D BACKGROUND ON MODELING AID EFFECTIVENESS

Extensive work has focused on analyzing the effectiveness of development aid on macroeconomic growth [10, 20, 31]. Even other works study the link between development aid on specific SDG outcomes. For example, the authors in [6] examine the effect of education aid on primary education using generalized method of moments regression. [33] estimate the effectiveness of development aid on 8 SDG outcomes (i.e., poverty ratio, extreme hunger, school enrollment, HIV prevalence, gender parity at school, access to water, child mortality, and maternal mortality) using a panel fixed effects regression. Yet, as we detail below, a novel methodological approach is needed for our task.

Previously, different methodological approaches have been applied in order to study the effectiveness of development aid. Several works draw upon field experiments (e.g., [11, 12]). Here, the objective is to provide confirmatory evidence whether there is indeed a causal effect (and of what magnitude) and thus to understand whether aid was generally effective. However, experiments involving development aid are costly and limited to micro-level analyses, that is, the sample is designed for a small geographic area and not across multiple countries, as needed for our task. Furthermore, standard randomized controlled trials with two treatment arms would be limited to binary treatments and not treatment-response curves and, without additional assumptions, estimates are limited to *average* effects and not *heterogeneous* effects.

Other works use observational data (e.g., [6, 33]). Yet, these works typically make strong modeling assumptions. First, they assume that the effect is linear in the aid volume. Second, the true treatment effect is typically underestimated due to the nature of the modeling approach, and, hence, the estimate is merely a lower bound. Third, the aforementioned works assume that the effectiveness of aid cannot vary across countries but that the effectiveness is instead identical for all countries. As such, these works estimate the *average* effect of aid and not the *heterogeneous* effect of aid. To address this, we develop a tailored method for predicting between-country heterogeneity in treatment–response curves. Importantly, we use machine learning to avoid overfitting and thus to generate predictions that generalize well across countries.

To support decision-making, one needs to answer the question *"if volume a is spent on development aid, what would be the most likely outcome on the SDGs?"*. To the best of our knowledge, one work [24] uses traditional machine learning for that purpose, where the authors combine LASSO regression with mathematical optimization to inform the optimal allocation of development aid. We have included the method in [24] as a baseline. Throughout our manuscript, we refer to it as linear model (LM) where we vary order and regularization term (i.e., linear regression, LASSO regression, or ridge regression). Different from the LM, our framework is designed as a more flexible, non-linear approach. On top of that, the LM suffers from an important limitation: it is based on traditional machine learning, and not causal machine learning. In particular, this method does not account for treatment selection bias, and, as a result, it may give unreliable predictions of treatment effect.

To address the above limitations, a tailored machine learning framework is needed in order to provide more reliable predictions of heterogeneous aid–response curves. Across all of our experiments, we find robust evidence that this baseline is outperformed by our proposed machine learning framework. Thereby, we address that the question *"if volume a is spent on development aid, what would be the most likely outcome on the SDGs?"* is inherently linked to counterfactual inference where hypothetical questions are answered with regard to which outcome to expect. In order to generalize well across countries, our framework further makes extensive use of recent innovations in statistics and machine learning.

E COMPARISON WITH EXISTING METHODS FOR TREATMENT EFFECT ESTIMATION

Here, we briefly review existing methods for treatment effect estimation and point to salient differences between them and our **CG-CT**.

Methods for binary treatments: Prior literature on counterfactual inference makes nowadays increasing use of machine learning. This then allows to predict heterogeneous treatment effects. Here, prior literature has been strongly focused on the setting with binary treatments [2, 21, 26, 43, 57]. However, this is different from our work where we have continuous treatments.

Methods for continuous treatments: There are comparatively few methods that consider continuous treatments [5, 22, 42]. Here, key methods are as follows:

- *Generalized propensity score (GPS)* [22]. The GPS is a two-step estimation procedure. In the first step, conditional distribution of treatment given covariates is modeled under Gaussian distribution assumption. The estimated parameters are then used to compute the estimate for GPS. In the second step, conditional distribution of outcome given treatment and GPS is modeled as 2nd-degree polynomial regression with an interaction term between treatment and GPS.
- *Dose-response network (DRNet)* [42]. To predict heterogeneous treatment–response curves, DRNet provides a multi-layer neural network with a shared representation, multiple treatment "heads" (for each possible treatment), and multiple dosage "heads" within each treatment "head" (for splitting respective dosage interval of each treatment). The aim of the architecture with multiple treatment and dosage "heads" is to avoid treatment and dosage information being lost within representation layers of the neural network.
- *SCIGAN* [5]. SCIGAN is a method based on generative adversarial network (GAN). SCIGAN is motivated by the work of [57] where counterfactual generator is used to generate potential outcomes of different treatment assignments. In particular, SCIGAN uses the generator which takes observed data (i.e., observed outcome, treatment, dosage, and covariates), random treatment, random dosage, and random noise to generate a potential outcome for a random treatment and dosage (sampled uniformly from given treatments and dosages). The aim of the generator is to produce potential outcomes that 'fool' the discriminator which discriminates between outcomes of observed treatment and dosage (i.e., the factual outcome), and outcomes of random treatments

and dosages (i.e., counterfactual outcomes). Once the counterfactual generator is trained, an inference network (i.e., an artificial neural network) that predictes heterogeneous treatment–response curves is trained on combined original and generated data.

- *Methods based on (semiparametric) statistical theory* [8, 15, 27, 34]. There are several methods for obtaining *asymptotic* (large-sample) convergence guarantees and valid confidence intervals. Here, methods roughly fall into two main categories: (i) methods based on the efficient influence function of the dose-response curve, and (ii) methods based on Neyman orthogonal losses (double machine learning). In the context of (i), influence function-based methods aim at correcting for asymptotic plugin bias that occurs from estimation errors in the nuisance functions. Examples include the estimator proposed in [27], which performs a pseudo-outcome regression, and VCNet [34], which employs a functional targeted regularization. However, existing methods from (i) only consider *average* but not *individualized* dose-response functions. In the context of (ii), double machine learning-based methods such as [15] and [8] construct Neyman orthogonal loss functions, yet which are locally insensitive to small estimation errors in the nuisance functions. These methods rely on sample splitting to guarantee fast (asymptotic) convergence rates [7]. Due to these characteristics, we found them not to be effective for our work and instead opt for our machine learning framework called **CG-CT**.

Note that both DRNet and SCIGAN are designed for multiple treatments and a dosage interval for each treatment. However, since we only have one treatment (i.e., development aid), we have implemented DRNet without multiple treatment "heads", and SCIGAN without treatment discriminator.

Differences to our work: In comparison with the above methods, our **CG-CT** offers three key advantages in our HIV setting. (1) Our balancing autoencoder uses adversarial learning to learn a representation that is not predictive of the treatment (thereby addressing treatment selection bias) while simultaneously reducing the dimension of covariate space that we use to control for confounding. Removing selection bias is especially important in our high-dimensional *and* low sample setting, which is not accounted for by other baselines. In particular, semiparametric theory-based methods rely on sample splitting to provide guarantees for *asymptotic* convergence rates [7, 15], but sample splitting may harm the estimation performance in low sample settings [9]. (2) SCIGAN as a GAN-based method requires a comparatively large data sample to provide a good fit. This is evidenced by poor performance of SCIGAN in our HIV setting. Hence, our novel counterfactual generator offers a better solution for generating counterfactual outcomes in small sample settings. (3) Our choice of the inference model to be GPS over more complex models (e.g., DRNet) has proven to be more robust in a small sample size setting. This is expected given that DRNet also has a rather large model architecture because of shared representation and multiple dosage "heads" that require more data for a good fit.

F HYPERPARAMETER TUNING

To ensure a fair comparison, we use the same training procedure and hyperparameter tuning (where applicable) for both **CG-CT** and the baselines. In particular, the GPS baseline is identical to the GPS inside our **CG-CT**; that is, they both build upon the same polynomial regression. Hence, this has an important implication: all performance improvements of **CG-CT** over GPS must be attributed to the fact that our **CG-CT** handles the data in a better way (i.e., by addressing high-dimensional covariates and treatment selection bias). This is further confirmed empirically in our ablation studies (see Supplement L below), where we combine the balancing autoencoder and the counterfactual generator with other inference models. Therein, we demonstrate that our **CG-CT** achieves consistent performance gains across all inference models.

Hyperparameters are tuned via cross-validation with a 80/20 split. Table S2 shows the list of hyperparameters of both our **CG-CT** and the baselines. The hyperparameters and ranges for the baselines (i.e., a linear model (LM) and an artificial neural network (ANN)) are standard. For dose-response network (DRNet) and SCIGAN, we have used similar hyperparameters as proposed in [5, 42]; however, we adjusted some of the ranges to accommodate the much smaller sample size in our study (i.e., the layer size and batch size are adjusted based on the number of observations n). Also, we vary additional hyperparameters such as learning rate and dropout rate.

Importantly, when implementing **CG-CT** with baselines as inference models (in our main analysis and in supplements), we use the same hyperparameter tuning for the inference model as for the respective baseline. This ensures fair comparison between the **CG-CT** and the baselines.

Table S2: Hyperparameters and search ranges for CG-CT and baselines.

Method	Hyperparameter	Search range
Linear model (LM)	Order	0, 1, 2
	Regularization	0.05, 0.1, 0.5, 1, 5
Artificial neural network (ANN)	Layer size	53, 27, 14
	Learning rate	0.001, 0.0005, 0.0001
	Dropout rate	0, 0.1, 0.2
	Number of epochs	100, 200, 300
	Batch size	22, 11, 6
Generalized propensity score (GPS) [22]	–	–
Dose–response network (DRNet) [42]	Layer size	53, 27, 14
	Representation layer size	22, 11, 6
	Learning rate	0.001, 0.0005, 0.0001
	Dropout rate	0, 0.1, 0.2
	Number of epochs	100, 200, 300
	Batch size	22, 11, 6
	Number of dosage intervals (E)	5
SCIGAN [5]	Layer size	53, 27, 14
	Learning rate	0.001, 0.0005, 0.0001
	Dropout rate	0, 0.1, 0.2
	Number of epochs	100, 200, 300
	Batch size	22, 11, 6
	Number of dosage samples	3, 5, 7
	Factual loss trade-off (λ)	1
CG-CT	Layer size	14, 10, 7
	Representation layer size	10, 7, 4
	Learning rate	0.001, 0.0005, 0.0001
	Dropout rate	0, 0.1, 0.2
	Number of epochs	100, 200, 300
	Batch size	22, 11, 6
	Balancing parameter (θ)	0.05, 0.1, 0.5, 1, 5
	Regularization penalty (α)	0.05, 0.1, 0.5, 1, 5
	Number of dosage samples (m)	3, 5, 7

G COVARIATE DISTRIBUTIONS FOR DIFFERENT DEVELOPMENT AID VOLUMES

Here, we analyze the distribution of covariates (i.e., country characteristics that we control for) for different volumes of development aid in order to asses the presence of treatment-dependent covariate shifts in the observational data. Given that our treatment variable is continuous, we split our data in three groups (low, medium, and high) by using the empirical distribution of the observed development aid and the respective quantiles such that each group has the same probability mass. Hence, the group with (i) *low* volume of development aid contains all observations where development aid was below the $\frac{1}{3}-$quantile; the group with (ii) *medium* volume of development aid contains all observations where development aid was between the $\frac{1}{3}-$quantile and the $\frac{2}{3}-$quantile; and the group with (iii) *high* volume of development aid contains all observations where development aid was above the $\frac{2}{3}-$quantile.

Table S3 shows the mean and the standard deviation for each covariate across three treatment groups, i.e., low, medium, and high volume of development aid. We observe that treatment-dependent covariate shifts are present in the observational data, particularly among wealth indicators (e.g., GDP per capita, access to electricity) and health indicators (e.g., maternal mortality, infant mortality, life expectancy). Therefore, addressing the treatment selection bias through, e.g., the use of our balancing autoencoder, is crucial to ensure a reliable predictions of the heterogeneous treatment effect of development aid.

Table S3: Analysis of covariate distributions for different volumes of development aid. Reported are the summary statistics for the covariates in terms of the mean and the standard deviation (SD: standard deviation). The summary statistics are arranged across three treatment groups depending on the volume of development aid, i.e., low, medium, and high.

Treatment group	**Low**	**Medium**	**High**
Development aid range (in USD millions)	[0.003 – 2.649]	[2.649 – 15.301]	[15.301 – 559.897]
Covariates	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**
GDP per capita PPP (in USD thousands)	14.13 ± 6.60	7.34 ± 5.83	5.15 ± 4.27
GDP growth (in %)	3.69 ± 4.77	3.70 ± 2.73	4.05 ± 2.79
Foreign direct investment (in USD billions)	3.40 ± 6.26	3.57 ± 11.57	2.90 ± 7.78
Consumer price index - inflation (in %)	5.34 ± 5.84	5.20 ± 5.68	12.71 ± 31.11
Unemployment (in %)	8.76 ± 5.85	7.23 ± 5.48	7.50 ± 7.28
Population (in millions)	18.98 ± 28.40	29.39 ± 49.55	76.61 ± 226.20
Fertility (number of births per woman)	2.46 ± 0.76	3.38 ± 1.45	3.92 ± 1.27
Maternal mortality (number of deaths per 100,000 live births)	80.77 ± 72.43	263.71 ± 285.31	379.14 ± 268.43
Infant mortality (number of deaths per 1,000 live births)	17.91 ± 12.54	30.89 ± 20.86	44.04 ± 18.01
Life expectancy (in years)	73.49 ± 4.25	68.74 ± 6.52	63.25 ± 5.59
School enrolment ratio (primary education)	1.02 ± 0.09	0.99 ± 0.12	1.09 ± 0.15
Prevalence of undernourishment (in % of population)	9.13 ± 8.43	14.64 ± 13.13	17.94 ± 11.45
Access to electricity (in % of population)	94.55 ± 9.58	74.32 ± 29.98	52.62 ± 28.76
Incidence of tuberculosis (number of cases per 100,000 people)	92.58 ± 133.42	133.27 ± 122.45	261.03 ± 169.08

H VISUALIZATION OF BALANCING REPRESENTATION

We now provide explanatory insights into how the balancing representation in our **CG-CT** works. To this end, we plot the covariates for the country characteristics with vs. without balancing representation (i.e, with vs. without applying the balancing autoencoder in **CG-CT**).

Figure S2 shows 2D representations of covariates using t-SNE [53] (Figure S2a) and isomap [44] (Figure S2b). We illustrate representations in 2D for the original covariates (left) and for the balanced representation of covariates learned by the balancing autoencoder (right). We examine whether the balancing autoencoder induces treatment balance in the covariate representations. For this, we split the data in two treatment groups by median (shown in Figure S2 as red and blue).

Overall, we observe that the balancing autoencoder produces more homogeneous covariate representation with respect to treatment as compared to original covariates, thereby successfully addressing treatment selection bias.

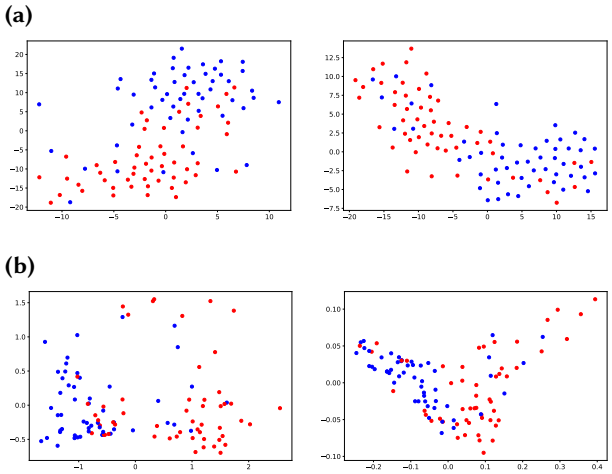


Figure S2: Illustration of balancing representations for country characteristics (covariates). We visualize the representations of covariates in 2D across different treatment groups, which we color in red and blue (by splitting the original treatment according to whether the value is above or below the median). On the left, we show original covariates and, on the right, the balanced representation of covariates obtained using the balancing autoencoder. In (a), we use t-SNE, and, in (b), we use isomap for visualization.

I EXAMINING THE VALIDITY OF UNDERLYING ASSUMPTIONS

Here, we conduct additional analyses to examine the validity of the underlying assumptions for identifiability of treatment effects, namely, positivity and ignorability.

Positivity assumption: The positivity assumption states that, for any covariates $X = x$, the probability of treatment $A = a$ is positive for every $a \in \mathcal{A}$. Formally, we have,

$$0 < p(A = a \mid X = x) < 1, \forall a \in \mathcal{A}, \text{if } p(x) > 0. \tag{10}$$

Validating whether the positivity assumption holds is generally impossible in real-world applications. However, we can use observational data to examine whether the probability estimates of a treatment given covariates lie in the range in which we predict the treatment effect, and should thus ensure reliable decisions. In our downstream decision-making problem where we optimize aid allocation, the range of aid volumes that we consider (i.e., the treatment) can vary from zero to $A_{\max} + \hat{\sigma}_A$, where $A_{\max}$ is the maximal value of observed aid for a single country, and $\hat{\sigma}_A$ is the estimated standard deviation for the aid variable A. Now, we examine the estimated probability of development aid over this range for six example countries that we have in the main paper. For this, we use a simple linear regression model to estimate the conditional mean of development aid given covariates and the standard deviation of the error term under a Gaussian distribution. We use the year 2016 for estimating the respective moments of the Gaussian distribution, and the year 2017 for evaluating the estimated probability curves. The estimated probability curves in Fig. S3 show that aid values over the given range are plausible for different countries, meaning that severe extrapolation is unlikely when predicting aid effects over this range. Hence, it is reasonable to expect that the positivity assumption is fulfilled.

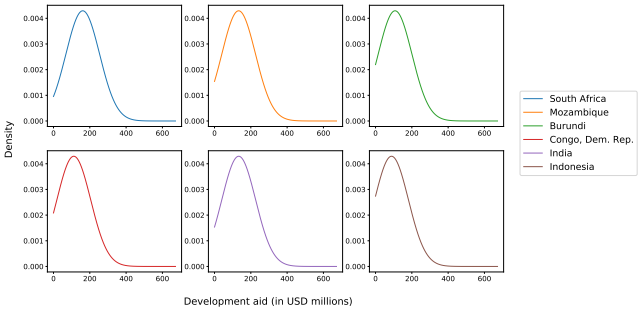


Figure S3: Estimated probability curves of development aid given country characteristics for the six example countries as in the results of the main paper: South Africa, Mozambique, Burundi, Congo, India, and Indonesia. The x-axis is set to range $[0, A_{\max} + \hat{\sigma}_A]$, where $A_{\max}$ is the maximal observed development aid and $\hat{\sigma}_A$ is the estimated standard deviation of development aid in 2017.

Ignorability assumption: The ignorability assumption states that potential outcomes $Y(a)$ are independent of the treatment variable A given covariates X, for every $a \in \mathcal{A}$. In other words, all variables that affect both treatment A and potential outcomes $Y(a)$ are measured as part of the covariates X. Therefore, ignorability is also referred to as "no hidden confounders" assumption. Formally, we have

$$Y(a) \perp\!\!\!\perp A \mid X = x \ \forall a \in \mathcal{A}.$$

Validating whether ignorability assumption holds is generally impossible in practice. This is a widespread problem in causal inference when predicting treatment effects from observational data. However, we can examine the robustness of the treatment effect estimate when adding additional variables regarding country characteristics in order to observe whether these variables are potentially hidden confounders or whether the results are robust. Specifically, we examine whether the aid effect coefficients α_1, α_2, and α_5 from the final stage of the GPS model change when a variable is added to country characteristics. To this end, we first estimate the probability density of estimates of these coefficients (i.e., $\hat{\alpha_1}, \hat{\alpha_2}$, and $\hat{\alpha_5}$) from our main analysis (i.e., when using only country characteristics as covariates) by performing the estimation over 10 runs. Then, we estimate the same coefficient when we add a new variable to the country characteristics and examine whether the estimated coefficients that represent the aid effect changes substantially. Here, we perform the analysis with the following additional variables: (i) past aid volume of a country from a year before; (ii) mean past aid volume of neighbor countries from a year before; (iii) past HIV infection rate in a country from a year before; and (iv) mean past HIV infection rate of neighbor countries from a year before. Thereby, we control for potential temporal dynamics and/or spillover effects across countries, which may be additionally relevant. The results in Fig. S4 show that predicted aid effect remain robust when adding any of these four covariates. This demonstrates that our set of country characteristics is sufficient to address potential confounding effects of these covariates. In the following section, we additionally provide a causal sensitivity analysis which shows that, even in the presence of unobserved confoudning, our results remain robust.

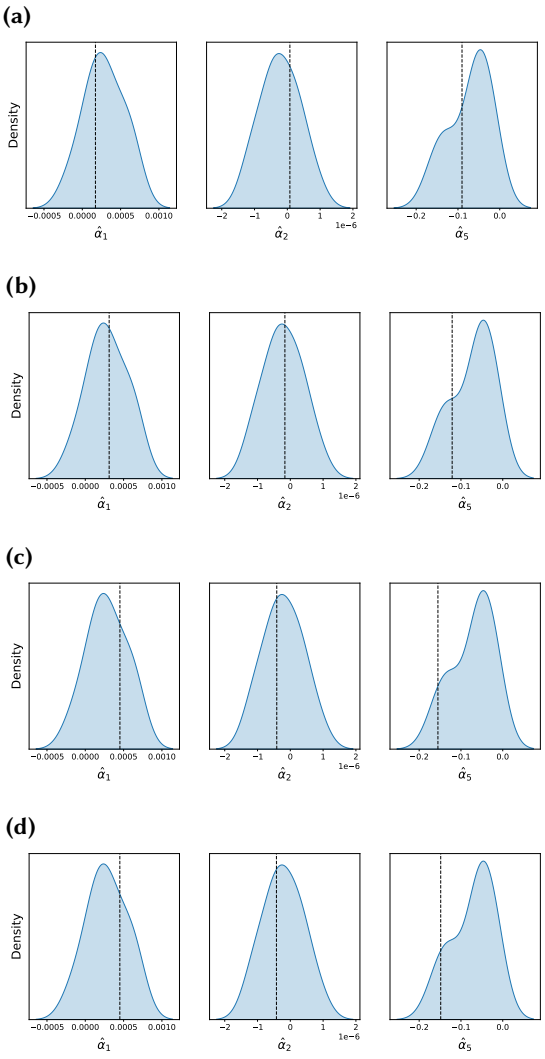


Figure S4: Robustness of predicted aid effect when adding additional control variables. Shown are the estimated density of aid effect coefficients $\hat{\alpha_1}, \hat{\alpha_2}$ and $\hat{\alpha_5}$ from the main analysis (i.e., when using only country characteristics as covariates) over 10 runs (blue area), and the observed value of the respective coefficient when adding an additional control variable (dashed black line). Here, we compare: (a) adding past aid volume of a country, (b) adding mean past aid volume of neighbor countries, (c) adding past HIV infection rate of a country, and (d) adding mean past HIV infection rate of neighbor countries.

J SENSITIVITY ANALYSIS FOR UNOBSERVED CONFOUNDING

We perform a sensitivity analysis of our results with respect to possible violations of the ignorability assumption. Thereby, we demonstrate that our results are robust to a certain level of unobserved confounding. To do so, we leverage recent advances in sensitivity analysis and use the method from Jesson et al. [25], which is state-of-the-art for continuous treatment. Jesson et al. use a continuous marginal sensitivity model (CMSM) that relaxes the ignorability assumption and then derive bounds around the treatment–response curve. For potential unobserved confounders U, the CMSM [25] assumes

$$\frac{1}{\Gamma} \leq \frac{f(a \mid x, u)}{f(a \mid x)} \leq \Gamma, \tag{11}$$

where $f(a \mid x, u)$ and $f(a \mid x)$ denote the conditional densities of the random variable A given $X = x$, $U = u$, or $X = x$, respectively. Here, the sensitivity parameter Γ controls the strength of allowed unobserved confounding.

To apply the method from [25], we need to predict the treatment–response curve $\mathbb{E}[Y \mid A = a, X = x]$ and the conditional outcome distribution $p(Y = y \mid A = a, X = x)$, which we model as a normal distribution with mean equal to $\mathbb{E}[Y \mid A = a, X = x]$ and constant variance. We use **CG-CT** to predict the treatment–response curve. We use different values of Γ that are informed by domain knowledge in that already several of the observed country covariates such as GDP, population size, etc. are known, important drivers of HIV infection rates. The results are shown in Fig. S5 and confirm that our predicted treatment–response curves remain stable under small violations of the ignorability assumption. Crucially, unobserved confounding even up to $\Gamma = 1.25$ cannot explain away our predicted treatment effect.

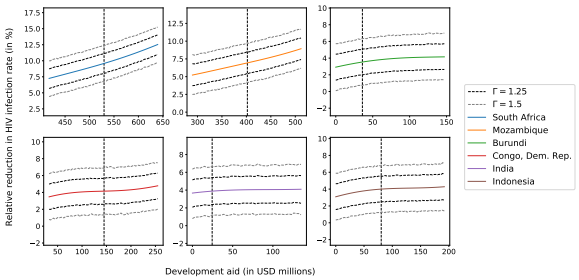
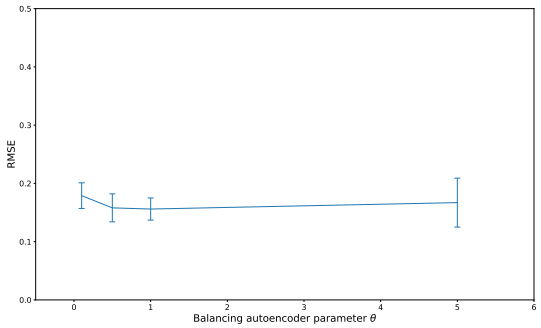


Figure S5: Bounds on predicted treatment–response curves for the six example countries as in the results of the main paper: South Africa, Mozambique, Burundi, Congo, India, and Indonesia. Vertical dashed line denotes the actual volume of development aid as observed in 2017. The x-axis is set to $A_{\text{obs}} \pm \hat{\sigma}_A$ (with a cut-off at zero to prevent negative values for Burundi and India), where A_{obs} is the observed development aid and $\hat{\sigma}_A$ is the estimated standard deviation of development aid in 2017. Shown are the predicted treatment–response curve (colored line) and sensitivity bounds for two levels of the strength of allowed unobserved confounding Γ (dashed black lines).

K SENSITIVITY ANALYSIS FOR HYPERPARAMETERS

As a sensitivity analysis, we evaluate the performance of our **CG-CT** in predicting treatment–response curves when varying hyperparameters. Here, our intention is to show that the performance remains at a similar level across various choices, which thereby adds to the robustness of our proposed method. Specifically, we vary two hyperparameters: the balancing parameter θ (in Figure S6a), and the number of generated counterfactual outcomes m (in Figure S6b). The results show that the performance of our **CG-CT** remains robust with respect to both hyperparameters.

(a)



(b)

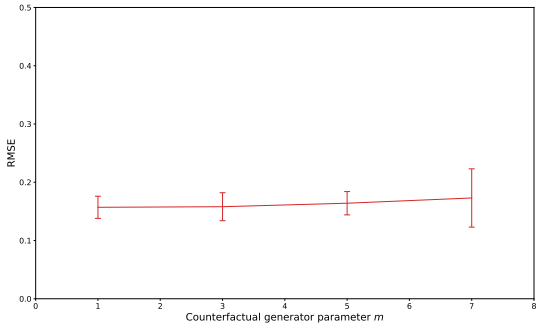


Figure S6: Sensitivity analysis to different hyperparameters. We visualize the performance of our CG-CT in predicting treatment–response curves using semi-synthetic data when varying the following two hyperparameters: in (a) the balancing parameter θ, and in (b) the number of counterfactual outcomes generated per data point m. We observe that the performance remains robust with respect to these two hyperparameters. Whiskers show the standard deviation estimates over 10 runs.

L SENSITIVITY ANALYSIS FOR CG-CT COMPONENTS

We now assess the contribution of different components in our **CG-CT** to the overall performance. To this end, we perform experiments (i) where we use other inference models, (ii) where we toggle the balancing autoencoder and the counterfactual generator on/off, and (iii) where we repeat the previous ablation study with other inference models. Hyperparameter tuning was done following the above for all experiments to allow for a fair comparison. The results are shown in Table S4. We discuss the findings in the following.

For (i), we vary the inference models as follows: dose-response-network (DRNet) [42], artificial neural network (ANN), linear model (LM), and generalized propensity score (GPS) [22]. The latter is used in our main analysis as it is the default in our **CG-CT**. The rationale was that it has a fairly parsimonious structure, which reduces the risk of overfitting while it offers the flexibility to handle various degrees of nonlinearities. We find that our **CG-CT** also works with other inference models. In fact, it consistently outperforms the baselines from the main paper, regardless of the underlying inference model. This thus confirms that our methodological innovations achieve consistent performance gains. Moreover, we find that our preferred choice for the inference model (GPS) as implemented in **CG-CT** is beneficial: it offers a superior performance with comparatively small variance.

For (ii), we toggle the balancing autoencoder and the counterfactual generator on/off. Here, we find that the performance gains are a result of combining both the balancing autoencoder and the counterfactual generator (as implemented in **CG-CT**). We also performed ablation studies with both components separately (i.e., only the balancing autoencoder or only the counterfactual generator). However, the experiments with just one of the two components did not lead to performance improvements in our **CG-CT**. Hence, the combination of *both* the balancing autoencoder *and* the counterfactual generator is important for achieving a state-of-the-art performance.

For (iii), we repeat the above ablation studies with other inference models. Here, we confirm that performance gains are achieved when combining *both* the balancing autoencoder *and* the counterfactual generator (whereas having only one of the two does not lead to performance improvements). Therefore, we find that our methodological innovations (i.e., the combination of both the balancing autoencoder and the counterfactual generator) leads to consistent improvements across all inference models. This corroborates the robustness of our contributions and shows the value of our methodological innovation for other inference methods.

Table S4: Sensitivity analysis for model components. We show the results of experiments with semi-synthetic data with different inference models and with different model components. We report square root of the mean integrated squared error (mean ± standard deviation averaged over 10 runs). Specifically, we vary the model components as follows: (i) without the balancing autoencoder and without the counterfactual generator (Base); (ii) only with the balancing autoencoder (BAE); (iii) only with the counterfactual generator (CF-GEN); and (iv) with both the balancing autoencoder and the counterfactual generator (CG-CT).

	Inference models			
	DRNet	ANN	LM	GPS
Base	0.281 ± 0.043	0.234 ± 0.030	0.209 ± 0.000	0.205 ± 0.000
BAE	0.303 ± 0.062	0.319 ± 0.105	0.192 ± 0.017	0.214 ± 0.009
CF-GEN	0.230 ± 0.032	0.176 ± 0.021	0.246 ± 0.003	0.210 ± 0.003
CG-CT	**0.181 ± 0.048**	**0.173 ± 0.018**	**0.159 ± 0.042**	**0.158 ± 0.024**

Lower = better

M EXPERIMENTS WITH OTHER TIME FRAMES

In our main analysis, we use HIV data from the year 2016 for learning, and the data from the year 2017 for evaluation. Here, we show robustness of our results in a different time frame, where we use HIV data from the year 2015 for learning and data from the year 2016 for evaluation. We run experiments with both semi-synthetic and real-world data, with the same baselines, performance metrics, and semi-synthetic data generation procedure as in our main analysis. Our results (in Table S5) show that our conclusions from the main analysis remain unchanged, i.e., our **CG-CT** offers state-of-the-art performance in experiments with both semi-synthetic and real-world data.

Reassuringly, we remind that we also experimented with multi-year data for learning; however, this did not led to significant out-of-sample performance improvements. Here, one explanation is that allocation practices are subject to changes over time (e.g., changes in priorities, changes in programs, funding mix) [59].

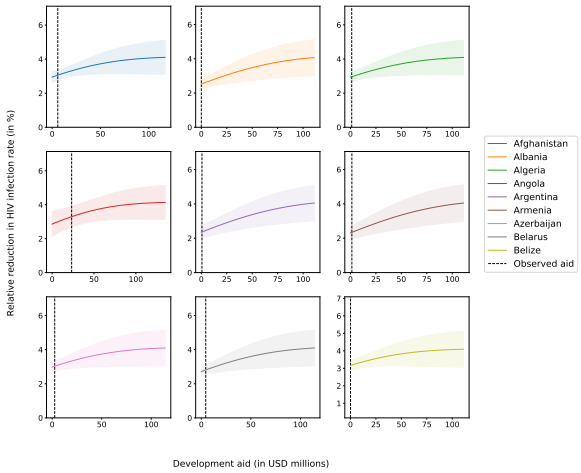
Table S5: Results of experiments with other time frames. We show the results of experiments with semi-synthetic and real-world data when using the year 2015 for learning, and the year 2016 for evaluation. For experiments with semi-synthetic data, we report the square root of the mean integrated squared error, and for experiments with real-world data, we report the root mean squared error (mean ± standard deviation averaged over 10 runs).

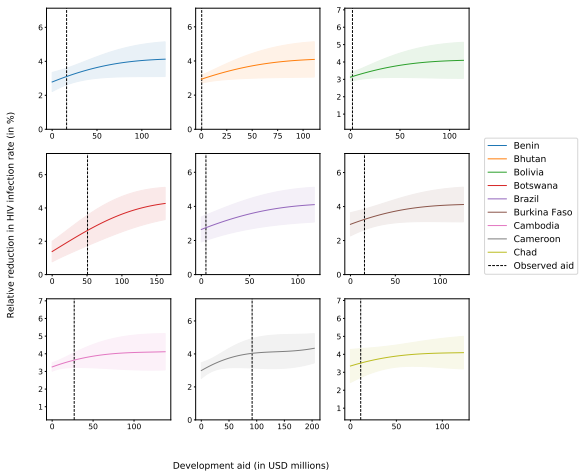
Data	Method					
	SCIGAN	DRNet	GPS	ANN	LM	**CG-CT**
Semi-synthetic	8.425 ± 3.629	0.255 ± 0.046	0.187 ± 0.000	0.221 ± 0.042	0.237 ± 0.000	**0.173 ± 0.023**
Real-world	2.397 ± 2.336	0.109 ± 0.019	0.090 ± 0.000	0.098 ± 0.010	0.087 ± 0.000	**0.087 ± 0.002**

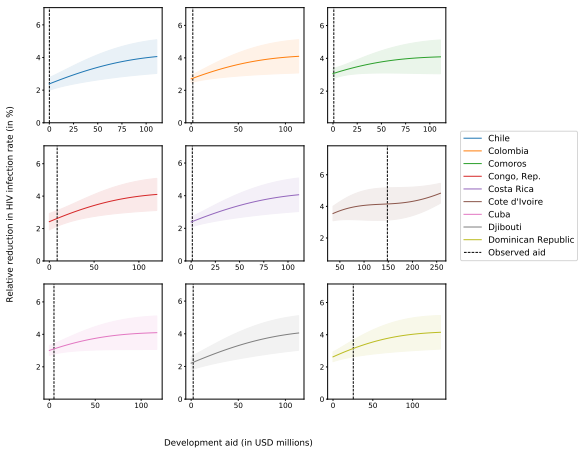
Lower = better

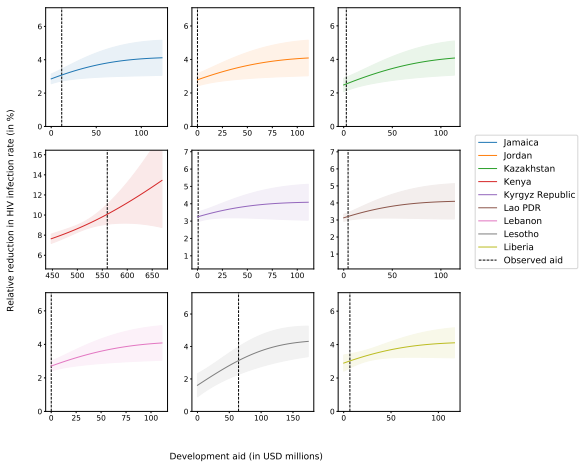
N PREDICTED AID–RESPONSE CURVES FOR DIFFERENT COUNTRIES IN 2017

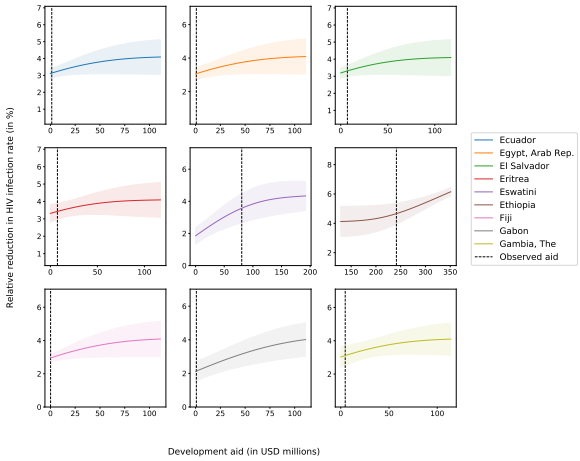
Here, we show the predicted aid–response curves for all remaining countries that are not shown in the main paper. The vertical dashed line denotes the actual volume of development aid as observed in 2017. The x-axis is set to $A_{\text{obs}} \pm \hat{\sigma}_A$ (with a cut-off at zero to prevent negative values), where A_{obs} is the observed development aid and $\hat{\sigma}_A$ is the estimated standard deviation of development aid in 2017. Shown are both point predictions (thick line) and the standard deviation over 10 runs (shaded area).

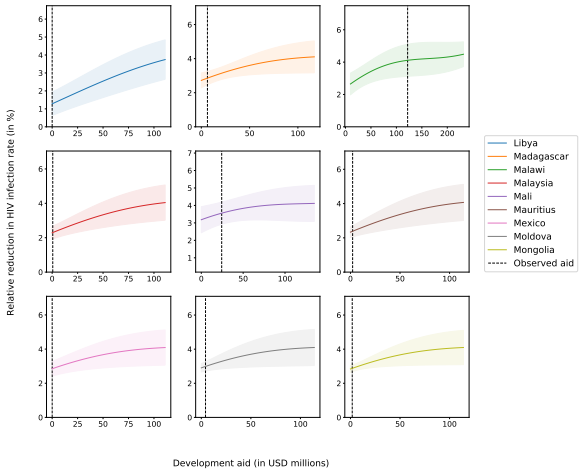


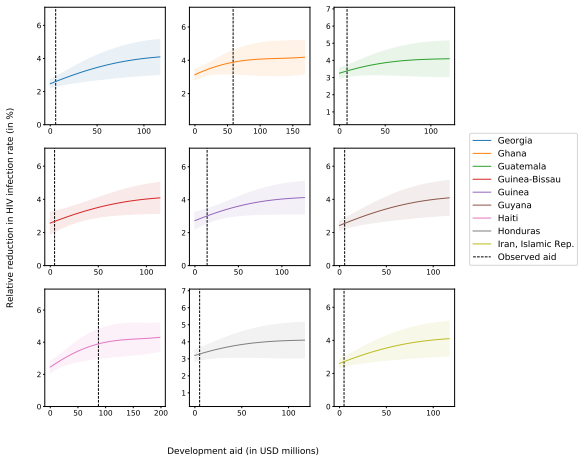


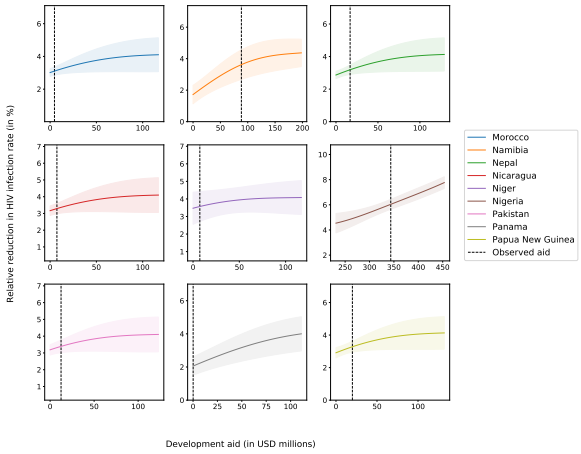

Relative reduction in HIV infection rate (in %)
Development aid (in USD millions)
Chile
Colombia
Comoros
Congo, Rep.
Costa Rica
Cote d'Ivoire
Cuba
Djibouti
Dominican Republic
Observed aid

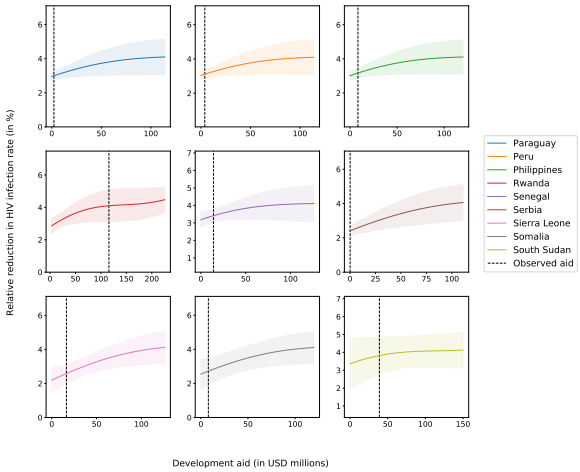

Relative reduction in HIV infection rate (in %)
Development aid (in USD millions)
Jamaica
Jordan
Kazakhstan
Kenya
Kyrgyz Republic
Lao PDR
Lebanon
Lesotho
Liberia
Observed aid

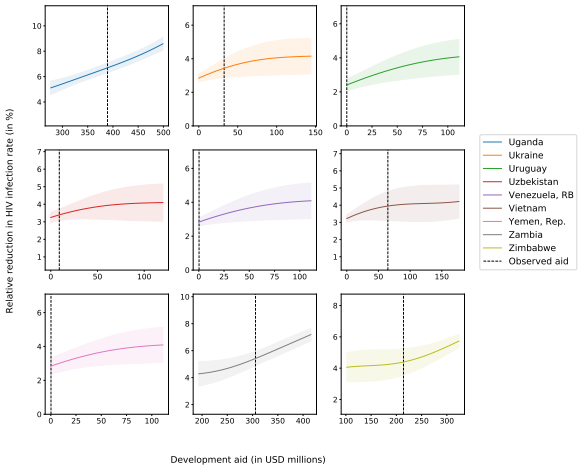

Relative reduction in HIV infection rate (in %)
Development aid (in USD millions)
Ecuador
Egypt, Arab Rep.
El Salvador
Eritrea
Eswatini
Ethiopia
Fiji
Gabon
Gambia, The
Observed aid

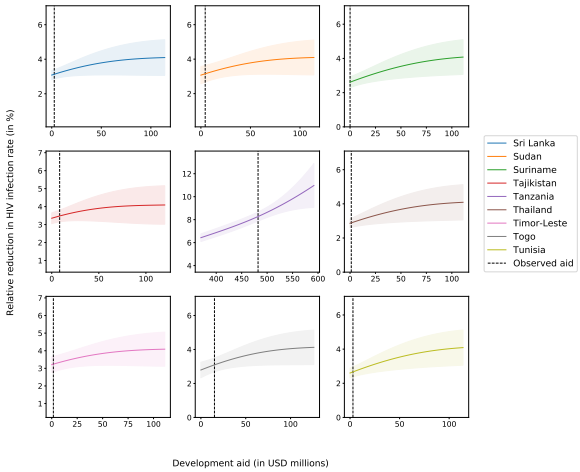

Relative reduction in HIV infection rate (in %)
Development aid (in USD millions)
Libya
Madagascar
Malawi
Malaysia
Mali
Mauritius
Mexico
Moldova
Mongolia
Observed aid


Relative reduction in HIV infection rate (in %)
Development aid (in USD millions)
Georgia
Ghana
Guatemala
Guinea-Bissau
Guinea
Guyana
Haiti
Honduras
Iran, Islamic Rep.
Observed aid


Relative reduction in HIV infection rate (in %)
Development aid (in USD millions)
Morocco
Namibia
Nepal
Nicaragua
Niger
Nigeria
Pakistan
Panama
Papua New Guinea
Observed aid


Relative reduction in HIV infection rate (in %)
Development aid (in USD millions)
Paraguay
Peru
Philippines
Rwanda
Senegal
Serbia
Sierra Leone
Somalia
South Sudan
Observed aid


Relative reduction in HIV infection rate (in %)
Development aid (in USD millions)
Uganda
Ukraine
Uruguay
Uzbekistan
Venezuela, RB
Vietnam
Yemen, Rep.
Zambia
Zimbabwe
Observed aid


Relative reduction in HIV infection rate (in %)
Development aid (in USD millions)
Sri Lanka
Sudan
Suriname
Tajikistan
Tanzania
Thailand
Timor-Leste
Togo
Tunisia
Observed aid