We would like to thank all reviewers for their valuable feedback which has helped us improve the paper!

Reviewer 1: ■ **Dataset descriptions:** Please note that Appendix I contains details about the datasets used in terms of patient features, possible interpretations of treatments and of patient outcomes. We also provide links to the publicly available datasets. Upon acceptance, we will release the code for the model and for the semi-synthetic data generation.

Reviewer 2: ■ **Novelty:** Our method only draws inspiration from [6] in terms of using a GAN framework to learn counterfactual outcomes. Nevertheless, to handle continuous interventions, we propose a novel hierarchical discriminator architecture. We also provide theoretical results, which are lacking from [6] to show that the proposed GAN framework can indeed learn the distribution of the counterfactual outcomes. Finally, we introduce a new semi-synthetic data simulation that can be used to benchmark causal inference methods for estimating the effects of continuous interventions.

■ **Sample efficiency:** Our remark regarding sample efficiency was perhaps a bit offhand. Experimentally, we showed that SCIGAN works for a few thousand samples (using the MIMIC dataset with 1920 training samples). We had not investigated how SCIGAN performs below this number. We have now performed a further experiment to evaluate model performance in terms of sample efficency. For the MIMIC dataset, in Table 1 we report evaluation metrics for training SCIGAN with different number of training samples N and evaluating on the same test set.

	$\sqrt{\text{MISE}}$	$\sqrt{\text{DPE}}$	$\sqrt{\text{PE}}$
$N = 100$	31.12 ± 63.39	7.72 ± 2.57	18.94 ± 29.07
$N = 500$	13.36 ± 10.46	4.07 ± 1.92	2.63 ± 0.94
$N = 1000$	3.80 ± 1.04	2.46 ± 1.75	1.03 ± 1.13
$N = 1500$	2.95 ± 0.37	0.70 ± 0.17	0.63 ± 0.12
$N = 1920$	2.09 ± 0.12	0.51 ± 0.05	0.32 ± 0.05

Table 1: Sample efficency analysis for MIMIC. Metrics are reported as Mean ± Std.

■ **PCA for GPS model:** PCA is only used for the GPS model for TCGA and News datasets, which contain a large number of features, to reduce computational complexity. Since GPS is a linear method, using PCA as a pre-processing step helps avoid problems with co-linear features. We used a publicly available implementation for GPS based on the causaldrf package in R. After re-running GPS without PCA on News we obtained similar results to the ones in Table 3 in the paper: 6.03 ± 0.01 ($\sqrt{\text{MISE}}$), 6.83 ± 0.01 ($\sqrt{\text{DPE}}$) and 22.56 ± 0.03 ($\sqrt{\text{PE}}$). ■ **Code and data:** We will release the code for the model and semi-synthetic data generation upon acceptance. ■ **Calibration:** We did not consider calibration - though it would certainly be an interesting future research direction. We would note, though, that [R1] is not about improving the generator but rather gleaning a useful discriminator from the training procedure (which would normally result in a degenerate discriminator), which could be used at test-time to evaluate the generated response-curves. We will add discussion about this in the conclusion. ■ **Hierarchical discriminator:** The term hierarchical refers to the fact that there are 2 levels to our discrimination procedure - (1) determine the factual treatment; (2) determine the factual dosage given the factual treatment. In contrast with the term ensemble which would typically refer to several models performing the same task, we have different models performing different tasks. ■ **Permutation invariance and equivariance:** We use permutation invariance and equivariance because we are fundamentally dealing with dose-response curves, which are themselves functions. To treat these as functions, we treat them as sets of points of the form (input, output). For this reason we use permutation invariance and equivariance - so that the networks act as functions on sets (rather than functions on vectors which would be the case without the in-/equi-variance).

Reviewer 3: ■ **Presentation and notation:** Please note that the problem we are aiming to solve requires complex notation due to the fact that we are handling treatments with continuous dosage. Moreover, our choice of architecture in terms of the hierarchical discriminator also needs complex notation. Unfortunately, we do not feel that the notation can be simplified much. Appendix B contains a table for all of our notation and we will work further to improve the presentation and notation in the revised manuscript. ■ **Details in appendices:** Due to the page limit for the conference, it was not possible to add all of the details in the main paper. We have tried to keep as much information as possible in the main paper, which involved many tough decisions about what was best placed in the main paper and what could be placed in the appendix. ■ **Stopping criteria for the GAN network:** For all experiments with SCIGAN, we used 5000 training iterations for the GAN network. This number of training iterations was chosen to ensure convergence of the generator loss, discriminator loss, as well as of the supervised loss. We will include details about the number of training iterations used in the paper. ■ **Issues with GAN training:** We have not encountered gradient vanishing problems when training our SCIGAN. It is not clear to us how the problem of mode collapse would even present itself in this setting as we are not discriminating between entirely real and entirely fake samples. ■ **Evaluation on real data and real-world applicability:** In real datasets, we only observe the outcome for the patient for a specific setting of the treatment and the dosage. The counterfactual outcomes, i.e. the patient outcomes under different possible interventions, cannot be observed. This is why it is not possible to use real data to evaluate how well the methods can estimate the entire dose-response curve for each patient. However, this does not mean that this method cannot be deployed in real world environments. In this regard, the problem is no different to the very well studied problem of treatment effect estimation for a binary/categorical treatment [6, 16, R2] for which there is a *wealth* of existing literature containing many examples of real-world applications. Evaluation on semi-synthetic data is standard for causal inference methods.

[R1] Dai, Zihang, et al. "Calibrating energy-based generative adversarial networks." preprint arXiv:1702.01691 (2017).
[R2] Jennifer L. Hill, "Bayesian nonparametric modeling for causal inference." JCGS, 2011