Sequential Treatment Effect Estimation with Variational Transformers: Application to COVID-19 Infection Clusters

Jinho Kang[1] , **Sungjun Lim**[2] , **Hojun Park**[3] , **Jaehun Jung**[3*] , **Jiyoung Jung**[1] , **Kyungwoo Song**[2*]
[1]Department of Artificial Intelligence, University of Seoul
[2]Department of Applied Statistics, Department of Statistics and Data Science, Yonsei University
[3]Artificial Intelligence and Big-Data Convergence Center, Gil Medical Center, Gachon University College of Medicine
{bubble3jh, jyjung}@uos.ac.kr, {lsj9862, kyungwoo.song}@yonsei.ac.kr, bacojun127@gmail.com, eastside1st@gmail.com

Abstract

Recent research focuses on integrating causal inference to enhance the explainability and robustness of deep learning models, especially for sensitive data like medical records. Existing methods often struggle with unobserved confounders and fail to incorporate exogenous variables. To address these issues, we propose a novel approach within Structural Causal Models (SCMs) that considers both unobserved confounders and includes exogenous variables to augment the dataset. We theoretically demonstrate that introducing exogenous variables enhances model robustness by linking it to the generalization bound. We also apply sequential treatment estimation through the known data-generating process to make the causal effect on potential outcomes more precise. We introduce the Causal Effect Variational Transformer (CEVT), a transformer-based causal model that incorporates our theoretical improvements into the Deep Neural Network (DNN) architecture, effectively capturing time-series data features. Using real COVID-19 infection cluster time-series data, we accurately estimate confirmed cases and infection duration while evaluating the treatment effect of government risk policies. Experimental results show that CEVT outperforms baseline models in estimation accuracy and remains robust in worst-case scenarios. The causality of treatment on outcomes is demonstrated by measuring treatment effects using domain knowledge and traditional metrics.

1 Introduction

There has been a notable increase in research endeavors aiming to incorporate causal inference for enhancing the explainability and robustness of deep learning models Deng *et al.* [2022]. This trend is particularly evident in the context of sensitive data, such as medical datasets, where the demand for such advancements is especially pronounced Sanchez *et al.* [2022]; Prosperi *et al.* [2020]. This surge in interest underscores the critical need to develop methodologies that not only predict but also provide insights into the underlying mechanisms of predictions, thereby ensuring both transparency and reliability in applications dealing with sensitive information.

To address this issue, we aim to analyze the impact of treatment on outcomes by mitigating the influence of confounders. In this study, we build upon the methodology of estimating latent confounder Z from previous work using a variational inference approach Louizos *et al.* [2017]. By capturing the interdependencies between Z and observed variables covariate X, treatment T, and potential outcome Y, the model extracts information about Z from proxy variables. We adapt this approach to our problem setting to infer the causal effect of treatment on outcome while mitigating confounding bias.

Furthermore, we assume the existence of an exogenous augmenting variable C in the given data, which is not estimable by the model. Under the assumption that C causally affects the latent confounder Z and the potential outcome Y, our model can more robustly learn the impact of treatment T on the potential outcome Y. We have proven this through the definition of a generalization bound and several assumptions, further demonstrating that appropriate augmentation can enhance the robustness of causal models.

Through the Data Generating Process (DGP) used in our real COVID-19 data experiments, we observe the direct treatment T_d, which causally affects the outcome, as well as the indirect treatment T_i. Here, T_i is a variable that influences T_d, representing an extension beyond the traditional treatment T. Assuming a causal path from T_i to T_d based on the known DGP, we have theoretically demonstrated that sequentially estimating the two causal variables T_i and T_d increases the probability of $P(Y|do(T))$. Furthermore, we validated the path from T_d to the outcome Y using causal discovery algorithms such as GES and LiNGAM on the data, as supported by prior research in the medical domain Chen *et al.* [2022]; Chuang *et al.* [2023]; Ghosh and Roy [2022] and detailed in Appendix B.3.

We propose the Causal Effect Variational Transformer, implemented in a DNN architecture, which incorporates sequential treatment effect estimation to actively leverage inductive bias and enhance training stability and considers exoge-

[*] Corresponding author

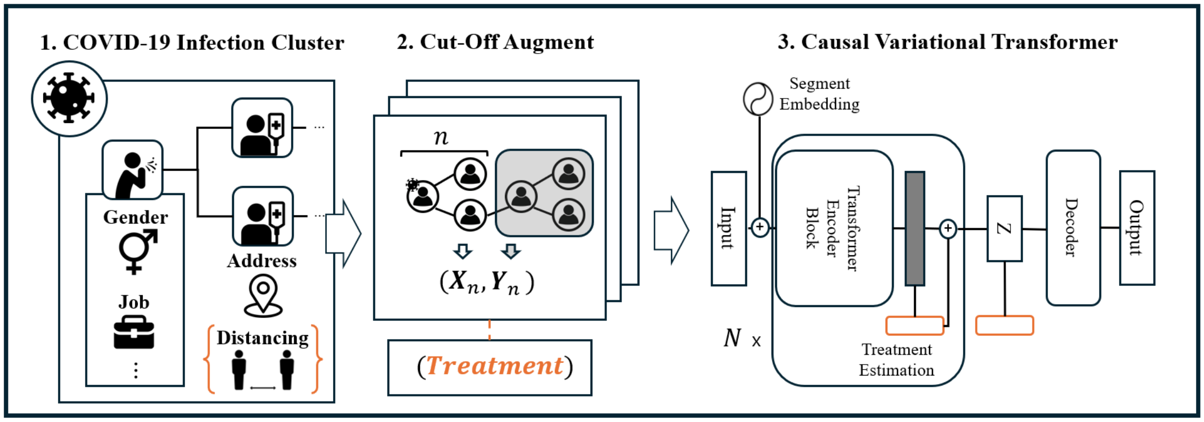


Figure 1: We collected COVID-19 infection cluster data along with various features and applied the Cut-Off Augmentation algorithm to generate a dataset with diverse feature-label pairs. From this dataset, government social distancing guidelines and risk indices were split as treatments. Subsequently, using a causal variational transformer that accounts for sequential treatments, we predicted the number of infections and the duration of each cluster.

nous augmentative variables to improve model robustness. Our theoretical analysis is validated with practical applications on two datasets: real-world COVID-19 infection cluster data and synthetic data that adhere to the assumed Structural Causal Model (SCM). Figure 1 presents an overall flowchart of our research's data collection, preprocessing, and cut-off augmentation and how these are applied to CEVT.

For the real-world COVID-19 dataset, governmental social distancing guidelines are set as the indirect treatment T_i and the risk index as the direct treatment T_d. When compared to conventional DNN baselines and Causal Neural Network baselines, the CEVT demonstrates superior performance in regression tasks predicting the number of confirmed cases in clusters and the duration of cluster persistence, achieving the best MAE and RMSE scores.

For the synthetic dataset, we can determine the ground truth of the relationships between variables and the treatment effects. Therefore, we compared the causal effect estimation performance of the baseline models and CEVT. We adapted the error metrics of Precision in the Estimation of Heterogeneous Effect (PEHE) and absolute error of Average Treatment Effect (ATE), which are commonly used in traditional causal models for binary treatment situations Louizos *et al.* [2017]; Shalit *et al.* [2017]; Shi *et al.* [2019], by modifying the formulas to fit the continuous treatment values. Compared to the baselines, CEVT performs the best for both $|ATE|$ and $PEHE$. Experimental results can be found in Appendix D

Our main contributions are as follows:

- We collect and appropriately preprocess and augment the COVID-19 dataset and theoretically prove that the Cut-Off augmentation technique can enhance the robustness of the learning model.
- We propose the Causal Effect Variational Transformer, a DNN architecture that incorporates sequential treatment effect estimation to leverage inductive bias and enhance training stability.
- We extend the $PEHE$ and $|ATE|$ metrics to measure causal effects for continuous treatment variables and experimentally demonstrate the estimation performance of CEVT using these adapted metrics.

2 Related Works

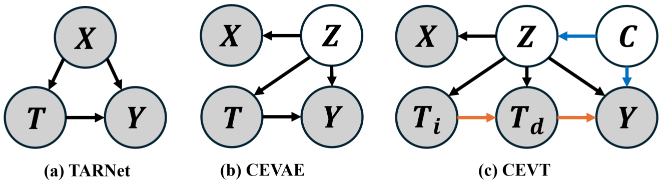
2.1 Causal Inference

Causal inference often involves estimating the causal effect using the Average Treatment Effect (ATE), which measures the expected difference in outcomes between treated and untreated groups Pearl [2022]; Wager and Athey [2018]; Johansson *et al.* [2016]. where $do(\cdot)$ denotes $do\text{-}calculus$ Pearl [2009], a formal framework used to derive causal effects from observational data by manipulating expressions involving interventions. Causal inference methods like Instrumental Variables (IV) Hartford *et al.* [2017]; Reiersøl [1945]; Angrist *et al.* [1996], Frontdoor Adjustment (FA) Zhang *et al.* [2024]; Xu *et al.* [2023], and Backdoor Adjustment (BA) Pearl [2009]; Louizos *et al.* [2017]; Israel *et al.* [2023] help determine the treatment effect on outcome. In detail, BA controls for confounders that affect both treatment and outcome, blocking non-causal paths. It marginalizes the confounders to isolate the direct causal effect.

On the one hand, proxy variables Wickens [1972]; Frost [1979] are often used when direct observation of a variable is impossible, as substitutes that are correlated with the unobservable variable. As the model and data get larger, proxy variables are promising ways to unmeasured confounders Montgomery *et al.* [2000]; Stock and Watson [2020]. However, no prior research on indirect treatment exists. We focus on situations with known causal paths and expect the inductive bias to tighten the estimate of $P(Y|do(T))$.

2.2 Causal Neural Network

Studies that estimate causal effects in DNNs are progressing Johansson *et al.* [2016]; Shalit *et al.* [2017]; Yoon *et al.* [2018]; Alaa and Van Der Schaar [2017]. TarNet Shalit *et al.* [2017] assumed covariate X as a confounder to estimate $P(Y|do(T))$. Dragonnet Shi *et al.* [2019], based on theoretical support, proposed an architecture to predict propensity



(a) TARNet **(b) CEVAE** **(c) CEVT**

Figure 2: Each subfigure represents the SCM assumed in previous studies and this study. (a) In the case of TARNet, the covariate X is assumed to be a confounding variable. (b) In CEVAE, X is assumed to be a proxy variable for the unobserved latent confounder Z. (c) In CEVT, a multi-treatment scenario indicated by the orange path is assumed to enhance the causal inference performance on y, and an augmentation exogenous variable indicated by the blue path is added to analyze the impact of augmentation through the Cut-Off algorithm on the robustness of the model.

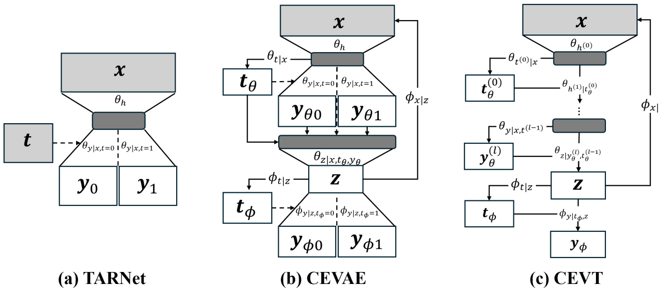
scores and conditional outcomes through an end-to-end process. CEVAE Louizos *et al.* [2017] used X as a proxy to estimate unobserved confounder Z with a VAE and then estimated $P(Y|do(T))$. These models assumed binary treatment and divided the NN's weight space accordingly, leading to structural limitations. First, even if a discrete variable causing the treatment is known, the T^2 weight parameter needs to integrate the binary treatment into the causal estimation. Second, they are not able to handle the continuous treatment properly. Our proposed CEVT, a modified version of CEVAE, overcomes these two weaknesses by predicting sequential treatment estimation.

3 Methodology

In Section 3, we introduce the structure and objective function of our proposed model, designed to accommodate multiple causal variables. Additionally, we describe the preprocessing steps and the attributes of the real-world COVID-19 dataset used to apply and validate our model.

3.1 Causal Effect Variational Transformer

We introduce our proposed method, the Causal Effect Variational Transformer (CEVT). CEVT assumes a novel problem setting through SCMs that were not considered in prior works. Figure 2 depicts the SCMs in previous studies and in CEVT. In (a) TARNet and (b) CEVAE, the simplest form of confounder situations is assumed, and the goal is to calculate the treatment effect under this assumption. Inspired by the Causal Effect Variational Autoencoder (CEVAE) Louizos *et al.* [2017], CEVT estimates the posterior distribution of latent confounders Z using a variational approach while assuming additional problems. This includes enhancing learning stability and performance through the orange causal path in Figure 2 and improving model robustness through the blue path in the figure. Detailed theoretical analysis of this approach is provided in Section 3.1.



(a) TARNet **(b) CEVAE** **(c) CEVT**

Figure 3: (a), (b), and (c) sequentially visualize the neural network structures of TARNet, CEVAE, and CEVT. In TARNet and CEVAE, the weight space of the neural network is divided based on binary treatment, and if multi-treatment is applied to the model, the weight space is exponentially partitioned. CEVT overcomes this by using an iterative conditioning method that passes through the transformer layer.

Furthermore, CEVT extends this idea by incorporating a novel Transformer encoder architecture Vaswani *et al.* [2017] and an iterative conditioning mechanism, enabling more effective estimation of individual treatment effects in the presence of multiple treatments. Unlike previous approaches that require separating parameters based on treatment assignment—resulting in a weight space divided by the square of the number of treatment variables, as seen in Figure 3, our proposed architecture stacks Transformer layers and employs an iterative conditioning method. This not only allows for efficient use of weights but also ensures greater stability in causal effect estimation. It theoretically enables infinite utilization of the inductive bias from the known DGP, leading to improved performance. For convenience and considering the SCM of real COVID-19 data, we defined two treatment variables and referred to them as direct treatment T_d and indirect treatment T_i.

When dealing with continuous treatments, the individual treatment effect (ITE) is defined as the expected change in the outcome for a unit change in the treatment while keeping the other treatment fixed. Formally, for a direct treatment t_d and an indirect treatment t_i, the ITE is given by:

$$\text{ITE}(x) := \frac{\partial}{\partial t_d} E[Y|X = x, \text{do}(T_d = t_d), T_i],$$

where x represents the observed covariates, y is the outcome, and $\text{do}(\cdot)$ denotes the do-operator for intervention Pearl [2009].

To estimate the causal effects from observational data in the presence of unobserved confounders, we assume a latent variable model following the causal graph in Figure 2-(c). The key idea is to simultaneously learn the unknown latent confounders Z and the causal effects by maximizing an objective function consisting of the evidence lower bound (ELBO) and auxiliary losses for the direct treatment T_d, indirect treatment T_i, and outcome Y.

Let L denote the ELBO:

$$\begin{aligned} L = \sum_{n=1}^{N} \mathbb{E}_{q_\phi(z_n|x_n,t_{d,n},t_{i,n},y_n)} \Big[& \log p_\theta(x_n, t_{d,n}, t_{i,n}|z_n) \\ & + \log p_\theta(y_n|t_{d,n}, t_{i,n}, z_n) \\ & + \log p(z_n) - \log q_\phi(z_n|x_n, t_{d,n}, t_{i,n}, y_n) \Big]. \end{aligned}$$

The ELBO consists of the log-likelihood of the covariates X, direct treatment T_d, and indirect treatment T_i given the latent confounders Z; the log-likelihood of the outcome Y given the treatments and latent confounders; the log prior probability of the latent confounders; and the negative log variational posterior of the latent confounders given the covariates, treatments, and outcome.

The overall objective function of CEVT, denoted as $\mathcal{L}_{\text{CEVT}}$, is defined as the sum of the ELBO and auxiliary losses:

$$\begin{aligned}\mathcal{L}_{\text{CEVT}} =& L + \frac{1}{N}\sum_{n=1}^{N}\Big[\log q_\phi(t_{i,n}|x_n) \\ &+ \log q_\phi(t_{d,n}|x_n, t_{d,n}) + \log q_\phi(y_n|x_n, t_{d,n}, t_{i,n})\Big].\end{aligned}$$

The auxiliary losses comprise the log-likelihood of the indirect treatment given the covariates, the log-likelihood of the direct treatment given the covariates and indirect treatments, and the log-likelihood of the outcome given the covariates and both treatments. These auxiliary losses serve as additional supervision signals that help guide the learning of the latent representations and improve the model's ability to estimate the causal effects accurately.

By maximizing the objective function $\mathcal{L}_{\text{CEVT}}$, CEVT learns to reconstruct the observed data, estimate the causal effects, and keep the latent space close to the assumed prior distribution.

The iterative conditioning process in CEVT can be described in more detail as follows: First, the input covariates X are passed through the first Transformer encoder layer $f_\theta^{(1)}$ to obtain the initial latent representation $Z^{(1)}$:

$$Z^{(1)} = f_\theta^{(1)}(X).$$

Next, the indirect treatment $\hat{T}_i$ is estimated using a linear projection layer g_{ϕ_i}:

$$\hat{T}_i = g_{\phi_i}(Z^{(1)}).$$

The estimated indirect treatment $\hat{T}_i$ is then added to the latent representation $Z^{(1)}$ and the original covariates X, and passed through the second Transformer encoder layer $f_\theta^{(2)}$ to obtain the updated latent representation $Z^{(2)}$:

$$Z^{(2)} = f_\theta^{(2)}(Z^{(1)} + \hat{T}_i + \beta X).$$

The direct treatment $\hat{T}_d$ is then estimated from $Z^{(2)}$ using another linear projection layer g_{ϕ_d}:

$$\hat{T}_d = g_{\phi_d}(Z^{(2)}).$$

The estimated direct treatment $\hat{T}_d$, along with the previously estimated indirect treatment $\hat{T}_i$, the updated latent representation $Z^{(2)}$, and the residual input $Z^{(1)}$, are added and passed through the third Transformer encoder layer $f_\theta^{(3)}$ to obtain the latent representation $Z^{(3)}$:

$$Z^{(3)} = f_\theta^{(3)}(Z^{(2)} + \hat{T}_d + \alpha\hat{T}_i + \beta Z^{(1)}).$$

The outcome $\hat{Y}$ is then estimated from $Z^{(3)}$ using a linear projection layer g_ψ:

$$\hat{Y} = g_\psi(Z^{(3)}).$$

The estimated outcome $\hat{Y}$, along with the latent representation $Z^{(3)}$ and the residual variables, are used to create $Z^{(4)}$ through an addition operation:

$$Z^{(4)} = f_\theta^{(4)}(Z^{(3)} + \hat{Y} + \alpha\hat{T}_d + \beta Z^{(2)}).$$

Here, γ, δ, and ϵ are coefficients for the residual connections, controlling the importance of the estimated treatments and the initial latent representation in the current conditioning step.

The latent representation $Z^{(4)}$ is then passed through the remaining Transformer encoder layers $f_\theta^{(5)}, \ldots, f_\theta^{(L)}$ to obtain the final latent representation Z':

$$Z' = f_\theta^{(L)}(\ldots(f_\theta^{(5)}(Z^{(4)}))\ldots).$$

Finally, the variational posterior $q_\phi(Z|X, T_d, T_i, Y)$ is modeled as a Gaussian distribution with mean μ_ϕ and variance σ_ϕ^2, which are parameterized by neural networks taking Z' as input:

$$q_\phi(Z|X, T_d, T_i, Y) = \mathcal{N}(Z; \mu_\phi(Z'), \sigma_\phi^2(Z')).$$

By incorporating the estimated outcome $\hat{Y}$ and the residual variables in the iterative conditioning process and passing the resulting latent representation through the remaining Transformer layers, CEVT can effectively capture the dependencies between the latent confounders, treatments, and outcome and learn a more accurate representation of the underlying causal structure. This iterative conditioning process is inspired by the conditioning approach in CEVAE. Instead of splitting the weight space for conditioning as done in CEVAE, CEVT enhances flexibility by iteratively adding condition embeddings. Combined with the variational approach for modeling the posterior distribution of the latent confounders using the final latent representation Z', allows CEVT to leverage the inductive bias from the known DGP.

Theoretical Analysis

In this section, we analyze the effect of data augmentation on the generalization performance of domain adaptation models from a theoretical perspective. We begin by stating a key proposition that provides an upper bound on the target domain error Zhao *et al.* [2018].

Consider a source domain D_S and a target domain D_T within the context of an out-of-distribution (OOD) generalization problem, where each domain is represented by a probability measure over the sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with $\mathcal{X} \subseteq \mathbb{R}^d$. In the source domain, we have a training dataset $S = \{z_i\}_{i=1}^n$, where $z_i \sim D_S$. The objective is to learn a model $f \in \mathcal{F}$ with parameters $\theta \in \Theta$, where $f : \Theta \times \mathcal{X} \to \mathbb{R}$, that can generalize effectively to the target domain.

Given a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$, the expected risk over D_S is defined as

$$L_S(f_\theta) \triangleq \mathbb{E}_{z \sim D_S}[\ell_\theta(z)],$$

and the empirical risk over S is

$$\widehat{L}_S(f_\theta) \triangleq \frac{1}{n}\sum_{z_i \in S}[\ell_\theta(z_i)],$$

where $\ell_\theta(z)$ is a shorthand notation.

The OOD generalization performance is measured by the gap between the target domain expected risk $L_T(f_\theta)$ and the source domain empirical risk $\widehat{L}_S(f_\theta)$. This gap can be decomposed into the in-domain generalization error, which is the gap between $\widehat{L}_S(f_\theta)$ and $L_S(f_\theta)$ in the source domain, and the out-of-domain distance, which is the discrepancy between D_S and D_T.

Proposition 1. (Zhao *et al.* [2018]) Given a hypothesis class $\mathcal{F}$ with a pseudo dimension of $\text{Pdim}(\mathcal{F}) = d'$ and considering two unlabeled empirical datasets $\widehat{D}_S$ and $\widehat{D}_T$ each of size n, it can be stated with at least a probability of $1 - \delta$ that the following holds true for every function f within the class $\mathcal{F}$:

$$L_T(f) \leq \widehat{L}_S(f) + \frac{1}{2} d_{\mathcal{F}\Delta\mathcal{F}}(\widehat{D}_T; \widehat{D}_S) + O\left(\sqrt{\frac{d'}{n}}\right).$$

Here, $d_{\mathcal{F}\Delta\mathcal{F}}(\widehat{\mathcal{D}}_T; \widehat{\mathcal{D}}_S)$ is defined as

$$2 \sup_{A_f \in \mathcal{A}_\mathcal{F}} \left| \mathbb{P}_{\widehat{\mathcal{D}}_S}[A_f] - \mathbb{P}_{\widehat{\mathcal{D}}_T}[A_f] \right|,$$

with $\mathcal{A}_\mathcal{F} = \{f(x) \oplus f'(x) \mid f, f' \in \mathcal{F}\}$, where $\oplus$ denotes the XOR operator. We consider two models trained with augmentation f_{aug} and trained with original dataset f_{ori}.

Theorem 1. Given Proposition 1 and assuming that the augmented dataset includes the original dataset and the unseen test dataset can cover a broader range of data, we have $D_{S_{\text{ori}}} \subset D_{S_{\text{aug}}} \subset D_T$. Furthermore, assuming that the model is sufficiently optimized such that $\widehat{L}_S$ is negligibly small, the following inequality holds:

$$\begin{aligned} &\widehat{L}_S(f_{\text{aug}}) + \frac{1}{2} d_\mathcal{F}\Delta_\mathcal{F}(\widehat{D}_T; \widehat{D}_{S_{\text{aug}}}) + O\left(\sqrt{\frac{d'}{n_{\text{aug}}}}\right) \\ &\leq \widehat{L}_S(f_{\text{ori}}) + \frac{1}{2} d_\mathcal{F}\Delta_\mathcal{F}(\widehat{D}_T; \widehat{D}_{S_{\text{ori}}}) + O\left(\sqrt{\frac{d'}{n_{\text{ori}}}}\right). \end{aligned}$$

This theoretical analysis suggests that data augmentation can improve generalization performance in domain adaptation settings. In the context of Section 3.2, a source domain with diverse cut-offs can achieve a lower generalization bound on the real-world test domain compared to one with a single cut-off. The augmented source dataset $D_{S_{\text{aug}}}$ with various cut-offs provides better coverage and representation of the test domain D_T than the original source dataset $D_{S_{\text{ori}}}$ with a single cut-off. Furthermore, Theorem 1 implies that the upper bound of the target domain generalization error for the model trained with augmented data, $L_T(f_{\text{aug}})$, is lower than or equal to that of the model trained with the original dataset, $L_T(f_{\text{ori}})$. This suggests that data augmentation can potentially improve domain generalization by reducing the discrepancy between the source and target domains and increasing the effective sample size. Detailed proof of Theorem 1 can be found in Appendix A.1.

Cut-Off	Clusters	Patients	Confirmed	Duration
1	445	1941	8.11 ± 8.90	3.10 ± 3.54
2	400	3274	7.03 ± 8.15	2.83 ± 3.25
3	323	3231	7.74 ± 8.55	3.20 ± 3.28
4	240	2593	8.45 ± 8.96	3.83 ± 3.53
5	180	2022	9.24 ± 9.73	4.46 ± 3.85
Overall	1588	13061	7.99 ± 8.81	3.41 ± 3.50

Table 1: The table shows the number of patients and clusters for each dataset with Cut-Off augmentation applied, as well as the mean and standard deviation of the confirmed and duration labels. In Section 4, all experiments except for the worst-case analysis were conducted using the Overall dataset, which includes all Cut-Offs.

Theorem 2. Let T_d be the direct treatment, T_i be the indirect treatment, and X be the observed covariates. Assuming that the indirect treatment T_i causally affects the direct treatment T_d, and both T_i and T_d share a common latent confounder Z. The following inequality holds:

$$P(Y|do(T_d = t_d), T_i, X) \geq P(Y|do(T_d = t_d), X),$$

where t_d is intervene value of T_d.

The theorem can be proved using the non-negativity property of KL Divergence, $D_{KL}(P(Z \mid T_i, X) \parallel P(Z \mid X)) \geq 0$. We obtain an inequality that suggests the log-likelihood of $P(Y|do(T_d = t_d), X)$ is greater when considering the indirect treatment T_i compared to when not considering it. This justifies the inclusion of the indirect treatment T_i, as it can lead to a more accurate estimate of the causal effect $P(Y|do(T_d = t_d), X)$. Detailed proof of Theorem 2 can be found in Appendix A.2.

3.2 Data Description and Preprocessing

Data Description

We collected data on 448 infection clusters, including 5,903 patients, from October 2020 to November 2021, a period of active COVID-19 epidemiological transmission. The dataset includes various features such as patient characteristics (e.g., gender, age, occupation) and cluster features (e.g., location type, reported address, government social distancing guidelines, and risk index). To comply with the definition of epidemiology, infection clusters were designed not to share the same patients Tupper *et al.* [2022]; Patiño-Galindo *et al.* [2017]; Lim *et al.* [2022].

Cut Off Augmentation

To generate corresponding labels for the number of confirmed cases and cluster duration for each preprocessed data point, we employed a cut-off augmentation method. Figure 4 illustrates an example of how the cut-off algorithm operates. Assuming the dotted box in the upper left represents the original cluster data, we sorted it into a sequence using relative dates based on cluster creation. In the cut-off stage, we applied masking to all parts of the cluster sequence except for the first day, making only the information of the corresponding patients accessible. The number of confirmed cases y and the duration d of the masked cluster, excluding the first day, were used as the target labels. We proceeded with the same approach from day 1 to day 5, and if the cut-off date exceeded the maximum length of the original cluster, it was prematurely excluded from the cut-off algorithm.

By assigning various information-rich cluster subsets and their corresponding different label sets to a single cluster data

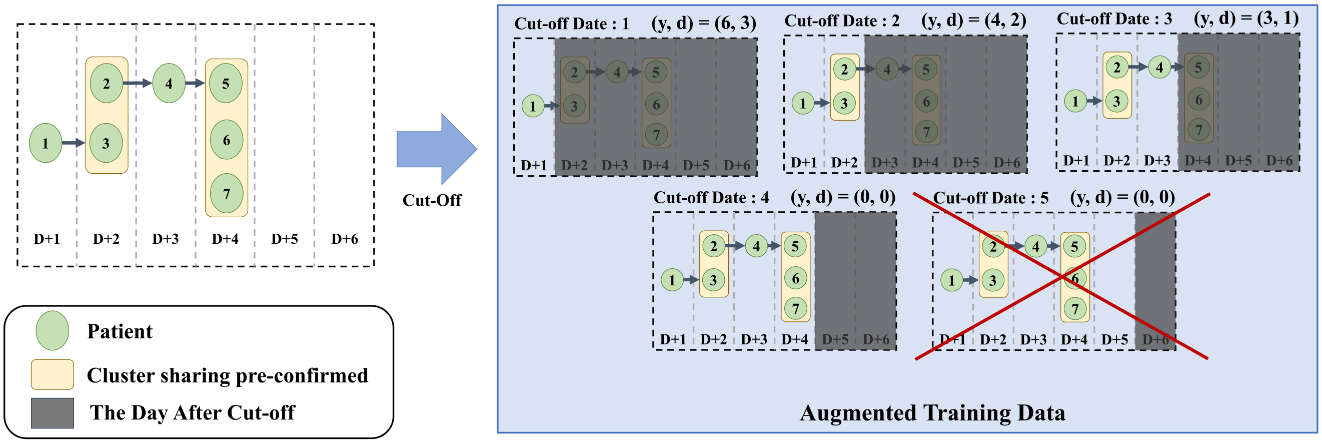

D+1
D+2
D+3
D+4
D+5
D+6
Cut-Off
Patient
Cluster sharing pre-confirmed
The Day After Cut-off
Cut-off Date : 1 (y, d) = (6, 3)
Cut-off Date : 2 (y, d) = (4, 2)
Cut-off Date : 3 (y, d) = (3, 1)
Cut-off Date : 4 (y, d) = (0, 0)
Cut-off Date : 5 (y, d) = (0, 0)
Augmented Training Data

Figure 4: The left side illustrates the original sequence data with patient nodes marked by dashed lines. By applying Cut-Off augmentation, a single sequence can be "cut off" into multiple possible combinations, allowing the observation of data subsets and the generation of diverse label sets accordingly. This enables the model to observe and learn from diverse data, and as will be shown in the theoretical analysis in Section 3.1, this augmentation positively impacts model robustness.

point, we enable the model to learn from a richer set of information. The application of the cut-off algorithm increases the number of clusters from 445 after preprocessing to 1,588, with each cluster having unique characteristics and label sets. Detailed statistics for each cut-off dataset can be found in Table 1. Considering the skewness of the label distribution, we made it possible to apply the Tukey transform as a hyperparameter.

4 Experiments

4.1 Experimental Details

We utilized the Adam optimizer Kingma and Ba [2014] with cosine annealing, splitting the data into train, valid, and test sets in an 8:1:1 ratio, and selected the model based on the sum of the best valid set MAE for the duration and confirmed label. The loss of CEVT can be broadly divided into three components: prediction loss, KLD loss, and reconstruction loss. We applied sigmoid scheduling to the lambda, the scaling factor of each loss component, to ensure the stability of posterior learning Bowman *et al.* [2015]. When dimensional reduction of the sequential Z from the transformer encoder was needed, we applied max pooling to extract features. Furthermore, we conducted experiments by adjusting the hyperparameters for the residual connections of x and t, as introduced in Section 3.1.

4.2 Prediction Accuracy

Our main task is to estimate the number of confirmed cases and the duration of clusters using the COVID-19 infection cluster data we collected. In doing so, we hypothesized that performing the task while considering the causal relationships known through prior research could lead to performance improvements. Table 2 shows the MAE and RMSE for each label. The table is divided into three sections. The top section of the table includes models suitable for regression in DNN, as well as baseline transformers that can effectively capture the time-series characteristics of clusters. We also experimented with iTransformer, a recent transformer variant. The next section of the table presents causal models that are easy to use with DNN. We used TARNet, Dragonnet, and CEVAE as baseline models for comparison. To ensure fairness in comparison with DNN baselines that cannot consider causal variables or with existing prior studies that can only consider a single causal variable, we added the missing treatment variables to the covariate X features and conducted the experiments. The table shows that CEVT achieved the lowest values for both the number of confirmed cases and cluster duration prediction, with MAEs of 1.154 and 1.896 and RMSEs of 1.883 and 3.217, proving superior to the baseline. CEVT's ability to outperform the benchmark models showcases its effectiveness in capturing causal relationships and leveraging them for improved predictions.

4.3 Causal Accuracy

In previous research within the binary treatment setting, the individual treatment effect (ITE) estimate and the associated causal metric were defined as follows for the treatment effect estimate of the hypothesis f for a unit:

$$\hat{\tau}_f(x) = f(x, 1) - f(x, 0).$$

Extending this definition, we consider continuous treatment values and redefine the ITE estimate for covariate x as:

$$\hat{\tau}_f(x) = \mathbb{E}\left[\frac{\partial}{\partial t} f(Y|X = x, do(T = t))\right].$$

This allows for a more flexible computation of treatment effects over a broader range, including both binary and continuous treatment values.

Based on prior research indicating a negative causal relationship between treatment and outcome Chen *et al.* [2022]; Chuang *et al.* [2023]; Ghosh and Roy [2022], we introduced a new metric called Causal Accuracy (CAcc) to evaluate the causal performance of the model further. CAcc is defined as follows:

$$CAcc = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left\{\hat{\tau}_f(x_i) < 0\right\},$$

Model	**MAE ↓**		**RMSE ↓**	
-	duration	confirmed	duration	confirmed
Linear	2.761	5.275	3.482	7.280
Ridge	2.941	5.241	3.705	7.870
MLP	1.338	2.964	2.043	4.565
Transformer	1.246	2.174	2.170	3.512
iTransformer	1.159	2.112	2.996	4.538
TARNet	1.559	3.452	2.268	5.316
Dragonet	1.284	2.898	2.018	4.520
CEVAE	1.387	3.048	2.437	4.921
CEVT	**1.154**	**1.896**	**1.883**	**3.217**

Table 2: MAE and RMSE for predicting cluster duration and the number of confirmed cases in the COVID-19 infection cluster dataset. CEVT outperforms all other baselines in both MAE and RMSE.

where $\mathbb{I}$ denotes the indicator function. For each data point, we intervene in the treatment and compare the estimated outcome values using the original treatment and the intervention treatment. If the intervention reveals a negative causal relationship between the treatment and outcome, we consider the model to have accurately learned the causal relationship, and we measure this using the CAcc metric. Table 3 presents the CAcc scores for each model. CEVT achieved the highest average CAcc of 0.686 for the Duration and Confirmed labels, demonstrating its ability to capture the intended causal relationships appropriately. Considering the abnormally extreme values of 0.983 and 0.0 for the Duration and Confirmed case metrics in the Linear model, it is evident that CEVT exhibits the most consistent performance across each label.

The experimental results highlight the importance of considering causal variables and demonstrate the superiority of CEVT in estimating the number of confirmed cases and cluster duration. By incorporating multiple treatment variables, employing an iterative conditioning mechanism, and utilizing Transformer layers, CEVT effectively captures the underlying causal structure and outperforms both non-causal baselines and existing causal models. The introduction of the CAcc further validates the ability of CEVT to learn the intended negative causal relationship between treatment and outcome, strongly supporting our hypothesis that considering causal relationships can significantly improve performance in the estimation task.

Model	**CAcc ↑**		
-	Duration	Confirmed	Average
Linear	**0.983**	0.000	0.491
Ridge	0.230	0.006	0.118
MLP	0.413	0.628	0.520
Transformer	0.385	0.374	0.380
iTransformer	0.553	0.624	0.588
TARNet	0.449	0.407	0.428
Dragonet	0.460	0.453	0.456
CEVAE	0.701	0.548	0.630
CEVT	0.736	**0.635**	**0.686**

Table 3: Causal Accuracy for the COVID-19 dataset. Prior knowledge indicates a negative causal relationship between the treatment and outcome. Accuracy was measured through interventions in the test dataset.

Model	**MAE ↓**		**RMSE ↓**	
-	duration	confirmed	duration	confirmed
Linear	3.180	5.615	4.092	8.677
Ridge	3.482	5.374	4.477	9.529
MLP	1.596	3.366	2.302	4.700
Transformer	1.908	3.296	3.470	7.660
iTransformer	2.166	4.104	3.201	9.361
TARNet	1.710	3.870	2.503	5.671
Dragonet	1.521	4.139	2.924	9.451
CEVAE	1.514	3.108	2.550	**4.398**
CEVT	**1.417**	**2.666**	**2.105**	5.169

Table 4: Evaluation results using only the test dataset with a cut-off of 1, representing the scenario with the least information. Even in the worst-case scenario, which is the most challenging for the model, CEVT demonstrates superior performance in most metrics.

4.4 Worst Case

Additionally, to validate the robustness of the models, in Table 4, we conducted experiments under the assumption that masking all sequences except for the first day in the cluster sequence represents the worst case scenario, where the model is expected to face significant difficulties due to the lack of information. For the most recent time-series baseline, iTransformer, the RMSE for the task of predicting confirmed cases increased by 2.06 times from 4.538 to 9.361 in the worst case. Consistent with the previous results, CEVT demonstrates the lowest MAE values of 1.417 and 2.666 and RMSE values of 2.105 at duration prediction, outperforming all other models. This indicates that the architecture of CEVT, which actively utilizes various information, is more advantageous in ensuring robustness. Among the models where the MAE and RMSE performance differ by less than 1.0 from the best-performing models, we selected the model with the highest CAcc, a causal metric we define in Eq. 1.

5 Conclusion

In this study, we collected and refined COVID-19 infection cluster data, augmented it with a cut-off algorithm, and proposed the CEVT model, which can theoretically incorporate infinite causal relationships. Our approach integrated inductive biases related to government distancing policies and risk indices and leveraged known causal relationships to enhance outcome estimation performance. By extending PEHE and $|\text{ATE}|$ metrics for continuous treatments and designing CEVT to handle multiple treatments, our model facilitated flexible causal effect analysis for both continuous and multiple treatment scenarios, demonstrated theoretically and experimentally using COVID-19 and synthetic data. While the model is specialized for time-series data, limiting its generalizability to multimodality data, future work could explore its applicability to a broader range of data types and domains. Extending CEVT to handle multimodal data is a valuable research direction. Despite these limitations, CEVT achieved significant results by defining new problems in causal inference and proposing solutions. By utilizing causal relationships and incorporating an expressive model structure with attention mechanisms, CEVT enhanced causal inference in complex real-world scenarios, improving the interpretability and insights into the model's decision-making processes.

References

Ahmed M Alaa and Mihaela Van Der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30, 2017.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Cathy WS Chen, Mike KP So, and Feng-Chi Liu. Assessing government policies' impact on the covid-19 pandemic and elderly deaths in east asia. *Epidemiology & Infection*, 150:e161, 2022.

Yu-Chuan Chuang, Kuan-Pei Lin, Li-An Wang, Ting-Kuang Yeh, and Po-Yu Liu. The impact of the covid-19 pandemic on respiratory syncytial virus infection: a narrative review. *Infection and Drug Resistance*, pages 661–675, 2023.

Zizhen Deng, Xiaolong Zheng, Hu Tian, and Daniel Dajun Zeng. Deep causal learning: representation, discovery and inference. *arXiv preprint arXiv:2211.03374*, 2022.

Peter A Frost. Proxy variables and specification bias. *The review of economics and Statistics*, pages 323–325, 1979.

Sujoy Ghosh and Saikat Sinha Roy. Global-scale modeling of early factors and country-specific trajectories of covid-19 incidence: a cross-sectional study of the first 6 months of the pandemic. *BMC Public Health*, 22(1):1919, 2022.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017.

Daniel Israel, Aditya Grover, and Guy Van den Broeck. High dimensional causal inference with variational backdoor adjustment. *arXiv preprint arXiv:2310.06100*, 2023.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Chaeyun Lim, Youngju Nam, Won Sup Oh, Sugeun Ham, Eunmi Kim, Myeonggi Kim, Saerom Kim, Yeojin Kim, and Seungmin Jeong. Characteristics of transmission routes of covid-19 cluster infections in gangwon province, korea. *Epidemiology & Infection*, 150:e19, 2022.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Mark R Montgomery, Michele Gragnolati, Kathleen A Burke, and Edmundo Paredes. Measuring living standards with proxy variables. *Demography*, 37(2):155–174, 2000.

Juan Ángel Patiño-Galindo, Manoli Torres-Puente, María Alma Bracho, Ignacio Alastrué, Amparo Juan, David Navarro, María José Galindo, Dolores Ocete, Enrique Ortega, Concepción Gimeno, et al. The molecular epidemiology of hiv-1 in the comunidad valenciana (spain): analysis of transmission clusters. *Scientific reports*, 7(1):11584, 2017.

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

Judea Pearl. Detecting latent heterogeneity. In *Probabilistic and causal inference: The works of judea pearl*, pages 483–506. 2022.

Mattia Prosperi, Yi Guo, Matt Sperrin, James S Koopman, Jae S Min, Xing He, Shannan Rich, Mo Wang, Iain E Buchan, and Jiang Bian. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375, 2020.

Olav Reiersøl. *Confluence analysis by means of instrumental sets of variables*. PhD thesis, Almqvist & Wiksell, 1945.

Pedro Sanchez, Antanas Kascenas, Xiao Liu, Alison Q O'Neil, and Sotirios A Tsaftaris. What is healthy? generative counterfactual diffusion for lesion localization. In *MICCAI Workshop on Deep Generative Models*, pages 34–44. Springer, 2022.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.

Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

James H Stock and Mark W Watson. *Introduction to econometrics*. Pearson, 2020.

Paul Tupper, Shraddha Pai, Caroline Colijn, et al. Covid-19 cluster size and transmission rates in schools from crowdsourced case reports. *Elife*, 11:e76174, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Michael R Wickens. A note on the use of proxy variables. *Econometrica: Journal of the Econometric Society*, pages 759–761, 1972.

Ziqi Xu, Debo Cheng, Jiuyong Li, Jixue Liu, Lin Liu, and Kui Yu. Causal inference with conditional front-door adjustment and identifiable variational autoencoder. *arXiv preprint arXiv:2310.01937*, 2023.

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International conference on learning representations*, 2018.

Congzhi Zhang, Linhai Zhang, and Deyu Zhou. Causal walk: Debiasing multi-hop fact verification with front-door adjustment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19533–19541, 2024.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31, 2018.

A Proof Of Theorem

A.1 Theorem 1

Theorem 1. Given Proposition 1 and assuming that the augmented dataset includes the original dataset and the unseen test dataset can cover a broader range of data, we have $D_{S_{\text{ori}}} \subset D_{S_{\text{aug}}} \subset D_T$. Furthermore, assuming that the model is sufficiently optimized such that $\widehat{L}_S$ is negligibly small, the following inequality holds:

$$\widehat{L}_S(f_{\text{aug}}) + \frac{1}{2} d_{\mathcal{F}}\Delta_{\mathcal{F}}(\widehat{D}_T; \widehat{D}_{S_{\text{aug}}}) + O\left(\sqrt{\frac{d'}{n_{\text{aug}}}}\right) \le \widehat{L}_S(f_{\text{ori}}) + \frac{1}{2} d_{\mathcal{F}}\Delta_{\mathcal{F}}(\widehat{D}_T; \widehat{D}_{S_{\text{ori}}}) + O\left(\sqrt{\frac{d'}{n_{\text{ori}}}}\right).$$

Proof of Theorem 1. The domain discrepancy between the target distribution D_T and the augmented source distribution $D_{S_{\text{aug}}}$ is defined as:

$$d_{\mathcal{F}}\Delta_{\mathcal{F}}(\widehat{D}_T; \widehat{D}_{S_{\text{aug}}}) = 2 \sup_{A_f \in A} \left| P_{\widehat{D}_T}[A_f] - P_{\widehat{D}_{S_{\text{aug}}}}[A_f] \right|.$$

Applying the triangle inequality, we have:

$$d_{\mathcal{F}}\Delta_{\mathcal{F}}(\widehat{D}_T; \widehat{D}_{S_{\text{aug}}}) \le d_{\mathcal{F}}\Delta_{\mathcal{F}}(\widehat{D}_T; \widehat{D}_{S_{\text{ori}}}) + 2 \sup_{A_f \in A_{\mathcal{F}}} \left| P_{\widehat{D}_{S_{\text{ori}}}}[A_f] - P_{\widehat{D}_{S_{\text{aug}}}}[A_f] \right|.$$

Given that $D_{S_{\text{ori}}} \subset D_{S_{\text{aug}}}$, we know:

$$\begin{aligned} & 2 \sup_{A_f \in A_{\mathcal{F}}} \left| P_{\widehat{D}_{S_{\text{ori}}}}[A_f] - P_{\widehat{D}_{S_{\text{aug}}}}[A_f] \right| \\ & \le 2 \sup_{A_f \in A_{\mathcal{F}}} \left| P_{\widehat{D}_{S_{\text{aug}}}}[A_f] - P_{\widehat{D}_T}[A_f] \right| \\ & = d_{\mathcal{F}}\Delta_{\mathcal{F}}(\widehat{D}_T; \widehat{D}_{S_{\text{aug}}}). \end{aligned}$$

Combining these results, we get:

$$d_{\mathcal{F}}\Delta_{\mathcal{F}}(\widehat{D}_T; \widehat{D}_{S_{\text{aug}}}) \le d_{\mathcal{F}}\Delta_{\mathcal{F}}(\widehat{D}_T; \widehat{D}_{S_{\text{ori}}}) + d_{\mathcal{F}}\Delta_{\mathcal{F}}(\widehat{D}_T; \widehat{D}_{S_{\text{aug}}}).$$

Subtracting $d_{\mathcal{F}}\Delta_{\mathcal{F}}(\widehat{D}_T; \widehat{D}_{S_{\text{aug}}})$ from both sides, we obtain:

$$0 \le d_{\mathcal{F}}\Delta_{\mathcal{F}}(\widehat{D}_T; \widehat{D}_{S_{\text{ori}}}).$$

Thus, we have:

$$d_{\mathcal{F}}\Delta_{\mathcal{F}}(\widehat{D}_T; \widehat{D}_{S_{\text{aug}}}) \le d_{\mathcal{F}}\Delta_{\mathcal{F}}(\widehat{D}_T; \widehat{D}_{S_{\text{ori}}}).$$

Assuming $\hat{L}_S$ is negligibly small and considering $n_{\text{aug}} \ge n_{\text{ori}}$, we can conclude that Theorem 1 holds.

A.2 Theorem 2

Theorem 2. Let T_d be the direct treatment, T_i be the indirect treatment, and X be the observed covariates. Assuming that the indirect treatment T_i causally affects the direct treatment T_d, and both T_i and T_d share a common latent confounder Z. The following inequality holds:

$$P(Y|do(T_d = t_d), T_i, X) \ge P(Y|do(T_d = t_d), X),$$

where t_d is intervene value of T_d.

Proof of Theorem 2. Using do-calculus, we express $P(Y \mid \text{do}(T_d = t_d), T_i, X)$ and $P(Y \mid \text{do}(T_d = t_d), X)$:

$$\begin{aligned} & P(Y \mid \text{do}(T_d = t_d), T_i, X) \\ & = \int P(Y \mid T_d, T_i, X, Z) P(Z \mid T_i, X)\, dZ, \\ & P(Y \mid \text{do}(T_d = t_d), X) \\ & = \int P(Y \mid T_d, X, Z) P(Z \mid X)\, dZ. \end{aligned}$$

Since KL divergence is always non-negative:

$$D_{KL}(P(Z \mid T_i, X) \parallel P(Z \mid X)) \ge 0.$$

This implies:

$$\begin{aligned} & \int P(Z \mid T_i, X) \log P(Z \mid X)\, dZ \\ & - \int P(Z \mid T_i, X) \log P(Z \mid T_i, X)\, dZ \le 0. \end{aligned}$$

By expanding and integrating, we obtain:

$$\begin{aligned} & \int P(Y \mid \text{do}(T_d = t_d), X) \log p(y \mid \text{do}(T_d = t_d), X)\, dy \le \\ & \int P(Y \mid \text{do}(T_d = t_d), T_i, X) \log p(y \mid \text{do}(T_d = t_d), T_i, X)\, dy. \end{aligned}$$

Thus:

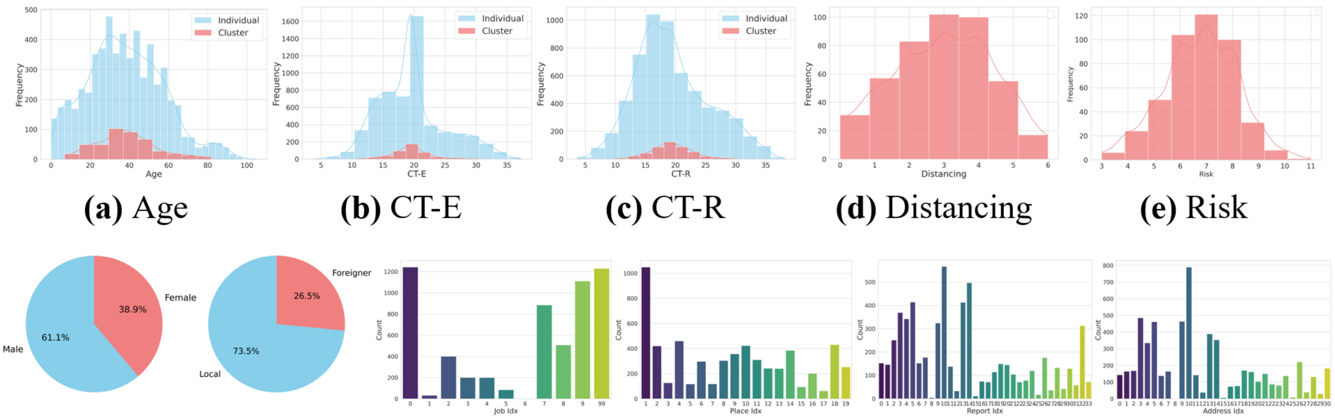
$$P(Y \mid \text{do}(T_d = t_d), T_i, X) \ge P(Y \mid \text{do}(T_d = t_d), X).$$

This shows that incorporating the indirect treatment T_i can lead to a more accurate estimation of $P(Y \mid \text{do}(T_d = t_d), X)$.

B Details of COVID-19 dataset

B.1 Ethics Considerations

This study was approved by the Institutional Review Board of Gachon University College of Medicine, Incheon, Republic of Korea (IRB No. GCIRB2021-434). The Ethics Committee of Gachon University College of Medicine granted a waiver of prior consent because it involved routinely collected medical data that were anonymized at all stages. The study was conducted ethically in accordance with the Declaration of Helsinki of the World Medical Association.



(a) Age **(b)** CT-E **(c)** CT-R **(d)** Distancing **(e)** Risk

(f) Gender **(g)** Local **(h)** Job **(i)** Place **(j)** Report **(k)** Address

Figure 5: The top row shows the distribution of continuous features: the light blue represents the distribution when counted on a patient-wise basis, and the red represents the distribution when counted on a cluster-wise basis. The bottom row shows the distribution of categorical features.

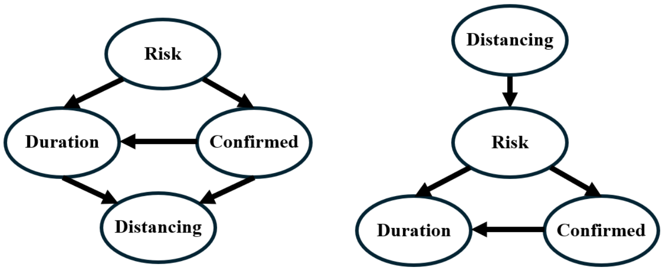
B.2 Dataset Preprocessing

The collected raw data has 46 columns, each providing various information about individual patients in the form of numerical, binary, or string values. However, considering the characteristics of the prediction task for the number of confirmed cases and cluster duration using machine learning models, certain data points were removed, including data irrelevant to the prediction, such as the name of the reported address, data containing future information, such as contact tracing information, and data with 25% missing values. Features such as address information, occupation information, and location type were converted to categorical indices. Missing values in the quarantine guidelines and risk index were replaced with other values within the same cluster while missing values in age were replaced with the average age of the location index. The onset date and confirmation date were converted to relative dates to facilitate personalized predictions for each cluster.

After preprocessing, we ultimately used the processed data as shown in Figure 4. Among the continuous data, we utilized features related to individual patients, such as Age, CT-R, and CT-E, and features related to clusters, such as Distancing and Risk. For categorical data, we employed features associated with individual patients, including Gender, is Local, and Job, as well as features linked to clusters, such as Place, Report, and Address. The distributions of the introduced data are visualized in Figure 5.

B.3 Causal Relationship

Based on prior research in the medical domain, we identified a negative causal relationship between government distancing guidelines, risk index, and the number of confirmed COVID-19 cases. To verify if our collected and refined data support this study, we analyzed the data using the GES and LiNGAM causal discovery algorithms. Figure 6 presents the causal DAGs from each algorithm, excluding covariates other than treatment and outcome. Assuming that the variables Confirmed and Duration constitute a single Y, the causal graph derived from LiNGAM in (b) aligns precisely with our assumed SCM. The causal graph revealed by GES in (a) shows that only the Risk variable, which we considered a Direct Cause, follows our assumption, although we already know from the DGP that Risk is generated through Distancing. Both figures consistently support the causal relationship between the Risk variable, which we assumed to be a Direct Cause, and the Outcome.



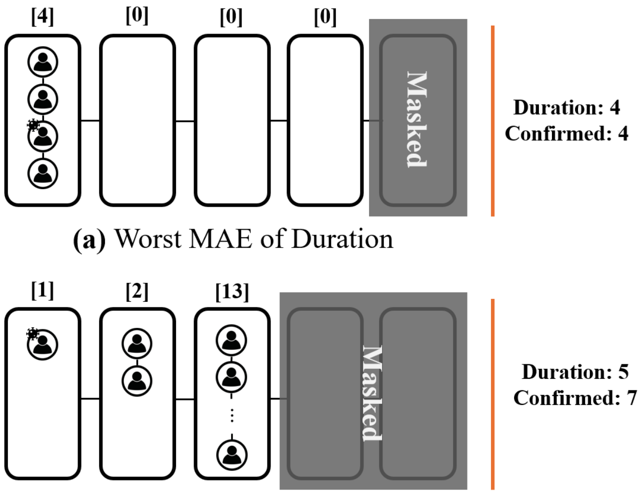
(a) GES **(b) LiNGAM**

Figure 6: Causal DAG discovered by Greedy Equivalence Search and the Linear Non-Gaussian Acyclic Model on COVID-19 dataset.

C Error Case Analysis

We analyzed the worst case of the baseline CEVAE model compared to CEVT. We compared the performance of our model on the data points where CEVAE received the lowest MAE loss and analyzed the characteristics of these data points, as shown in Figure 7. Figures 7-(a) and (b) respectively represent the worst cases for CEVAE in terms of duration and confirmed label.

In the case of (a), where there are four patients on the first day and no patients on the remaining observable days, it is a difficult type of data if the sequential characteristics are



(a) Worst MAE of Duration

(b) Worst MAE of Confirmed

Figure 7: Error Case analyzed based on CEVAE's MAE, (a), (b) Duration and Confirmed respectively.

not captured. For CEVAE, which cannot understand the positional information of the sequence, it cannot differentiate whether there are four patients on the first day or if there is approximately one patient over four days. This results in a significant difference in MAE between CEVAE and CEVT, which are 18.50 and 3.05, respectively.

In the case of (b), although there is a large amount of information available with 13 observed patients, the duration and confirmed label are 5 and 7, respectively, placing them in the top 16% of the label distribution, which qualifies them as OOD data. This makes it a challenging task for models that are not robust. The MAEs for CEVAE and CEVT are 29.09 and 16.51, respectively, indicating that CEVT can somewhat maintain performance even on OOD data points.

D Causl Effect Estimation

To validate the causal estimation performance of our model, we generated synthetic data that follows the assumed SCM. We created 1000 data points using the linear dependent dataset generation function provided by DoWhy library Sharma and Kiciman [2020]. In accordance with traditional approaches for evaluating the performance of causal inference models, we followed the metrics $|ATE|$ and $PEHE$ Louizos *et al.* [2017]; Shalit *et al.* [2017]; Shi *et al.* [2019]. However, instead of using the binary treatment $\hat{\tau}$ defined in eq. 1, we utilized the redefined $\hat{\tau}$ for the continuous setting, as defined in eq. 1, for measurement and evaluation.

For the population causal effect, we report the expected absolute error on the average treatment effect ($|ATE|$):

$$\epsilon_{|ATE|} = \frac{1}{n}\sum_{i=1}^{n} |\hat{\tau}_f(x_i) - \tau(x_i)|$$

where $\tau(x_i, t)$ is the ground truth ITE for unit i at treatment level t. To measure the accuracy of the individual treatment effect estimation, we use the expected Precision in Estimation of Heterogeneous Effect (PEHE):

Model	$\sqrt{\epsilon_{PEHE}}$ ↓	ϵ_{ATE} ↓
Linear	18.376	5.48
Ridge	34.767	10.495
MLP	18.497	5.309
Transformer	25.532	7.53
iTransformer	31.009	9.181
TARNet	486.56	17.965
Dragonet	479.806	17.781
CEVAE	22.872	6.497
CEVT	**17.168**	**4.289**

Table 5: Results of experiments conducted on a synthetic dataset. The errors for expected PEHE and ATE, metrics frequently used in prior studies for causal effect analysis, were analyzed.

$$\epsilon_{PEHE} = \frac{1}{n}\sum_{i=1}^{n} [\hat{\tau}_f(x_i) - \tau(x_i)]^2$$

The performance results are reported in Table 5, where CEVT achieves the best performance in both ϵ_{PEHE} and $|\epsilon_{ATE}|$. This demonstrates that CEVT can accurately estimate both individual treatment effects and the causal effect on the entire population.