; Wan et al., 2022), as they implicitly reveal the importance of data quality, which we focus on explicitly.

Additionally, the settings and assumptions in these works are different. Tucker and Joachims (2022) considers the problem of counterfactual predictions that is often the case in advertising. Wan et al. (2022) explores efficient data collection under safety constraints. In our work, data collection is not an option with freedom in the high-stake clinical setting. We apply DataCOPE as a demonstrative example to show how data quality affects OPE performance and use it as a proxy indicator. Our work

complements this line of literature by quantifying the quality of the dataset for a given policy in terms of uncertainty.

Jiang and Li (2016) extends the doubly robust method to the sequential setting. Its counterpart in the contextual bandit setting, DR, is already used as a backbone method in verifying the universal predictive power of DataCOPE.

The work by (Thomas et al., 2015) estimates a lower bound for OPE but relies on the assumption that the behavior policy adequately covers the action space, which can not hold in high-stakes applications like clinical guidelines. In contrast, our work focuses on the problem of evaluating a target policy with an imperfect dataset, and aims to identify which policies can be accurately evaluated on which subset of the data. Thus, our approach offers a different perspective on the problem of OPE.

The works (Taufiq et al., 2022; Zhang et al., 2022) use conformal prediction techniques to improve the accuracy and reliability of OPE estimates, but they still belong to the OPE algorithm class (algorithm-centric). In contrast, DataCOPE is not an OPE algorithm but rather an evaluation method that provides a data-centric perspective on the feasibility of OPE for a given target policy and dataset. It evaluates the quality of the data and the assumptions underlying OPE, which can inform the development of OPE algorithms as well as the interpretation of their results.

Udagawa et al. (2022) is a concurrent work that examines the issue of estimator selection. In contrast, our work places greater emphasis on the notion that for certain ill-defined problems, such estimator selection may be of little help since all estimators would perform similarly poorly in comparison to when they are matched with a more suitable dataset-target policy. This highlights the fundamental difference between a data-centric approach and an algorithm-centric approach to tackling OPE problems.

A.3 Benchmarking OPE Algorithms.

As with the problem of evaluating policies, actually evaluating the quality of OPE algorithms is similarly difficult. In recent years, the community has proposed various large-scale datasets for the purpose of benchmarking OPE algorithms, including Open-Bandit (Saito et al., 2020) and DOPE (Fu et al., 2021) benchmark OPE in the commercial recommendation and robotics settings. (Voloshin et al., 2019) empirically studies many current methods, and stress tests OPE algorithms in the RL setting. We note an important difference in healthcare that there are always human-based clinical guidelines, rather than machine-learning policies. Moreover, a central problem remains - given a new task for which OPE is required, there is no reliable way to estimate the quality of any value estimation - which brings us to our contribution.

A.4 Exploration

While exploration aims to minimize long-term regret, our work is focused on minimizing the error in off-policy policy value estimation. To illustrate this, consider a clinical trial where the purpose of introducing new treatments (i.e., exploration) is to improve long-term feedback through exploration. In contrast, our work focuses on answering the question of whether, given a past or future dataset of patient information, we can minimize the estimation error of an introduced new treatment. Therefore, we believe that our work can provide valuable insights into improving the accuracy of policy value estimation and complement the exploration strategies in the literature.

A.5 Active Learning and Active Data Collection

DataCOPE deals with a fundamentally different task from active learning (Lewis, 1995; Sebastiani and Wynn, 2000; Mussmann and Liang, 2018; Houlsby et al., 2011; Kirsch et al., 2019; Nguyen et al., 2022). The goal of active learning-based uncertainty sampling is to improve the model by selecting which samples to best label next (model-centric task). This is different from DataCOPE which is focused on an already given dataset, with the goal of characterizing the types of samples and their effect on OPE. In addition, the cost functions in acquiring new samples are always very clearly defined in active learning, yet our setting does not set explicitly trade-offs.

Our proposed metric could be extended for the purposes of data acquisition. The active learning literature has explored two main paradigms, Maximum Entropy Sampling (MES) or Uncertainty

Sampling, and Bayesian Active Learning by Disagreement (BALD), which aim to query samples with the maximum predictive entropy or maximize the mutual information between the observation and the model parameters, respectively. These approaches reflect model uncertainty and are linked to our notion of epistemic uncertainty, albeit measured differently. Our work could inspire future research on how to leverage our proposed metric for more principled data acquisition in OPE settings.

Different from previous active learning works or OPE data collection methods like (Tucker and Joachims, 2022), DataCOPE is used for evaluating datasets at the same size yet with different samples to demonstrate the importance of the match between dataset and target policy in OPE problems. Those different samples are collected by different strategies: the baseline one is collected through uniform sampling, while the other one is collected according to the epistemic uncertainty.

In our experiments, we are not actively recruiting new data points that have the highest epistemic uncertainty, nor can we be able to add new features to the task to decrease aleatoric uncertainty. What we intended to emphasize in this section is that DataCOPE is a hindsight descriptive tool that compares the quality of a dataset given a certain target policy.

DataCOPE is distinct from prior active learning literature or OPE data collection methods such as (Tucker and Joachims, 2022) in that DataCOPE assesses datasets of the same size but with varying samples to emphasize the significance of matching the dataset with the target policy in OPE problems. Different sampling strategies are employed to collect these diverse samples in Section C: the baseline approach employs uniform sampling, while the other method leverages epistemic uncertainty to collect samples. The takeaway message is that: datasets are not equal in evaluating the same target policy. A lower epistemic uncertainty leads to a generally improved performance among a batch of OPE algorithms. There is no doubt that advanced data-collecting algorithms like (Tucker and Joachims, 2022) can potentially be more efficient, though might not be practical in high-stake scenarios.

To sum up the difference: in our experiments, we do not actively recruit new data points that have the highest epistemic uncertainty, nor can we add new features to the task to reduce aleatoric uncertainty. The purpose of this section is to emphasize that DataCOPE is a retrospective descriptive tool that compares the quality of a dataset given a specific target policy.

B Using DataCOPE for Evaluating OPE Algorithms

We quantitatively demonstrate that the uncertainty components from DataCOPE can be used to estimate the OPE residuals with calibration across a range of OPE methods and therefore could be broadly used as a performance indicator when the ground truth is unknown.

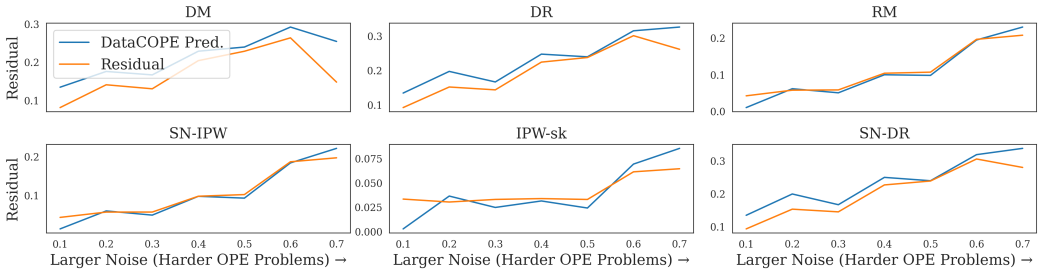

Figure 4: *DataCOPE works as an accurate proxy of the OPE residual.* DataCOPE is able to predict OPE residuals of different algorithms with calibration. Dataset: `Breast Cancer`.

Experiment Settings For each dataset, we run different OPE algorithms and are able to calculate the value estimation residual (error), comparing this to the output of DataCOPE, which has decomposed the uncertainties into aleatoric and epistemic components. The correlation between the OPE residuals and the uncertainties is then calculated and reported, noting that a strong correlation implies that the DataCOPE uncertainties have predictive power when it comes to determining if the OPE is accurate.

To test the impact of mismatched behavior and target policies, in the `Diabetes` dataset, we consider the impact of injecting different levels of noise during policy generation. With higher training noise,

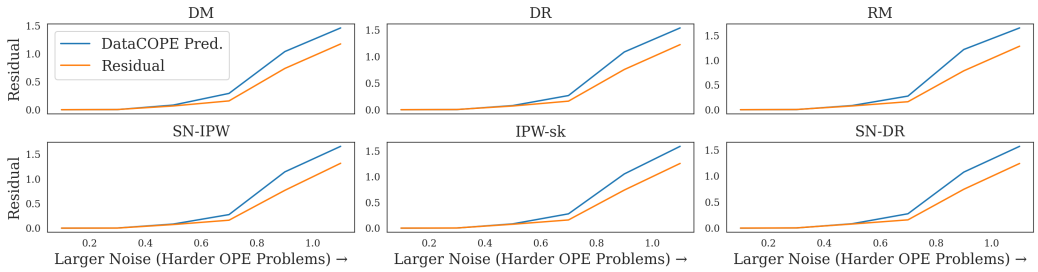

Figure 5: *DataCOPE works as an accurate proxy of the OPE residual.* DataCOPE is able to predict OPE residuals of different algorithms with calibration. Dataset: `Diabetes`.

the policies will be forced to rely on more general features and employ simpler heuristics - in this way the relative *complexity* of the target policy compared to the behavior policy *increases* allowing the difficulty of OPE given the mismatched behavior policy and target policy to be manipulated and explored.

Specifically, we test across a range of values of complexity and add Gaussian noise with variance $[0.1, 0.3, 0.5, 0.7, 0.9, 1.1]$ during behavior policy learning. We then perform calibration to predict the OPE residuals using DataCOPE according to Equation (9): we fit model h on the held-out training data, and dub such a model the calibrated OPE residual predictor, before again aiming to predict OPE residuals using our uncertainty decomposition.

Results Raw numerical results are reported in Table 1, which quantitatively shows the correlation between uncertainties obtained from DataCOPE and value estimation residuals of different OPE algorithms. DataCOPE in general performs well across all datasets and algorithms, with a strong correlation with the residuals.

Take-away: *Calibrated uncertainties decomposed by DataCOPE can be used to accurately predict the performance across a range of OPE algorithms.*

On the impact of complexity, Figure 4 and Figure 5 visualizes our experiment results, plotting both the prediction of DataCOPE, and the true OPE residual for multiple OPE methods as the complexity differential increases. DataCOPE is able to predict OPE residuals of various algorithms with high accuracy. In general, the deviations of predicted values from real residuals get higher as the complexity increases. Moreover, we empirically find DataCOPE tends to overestimate the OPE residual, lending its use as a performance lower-bound, being most suitable for high-stake application scenarios like healthcare.

Take-away: *As complexity differential between policies increases, the OPE performance decreases - an effect DataCOPE is able to predict and track well.*

C Matching Dataset and Target Policy with DataCOPE

So far we have shown that DataCOPE can tell if our OPE estimates will be any good, and for which individuals this will apply. We now move to show how the uncertainty components discovered by DataCOPE can be informative in comparing datasets for evaluating a certain target policy. We consider the scenario of the clinical setting where new policies need to be evaluated before verifying those policies in clinical trials. We show that DataCOPE can act as a performance indicator and be informative in the pursuit of efficient and accurate OPE.

Experimental Setup We use DataCOPE to compare two data collection strategies for OPE problems, aiming to show that the general OPE performance is aligned with epistemic uncertainty. Here, we synthetically generate a biased logged dataset by running biased and highly deterministic behavior policies π_b for collecting those data. Specifically, we explore the *coverage* of the behavioral policy by thresholding examples by their quantile values on certain features. We remove examples with the top 60% high values at the feature named *worst concave point* in the `Breast Cancer` dataset and remove examples with top 60% high values of the feature named *log of serum triglycerides level* in

the `Diabetes` dataset [1] during behavior policy generation. In the meantime, the target policy π_e is still generated with the full dataset, and hence there exists a behavior bias on the training examples out of the threshold.

We demonstrate the effectiveness of DataCOPE by experimenting with the two most straightforward data collection strategies: uniform sampling (e.g., uniformly recruiting patients in clinical trials) and uncertainty-guided sampling — according to the principle of *optimism in the face of uncertainty* (Kearns and Singh, 2002; Brafman and Tennenholtz, 2002) (i.e. picking those with the highest uncertainty).

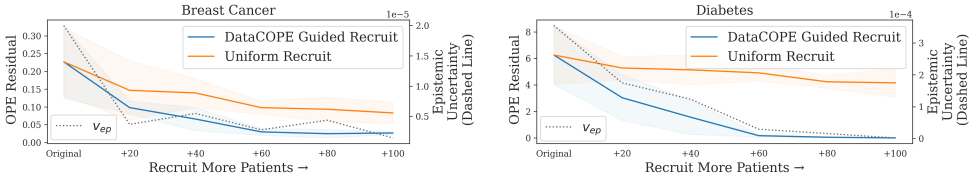

Figure 6: Increasing the number of examples according to epistemic uncertainty in the learning of π_b clearly improves OPE performance. While uniformly sampling new examples only improves OPE a little. DataCOPE identifies which dataset or data collection strategy is better for evaluating a given target policy.

Results In Figure 6, we report the changes in OPE residuals as newly recruited patients are added. Therefore, the aligned x-axis indicates the size of the dataset is the same — yet their quality can be different — due to different data collection strategies and discrepancies from the behavior policy. We report the epistemic uncertainty in OPE when recruitment is guided by DataCOPE. Curves are obtained by averaging over 6 benchmark algorithms. For both datasets, we find DataCOPE is able to identify a better data collection strategy according to the epistemic uncertainty.

Take-away: *DataCOPE is able to identify the target policy π_e's deviation from the behavior policy. The epistemic uncertainty component of DataCOPE informs the inherent difficulty of OPE problem in general as well as in sub-groups. The quality of the datasets for a specific target policy evaluation task can be compared according to DataCOPE.*

D Implementation Details

D.1 Code and Pseudo Code

Our code is open-sourced anonymously at

`https://anonymous.4open.science/r/Data-Centric-OPE-59E5/`.

Algorithm 1 DataCOPE

Input
Logged Dataset $D = \{x_i, a_i, r_i\}$, target policy π_e.
Output
h that Forecasts OPE Residual
1. Build Value Estimator
Optimize $\hat{q}_\theta$ with Equation (7) for non-binary reward estimation and classifiers with logits outputs otherwise.
2. Uncertainty Decomposition
Decompose the uncertainties in estimating $\hat{q}_\theta$ with Equation (8).
3. (Optional) Calibration
Quantify OPE residual according to Equation (9).
Return
Difficulty in estimating different examples

D.2 Hardware and Training Time

We experiment on a machine with 2 TITAN X GPUs and 32 Intel(R) E5-2640 CPUs. In general, the computational expense of DataCOPE is cheap. With our PyTorch-based implementation, de-

[1] these are selected as the most influential dimension for linear models to manifest the difference

composing the uncertainty components with DataCOPE using 100 neural network ensembles takes approximately 10 minutes to run. OPE algorithms take 2 to 10 minutes to run depending on their complexity and can be accelerated by using GPUs.

D.3 OPE Algorithms

Our implementation of OPE algorithms is based on the implementation of (Saito et al., 2020), open-sourced at `https://github.com/st-tech/zr-obp/tree/master/obp/ope`.

Different from (Saito et al., 2020), we do not assume the access of the behavior policy π_b, hence we build estimators of π_b based on the dataset by training classification or regression models with neural networks if a parameterized behavior policy is needed. e.g., in the case of IPW.

D.4 Hyper-Parameters for Neural Approximators

The hyper-parameters for neural-network-based approximators we used for MDN are reported in Table 4.

Table 4: Hyper-parameters of neural network approximators

Hyper-param	Choice
Network Layer	3
Activation Function	ReLU
Hidden Units	64
Training Epoch	100
Optimizer	Adam
Learning Rate	1e-3

D.5 Hyper-Parameters for DataCOPE

In DataCOPE, there are two hyper-parameters to be specified: the number of ensemble models used and the number of Gaussians in the MDN model in regression tasks. DataCOPE is robust to both hyper-parameters in our empirical study.

In our experiments, we find using 100 ensemble models is computationally affordable on tabular data. As it takes less than 10 minutes to run. In the meantime, we find using 10 ensemble models can provide decent performance.

The main objective of introducing MDN is to be able to quantitatively decompose the uncertainty in regression settings. Therefore, while the number of Gaussian mixtures used should depend on the inherent structure of data distribution, whether or not capturing the exact distribution is of less importance than being able to capture the uncertainty components. The number of Gaussians we use in our experiment is set to be 3 for all datasets. We empirically find using $3, 5, 10$ mixture of Gaussians does not clearly affect the performance.

E Experiment Details

E.1 Calibration

In practice the OPE residual $\bar{\xi}$ is always infeasible, such a difficulty leads us to introduce a practical substitute that estimates $\bar{\xi}$ without relying on that infeasible ground-truth policy value. We elaborate on the calibration step in this section.

We leveraged a cross-validation type method in the calibration step. To conform with typical procedures in supervised learning, we hold out a portion of the training data, where we have knowledge of their accurate instance-wise value (i.e., the return of the context-action pairs). By doing this, we are able to break down the uncertainties of the instance-wise value prediction. Subsequently, we use the reserved ground-truth value to calculate the OPE residuals of each OPE algorithm as the

objective, and uncertainties as the input of a regression model. Our implementation employs the bootstrap sampling method.

It is also worth mentioning that such a calibration step is optional, and DataCOPE can evaluates OPE problems without calibration:

In fact, our contribution and novelty of DataCOPE do not rely on such a calibration step. This is demonstrated in our experiments: The difficulty of OPE problems is correlated with our decomposed uncertainties. Such a decomposition, without calibration, is useful in that (1) it permits a comparison among datasets to answer the question of "which dataset is most appropriate in evaluating a certain given policy" (Section C) (2) it permits a hindsight interpretation of clinical guideline deployment and vulnerable group identification. (Section 4.2)

E.2 Synthetic Dataset

Policy Generation Following standard procedures in the OPE literature (Chu et al., 2011; Li et al., 2012; Agrawal and Goyal, 2013), we adapt supervised learning datasets into example logged bandit datasets where we know the true underlying generative process.

We use linear models for the behavior policy π_b, which is trained on the dataset examples $\{(x_i, y_i)\}_{i=1}^N$. Given context x_i with corresponding labels y_i we learn a policy to generate actions by minimizing the expected negative log-likelihood

$$\mathbb{E}_{x,y}\big[\mathrm{NLL}(\pi_b(x_i), y_i)\big], \tag{10}$$

under a predictive Gaussian/Bernoulli distribution for regression/classification respectively.

We then evaluate (x_i, a_i) with the help of ground-truth labels y_i. For classification tasks, the reward of action a is given as $r_i^a = \mathbf{1}(a_i = y_i)$, where $\mathbf{1}$ is the indicator function; while for regression tasks, the reward of action a_i is given by the coefficient of determination of the prediction. i.e., $r_i^a = 1 - \frac{u}{v}$, where u is the residual sum of squares $u = ||y - a||_2^2$ and v is the total sum of squares $v = ||y - \bar{y}||_2^2$, $\bar{y}$ denotes the mean of y. In this way, we can generate $\{(x_i, a_i, r_i^a)\}_{i=1}^N$ tuples as our dataset containing N examples.

Target Policy We also generate target policies using neural network models trained on the *same* training data $\{(x_i, y_i)\}_{i=1}^N$, but with injected noise (that will depend on the particular experiment, as π_e. The ground-truth performance of policy π_e is then given by $r_j^b = \mathbf{1}(b_j = y_j)$, where $b_j = \pi_e(x_j)$ - importantly this is available to us and thus allows for the evaluation of the quality of the OPE.

In particular, the off-policy evaluation problem is by definition estimating $\mathbb{E}[r_j^b]$, for $j = 1, ..., N$. As we have access to $y_j, j = 1, ..., N$, we can quantitatively evaluate our method and compare results with the ground-true policy values.

E.3 Organ Transplant

Data Description and Logged Contextual Bandit Formalism. We examine data from the Organ Procurement & Transplant Network (OPTN) (Leppke et al., 2013), which includes information on patients registered for liver transplants from 1995 to 2020. Our focus is on the policy of matching organs that become available to patients who are waiting for a transplant.

For each decision, the set of potential patients (action space) includes those on the waitlist at the time the organ becomes available, and the information considered for each patient (context) includes both the organ and the patient's characteristics.

Data Preparation. The OPTN dataset includes 308,912 patients who were either waiting for or had received a liver transplant. We eliminated patients who hadn't received a transplant, were under 18 or had a donor under 18, or had missing data for certain variables, leaving us with 31,045 patients. Consequently, we consider 8 features in total: *ABO Mismatch, Age, Creatinine, Dialysis, INR, Life Support, Bilirubin,* and *Weight Difference.*

The focus of our experiments is on patients who received organs prior to 2005, allowing us to divide the data into distinct phases based on evolving allocation guidelines. Specifically, 512 patients were

allocated organs under the Institute of Medicine recommendations before 2000, 1969 patients under the OPTN Final Rule between 2000 and 2002, and 2515 patients after the implementation of MELD between February 26th, 2002 and May 1st, 2003. Those data serve as the training set, with additional data containing 3914 patients collected between May 1st, 2003 and January 1st, 2005 used to evaluate OPE for MELD and DataCOPE's estimation of MELD.

We streamline the data following (Hüyük et al., 2022) and select patients out of their contemporary patients from the transplant waitlist for an organ to be allocated to. Hence formalizing the dataset as a logged contextual bandit task. For every coming organ, the allocation policy selects one from the waiting patients to allocate the organ and receives the patients' survival time as the reward.

F Additional Experiments

In order to validate that DataCOPE is effective and powerful, we further validate DataCOPE on other UCI datasets. We experiment with synthetic contextual bandit dataset generated from both classification tasks `Digits, Wine` and regression task `Boston`. Results are presented in Table 5. On all datasets, we find DataCOPE is able to act as a well-performing proxy of the OPE residual.

Table 5: DataCOPE is able to predict the difficulty of various OPE algorithms under various settings. The correlation between DataCOPE's two uncertainties and OPE residuals under different settings is high. The ablation studies show that uncertainty decomposition is important in predicting OPE performance. (Reported numbers: correlation for DataCOPE rows, and for the ablation studies we report the performance difference compared with DataCOPE, **higher is better**)

Dataset	Ablations	DM	DR	RM	SNIPW	IPWsk	SNDR
Wine	DataCOPE	0.801	0.873	0.915	0.918	0.829	0.895
	w/o v_{al}	-0.004	-0.018	-0.238	-0.193	-0.176	-0.045
	w/o v_{ep}	-0.049	-0.027	-0.037	-0.02	-0.019	-0.008
	w/o Decomposition	-0.015	-0.003	-0.09	-0.062	-0.057	-0.0
Digits	DataCOPE	0.912	0.907	0.963	0.996	0.899	0.957
	w/o v_{al}	-0.224	-0.22	-0.004	-0.028	-0.099	-0.126
	w/o v_{ep}	-0.3	-0.295	-0.0	-0.063	-0.154	-0.188
	w/o Decomposition	-0.241	-0.238	-0.002	-0.035	-0.111	-0.14
Boston	DataCOPE	0.805	0.787	0.786	0.779	0.776	0.787
	w/o v_{al}	-0.024	-0.062	-0.06	-0.057	-0.045	-0.068
	w/o v_{ep}	-0.071	-0.105	-0.117	-0.107	-0.094	-0.107
	w/o Decomposition	-0.07	-0.104	-0.117	-0.106	-0.094	-0.106

G Discussion on Limitations and Future Work

In our work, we introduce DataCOPE as a Data-Centric solution for evaluating OPE algorithms. However, it is important to note that while DataCOPE serves as an OPE performance proxy, it does not directly improve the performance of OPE algorithms. It addresses the question of the difficulty of OPE problems but does not provide solutions for solving these OPE problems more accurately (without obtaining more in-distribution data). Moreover, our exploration of dataset-policy matching is conducted in a post-hoc manner (i.e., passive). While it highlights the differences and importance of dataset-target policy alignment, it does not directly enable active data collection (i.e., active). This limitation is inherent to OPE problems themselves, such as in healthcare applications where collecting data on specific patient types is not feasible. However, to evaluate a certain target policy, comparison between available datasets is feasible and important.

In future work, several potential directions are promising: the most straightforward extension is to combine DataCOPE with off-policy improvement, however, this requires a trade-off between policy evaluation confidence and potential policy evaluation performance; similarly, DataCOPE can be combined with uncertainty-based exploration algorithms and be extended to the online setting, where data-centric policy learning can be studied; finally, DataCOPE can be linked with offline RL, where

multiple-step decision-making policies need to be evaluated and improved. We believe that all of these directions are essential and warrant further investigation, though they are beyond the scope of this paper.