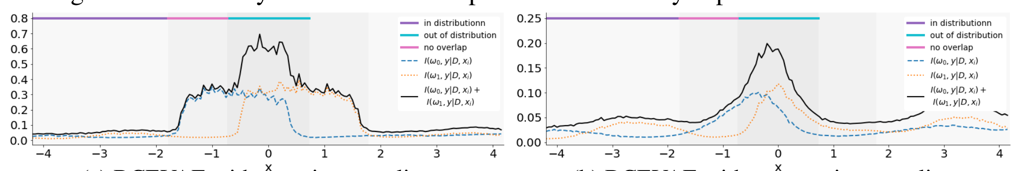
We thank reviewers for their useful feedback. We are encouraged that all reviewers recognize the relevance and importance of being uncertain when overlap does not hold or when covariate shift has occurred, and, moreover, by recognizing our "valuable" (R1), "novel" (R2) and "interesting" (R4) contributions, the "significant conceptual novelty" (R3) of our methods, and that our estimators "overwhelmingly allow methods to recognize samples that violate overlap or don't resemble training data" (R1). We are pleased that reviewers find our empirical results convincing, that they show significant improvement over baselines (R3) and are carried out through realistic (R3), rigorous (R1), and reproducible experiments. Finally, we appreciate that the majority of concerns are given as suggestions for improvement and we address the reviewers' insightful comments (noting that R2 understood the paper well–despite their low confidence score) in the following.

R1 "There is occasional exposition that I don't feel precisely captures causal inference..." You are correct that CATE is not identifiable under unmeasured confounding and that the reason we consider CATE (and not ITE) is that ITE is not identified without parametric assumptions. We will integrate this feedback and be more precise.

R1 "I couldn't really follow the section on negative sampling [...] does it produce better calibrated estimates of the variance of CATE?" Negative sampling is a method specifically for CEVAE, not for the other neural methods. Its effect is to increase the estimated uncertainty more sharply even when a point x is only slightly outside the region with good data coverage. We will clarify how this works. Epistemic uncertainty is plotted in black below:



(a) BCEVAE with negative sampling (b) BCEVAE without negative sampling

R1 "Eqn 4 reviews one notion of uncertainty for regression tasks. [What is its role?]" Eq. 4 is a typical measure for epistemic uncertainty in non-causal tasks. It is simply there to build basic background knowledge about epistemic uncertainty, which we later extend to the task at hand. The particular decomposition of the variance in eq. 4 is just the standard implementation. We agree that this can be clarified.

R2 "The paper makes the assumption of ignorability [...] This is unlikely to hold in practice" We agree with the reviewer that the assumption is unlikely to hold in practice and that our extension to Causal Effect Variational Autoencoders replaces this with a slightly weaker assumption. Addressing true hidden confounding is beyond the scope of this paper and we leave it to future work.

R2 "results in the paper seem to sweep over some threshold [[...] where or how this threshold should be set?" First, thank you for pointing out arxiv:1903.12220, it is relevant to this question and will be added to our lit survey. In general, setting the threshold will be a domain-specific problem that depends on the cost of type I (incorrectly recommending treatment) and type II (incorrectly withholding treatment) errors. We would appeal to domain experts to ascertain such costs. For example, in lung cancer screening (PMC4817217), the CT scan is a covariate, whether to obtain follow-up scans is the treatment, and death due to lung cancer is the outcome. Here, the cost of a type II error is much higher and would need to be accounted for in determining the threshold. In the diagnostic setting, thresholds have been set to satisfy public health authority specifications on sensitivity and specificity; e.g. for diabetic retinopathy detection (nature.com/articles/s41598-017-17876-z). When deployed, the treatment recommendations are given for novel individuals; therefore, thresholds will need to be determined using the data available at training time.

R3 Authors should add more detail to section 3 Currently we introduce the rejection policies in section 6 and give details in the appendix. Following the reviewer's suggestion, we will move this to section 3.

R3 "How can the epistemic uncertainty estimation formula (7) be applied to CEVAE, for which both w and z are random?" Great point, this should be clarified. Instead of just sampling the parameters w0,w1, we sample these parameters *and* z, independently. z must be sampled too as this is what the standard CEVAE does (see eq. 8).

R3 "the predictive uncertainty policy and the epistemic uncertainty policy appear to behave very similarly? Are there data / model scenario that we would prefer one over the other?" The predictive policy is shown for comparison; it is not part of our method. The policies behave the same when there is no aleatoric/label noise (like with MNIST) because then aleatoric uncertainty is zero. But when we do have aleatoric/label noise (like with IHDP) they can behave differently (predictive being worse). Whenever the task is to estimate CATE, we must use the epistemic uncertainty because the uncertainty in CATE is only epistemic (second r.h.s. term of eq. 7). The aleatoric uncertainty component would only matter for estimating the (unidentifiable) individual treatment effect (ITE).

R4 "the core value behind this work is an effort to handle no-overlap [...] Yet, it's a bit hard to say their proposed methodology is quite novel in terms of technicality." We agree that part of the contribution is an effort to handle the no-overlap situation and covariate shift by using the combination of modeling uncertainty and causal effect estimation, both in VAEs and with a plethora of other SOTA approaches in the causality field. To the best of our knowledge, we are the first to bring these methods together and the first to empirically show their value for this task. We believe that the fact that we arrive at these results by integrating existing methods does not take away from the significance or the novelty of our solution, as noted by R1, R2 and R3.