Counterfactually-Equivalent Structural Causal Modelling Using Causal Graphical Normalizing Flows

Sourabh Balgi SOURABH.BALGI@LIU.SE
STIMA, IDA, Linköping University, Sweden

Jose M. Peña JOSE.M.PENA@LIU.SE
STIMA, IDA, Linköping University, Sweden

Adel Daoud ADEL.DAOUD@LIU.SE
IAS, IEI, Linköping University, Sweden

Editors: J.H.P. Kwisthout & S. Renooij

Abstract

Recent research has highlighted the properties that deep-learning inspired causal models such as Deep-Structural Causal Model (Deep-SCM), Causal Autoregressive Flow (CAREFL) and Causal-Graphical Normalizing Flow (c-GNF) should exhibit to guarantee observational and interventional distribution equivalence with the true underlying causal data generating process (DGP), making them suitable for estimating average causal effect (ACE) or conditional ACE (CACE). However, for accurate individual-level causal effect (ICE) estimation and personalized treatment/public-policy formulation, it is crucial to ensure counterfactual equivalence between these models and the DGP. Firstly, we demonstrate that c-GNFs provide counterfactual equivalence under certain monotonicity assumption of the DGP, enabling precise ICE estimation and personalized treatment/public-policy analysis. Secondly, using this counterfactual equivalence of c-GNFs, we perform a counterfactual analysis and personalized public-policy analysis of the impact of International Monetary Fund (IMF) programs on child poverty using large-scale real-world observational data. Our results indicate a reduction in child poverty due to the IMF program at different personalization granularities. Our study also performs sensitivity analyses to assess potential threats to the unconfoundedness assumption and estimates ACE bounds and the E-value. This illustrates the potential of c-GNFs for causal and counterfactual inference in fields such as social, natural, and medical sciences.

Keywords: Causality; counterfactual reasoning; normalizing flows; social sciences.

1. Introduction

While many studies have enhanced our understanding of the factors that contribute to poverty among vulnerable groups, such as children, hindering their development (Banerjee and Duflo, 2011), there is a lack of knowledge on how to tailor public policies effectively to alleviate poverty for these vulnerable groups, especially during periods of macroeconomic instability (Halleröd et al., 2013; Kino et al., 2021). In other words, a crucial question is whether policymakers can customize policies to each child's specific circumstances rather than implementing a single policy for the entire population. This idea of personalization follows from a policy vision of tailoring interventions to an individual's needs and context for optimal outcome achievement. Specifically, Banerjee and Duflo (2011) state,

> "... We have to abandon the habit of reducing the poor to cartoon characters and take the time to really understand their lives, in all their complexity and richness."

For instance, the International Monetary Fund (IMF) is a powerful international organization responsible for promoting global macroeconomic stability. However, the impact of IMF programs on children is a subject of debate (Daoud et al., 2017; Daoud and Reinsberg, 2018; Daoud et al., 2019). One of the reasons is that, currently, government officials primarily rely on an individual's income and similar characteristics to identify vulnerable populations for social welfare programs. However, these programs often follow a simple rule: Eligible individuals receive a fixed "one-size-fits-all" policy, while ineligible individuals receive no policy support. Although this approach is transparent and applies the best average solution for the population (average causal effect or ACE) or group (conditional ACE or CACE), it lacks adaptability to individuals' specific needs, which is crucial for effectively addressing poverty, health issues, and other social problems (Potash et al., 2015; Ghani, 2018; Ye et al., 2019; Shiba et al., 2021b,a; Kino et al., 2021). To combat these challenges, policymakers need methods that can personalize public policies (Tabar et al., 2022). This requires moving beyond ACE/CACE estimation and employing counterfactual inference (Pearl, 2009). This enables personalized public policy analysis (P^3A), tailoring policies to the specific contexts of each individual child, also known as individual causal effect or ICE.

The main reason for the above mentioned lack of personalized policy making is that it requires answering counterfactual questions at a desired personalized granular level. Compared with causal questions, counterfactual questions are "*what if ..?*" questions that need a finer understanding/assumption of the underlying data generation process (DGP) or structural causal model (SCM) to be answered (Haavelmo, 1943; Goldberger, 1972; Ploch et al., 1975; Fienberg and Duncan, 1975; Pearl, 2009; Matsueda, 2012; Peters et al., 2017). Typically, the true DGP (which is unknown in most real-world applications) or at least a counterfactually equivalent SCM to the true DGP is requiered to answer counterfactual queries (Mooij et al., 2016; Peters et al., 2017). In other words, while there are machine learning algorithms for estimating ACE and CACE (Shalit et al., 2017; Künzel et al., 2019; Athey et al., 2019; Wodtke, 2020), there is a lack of similar methods for estimating ICE, as it involves modeling and estimating the unobserved causes that impact individual-level outcomes. Recent developments in deep-learning based generative models for causal inference, such as Neural Causal Model (NCM) (Xia et al., 2021), Deep-SCM (Pawlowski et al., 2020), Causal Autoregressive Flow (CAREFL) (Khemakhem et al., 2021), Causal-Graphical Normalizing Flow (c-GNF) (Balgi et al., 2022), and Causal Normalizing Flow (CNF) (Javaloy et al., 2023), have demonstrated the applicability of deep-learning based methods for causal effect estimation. Especially, c-GNFs stand out as they allow counterfactual inference since they are invertible by construction and, thus, they enable the estimation of individually specific unobserved causes.

Our current work makes the following main contributions:

1. We show that under certain monotonicity assumption of the underlying true DGP, c-GNFs are counterfactually equivalent to the unknown underlying causal system, enabling exact individual-level counterfactual analysis for personalized treatment/public-

policy prediction. Since the most common noise models in the literature, such as additive noise models (Hoyer et al., 2008) and post-nonlinear causal models (Zhang and Hyvärinen, 2009), are monotonic transformations of the respective noises, they satisfy our monotonicity assumption. Therefore, c-GNFs satisfy counterfactual equivalence under the assumption of such noise models and hence they are applicable for individual level counterfactual analysis and personalized treatments.

2. Equipped with the counterfactual equivalence, we apply c-GNFs to a large-scale real-world dataset on child poverty in IMF programs, recovering a counterfactually equivalent SCM using c-GNFs and thus conducting counterfactual analysis at different levels of personalization: ACE at the global level, CACE at the country level, and ICE at the individual (child) level. Our analysis achieves this personalization by combining micro (i.e., child and family conditions) and macro factors (i.e., politics, economy, and society) about the families and countries in which children reside when a government introduces an IMF program. This personalized approach aligns with the policy vision of tailoring interventions to individuals' needs and contexts.

The structure of our paper is as follows. Firstly, in Section 2, we introduce the necessary notations and define the problem. We also define the causal estimand of interest at different levels of personalization, i.e., the ACE, CACE, and ICE. In Section 3, we demonstrate that the c-GNF models an SCM that is counterfactually equivalent to the underlying causal DGP when certain monotonicity assumption holds. In Section 4, we present the experimental setup, followed by the analysis of the results, and a discussion on the assumptions and limitations. We also include a sensitivity analysis for the unconfoundedness assumption. Finally, in Section 5, we conclude the paper by summarizing our key contributions and highlighting their significance.

2. Problem Definition

Figure 1(a) presents the DAG of the social system under study in this work as derived by domain experts (i.e., social scientists) and prior work (Daoud and Johansson, 2024). In the figure, the green node `IT_prog_cgn` denotes the cause/policy/treatment of interest, and the blue node `CP_degree` denotes the effect/outcome of interest. Pink nodes denotes the observed confounders (i.e., causes of both the treatment and the outcome), and the blue nodes denote effect modifiers (i.e., causes of the outcome but not of the treatment).

In the social system under study, the degree of the child poverty `CP_degree` is calculated as the sum of seven binary individual dimensions of poverty: (i) education, (ii) health, (iii) information, (iv) malnutrition, (v) sanitization, (vi) shelter, (vii) water, resulting in an aggregate degree of child poverty between 0 and 7 with 0 indicating no poverty and 7 indicating severe poverty. The corresponding nodes are prefixed by `CP` in Figure 1(a). The set of macroeconomic factors confounding the causal effect of the IMF program include economic factors (prefixed by `EC`), political factors (prefixed by `PO`), political will (prefixed by `PW`) and fiscal factors (prefixed by `PS`). These factors respectively span a country's level of economic development/inflation/trade, democracy/laws/corruption/war, political will/motivation to implement IMF programs and institutional arrangements/spending. Finally, the set of microeconomic and family living condition factors acting as effect modifiers include factors

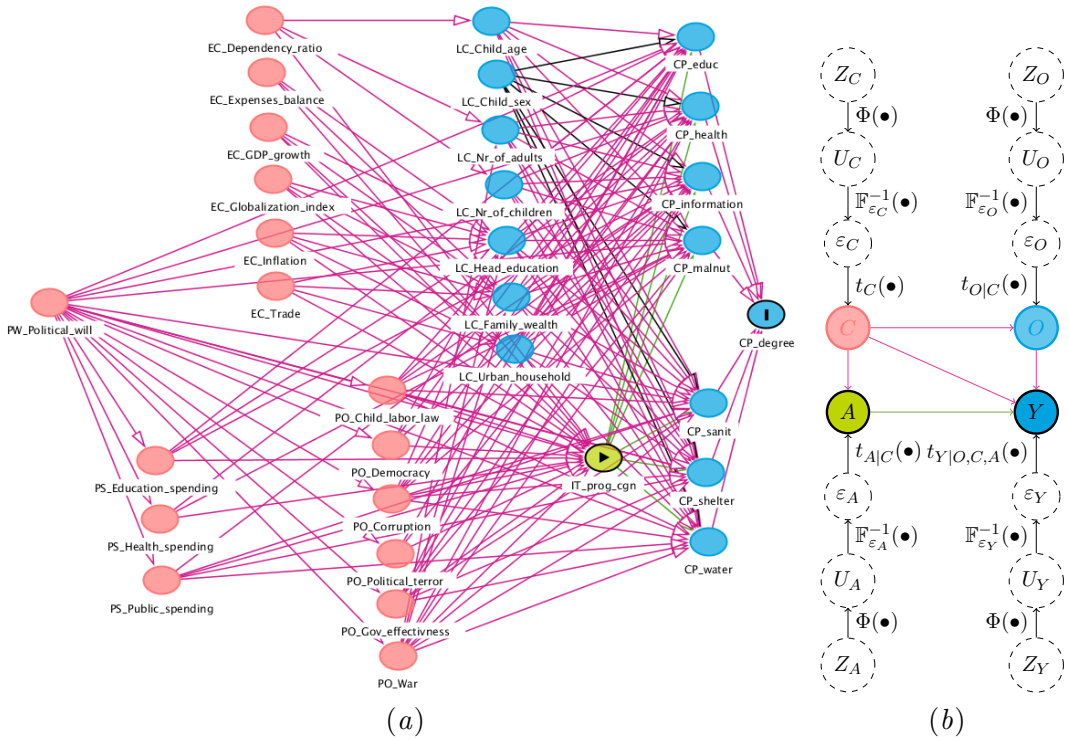


Figure 1: (a) Hypothesized DAG with the 32 observed variables (treatment, outcome and confounders/covariates) for the IMF causal system. (b) Conceptual DAG $\mathcal{G}$ for the IMF causal system. The solid-lined and the dash-lined nodes respectively indicate observed and unobserved variables.

such as child's age, sex and family wealth. These are prefixed by `LC`. Although these effect modifiers are unnecessary for adjusting given the macro-economic confounders, as we will explain later, they are critical to identify potential outcomes and personalized optimal treatment.

Figure 1(b) represents a conceptualization/abstraction of the DAG in Figure 1(a). In the conceptual DAG, A denotes the cause/policy/treatment of interest and Y denotes the effect/outcome of interest. The node C encompasses all the macroeconomic factors, the node O encompasses all the microeconomic and family living condition factors. The dashed nodes and arrows will be explained later.

As mentioned above, the DAG introduced in this section has previously been considered by Daoud and Johansson (2024) for heterogeneous treatment effect identification and estimation by means of generalized random forests (GRFs) (Athey et al., 2019). However, the authors made the following simplifying assumption for computational ease: They considered an indicator of child poverty obtained by thresholding the actual degree of child poverty so that degrees 2 to 7 indicate poor and degrees 0 and 1 indicate not poor. This assumption changes the expressions of ICE/CACE/ACE leading to wrong conclusions as we detail in Appendix A. Moreover, GRFs do not model the unobserved causes that are crucial for the counterfactual equivalence necessary for the ICE estimation, thus limiting P^{3}A that is of

interest to us. Also, GRFs lack the possibility to make the assumption of a detailed DAG as in Figure 1(*a*), which is also necessary for counterfactual equivalence. Our work has not such limitations.

2.1. Causal Estimands of Interest

Let $Y{:=}t_Y(A,C,O,\varepsilon_Y){=}t_{Y|A,C,O}(\varepsilon_Y)$ be the true DGP or structural equation for the outcome of interest Y in Figure 1(*b*) as a function of all its observed $\{A,C,O\}$ and unobserved ε_Y causes/parents such that the potential outcome under intervention $A{:=}a$ is $Y_a(C,O,\varepsilon_Y){:=}t_{Y|a,C,O}(\varepsilon_Y)$. Then, our causal estimands of interest are formally defined as follows.

$$\mathbf{ICE}(C,O,\varepsilon_Y){=}Y_1(C,O,\varepsilon_Y){-}Y_0(C,O,\varepsilon_Y) \tag{1a}$$
$$\mathbf{CACE}(X{=}x){=}\mathbb{E}_{\{C,O,\varepsilon_Y\}\setminus X=x}[\mathbf{ICE}(C,O,\varepsilon_Y)] \tag{1b}$$
$$\mathbf{ACE}{=}\mathbf{CACE}(\{\emptyset\}){=}\mathbb{E}_{\{C,O,\varepsilon_Y\}}[\mathbf{ICE}(C,O,\varepsilon_Y)] \tag{1c}$$

That is, the difference of the potential outcomes under interventions $A{:=}0,1$ provides the ICE in Eq. (1a), and marginalizing/averaging the ICEs over the respective population of interest provides the CACE in Eq. (1b) and ACE in Eq. (1c). The term 'average' in the ACE/CACE corresponds to the expectation over the non-conditioning causes $\{C,O,\varepsilon_Y\}\setminus X{=}x$ where $X{=}x$ denotes the conditioning set. Therefore, in the personalization granularity spectrum, we have that the ICE offers the highest personalization and the ACE that offers no personalization at the extremes, with the CACE offering conditional/intermediate personalization. Even though *do*-calculus (Huang and Valtorta, 2006; Shpitser and Pearl, 2006; Pearl, 2012) identifies the non-parametric expression for the ACE or CACE based on the interventional distribution expressions, *do*-calculus cannot be used to estimate the ICE as the knowledge of the structural equation or the exogenous noise of the SCM ε_Y for the outcome of interest Y is unobserved/unknown.

Our main objective in this work is to estimate the causal effect of the treatment (IMF program) on child poverty at any given population level or granularity, and personalize treatments at the desired granularity. Depending on the granularity, the instantiation of the causal effect of interest might vary, i.e., the ACE defines the causal effect when the granularity is the entire population level (Global-South), the CACE defines the causal effect when granularity is a sub-population level (e.g., country, age, gender, etc.), and the ICE defines the causal effect when personalization granularity is the individual child level. Since the IMF program in the observational dataset in our experimental application is personalized at the country level, country-wise CACE is considered for intermediate granularity level.

3. Counterfactual Equivalence with Causal-Graphical Normalizing Flows

In this section, we establish a condition under which a c-GNF is counterfactually equivalent to the underlying true SCM. Although we focus on the DAG in Figure 1(*b*), the result holds for any DAG.

Any given observed variable $X \in \{O, C, A, Y\}$ in Figure 1(*b*) may be expressed as a SCM with functional mechanism/transformation t_X, its observed parents/causes X_{pa}, and unobserved parent/cause ε_X, i.e.,

$$X := t_X(X_{pa}, \varepsilon_X) = t_{X|X_{pa}}(\varepsilon_X) \ . \tag{2a}$$

The unobserved/unknown cause ε_X in Eq. (2a) may follow any unknown arbitrary distribution that we may represent in terms of an uniform random variable U_X in $[0, 1]$ as $\varepsilon_X := \mathbb{F}^{-1}_{\varepsilon_X}(U_X)$, where $\mathbb{F}_{\varepsilon_X}$ denotes the cumulative distribution function (CDF) of ε_X. Hence, substituting for ε_X in Eq. (2a), we have

$$X := t_{X|X_{pa}}(\mathbb{F}^{-1}_{\varepsilon_X}(U_X)) \ . \tag{2b}$$

The uniform random variable U_X can further be represented in terms of the standard normal random variable Z_X, i.e., $U_X := \Phi(Z_X)$ where Φ denotes the CDF of the standard normal variable Z_X. Hence, substituting for U_X, the SCMs in Eqs. (2a)-(2b) may be equivalently written as

$$X := t_{X|X_{pa}}(\mathbb{F}^{-1}_{\varepsilon_X}(\Phi(Z_X))) \ . \tag{2c}$$

A graphical representation of the equation above can be seen in Figure 1(*b*). As Eqs. (2a)-(2c) are simple substitutions, they represent observationally, interventionally and counterfactually equivalent SCMs, which implies that the causal and counterfactual effects of these SCMs are the same.

We represent the SCM in Eq. (2c) as

$$X := \mathbb{T}_{X|X_{pa}}(Z_X) \ , \tag{3}$$

and we refer to it as encapsulated-SCM because it encapsulates the underlying true DGP.

Given that $\mathbb{F}^{-1}_{\varepsilon_X}$ and Φ are monotonic functions, suppose that $t_{X|X_{pa}}$ is also a monotonic transformation. Then, from the results of compositions of monotonic functions (Lorch and Newman, 1983), the composition $t_{X|X_{pa}} \circ \mathbb{F}^{-1}_{\varepsilon_X} \circ \Phi$ in Eq. (2c) is also monotonic and unique. Hence, $\mathbb{T}^{-1}_{X|X_{pa}}$ is an invertible transformation of an arbitrarily distributed observed variable X into a standard normal variable Z_X, i.e. a normalizing flow or c-GNF since we are in a causal setting (Balgi et al., 2022). Thus, a c-GNF represents a counterfactually-equivalent SCM to the underlying true SCM, under the assumption of a monotonic DGP $t_{X|X_{pa}}$. As the most common noise models in the literature, such as additive noise models (Hoyer et al., 2008) or post-nonlinear causal models (Zhang and Hyvärinen, 2009), are monotonic transformations of the respective noises, c-GNFs satisfy counterfactual equivalence for those models. Nasr-Esfahany et al. (2023) has derived similar results under similar conditions.

For compact notation, the encapsulated-SCM $\mathbb{T}_{X|X_{pa}} : Z_X \to X$ for the random variable X in Eq. (3) is extended to the random vector in the DAG in Figure 1(*b*) as $\mathbb{T}_{\mathcal{G}} : \mathbf{Z} \to \mathbf{X}$.

We model the c-GNF $\mathbb{T}_{\mathcal{G}}^{-1}$:$\mathbf{X}\rightarrow\mathbf{Z}$ as a deep-neural-network (DNN), specifically, as a Unconstrained Monotonic Neural Network (UMNN) (Wehenkel and Louppe, 2019) and the graphical conditioner from GNFs (Wehenkel and Louppe, 2021) for appropriately conditioning the transformation with the respective parent variables as defined in $\mathcal{G}$. The UMNN (Wehenkel and Louppe, 2019) transformer, a strictly monotonic integration based transformer ensures that $\mathbb{T}_{\mathcal{G}}^{-1}$ and $\mathbb{T}_{\mathcal{G}}$ are monotonic. Hence, c-GNFs are able to universally model any arbitrary data distribution and thus the underlying non-linear DGP (Huang et al., 2018). The DNN parameterised c-GNF $\mathbb{T}_{\mathcal{G}}^{-1}(\bullet;\theta)$ is trained as any other normalizing flow by maximizing the log-likelihood of the observational training dataset $\{\mathbf{X}^{\ell}\}_{\ell=1}^{N_{train}}$ (Kobyzev et al., 2021; Papamakarios et al., 2021; Wehenkel and Louppe, 2021), expressed as

$$\mathcal{LL}(\theta)=\sum_{\ell=1}^{N_{train}}\log\Big(f_{\mathbf{X}}(\mathbf{X}^{\ell};\theta)\Big)\quad, \tag{4}$$

$$f_{\mathbf{X}}(\mathbf{X}^{\ell};\theta)=f_{\mathbf{Z}}\left(\mathbb{T}_{\mathcal{G}}^{-1}(\mathbf{X}^{\ell};\theta)\right)*\left|\det\left(J_{\mathbb{T}_{\mathcal{G}}^{-1}(\mathbf{X}^{\ell};\theta)}(\mathbf{X}^{\ell})\right)\right|, \tag{5}$$

where θ denotes the parameters of the UMNN transformer and the graphical conditioner, which are optimized using minibatch stochastic gradient descent. From the construction of normalizing flows, the joint probability density function of a multivariate standard normal $f_{\mathbf{Z}}(\mathbf{Z}^{\ell})$ and the determinant of the Jacobian of the c-GNF $\det(J_{\mathbb{T}_{\mathcal{G}}^{-1}(\mathbf{X}^{\ell};\theta)}(\mathbf{X}^{\ell}))$ are computationally efficient to obtain, resulting in computationally efficient training of the c-GNF $\mathbb{T}_{\mathcal{G}}^{-1}$:$\mathbf{X}\rightarrow\mathbf{Z}$.

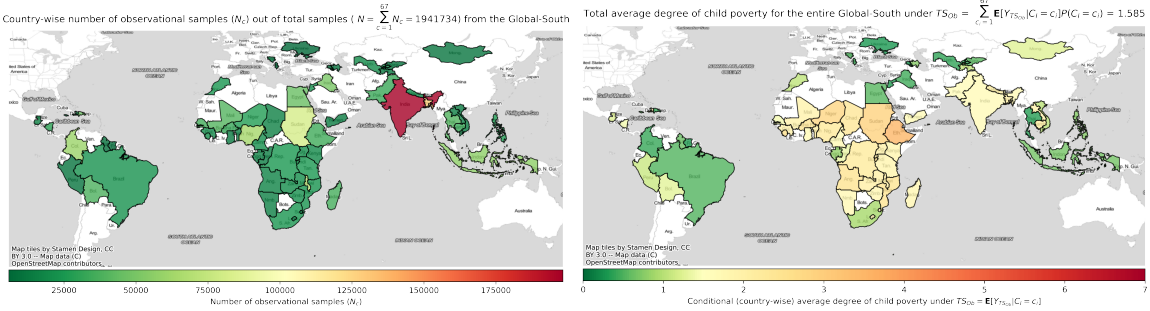
The c-GNF $\mathbb{T}_{\mathcal{G}}^{-1}$ or the corresponding SCM $\mathbb{T}_{\mathcal{G}}$ are invertible by construction, facilitating computationally efficient counterfactual inference using Pearl's first law of causal inference (Pearl, 1999, 2009). Specifically, let Y:=$\mathbb{T}_Y(A,C,O,Z_Y)$=$\mathbb{T}_{Y|A,C,O}(Z_Y)$ be the counterfactually-equivalent SCM of the outcome of interest Y, identified as explained above under the monotonicity assumption of the underlying true DGP of Y, i.e., Y:=$t_{Y|A,C,O}(\varepsilon_Y)$. Pearl's first law of causal inference provides three steps to identify the unknown potential outcome Y_a^{ℓ} of the ℓ^{th} individual under the intervention A:=a that is necessary to estimate the ICE, CACE and ACE in Eqs. (1a)-(1c):

(i) Abduction step: For the set of N individual observational samples $\{(y^{\ell},a^{\ell},c^{\ell},o^{\ell})\}_{\ell=1}^{N}$ from the given dataset, their respective individual unobserved exogenous noises $\{z_Y^{\ell}\}_{\ell=1}^{N}$ are identified using z^{ℓ}=$\mathbb{T}_{Y|a^{\ell},c^{\ell},o^{\ell}}^{-1}(y^{\ell})$.

(ii) Action step: The structural equation corresponding to the treatment/interventional variable A in Eq. (2c) is replaced with the desired interventional value A:=a to obtain the mutilated structural equation Y_a:=$\mathbb{T}_Y(a,C,O,Z_Y)$=$\mathbb{T}_{Y|a,C,O}(Z_Y)$.

(iii) Prediction step: The counterfactual $Y_a^{\ell}(C^{\ell},O^{\ell},Z_Y^{\ell})$ necessary for the ICE estimation in Eq. (1a) for the ℓ^{th} individual sample with respective observed causes $\{c^{\ell},o^{\ell}\}$ and unobserved cause $\{z_Y^{\ell}\}$ is computed with the following mutilated structural equation Y_a:=$\mathbb{T}_Y(a,C,O,Z_Y)$=$\mathbb{T}_{Y|a,C,O}(Z_Y)$ as Y_a^{ℓ}:=$\mathbb{T}_Y(a,c^{\ell},o^{\ell},z_Y^{\ell})$.

The ICE in Eq. (1a) for the ℓ^{th} individual is estimated as ICE^{ℓ}=Y_1^{ℓ}−Y_0^{ℓ}. The average causal effects such as the CACE in Eq. (1b) and ACE in Eq. (1c) are estimated using Monte-Carlo expectation estimation by averaging the ICE of all the individuals is the (sub)population of interest. That is, averaging the ICE of all the individuals in a country estimates the CACE for the specific conditioning country, while averaging over the entire



(a) (b)

Figure 2: (a) Country-wise distribution of the child population in the IMF dataset. (b) Average degree of child poverty in the factual world, i.e., under TS_{Ob}.

population estimates the ACE of the entire population. Note that it is sufficient to model counterfactually equivalently the structural equations of the variables on the causal paths from treatment to outcome, while the rest are irrelevant. In our DAG in Figure 1(a), this corresponds to the seven binary `CP` prefixed variables because, recall, the outcome `CP_degree` is just the sum of them.

4. Experimental Results

4.1. Experimental Setup and IMF Dataset Details

We consider the IMF child poverty dataset used by Daoud and Johansson (2024), Daoud et al. (2017), and Halleröd et al. (2013) with 1,941,734 children under the age of 18, residing in 67 countries from the Global-South. The benefit of this dataset is that it represents 2.8 billion children (50%) of the world's population by the year 2000. Due to the sensitive nature and the accompanying ethical considerations, the IMF dataset is not publicly available but it can be requested upon from the original authors. Figure 2(a) shows the country-wise distribution of the number of children in the dataset. Figure 2(b) shows the country-wise average degree of child poverty as observed in the dataset.

Since DNNs are prone to overfit, we split the data into 1,922,316 training, 9,709 validation and 9,709 test samples (as few thousand samples for the validation and test sets are typical in social sciences). We use only the training set samples for training and use the held-out validation set for early stopping to get the model with best validation loss. We further evaluate the generalization of the best validation loss model on the held-out test set. We use fully-connected layers with [40, 30, 20] hidden units for the graphical conditioner and fully-connected layers with [15, 10, 5] hidden units for the UMNN transformer. We consider the PyTorch (Paszke et al., 2017) baseline codes from c-GNFs[1] (Balgi et al., 2024, 2022), UMNNs (Wehenkel and Louppe, 2019), and GNFs (Wehenkel and Louppe, 2021) setting AdamW (Loshchilov and Hutter, 2019) optimizer with learning-rate=$3e{-}4$ and a batch-size of 1024 (4GB of GPU memory) for all our experiments. We rerun our experiments with five random seeded initializations to assess the robustness of the training and

1. https://github.com/cGNF-Dev

the model counterfactual predictions, and we report box-plots to show the variability of the results across simulations.

For any given personalization granularity, we aim to identify the optimal treatment to prescribe. In our analysis, we consider five treatment strategies at three personalization granularities. The granularities are Global-South (no personalization), country (intermediate personalization), and individual child (finest personalization) levels. The five treatment strategies are described as follows: (i) TS_{Ob} : The IMF program is encouraged (A:=1) or discouraged (A:=0) for the entire country based on the observed treatment, i.e., the countries treated get encouraged and the rest do not. (ii) TS_1 : The IMF program is encouraged (A:=1) for the entire Global-South. Then, there is no personalization at any granularity. (iii) TS_0 : The IMF program is discouraged (A:=0) for the entire Global-South. Then, there is no personalization at any granularity. (iv) TS_C : The IMF program is encouraged (A:=1) or discouraged (A:=0) or neither for the entire country based on the country's CACE in Eq. (1b). Then, there is personalization at country level. (v) TS_I : The IMF program is encouraged(A:=1) or discouraged (A:=0) or neither based on the child's ICE in Eq. (1a). Then, there is personalization at child level.

4.2. Analysis of the Experimental Results

Figure 3(a) indicates the country-wise average degree of child poverty in the counterfactual world under TS_1, i.e., all the countries in the Global-South receive IMF program. Similarly, Figure 3(b) indicates the country-wise average degree of child poverty in the counterfactual world under TS_0, i.e., none of the countries in the Global-South receives IMF program. Comparing Figures 3(a) and 3(b) indicates that the IMF program in expectation is beneficial in the poverty reduction when countries are prescribed the IMF program over not prescribing the IMF program. Figure 3(c) shows the boxplots of the differences of the counterfactual worlds under TS_1 and TS_0 indicating the CACE, over five random seeded simulations plotted along with the zero-ACE line for reference. Figure 3(c) reconfirms the beneficial nature of the IMF program in the Global-South across multiple simulations as the average degree of the child poverty is observed to be reduced by 1.2 ± 0.24 degrees. In Appendix A, we additionally present and discuss the degree-wise statistics of child-poverty across treatment strategies aggregated over five random seeded simulations.

Figure 3(d) indicates the country-wise average degree of child poverty in the counterfactual world under TS_C, i.e., there is optimal treatment personalization at country-level via the optimal treatment identified based on the CACE in Figure 3(c). Figure 3(e) indicate the country-wise average degree of child poverty in the counterfactual world under TS_I, i.e., there is optimal treatment personalization at individual child-level via the optimal treatment identified based on the ICE. From Figures 2(b) and 3, the treatment strategies can be sorted in the decreasing order of their expected degree of child poverty $\mathbb{E}[Y_{TS_a}]$ as follows: $TS_0 > TS_{Ob} > TS_1 > TS_C > TS_I$. This indicates that the IMF program is beneficial for the Global-South ($TS_0 > TS_{Ob} > TS_1$). The personalization at country level due to the treatment strategy TS_C represents a significant reduction in child poverty over the 'one-size-fits-all' treatment strategy TS_1 and the sub-optimal naturally observed treatment strategy TS_{Ob}. Moreover, personalization at the individual child level also exhibits further

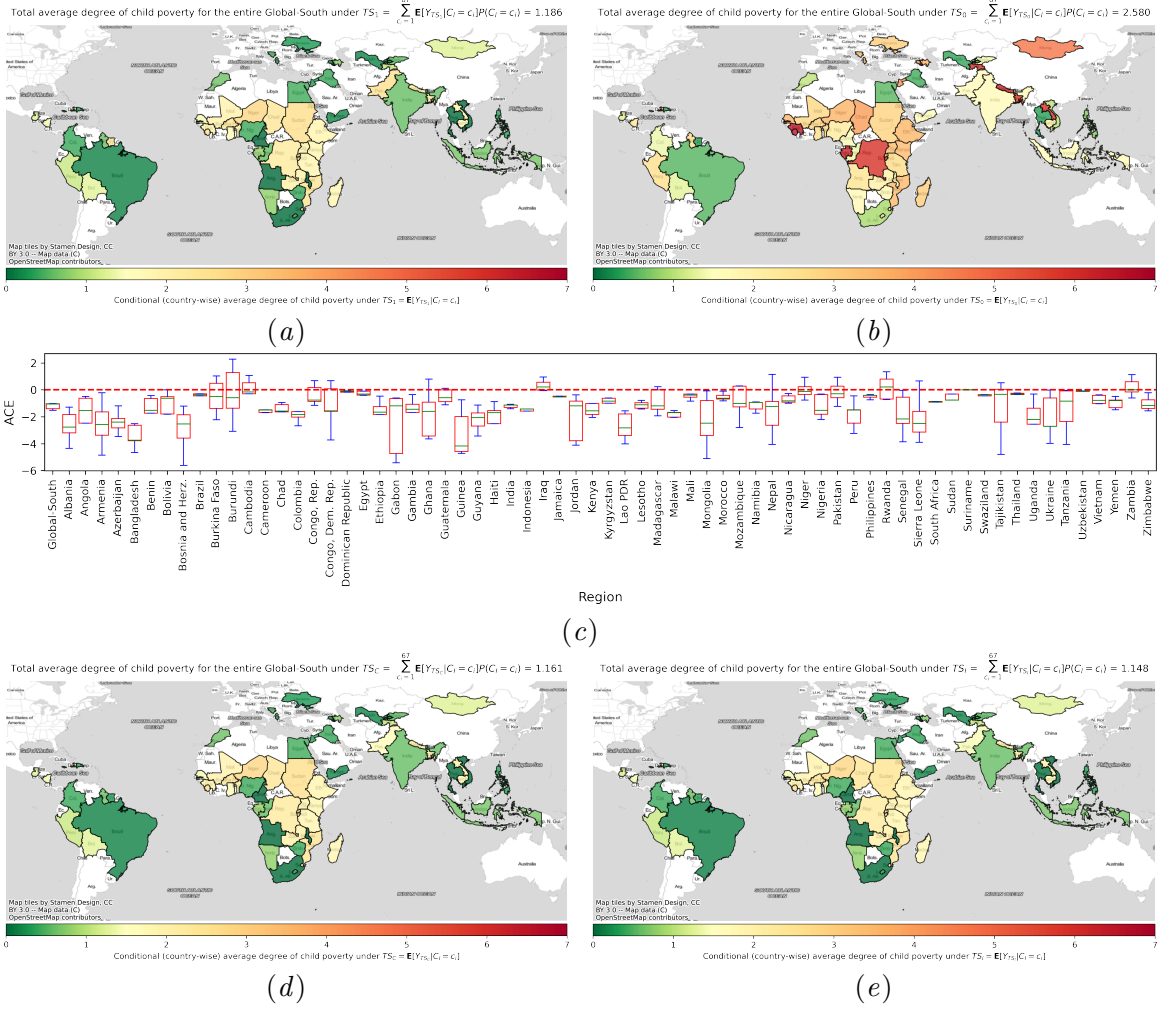


Figure 3: (a) Average degree of child poverty in the counterfactual world under TS_1. (b) Average degree of child poverty in the counterfactual world under TS_0. (c) Boxplots of the differences of the counterfactual worlds under TS_1 and TS_0 indicating the CACE, over five random seeded simulations plotted along with the zero-ACE line for reference. (d) Average degree of child poverty in the counterfactual world under TS_C. (e) Average degree of child poverty in the counterfactual world under TS_I.

benefits ($TS_C > TS_I$). Our results indicate the benefits of performing P^{3}A to effectively combat social ills in contrast to the 'one-size-fits-all' approaches.

4.3. Assumptions and Limitations

Performing counterfactual inference at the top rung in Pearl's ladder of causation necessitates the assumption of a causal DAG, which requires domain knowledge expertise (Pearl, 2009; Pearl and Mackenzie, 2018). Furthermore, it is necessary that the SCM is known and is counterfactually-equivalent to the underlying causal DGP, which requires additional

assumptions on the functional forms of the DGP. In this work, we showed the counterfactual equivalence of the DGP with the c-GNF, under the assumption of monotonicity of the observed effects with respect to their respective unobserved causes, an assumption satisfied by the most well studied and widely assumed models such as such as additive noise models (Hoyer et al., 2008) and post-nonlinear causal models (Zhang and Hyvärinen, 2009). Unconfoundedness or no unobserved confounders is also assumed in our work, as the DAG is expressed only respect to the observed confounders/variables. We present below a sensitivity analysis of our results to assess the impact of violations of the unconfoundedness assumption. The results obtained confirm our previous conclusions.

Causal effect identification from observational studies inevitably presumes certain unconfoundedness assumption in the DAG, e.g., between the treatment and the outcome in the case of back-door adjustment (Pearl, 2009) or between the treatment's parent and the outcome in the case of instrumental variables (Angrist et al., 1996) or between the treatment's child and the outcome in the case of the front-door criterion (Robins and Greenland, 1992; Pearl, 2014). However carefully one formulates the DAG, it is probable that some confounding variables might be missing. While unconfoundedness is an untestable assumption (Rubin, 1990; Robins and Hernán, 2008), sensitivity analysis enables scholars to evaluate the influence of such unmeasured confounding (Cornfield et al., 1959; Sjölander, 2020). Hence, we present here two sensitivity analyses: Assumption-free (AF) bounds (Robins, 1989; Manski, 1990) and the E-value (VanderWeele and Ding, 2017; Mathur et al., 2018).[2] None of the analyses makes use of our monotonicity assumption.

Unlike for binary outcomes, the AF bounds are not available for non-binary outcomes such as the categorical total degree of child poverty (degree 0: no poverty to degree 7: severe poverty). Since the degree of child poverty is formulated as the sum of seven binary individual dimension of child poverty, we may identify the AF bounds for each binary individual dimension of child poverty and extend the AF bounds to the total degree of child poverty as the sum of the respective lower and upper bounds. From the observational dataset, the AF lower and upper bounds for the seven binary individual dimensions of child poverty are identified as (i) education: $[-0.4843, 0.5157]$, (ii) health: $[-0.5009, 0.4991]$, (iii) information: $[-0.4893, 0.5107]$, (iv) malnutrition: $[-0.5111, 0.4889]$, (v) sanitization: $[-0.5012, 0.4988]$, (vi) shelter: $[-0.4756, 0.5244]$, (vii) water: $[-0.4738, 0.5262]$. Note that since the binary AF bounds are of width 1, and the total degree of child poverty is a sum of seven binary variables, the AF bounds for the total degree of the child poverty is of width 7. Our ACE estimate of -1.2 ± 0.24 is indeed within the bounds.

Further, the E-value (VanderWeele and Ding, 2017; Mathur et al., 2018) for the ACE estimate of -1.2 under the observed confounding is obtained as 5.47, i.e., a significant measure of unobserved confounding is required over the observed confounding to explain away the ACE of -1.2. This high E-value indicates that to explain away the identified ACE, it is required that the unobserved confounding to be significant. Based on the expertise of domain professionals, it is considered improbable that such strong unobserved confounding exists.

2. https://www.evalue-calculator.com/evalue/

5. Conclusion

In this article, we presented the monotonicity assumption of the underlying true causal data generating process necessary for counterfactual equivalence to the SCM modeled using normalizing flows, particularly c-GNFs. Utilizing this counterfactual equivalence of c-GNFs, we showcased a real-world application to draw counterfactual insights on the impact of the IMF program on child poverty from the observational data. Our findings in terms of ACE indicated the IMF program to be beneficial for the Global-South, in expectation. The sensitivity analysis via assumption-free bounds and the E-value confirmed our conclusions. Unlike the traditional 'one-size-fits-all' treatment, we proposed an empirical framework to formulate and identify personalized treatment strategies at different population granularity levels by computing the ACE, country-wise CACE, and child-wise ICE. Though we demonstrated the personalized treatment strategy formulation framework in a social science setting, the applicability of c-GNFs to the field of medical science for personalized medicine also follows.

References

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of Causal Effects using Instrumental Variables. *Journal of the American Statistical Association (JASA)*, 91(434): 444–455, 1996.

S. Athey, J. Tibshirani, and S. Wager. Generalized Random Forests. *The Annals of Statistics*, 47:1148–1178, 2019.

S. Balgi, J. M. Peña, and A. Daoud. Personalized Public Policy Analysis in Social Sciences Using Causal-Graphical Normalizing Flows. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 11810–11818, 2022.

S. Balgi, A. Daoud, J. M. Pena, G. T. Wodtke, and J. Zhou. Deep Learning With DAGs. *arXiv preprint arXiv:2401.06864*, 2024.

A. V. Banerjee and E. Duflo. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. Public Affairs, 2011.

S. R. Cole and C. E. Frangakis. The Consistency Statement in Causal Inference: A Definition or An Assumption? *Epidemiology*, 20(1):3–5, 2009.

J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions. *Journal of the National Cancer Institute*, 22(1):173–203, 1959.

A. Daoud and F. D. Johansson. The Impact of Austerity on Children: Uncovering Effect Heterogeneity by Political, Economic, and Family Factors in Low- and Middle-income Countries. *Social Science Research*, 118:102973, 2024.

A. Daoud and B. Reinsberg. Structural Adjustment, State Capacity and Child Health: Evidence From IMF Programmes. *Epidemiology*, 48(2):445–454, 2018.

A. Daoud, E. Nosrati, B. Reinsberg, A. E. Kentikelenis, T. H. Stubbs, and L. P. King. Impact of International Monetary Fund Programs on Child Health. *Proceedings of the National Academy of Sciences (PNAS)*, 114(25):6492–6497, 2017.

A. Daoud, B. Reinsberg, A. E. Kentikelenis, T. H. Stubbs, and L. P. King. The International Monetary Fund's Interventions in Food and Agriculture: An Analysis of Loans and Conditions. *Food Policy*, 83:204–218, 2019.

S. Fienberg and O. D. Duncan. Introduction to Structural Equation Models. *Journal of the American Statistical Association (JASA)*, 72:485, 1975.

R. Ghani. Data Science for Social Good and Public Policy: Examples, Opportunities, and Challenges. In *ACM Special Interest Group in Information Retrieval (SIGIR)*, pages 3–3, 2018.

A. S. Goldberger. Structural Equation Methods in the Social Sciences. *Econometrica*, 40: 979–1001, 1972.

T. Haavelmo. The Statistical Implications of a System of Simultaneous Equations. *Econometrica*, 11:1–12, 1943.

B. Halleröd, B. Rothstein, A. Daoud, and S. Nandy. Bad Governance and Poor Children: A Comparative Analysis of Government Efficiency and Severe Child Deprivation in 68 Low-and Middle-income Countries. *World Development*, 48:19–31, 2013.

P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear Causal Discovery with Additive Noise Models. *Neural Information Processing Systems (NeurIPS)*, 21, 2008.

C. Huang, D. Krueger, A. Lacoste, and A. C. Courville. Neural Autoregressive Flows. In *International Conference on Machine Learning (ICML)*, pages 2083–2092, 2018.

Y. Huang and M. Valtorta. Pearl's Calculus of Intervention is Complete. In *Uncertainty in Artificial Intelligence (UAI)*, 2006.

A. Javaloy, P. Sánchez-Martín, and I. Valera. Causal Normalizing Flows: From Theory to Practice. In *Neural Information Processing Systems (NeurIPS)*, 2023.

I. Khemakhem, R. P. Monti, R. Leech, and A. Hyvärinen. Causal Autoregressive Flows. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3520–3528, 2021.

S. Kino, Y.-T. Hsu, K. Shiba, Y.-S. Chien, C. Mita, I. Kawachi, and A. Daoud. A Scoping Review on the Use of Machine Learning in Research on Social Determinants of Health: Trends and Research Prospects. *Social Science & Medicine - Population Health (SSM-PH)*, 15:100836–100855, 2021.

I. Kobyzev, S. Prince, and M. Brubaker. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):3964–3979, 2021.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for Estimating Heterogeneous Treatment Effects using Machine Learning. *Proceedings of the National Academy of Sciences (PNAS)*, 116:4156–4165, 2019.

L. Lorch and D. J. Newman. On the Composition of Completely Monotonic Functions and Completely Monotonic Sequences and Related Questions. *J. London Math. Soc.(2)*, 28 (1):31–45, 1983.

I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

C. F. Manski. Nonparametric Bounds on Treatment Effects. *American Economic Review*, 80(2):319–323, 1990.

M. B. Mathur, P. Ding, C. A. Riddell, and T. J. VanderWeele. Website and R Package for Computing E-values. *Epidemiology*, 29(5):e45, 2018.

R. L. Matsueda. *Key Advances in the History of Structural Equation Modeling*. The Guilford Press, 2012.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing Cause From Effect Using Observational Data: Methods and Benchmarks. *Journal of Machine Learning Research (JMLR)*, 17(1):1103–1204, 2016.

A. Nasr-Esfahany, M. Alizadeh, and D. Shah. Counterfactual Identifiability of Bijective Causal Models. In *International Conference on Machine Learning (ICML)*, pages 25733–25754, 2023.

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research (JMLR)*, 22(57):1–64, 2021.

A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic Differentiation in PyTorch. *NeurIPS Workshops*, 2017.

N. Pawlowski, D. C. de Castro, and B. Glocker. Deep Structural Causal Models for Tractable Counterfactual Inference. In *Neural Information Processing Systems (NeurIPS)*, 2020.

J. Pearl. Probabilities of Causation: Three Counterfactual Interpretations and Their Identification. *Synthese*, 121(1):93–149, 1999.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2009.

J. Pearl. The *do*-Calculus Revisited. In *Uncertainty in Artificial Intelligence (UAI)*, page 3–11, 2012.

J. Pearl. Interpretation and Identification of Causal Mediation. *Psychological Methods*, 19 (4):459, 2014.

J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

D. R. Ploch, A. S. Goldberger, and O. D. Duncan. Structural Equations Models in the Social Sciences. *Social Forces*, 54:503, 1975.

E. Potash, J. Brew, A. Loewi, S. Majumdar, A. Reece, J. Walsh, E. Rozier, E. Jorgenson, R. Mansour, and R. Ghani. Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning. In *ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)*, pages 2039–2047, 2015.

J. M. Robins. The Analysis of Randomized and Non-randomized AIDS Treatment Trials using a New Approach to Causal Inference in Longitudinal Studies. *Health Service Research Methodology: A Focus on AIDS*, pages 113–159, 1989.

J. M. Robins and S. Greenland. Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology*, pages 143–155, 1992.

J. M. Robins and M. A. Hernán. Estimation of the Causal Effects of Time-varying Exposure. *Longitudinal Data Analysis (LDA)*, pages 553–599, 2008.

D. B. Rubin. Formal Mode of Statistical Inference for Causal Effects. *Journal of Statistical Planning and Inference (JSPI)*, 25(3):279–292, 1990.

U. Shalit, F. D. Johansson, and D. Sontag. Estimating Individual Treatment Effect: Generalization Bounds and Algorithms. In *International Conference on Machine Learning (ICML)*, pages 3076–3085, 2017.

K. Shiba, A. Daoud, H. Hikichi, A. Yazawa, J. Aida, K. Kondo, and I. Kawachi. Heterogeneity in Cognitive Disability After a Major Disaster: A Natural Experiment Study. *Science Advances*, 2021a.

K. Shiba, J. M. Torres, A. Daoud, K. Inoue, S. Kanamori, T. Tsuji, M. Kamada, K. Kondo, and I. Kawachi. Estimating the Impact of Sustained Social Participation on Depressive Symptoms in Older Adults. *Epidemiology*, 2021b.

I. Shpitser and J. Pearl. Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1219–1226, 2006.

A. Sjölander. A Note on a Sensitivity Analysis for Unmeasured Confounding, and the Related E-value. *Journal of Causal Inference (JCI)*, 8(1):229–248, 2020.

M. Tabar, W. Jung, A. Yadav, O. W. Chavez, A. Flores, and D. Lee. Forecasting the Number of Tenants At-Risk of Formal Eviction: A Machine Learning Approach to Inform Public Policy. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5178–5184, 2022.

T. J. VanderWeele. Concerning the Consistency Assumption in Causal Inference. *Epidemiology*, 20(6):880–883, 2009.

T. J. VanderWeele and P. Ding. Sensitivity Analysis in Observational Research: Introducing the E-value. *Annals of Internal Medicine*, 167(4):268–274, 2017.

A. Wehenkel and G. Louppe. Unconstrained Monotonic Neural Networks. In *Neural Information Processing Systems (NeurIPS)*, pages 1545–1555, 2019.

A. Wehenkel and G. Louppe. Graphical Normalizing Flows. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 37–45, 2021.

G. T. Wodtke. Regression-Based Adjustment for Time-Varying Confounders. *Sociological Methods & Research (SMR)*, 49(4):906–946, 2020.

K. Xia, K. Lee, Y. Bengio, and E. Bareinboim. The Causal-Neural Connection: Expressiveness, Learnability, and Inference. In *Neural Information Processing Systems (NeurIPS)*, pages 10823–10836, 2021.

T. Ye, R. Johnson, S. Fu, J. Copeny, B. Donnelly, A. Freeman, M. Lima, J. Walsh, and R. Ghani. Using Machine Learning to Help Vulnerable Tenants in New York City. In *ACM Computing and Sustainable Societies (COMPASS)*, pages 248–258, 2019.

K. Zhang and A. Hyvärinen. On the Identifiability of the Post-Nonlinear Causal Model. In *Uncertainty in Artificial Intelligence (UAI)*, 2009.

Appendix A. Additional Degree-wise Child Poverty Results and Analysis

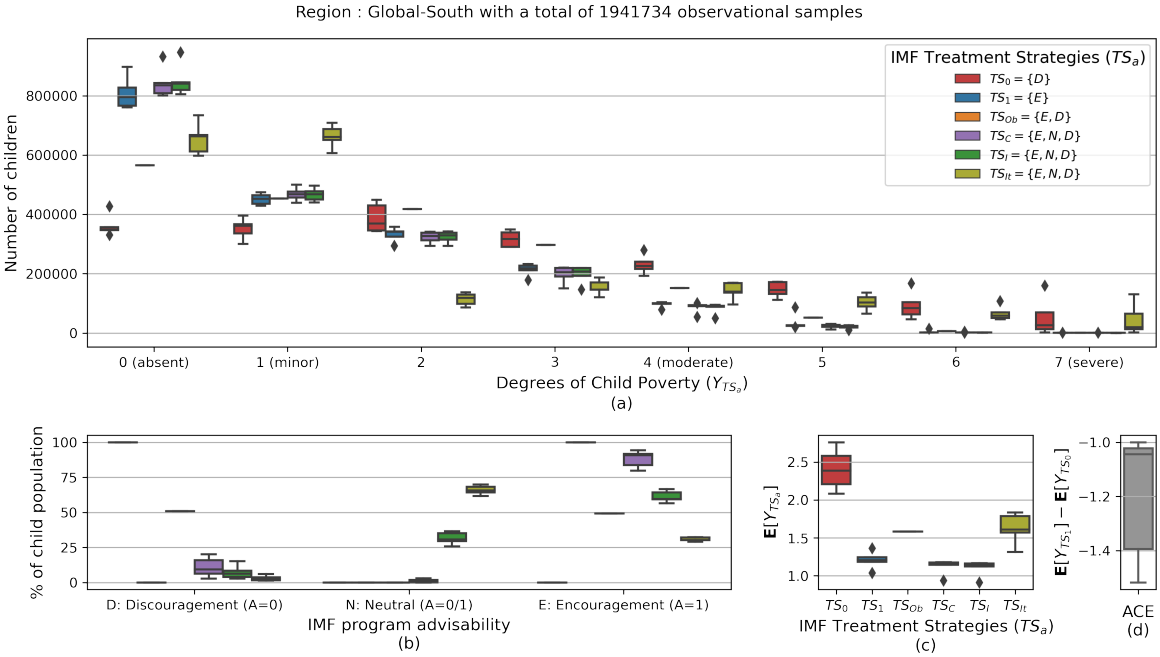


Figure 4: Boxplots over five random seeded simulations to assess c-GNF robustness across multiple simulations. (a) Number of samples in each of the 8 possible degrees of child poverty across different treatment strategies. (b) Advisability of the IMF program for each treatment strategy. (c) Average degree of child poverty for each treatment strategy. (d) The ACE estimate between treatments TS_1 and TS_0.

Figure 4(a) shows the child poverty degrees for the entire Global-South across multiple treatment strategies for multiple simulations. More importantly, Figure 4(b) shows the IMF program advisability statistics for the different treatment strategies. TS_I provides the finest personalization at the individual child level and we observe that, for this strategy on 1,941,734 children from the Global-South, the IMF program is harmful (hence, discouraged) for 7.25±4.89%, beneficial (hence, encouraged) for 61.22±4.08%, and neutral for the rest. Contrast to TS_I, TS_C provides intermediate personalization at country level and we observe that for this strategy, the IMF program is discouraged for 10.8±7.1%, encouraged for 88.1±6.04%, and neutral for the rest. With TS_C, since personalization is at the country level, we see an increase in the population that are discouraged and encouraged irrespective of the IMF programs being individually beneficial/harmful/neutral.

Figure 4(c) shows the variation of the average degree of child poverty for all the treatment strategies obtained from averaging Figure 4(a). From Figure 4(c), the treatment strategies can be sorted in the decreasing order of their expected degree of child poverty $\mathbb{E}[Y_{TS_a}]$ as $TS_0 > TS_{It} > TS_{Ob} > TS_1 > TS_C > TS_I$. This indicates that the IMF program is beneficial for the Global-South ($TS_0 > TS_{Ob} > TS_1$). Figure 4(c) shows that the personalization at the country level due to the treatment strategy TS_C (purple) represents a significant reduction in child poverty over the 'one-size-fits-all' treatment strategy TS_1 (blue) and the

sub-optimal naturally observed treatment strategy TS_{Ob} (orange). Moreover, personalization at the individual child level also exhibits further benefits ($TS_C > TS_I$). Note that the zero variance of the TS_0, TS_1 and TS_{Ob} is due to the fact that the respective treatments are held constant for these strategies. Figure 4(d) reconfirms the beneficial nature of the IMF program across multiple simulations as the average degree of the child poverty is observed to be reduced by 1.2±0.24 degree. Our results indicate the benefits of performing P^{3}A, to effectively combat social ills in contrast to the 'one-size-fits-all' approaches.

Note that for $A{:=}a^\ell$, we have $Y^\ell_{a^\ell}{=}y^\ell$ and this is referred as the consistency of potential outcomes, i.e., the potential outcome $Y^\ell_{a^\ell}$ under the factual treatment $A{:=}a^\ell$ is the same as the factual outcome y^ℓ (Cole and Frangakis, 2009; VanderWeele, 2009). The invariability of the boxplot of TS_{Ob} in Figure 4 (i.e., naturally observed treatment) validates the consistency assumption of the c-GNF model, as the same outcome is expected for the same observed treatments across multiple simulations.

Along with the five treatment strategies, we additionally consider the treatment strategy TS_{It} where the IMF program is encouraged ($A{:=}\,1$) or discouraged ($A{:=}0$) or neither based on the child's ICE with indicator thresholding of Y, i.e., $1_{[Y_a(C,O,Z_Y)>1]}$ from Daoud and Johansson (2024), i.e., degrees 2 to 7 indicate poor and degrees 0 and 1 indicate not poor, resulting in the modified ICE expression $\mathbf{ICE}_t(C,O,Z_Y){=}1_{[Y_1(C,O,Z_Y)>1]}{-}1_{[Y_0(C,O,Z_Y)>1]}$. Then, there is personalization at child level but based on the binarized Y as opposed to the actual degree Y. Our findings indicate the importance of considering all the seven poverty degrees in the analysis in contrast to Daoud and Johansson (2024), because the IMF program may help children move from severe to moderate poverty if not totally from severe to minor/absent poverty. This critical intra-degree improvement may get obscured if a binary indicator of poverty (poor vs non-poor) is used, leading to the erroneous conclusion that personalization of the IMF program is irrelevant. This fundamental oversight from over-simplification by grouping poverty degrees 2 to 7 neglects the improvements within the group, e.g., 7 to 2 is a significant improvement for which the IMF program will be rightly encouraged in TS_I. However, TS_{It} considers no change in the indicator of the poverty and hence wrongly assumes that the IMF program is neutral for resource optimization. In other words, TS_{It} wrongly values a change of degree from 2 to 1 more than a change of degree from 7 to 2. This is specifically seen in the advisability plots of TS_{It} in Figure 4(b) where, compared to TS_I, there is a preferential increase in the neutral advisability over the encouragement/discouragement alternatives. This important finding of ours reinforces the need of the radical rethinking proposed in Banerjee and Duflo (2011) that suggests one should not resort to over-simplification.