SCouT: Synthetic Counterfactuals via Spatiotemporal Transformers for Actionable Healthcare

BHISHMA DEDHIA, ROSHINI BALASUBRAMANIAN, and NIRAJ K. JHA, Princeton University, USA

The synthetic control method has pioneered a class of powerful data-driven techniques to estimate the counterfactual reality of a unit from donor units. At its core, the technique involves a linear model fitted on the pre-intervention period that combines donor outcomes to yield the counterfactual. However, linearly combining spatial information at each time instance using time-agnostic weights fails to capture important inter-unit and intra-unit temporal contexts and complex nonlinear dynamics of real data. We instead propose an approach to use local spatiotemporal information before the onset of the intervention as a promising way to estimate the counterfactual sequence. To this end, we suggest a Transformer model that leverages particular positional embeddings, a modified decoder attention mask, and a novel pre-training task to perform spatiotemporal sequence-to-sequence modeling. Our experiments on synthetic data demonstrate the efficacy of our method in the typical small donor pool setting and its robustness against noise. We also generate actionable healthcare insights at the population and patient levels by simulating a state-wide public health policy to evaluate its effectiveness, an in silico trial for asthma medications to support randomized controlled trials, and a medical intervention for patients with Friedreich’s ataxia to improve clinical decision making and promote personalized therapy (code is available at https://github.com/JHA-Lab/scout).

CCS Concepts: • Computing methodologies → Learning latent representations; • Applied computing → Life and medical sciences;

Additional Key Words and Phrases: Causal inference, precision medicine, randomized trials, synthetic control, transformers

ACM Reference format:
Bhishma Dedhia, Roshini Balasubramanian, and Niraj K. Jha. 2023. SCouT: Synthetic Counterfactuals via Spatiotemporal Transformers for Actionable Healthcare. ACM Trans. Comput. Healthcare 4, 4, Article 23 (October 2023), 28 pages. https://doi.org/10.1145/3617180

B. Dedhia and R. Balasubramanian contributed equally to this research.
This work was supported by NSF grant CCF-2203399. The experiments reported in this article were performed on the computational resources managed and supported by Princeton Research Computing at Princeton University.
Authors’ addresses: B. Dedhia (Corresponding author) and N. K. Jha, Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08540; e-mails: [bgedhia, jha]@princeton.edu; R. Balasubramanian, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540; e-mail: roshinib@alumni.princeton.edu.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. 2637-8051/2023/10-ART23 $15.00
https://doi.org/10.1145/3617180

ACM Transactions on Computing for Healthcare, Vol. 4, No. 4, Article 23. Publication date: October 2023.
1 INTRODUCTION

What is spoken of the unchanging or intelligible must be certain and true; but what is spoken of the created image can only be probable; being is to becoming what truth is to belief.

Plato’s timeless allegory of the Cave is a classical philosophical thought experiment that routinely comes up in discussions of how humans perceive reality and whether there is any higher truth to existence. A group of people live chained to the wall of a cave all their lives, facing a blank wall and watching shadows projected on the wall from objects passing in front of a fire behind them, until one one of them is freed. This prisoner imagines what it would be to look around, only to discover the real nature of the world they have perceived through the shadows thus far. The pursuit of this higher truth has enabled us with the ability to be unshackled from the perceived reality and mathematically reason about alternate, imagined perspectives, an ability coined as counterfactual reasoning. Such reasoning abilities form the hallmark of an agent operating in a dynamic environment and allows it to reliably manipulate the world by ascertaining the possible counterfactuals of potential interventions. Examples of such agents and interventions are ubiquitous in healthcare: policy-makers enact laws to improve public health, healthcare systems iteratively improve the quality of care they provide, and physicians determine medical treatment plans for their patients. In each of the previous cases, to make a decision, the physicians and policy-makers acting as agents need to evaluate the ability of each intervention option in the form of medications and treatment policies to yield the desired results.

Often, there is interest in reliably estimating the effect of an intervention at both the population and individual levels. Our understanding of the genetic basis of diseases has accelerated in recent years and ushered in the era of precision medicine, a new paradigm in therapeutics that advocates for personalized treatments. Even patients diagnosed with the same disease may benefit from individualized therapies [27]. Hence, tailored clinical approaches based on insight extracted from large amounts of data—including clinical, genetic, and physiological information—are beneficial, as depicted in Figure 1. Public health officials also use demographic and economic data to assess possible interventions and make effective decisions.

Randomized Controlled Trials (RCTs) are the current gold standard for evaluating the effect of a treatment. In an RCT, participants are randomly assigned to either an intervention arm or a control arm. The outcome variables of the exposed and control groups are then compared to estimate the treatment effect. This approach is population-based and forces physicians to adopt a one-size-fits-all treatment that is prone to error. Moreover, RCTs raise a variety of issues around ethics and cost [38]. In contrast to traditional population-based methods of evaluating treatments, precision medicine encourages personalized approaches to “delivering the right treatments, at the right time, every time to the right person,” a phrase famously recited by former U.S. President Barack Obama. In the absence of RCTs, social scientists rely on comparative case studies to estimate the effect of an intervention, in which a group of ‘donor’ units is collected that resemble the unit of interest prior to the onset of intervention. Then the evolution of the donors produces a counterfactual for the unit of interest. However, finding the donors is empirically driven and the existence of a good match is not guaranteed.

Abadie et al. [3, 4] pioneered a powerful data-driven approach to aggregate observational data from several donor control units to compute the Synthetic Control (SC) of a target unit that receives treatment. They posited that a combination of several unaffected units provides a better match to the affected unit than any unaffected unit alone. Hence, the SC method assigns a non-negative weight to each donor. Such weights sum to 1.0 such that a convex combination of these units best estimates the treated unit prior to the intervention. After that, the post-intervention donor combination yields the desired counterfactual. Prior works have relaxed the convex constraints on the vanilla SC estimator and explored regression-based estimators to compute the assigned weights [20, 24]. Moreover, recent works [6–8] use matrix estimation methods to make SC more robust to noise and missing data. Similarly, Synthetic Intervention (SI) for a target control unit can be estimated by...
Fig. 1. An individual’s genetic, physiological, and clinical data can be used to synthesize their counterfactuals under multiple interventions using donor units.

Fig. 2. Synthetic control. (a) The vanilla SC method generates control via a convex interpolation of the donors across temporal slices using time-agnostic weights. (b) We view the problem through the lens of a Seq2Seq modeling problem and use local spatiotemporal context while predicting control.

selecting donor units that undergo the intervention [5]. Whereas SC can generate longitudinal trajectories of patient progress with no intervention exposure, SI can simulate trajectories under a medical intervention. The class of linear SC methods essentially solves an optimization problem for mapping donors to the target across temporal slices using universal, time-agnostic weights. Although approaches based on linearly combining donor outcomes to estimate the counterfactual yield an agile model, they overlook both the inter-unit and intra-unit temporal context while making predictions, as illustrated in Figure 2. These additional relations are essential to take into account because the evolution dynamics of a unit may lead to long-range temporal dependencies that can extend into other units and cause co-movement. We formalize this notion in Section 3.1.

In this work, we are interested in the point treatment setting where the counterfactual is estimated after an intervention is administered once to a target unit at some time instance. We recast the problem of counterfactual prediction as a Sequence-to-Sequence (Seq2Seq) mapping problem where the pre-intervention spatiotemporal
context of every unit is mapped into the post-intervention sequence of the target. From a modeling standpoint, we introduce a novel Transformer-based model [40] that uses self-attention to learn complex dynamics. At its core, a Transformer is a generalized differentiable computer that is expressive in the forward pass, easily optimizable via back-propagation, and highly parallel computationally. Thus, it can be used as an efficient universal compute engine wrapped around by modality-specific encoder and decoder heads to model diverse modalities simultaneously. Prior counterfactual estimation methods hinge on the provision of simple vector covariates but are unamenable to complex inputs like images and audio, whereas Transformers have shown remarkable modeling properties in these areas. Thus, this choice of model opens up the avenue to do counterfactual estimation on all kinds of medical data. In Section 5.4, demonstrate our framework on ubiquitously occurring discrete classes, a simple abstraction that no prior works have tackled. Due to the spatial component of the problem, we modify the decoder attention map to enable bidirectional attention spatially while still maintaining temporal causality. Furthermore, we inject spatial and temporal embeddings to help the model learn the input token order and add an additional target embedding to separate the target and donor units. From a computational perspective, whereas SC estimates rely on a relatively small number of donors, Transformer models traditionally require large amounts of data to be trained from scratch. To mitigate this problem, we formulate a novel self-supervision-based pre-training task for the Transformer model. The Transformer-based estimator captures rich nonlinear function classes, accounts for irregular observations, and is easy to train using backpropagation. It is therefore a practical non-parametric approach for modeling complex real-life data compared to linear SC estimators whose assumption of time-invariant correlation across units fails for weakly coupled dynamical systems. Our experiments show that the model outperforms prior state-of-the-art estimators on synthetic datasets for different donor pool sizes and remains robust to noise. Running ablation studies on the pre-training step reveals that it is critical to learning a good estimator. We study the effectiveness of an anti-tobacco law passed in California and probe the attention layer of the Transformer model to generate a rich description of spatiotemporal donor weights. We also focus on the individual level and model the control counterfactual of patients in a randomized trial on drugs for childhood asthma. Finally, we simulate synthetic counterfactuals under a calcitriol supplement intervention for a Friedreich’s Ataxia (FA) patient.

2 RELATED WORK

Our work relates most closely to that of Abadie and Gardeazabal [4], who first introduced the SC method to measure the economic ramifications of terrorism on Basque Country in a panel data setting. They took observational data from other Spanish regions and assigned them convex weights to generate an interpolated SC. They also used their method to analyze the Proposition 99 Excise Tax effect on California’s tobacco consumption [3]. Since then, this method to estimate counterfactuals has gained a lot of popularity and has been used in the social sciences to study legalized prostitution [17], immigration policy [12], taxation [26], organized crime [32], and various other policy evaluations. Other notable works that use regression-based techniques to estimate the control using observed data from other units are the panel-data-based estimation method introduced by Hsiao et al. [24] and the additive-difference-based estimator proposed by Doudchenko and Imbens [20]. Recently, Robust Synthetic Control (RSC) [7] was proposed as a robust variant of the SC method. It sets up the problem as an unconstrained least squares regression problem with sparsity terms and therefore extrapolates the control from the donors. In addition, the authors introduce a low-rank decomposition denoising step to handle missing data and noise that are ubiquitous in real-world data. Multi-Dimensional Robust Synthetic Control (mRSC) [6] extends RSC and proposes an algorithm to incorporate multiple covariates into the model. Athey et al. [8] pose the SC problem as estimating missing values in a matrix. For an extensive literature review of SC, see the work of Abadie [1]. Qian et al. [33] propose a neural network based estimator that learns a unit-specific representation and thereafter uses them to find donor weights. Concomitantly to us, Melnychuk et al. [29] propose a Transformer-based architecture for counterfactual estimation. The linear SC
estimators have been generalized to generate synthetic counterfactuals under any intervention [5]. Most of the preceding works make the strong parametric assumption that the underlying data-generating process is a linear factorized model of a time-specific and unit-specific latent. Therefore, the proposed models combine donor outcomes across temporal slices and forgo any temporal context. Comparative case studies have been used historically to estimate the control [15, 16]. However, such works have not formalized the selection of the donor pool.

Orthogonal to SC, structural time series methods under the stringent unconfoundedness assumption use temporal regularities across control paths to extrapolate the pre-treatment trajectory to the counterfactual outcome of the treated unit [2, 25, 36]. Structural models need many donor paths and are not amenable to settings like healthcare where the donor pool is small. Difference in Difference (DiD) methods have also been used in healthcare applications [43], but they rely on the strong assumption that the slopes of observed post-treatment units and control units diverge only because of the treatment such that any difference in slopes can be attributed to the intervention. Thus, non-parallel trends present a major challenge to the validity of this model. Besides, there exists a rich literature on non-SC machine learning driven counterfactual estimation. Neural network based techniques have been used to estimate the individual treatment effect in both static [37, 45] and longitudinal settings [11].

Although the importance of disease progression models is widely recognized, there are few works that apply advanced modeling approaches to the task. Many methods use basic statistics and rely on a large amount of domain knowledge. More sophisticated probabilistic or statistical models exist but tend to be exclusively applicable to a specific target disease. Using a more general framework, Wang et al. [41] predict the stage of chronic obstructive pulmonary disease progression with a graphical model based on the Markov jump process. One Transformer model under the broad umbrella of disease progression models has been proposed; Nguyen et al. [31] predict the development of knee osteoarthritis with a Transformer model designed for multi-agent decision making but rely on imaging data. In addition, most neural network approaches to disease progression only predict one to a few metrics of disease, highlighting an area that requires further research.

Many works use neural networks as Seq2Seq models for machine translation using feed-forward networks [10], recurrent neural networks [30], and long short-term memories [39]. Transformer-based models have achieved state-of-the-art results on tasks that span natural language processing [40] and, more recently, computer vision [19]. They are pre-trained on large corpora and then fine-tuned for a specific task, with the most notable works being BERT [18], which uses self-supervision-based denoising tasks, and GPT [14, 34], which uses a language-modeling-based pre-training task. Our work adds to these works by proposing a model to map spatiotemporal data and introducing a self-supervision-based pre-training task that supports the mapping.

3 BACKGROUND
In this section, we outline the key assumptions and formally set up the problem of generating synthetic counterfactuals using donor units.

3.1 Assumptions
The problem consists of a ‘target’ unit and N ‘donor’ units with N typically small. A unit can refer to an aggregate of a population segment or an individual. N donor units undergo the intervention we are interested in at the same time instance. Without loss of generality, we assume that control assignment is also a type of intervention and finding the counterfactual under control reduces to the SC problem. For each unit, panel data are available for a time period T and the covariate in each panel can have missing data. We make standard assumptions about the units:
(1) **Stable Unit Treatment Value Assumption (SUTVA):** The potential outcomes for any unit do not vary with the treatments assigned to other units, and therefore there are no spillover effects.

(2) **No anticipation of treatment:** This implies that the units cannot anticipate the assignment of treatment *a priori*, and therefore the assignment has no effect on the pre-intervention trajectory.

(3) **Exchangeability (no unmeasured confounders):** This assumption implies that units that are exposed and unexposed have the same potential outcomes on average, and hence in the case of an observational study, a control group can be reliably used to measure the counterfactual of the treatment group.

(4) **Positivity (common support):** This necessitates that covariates in the control and the treatment group have overlapping distributions (common support), in the absence of which it is impossible to understand the causal effect of a treatment because a subset of the population is always entirely left treated or untreated.

**Data Generation Process.** The underlying data generation of a unit $i$ consists of unit-specific latent $\theta_i$ and time-indexed latents $\rho_t$. The observation of all covariates of unit $i$ at time $t$, $y_{it}$, depends on all prior temporal latents and is given by

$$ y_{it} = m_{it} + \epsilon_{it}, \quad m_{it} \sim N(0, \sigma^2), $$

$$ m_{it} = f(\theta_i, \rho_t, \ldots, \rho_1). $$

Function $f$ is implicitly learned by the Transformer that can capture a wide class of nonlinear autoregressive dynamics. This generalized model subsumes the models popularly assumed in the literature. Abadie and Gardeazabal [4] assume the following model for the data generation process:

$$ y_{it} = \rho_t^T X_i + \epsilon_{it}. $$

Here, $\rho_t$ denotes the latent at time $t$ and $X_i$ denotes the observed covariates of unit $i$. Setting $f$ to the linear factor function and $\theta_i = X_i$ in Equation (2) yields the model defined by Equation (4). Similarly, Qian et al. [33] assume that

$$ y_{it} = \rho_t^T C_i + \epsilon_{it}, $$

where $C_i$ is a latent unit-specific representation. It is straightforward to see that Equation (2) captures this model as well. Amjad et al. [6, 7] assume a non-autoregressive, nonlinear $f$ that is once again a special case of our assumption. Next, we set up the problem formally.

### 3.2 Problem Setup

The synthetic counterfactual problem uses a ‘target’ unit and $N$ ‘donor’ units. For SC, the target unit $Y$ is exposed to a treatment or intervention, and the donor units are assumed to be unexposed and, therefore, natural controls. Thus, we are interested in the counterfactual of the target under the control assignment. On the contrary, donors are exposed to the intervention at a time instance and are natural controls prior to that instance for computing SI. Therefore, SC and SI are dual problems and simply differ in the type of intervention the donors receive at the intervention instance.

Each unit is $T$ time units long, and at each timestep, the sequence has $K$ covariates (e.g., Gross Domestic Product, literacy rate, or physiological data), where each covariate may have missing values. Let each element in the target sequence $Y \in \mathbb{R}^{T \times K}$ be denoted by $y^{t,k}$, where $t$ denotes the time instance and $k$ denotes the covariate. Without loss of generality, assume that the covariate of interest is at $k = 1$, and let $T_0$ be the time of intervention. The target sequence $Y = [Y^-, Y^+]$ is thus divided into a pre-intervention sequence $Y^- \in \mathbb{R}^{T_0 \times K}$ and a post-intervention sequence $Y^+ \in \mathbb{R}^{(T - T_0) \times K}$. Let the tensor $X \in \mathbb{R}^{N \times T \times K}$ represent the data from $N$ donor units and $x^{i,t,k}$ represent each element of the donor tensor $X$, where $i$ is the donor, $t$ is the time instance, and $k$ is the covariate. Let $X = [X^-, X^+]$, where $X^- \in \mathbb{R}^{N \times T_0 \times K}$ and $X^+ \in \mathbb{R}^{N \times (T - T_0) \times K}$ denote the pre-intervention and post-intervention data of the donors, respectively. The SC problem involves learning a predictor $f^*_\theta(\cdot)$ for
the post-intervention control of the covariate of interest, \( \hat{y}_{t}^{T} \), \( \forall t \in [T_0 + 1, \ldots, T] \), using the donor data \( X \) and the pre-intervention target sequence \( Y^- \). More formally,

\[
\hat{y}_{t}^{T} = f^T_{\theta}(X, Y^-), \quad \forall t \in [T_0 + 1, \ldots, T].
\]

### 3.3 Transformers

Transformers [40] were proposed to model sequential data. They have achieved state-of-the-art results on several natural language processing tasks. The model consists of stacked layers, where each layer has a self-attention module followed by a feed-forward module, and each module has residual connections [23] and layer normalization [9]. The self-attention module learns powerful internal representations that capture important semantic and syntactic associations across input tokens [28]. This module decomposes each token into a query (\( Q \)), key (\( K \)), and value (\( V \)) vector, and uses these vectors to aggregate global information at each sequence position. For finer details of this architecture, we refer readers to the original article.

### 3.4 Benchmarks

In this section, we describe prior methods to estimate counterfactuals for longitudinal outcomes under a point treatment setting. Since the introduction of the SC technique [4], several modifications to it have been proposed. We describe several state-of-the-art estimators that we benchmark our method against.

#### 3.4.1 Robust Synthetic Control

The RSC method [7] assumes a linear model and introduces a denoising step to the SC method. Furthermore, it uses regularized regression to find the optimal weights \( \beta^* \). Let \( \hat{p} \) be the fraction of missing data in \( X \). The authors replace missing data in \( X \) with 0 and estimate a low-rank matrix \( \hat{M} \) from \( X \) as follows:

\[
X = \sum_{i} u_i \cdot v_i^T, \quad S = \{i : s_i \geq \mu\},
\]

\[
\hat{M} = \frac{\sum_{i \in S} s_i u_i v_i^T}{\max(1 - \hat{p}, \frac{1-\mu}{\mu})}.
\]

The preceding threshold \( \mu \) is a hyperparameter. Let \( \hat{M} = [\hat{M}^-, \hat{M}^+] \), where \( \hat{M}^- \) and \( \hat{M}^+ \) denote the pre-intervention and post-intervention data, respectively. The optimal weights \( \beta^* \) are computed using the LASSO regularized regression as follows:

\[
\beta^* = \arg\min_{\beta} \|Y^- - M^-T \beta\|^2 + \eta \|\beta\|_1.
\]

The counterfactual is predicted through

\[
\hat{Y} = \hat{M}^+T \beta^*.
\]

#### 3.4.2 Multi-Dimensional Robust Synthetic Control

Amjad et al. [6] generalize RSC to incorporate multiple additional covariates to predict the covariate of interest. At a high level, they flatten their donor tensor \( X \in \mathbb{R}^{N \times T \times K} \) to a matrix \( Z \in \mathbb{R}^{N \times KT} \). After that, they run the denoising step from RSC on \( Z \) and perform a covariate-based reweighting on the resultant low-rank matrix and target matrix \( Y \). In their final step, they run regression on the pre-intervention data of the reweighted matrices.

#### 3.4.3 Matrix Completion Using Nuclear Norm Minimization

With regard to Matrix Completion using Nuclear Norm Minimization (MC-NNM), Athey et al. [8] estimate the counterfactual by treating them as missing
values of a matrix. Let \( Y \in \mathbb{R}^{T \times (N+1)} \) denote the outcome matrix of the donor and target units. Then the underlying structure assumed on the matrix \( Y \) by the authors is

\[
Y = L^* + \epsilon,
\]

\[
L^* = \hat{L} + \hat{\tau} \mathbf{1}_T^T + \mathbf{1}_N \hat{\Delta}^T.
\]

Here, \( L^* \) denotes the true mean matrix, \( \hat{\tau} \) models the fixed temporal effect, \( \hat{\Delta} \) accounts for the unit-wide fixed effect, and \( \mathbf{1} \) indicates a vector of ones. Thereafter, the true matrix \( L^* \) is recovered by setting up the following optimization problem with a penalty on the nuclear norm.

\[
(\hat{L}, \hat{\tau}, \hat{\Delta}) = \arg \min_{L, \tau, \Delta} \|Y - L - \tau \mathbf{1}_T^T - \mathbf{1}_N \Delta^T\|_F + \lambda \|L\|_N
\]

3.4.4 Sync-Twin. Qian et al. [33] propose a learning algorithm to learn a deep representation \( c_i \) for the units using pre-intervention covariates. They assume the following factorized data generation model:

\[
Y_{it} = q_t^T c_i + \epsilon_{it},
\]

where \( q_t \) denotes the temporal factors. Thereafter, \( c_i \) is learned by setting up the following reconstruction task over the covariates:

\[
\hat{c}_i = \arg \min_{c_i} \| (\hat{X}_i - X_i) \odot M \|.
\]

Here, \( M \) is a mask for missing covariates. Since the underlying model is linear in \( c_i \), the donor weights are calculated by regressing in the representation space \( C \):

\[
\beta^* = \arg \min_{\beta} ||c_{\text{treated}} - \hat{C}^T \beta||^2.
\]

4 SPATIOTEMPORAL SEQ2SEQ MODELING

In our work, we view the problem through the lens of Seq2Seq modeling. For the intervention instance \( T_0 \), we denote the counterfactual of the target until time \( t \) as \( \hat{Y}^t = [\hat{y}_0, \hat{y}_{t_0+1}, \ldots, \hat{y}_t] \) and similarly the post-intervention of the donor data until \( t \) as \( X^{*t} \in \mathbb{R}^{N \times (T - t_0) \times K} \). Then we model the distribution of the synthetic counterfactual as

\[
P(\hat{Y} | X, Y^-) = \prod_{t = t_0 + 1}^T P_0(\hat{y}_t | \hat{Y}^{t-1}, X^{*t}, X^-, Y^-).
\]

We use an encoder-decoder model, where the encoder \( f_a(.) \) computes a hidden representation \( \mathcal{V} \) for the pre-intervention data \( (Y^-, X^-) \) and passes it to the decoder \( g_b(.) \) that uses the representation and the post-intervention donor data \( X^{*t} \) to autoregressively generate the control \( \hat{y}_t ^t \):

\[
\mathcal{V} = f_a(X^-, Y^-),
\]

\[
P_0(\hat{y}_t | \hat{Y}^{t-1}, X^{*t}, X^-, Y^-) = g_b(\hat{Y}^{t-1}, X^{*t}, \mathcal{V}).
\]

4.1 Architecture

An overview of our Transformer model is given in Figure 3. It encodes pre-intervention data of temporal length \( l^- \) and decodes it into post-intervention data of temporal context \( l^+ \). The pre-intervention data of the target and the donors can be represented by the tensor \( Z^- = [Y^-, X^-] \in \mathbb{R}^{(N+1) \times l^- \times K} \). Similarly, the post-intervention data are represented by \( Z^+ = [\hat{Y}; X^+] \in \mathbb{R}^{(N+1) \times l^+ \times K} \). The standard Transformer receives a 1D sequence of input tokens. Hence, we flatten \( Z^- \) and \( Z^+ \) into a 1D sequence \( Z_{flat}^- \in \mathbb{R}^{(N+1)l^- \times K} \) and \( Z_{flat}^+ \in \mathbb{R}^{(N+1)l^+ \times K} \), respectively. We
then project each token into the hidden dimension $D$ of the Transformer via a trainable linear weight $W_e \in \mathbb{R}^{K \times D}$ to give sequences $E^- \in \mathbb{R}^{(N+1)T \times D}$ and $E^+ \in \mathbb{R}^{(N+1)T' \times D}$. More precisely,

$$E^- = Z_{flat} W_e \left[ x^{i:t} W_e; \cdots x^{N,t} W_e; y^{i:t} W_e \right]_{t=T_0-1}^{T_0},$$

$$E^+ = Z_{flat} W_e \left[ x^{i:t} W_e; \cdots x^{N,t} W_e; \hat{y}^{i:t} W_e \right]_{t=T_0+1}^{T_0+T'},$$

4.1.1 Positional and Target Embeddings. Each token in the sequence belongs to one of $N + 1$ tokens, one of $T$ time instances, and either a donor or the target unit. We therefore inject a learnable spatial embedding $E_{\text{spatial}}(.) \in \mathbb{R}^{(N+1)T \times D}$, time embedding $E_t(.) \in \mathbb{R}^{T \times D}$, and a target embedding $E_{\text{target}}(.) \in \mathbb{R}^{T \times D}$ to enable the model to differentiate between spatiotemporal positions of the token in the data matrix as well as separate donors from the target unit. The resultant encoder input $H^-$ and the decoder input $H^+$ sequences are obtained as follows:

$$H^- = E^- + E_{\text{spatial}}(E^-) + E_t(E^-) + E_{\text{target}}(E^-),$$

$$H^+ = E^+ + E_{\text{spatial}}(E^+) + E_t(E^+) + E_{\text{target}}(E^+).$$

4.1.2 Encoder. We use the vanilla bidirectional Transformer encoder with a latent dimension $D$ consisting of $l$ stacked identical layers. The encoder processes the input sequence $H^-$ and outputs a sequence of representations.
\[ \mathcal{V} \] over the input. A key \( K \) and value \( V \) vector are computed over each of the tokens in \( \mathcal{V} \) and passed to the decoder.

### 4.1.3 Decoder

The decoder is tasked with autoregressively generating the counterfactual \( \hat{Y} \), given the encoder output \( \mathcal{V} \) and sequence \( H^+ \). Our decoder design mostly follows the vanilla Transformer decoder with \( l \) stacked identical layers. Each layer has two kinds of attention, namely a causal self-attention module that operates over the decoder hidden states and the ‘encoder-decoder’ attention module that operates over the joint representation of the encoder and decoder. However, we do modify the causal attention mask used in the self-attention module to account for the tokens that lie on the same temporal slice but differ spatially. In other words, we enforce temporal causality but allow bidirectionality spatially. The modified causal mask for spatiotemporal data is illustrated in Figure 4. We then project the hidden state of SC via a linear weight \( W_d \in \mathbb{R}^{D \times 1} \) to obtain the prediction.

Algorithm 1 lays out the skeleton of our model.

**ALGORITHM 1:** Transformer \((X^-, X^+, Y^-, Y^+) \rightarrow \hat{Y}\)

```plaintext
Require: Encoder, Decoder, \( We, W_d, E_{Spatial}, E_t, E_{Target} \)
\[ Z_{flat}^- = flatten(\text{concat}(X^-, Y^-)) \] \( \rightarrow (N + 1)l^- \times K \)
\[ Z_{flat}^+ = flatten(\text{concat}(X^+, Y^+)) \] \( \rightarrow (N + 1)l^+ \times K \)
\[ E^- = Z_{flat}^- W_e \] \( \rightarrow (N + 1)l^- \times D \)
\[ E^+ = Z_{flat}^+ W_e \] \( \rightarrow (N + 1)l^+ \times D \)
\[ H^- = E^- + E_{Spatial}(E^-) + E_t(E^-) + E_{Target}(E^-) \]
\[ H^+ = E^+ + E_{Spatial}(E^+) + E_t(E^+) + E_{Target}(E^+) \]
\[ \mathcal{V} = \text{Encoder}_e(H^-) \]
\[ Z = \text{Decoder}_d(H^+, \mathcal{V}) \]
\[ \hat{Z} = Z_{target} \] \( \rightarrow \text{Target Hidden States} \)
\[ \hat{Y} = \hat{Z} W_d \] \( \rightarrow l^+ \times K \)
```

Fig. 4. Comparison of the Vanilla decoder mask that enforces temporal causality and a modified causal mask (a) and that additionally allows spatial bidirectionality (b).
4.2 Data Preprocessing
We replace missing data with 0 and scale each covariate in the data between 0 and 1 before processing by the Transformer. Motivated by mRSC [6], we apply a low-rank transformation to the data tensor $X \in \mathbb{R}^{N \times T \times K}$ when data are missing. We flatten $X$ to $X_{flat} \in \mathbb{R}^{N \times TK}$ and retain the top $m$ singular values of $X_{flat}$.

4.3 Pseudo-Counterfactual Prediction Pre-Training
Transformer-based language models are usually pre-trained on an unsupervised task [18, 34] to learn high-capacity representations that help boost downstream performance on discrimination tasks. In the same way, we pre-train the model on donor data to reliably reconstruct the counterfactual, the ground truth of which is known. In each training iteration, we sample a donor unit $X_i \in \mathbb{R}^{1 \times T \times K}$, $i \in [1, \ldots, N]$ and an intervention time $T' \in [1, \ldots, T]$. We treat the sampled donor as a pseudo-target and task our model with generating its post-intervention counterfactual. Let $X^-_{\overline{i}} \in \mathbb{R}^{(N-1) \times l^- \times K}$ and $X^+_{\overline{i}} \in \mathbb{R}^{(N-1) \times (l^+ - l^-) \times K}$ denote the pre-intervention and post-intervention donor data excluding the sampled donor, respectively, and let $\hat{x}^-_i \in \mathbb{R}^{(N-1) \times (t - T') \times K}$ denote the post-intervention control of the pseudo-target until time $t$ and $\hat{x}^+_i$ be the control at instance $t$. The objective of the model is to maximize the following log-likelihood:

$$L(\alpha, \phi) = \sum_{t=T+1}^{T'+l^+} \log(P(\hat{x}^+_i | \hat{x}^{-1}_i, X^+_{\overline{i}}, X^-, \alpha, \phi)).$$

(24)

We use the teacher forcing algorithm [42] while training and assume a Gaussian model for the likelihood that reduces the loss function to squared error.

4.4 Fine-Tuning
Fine-tuning proceeds by fitting the model on the pre-intervention data $Y^-$ of the target unit using the pre-intervention donor data $X^-$. In each fine-tuning iteration, we sample a time instance $T'$ in $[1, \ldots, T_0]$ for use as the pseudo-intervention instance. Then, we treat $X^-$ and $Y^-$ as the data in hand and divide it into pseudo-pre-intervention and pseudo-post-intervention data to predict the pseudo-post-intervention of the target.

4.5 Inference
We generate the SC $\hat{Y}$ in a sliding window fashion, where we start at $T_0$ and generate post-intervention data of temporal length $l^+$ each time. The generated control is used as pre-intervention data for the subsequent time steps. Algorithm 2 shows the pseudo-code for training and inference.

4.6 Visualization
A key property of the simple linear model assumed by the vanilla SC method is its interpretability in the form of donor weights. Modeling the spatiotemporal context helps us visualize the donor contributions across different time instances. Transformers are effective at long-range credit assignment, and the attention score $a_{ij}$ in the attention matrix $A$ of the Transformer specifies the proportion with which a token $i$ attends to token $j$. We probe the self-attention layers of the decoder and extract the attention scores of various donors while predicting the target token. Although not entirely dispositive of the donor significance, this can help domain experts uncover the relations between targets and important donors.

We qualitatively compare our proposed approach against prior techniques in Table 1 with respect to four features: (1) nonlinearity (whether the model captures nonlinear dynamics), (2) missing data (whether the proposed algorithm considers missing data values for the outcome variable and covariates), (3) temporal context (whether
ALGORITHM 2: Training and Inference

Require: Transformer(α, ϕ), Learning rate λ
» Pre-training(X, iters)
for j = 1 to iters do
    i = random.sample(N)
    T₀ = random.sample(T)
    Y = X[i]
    X' = X.remove(i)  
    Y⁻, Y⁺ = Y [T₀ - l⁻ : T₀], Y [T₀ : T₀ + l⁺]
    X⁻, X⁺ = X'[i : T₀ - l⁻ : T₀], X'[i : T₀ : T₀ + l⁺]
    preds = Transformer(X⁻, X⁺, Y⁻, Y⁺)
    L = mean ((Y⁺ - preds)²)
    α ← α - λ∇α L, \phi ← \phi - λ∇ϕ L
end for
» Fine-tuning(X, Y, T₀, iters)
for j = 1 to iters do
    T' = random.sample(T₀)
    Y' = Y [i : T₀]
    X' = X [i : T₀]
    Y⁻, Y⁺ = Y' [T' - l⁻ : T'], Y' [T' : T' + l⁺]
    X⁻, X⁺ = X'[i : T' - l⁻ : T'], X'[i : T' : T' + l⁺]
    preds = Transformer(X⁻, X⁺, Y⁻, Y⁺)
    L = mean ((Y⁺ - preds)²)
    α ← α - λ∇α L, \phi ← \phi - λ∇ϕ L
end for
» Inference(X, Y, T₀) → Ỹ
X⁻ = X [i : T₀ - L⁻ : T₀]
Ỹ = []
for l = 1 to l⁺ do
    X⁺ = X [i : T₀ + l]
    preds = Transformer(X⁻, X⁺, Y⁻, Ỹ)
    Ỹ = Ỹ.append(prefs[i])
end for
return Ỹ

the model implicitly or explicitly considers the prior context), and (4) multimodal inputs (whether multimodal data like images, audio, and discrete classes can be processed by the proposed approach).

5 EXPERIMENTS
We demonstrate the efficacy of our method through quantitative and qualitative experiments. We give the training details, additional results, and hyperparameters for each experiment in the appendix.

5.1 Synthetic Data
Counterfactual estimators are difficult to evaluate on real-life data since the ground truth counterfactual is unknown. Hence, we benchmark our method on synthetic data where the underlying data generation process is
known \textit{a priori}. We set up an SC problem here by constraining the donors to be controls. The ‘mean’ synthetic data \( M \) is generated using a latent factor model. \( M \in \mathbb{R}^{(N+1) \times T \times 2} \) has two covariates, and the first row represents the target data with the remaining tensor representing donor data. Each row \( i \) and column \( j \) is assigned a latent factor \( \theta_i \) and \( \rho_j \), respectively, and covariate \( k \) is generated using \( f_k(\theta_i, \rho_j) \). We use
\begin{align*}
    f_1(\theta_i, \rho_j) &= \theta_i + \frac{\rho_j \times \theta_i}{T} \exp \left( \frac{\rho_j}{T} \right) + \cos \left( 2F_1 \frac{\pi}{180} \right) \\
    &\quad + \sin \left( F_1 \frac{\pi}{180} \right) + \cos \left( 2F_2 \frac{\pi}{180} \right) + \sin \left( F_2 \frac{\pi}{180} \right),
\end{align*}
\begin{align*}
    f_2(\theta_i, \rho_j) &= \frac{10}{1 + \exp(-\theta_i - \frac{\rho_j}{7} - 0.7\frac{\theta_i \rho_j}{T})}.
\end{align*}
Here, \( F_1 = (\rho_j \mod 360) \) and \( F_2 = (\rho_j \mod 180) \). Thereafter, we add noise \( \epsilon \sim N(0, \sigma^2) \) to each entry in \( M \) to obtain the observed matrix \( O \). Given an intervention instance \( T_0 \), we task the estimator to reconstruct the post-intervention (\( t > T_0 \)) mean sequence of the first covariate from noisy observations. Note that our data generation process is similar to that of Amjad et al. [6], with the critical distinction being that the mean target sequences are generated from a high-rank latent factor model instead of being linearly extrapolated from donor units composed from a low-rank latent factor model. This makes the task inherently ‘harder.’ For all the experiments run on synthetic data, we set \( T = 2,000, T_0 = 1,600, \) and \( \rho_j = j, \) and \( \theta_i \) is uniformly sampled from \([0, 1]\). Our proposed method is compared against all the benchmark methods mentioned in Section 3.4, using the \textbf{Root Mean Squared Error (RMSE)} between the estimates and the true post-intervention mean. In addition, to this we also learn a simple Transformer-based decoder baseline to perform time series modeling on the donor trajectories. This baseline therefore corresponds to the extreme case of zero donors used to extrapolate the counterfactual under the modeling framework.

5.1.1 Robustness against Noise. Next, we investigate the robustness of our method against noise in the observed data. We set \( N = 10 \) and generate three synthetic datasets with varying noise levels \( \sigma^2 \in \{0.5, 1, 2\} \). The RMSE scores are reported in Figures 5 and 6 and show the SC estimates. The Transformer baseline performs the worst, which is expected given traditional ‘Transformers’ data inefficiency. This also indicates the difficulty of using simple time series extrapolation with few donors and shows the benefits of directly modeling the donor outcomes to perform few-shot counterfactual estimation. As noise increases, our predictions remain robust and lie close to the true mean without using any denoising step. Moreover, pre-training is essential for helping our model learn a useful representation of the underlying latent model, thus in making accurate predictions.
Fig. 5. Comparison between post-intervention RMSE of our approach and prior estimators using observations with varying noise levels. Our approach surpasses or performs at par with these techniques at all levels of noise. Moreover using donors and therefore spatiotemporal transformers (yellow) gives 6 to 20× reduction in error over simple temporal modeling of control path using Transformers (purple). Error bars are plotted at 90% confidence intervals.

Fig. 6. Estimates of the post-intervention mean of the target from noisy observations for $N = 10$ and varying levels of noise. The vertical black line indicates the intervention instance.

5.1.2 Effect of the Donor Pool Size. From Figure 7, we gain insight into the effect of the donor pool size on the accuracy of our prediction. Synthetic counterfactual problems often tend to have a small number of donors—for example, few states tend to experience similar policies and rare diseases afflict a small number of the population. Hence, we set $N \in [5, 10, 15, 25, 50, 100]$ with $\sigma^2 = 1$. Our estimator achieves a lower RMSE for most
Fig. 7. Post-intervention RMSE of various estimators with varying donor pool size. Our approach performs competitively for most donor pool sizes with the pre-training step being crucial to achieving an accurate prediction. Using spatiotemporal transformers (yellow) reduces error over a traditional Transformer (purple) by 6 to 19×. Plotted at 90% confidence intervals.

We evaluate our approach on real-life datasets at both population and individual levels spanning public health, randomized trials, and rare diseases.

5.2 Analyzing Public Health Policy: California Proposition 99

In 1988, California became the first in the United States to pass a large anti-tobacco law, Proposition 99, which hiked the excise tax on cigarette sales by 25 cents. To evaluate the effectiveness of this law, we construct a synthetic California without Proposition 99 and measure the per capita cigarette sales in this synthetic unit. The donor pool consists of 38 control states where no significant policies for tobacco control were introduced and additional covariates like beer consumption, population, and income are included. Figure 9 compares the California counterfactual prediction of our method and various baselines, all showing that the cigarette sales would have been higher in the absence of the law. We infer that per-capita cigarette sales fell by 45 packets.

\(^1\)mRSC yields a poor pre-intervention fit and has been omitted. Sync-Twin is not amenable to missing outcome values and is omitted for real data evaluations.
in real California toward the year 2000 compared to synthetic California. Moreover, our model makes the intuitive prediction that in the absence of the law, cigarette sales in California would converge to the national average. As explained in Section 4.6, attention scores of the model can be used to extract the contribution of the donors in making the counterfactual prediction. These weights across the donor space and time are illustrated in Figure 10, indicating the combination that best reproduces the outcome before the passage of Proposition 99.

5.3 SCoUT for In Silico Randomized Trials: Drug Trials for Childhood Asthma

At the patient level, the synthetic counterfactual is a dynamic, virtual representation of the human being over time and enables applications for in silico clinical trials. As an example, we look at CAMP (the Childhood Asthma Management Program) [22], an RCT designed to study the long-term pulmonary effects of three treatments (budesonide, nedocromil, and placebo) on children with mild-to-moderate asthma. The trial’s placebo arm
Fig. 9. Comparison of methods on synthetic vs. true California. We infer that the passage of Proposition 99 reduced per-capita cigarette sales from the gap between observed California (yellow) and the synthetic counterfactual of California (green). Our model also predicts that in the absence of Proposition 99, cigarette sales in California would slowly converge to the national average (dotted line).

contains anonymized longitudinal data of 275 patients with more than 20 spirometry measurements per patient. Pre-Bronchodilator Forced Expiratory Volume to Forced Vital Capacity (PreFF) ratio is a vital metric of lung capacity in asthma patients that measures volume of air that an individual can exhale during a forced breath prior to the usage of a bronchodilator. Here, we model the control arm of the RCT and predict the PreFF of a target patient using the other placebos as donors. We consider two settings, where pre-intervention lengths are 35% and 75% of the total trajectory. One at a time, we set one patient as the target unit and consider the others as donors. In this fashion, we model the control paths of the first five patients in the placebo arm, the ground truths available to us. The average RMSE across these patients is reported in Figure 11. We plot the counterfactual estimates for patients in Figure 12. Our method generates a reliable control along with MC-NNM, whereas estimates produced by RSC and mRSC are highly biased. We posit that local spatiotemporal mapping is reasonably effective in controlling for time-dependent confounders, whereas linear estimators that assign time-agnostic weights suffer significantly.

5.4 Me, My Doctor, and My Digital Twin: A Case Study on FA
FA is a fatal degenerative nervous system disorder with no cure. As a progressive disease with a wealth of available data and a growing number of potential therapeutic interventions, FA is one of the many genetic diseases that can benefit from precision medicine. We use clinical data collected during the FA-COMS (FA Clinical Outcome Measure Study) cohort [35], a natural history study involving yearly assessments of a core set of clinical measures and quality of life assessments, and maintained by the Critical Path Institute. A total of 163 metrics were included in our model based on clinical relevance and minimal missing data. Each metric can be classified as (1) either longitudinal or static and (2) binary, ordinal, categorical, or continuous. Prior works only consider continuous data and, unlike the generalized SCouT framework, are therefore not amenable to categorical data. A recent study found that calcitriol, the active form of vitamin D, is able to increase frataxin
Fig. 10. Figures (a), (b), and (c) on the left visualize donor attentions for the three layers of the Transformer. Attention is sparse and spread across specific donors, with the sparsity becoming more pronounced in the deeper layers. For each attention distribution map, we extract the top 5 donors and plot them against our synthetic counterfactual on the right. We note that New Hampshire, which has highest cigarette sale per capita among all the states, is consistently used across all attention layers. However, the model selects specific donors at every layer to iteratively refine its counterfactual.
levels and restore mitochondrial function in cell models of FA. FA is caused by a deficiency of frataxin. Accordingly, calcitriol supplements could potentially improve health outcomes for FA patients [13]. We explore the effect of calcitriol as an example of a synthetic medical intervention. Our resulting dataset tailored for the calcitriol SI analysis included $N = 21$ donor units who indicated that they were taking a calcitriol supplement at some point during the study and had sufficient pre-intervention data for training (see Appendix D.2 for data-formatting details). The dataset was aligned such that every donor patient began taking the supplement at $T_0 = 8$. A few patients never indicated that they had taken the supplement during the study period; hence, we set these patients as the target unit and simulated the counterfactual situation under which the patients do receive treatment at $T_0$. It is important to note that unlike RCTs, where we evaluate the effect of a treatment on the average individual, SI presents the predicted treatment effect on an individual patient, in line with the goals of precision medicine. Imagine an FA patient walking into a clinic, where the physician has access to data from patient’s previous health records and is deciding whether or not to recommend a calcitriol supplement. The physician can use the SI technique to generate a counterfactual under this treatment for any desirable metric, as we demonstrate next. We evaluated the patients on several rating scales commonly used for FA patients [21]:

- FARSn (FA Rating Scale neurologic examination) [21] is one of the most used tests to evaluate an FA patient’s disease progression. It involves scoring on several subscales broadly divided into bulbar, lower limb, upper limb, peripheral nervous system, and upright stability functionalities.
- The 9-HPT (the Nine-Hole Peg Test) is a quantitative measure of finger dexterity and is conducted on both the dominant and non-dominant arm, in that order. The patient is told to pick up and place pegs into open holes on a board in front of them.
- The ADL (Activities of Daily Living) Scale is widely used to rank adequacy and independence in basic tasks that a person could expect to encounter every day, including grooming, dressing, walking, and drinking.

We compare the predictions under the counterfactual scenario where the patients begin to take a calcitriol supplement at time $T_0$ against the true observed values for the patients with no intervention. We included metrics from the FARSn, 9-HPT, and ADL exams that demonstrated the counterfactual under calcitriol use for the
Fig. 12. Estimates of PreFF for participants in the placebo group using other units in the placebo group as donors. The vertical black line indicates intervention instance. Our model reliably generates SC for both pre-intervention periods.

patients. Figures 13 through 15 show the predictions for the FARSn, ADL, and 9-HPT scales, respectively. Most counterfactuals seem to suggest that calcitriol use would have been beneficial with respect to all metrics highlighted in the following.

Across specific metrics and select patients, comparing the calcitriol counterfactual and observed patient values also predicted little benefit from the supplement or were otherwise difficult to extract qualitative meaning from.
Finally, none of the comparisons indicated that calcitriol would allow further disease progression in the patients, which is comforting evidence that, at worst, calcitriol has little effect. At best, the supplement could greatly improve the quality of life for a patient with FA.

6 CONCLUSION

SC-based methods paved the way for more effective techniques to estimate the counterfactual of a unit from several donor units. However, these methods are still restricted because they cannot capture inter-unit and intra-unit temporal contexts and are parametrically constrained to model linear dynamics. We introduced a Transformer-based Seq2Seq model that leverages local spatiotemporal relations to estimate the counterfactual sequence through multimodal inputs better and overcome these limitations. Experiments on synthetic and real-world data demonstrate that our model has high predictive power even in noisy data and donor pools of varying sizes. Explicitly modeling donor units is a strong inductive bias and significantly enhances predictive power compared to time series modeling. Our method leverages the Transformer model’s credit assignment ability to analyse the first-of-its-kind temporal evolution of donor contribution. We also demonstrated the applicability of our Transformer-based model in simulating counterfactual scenarios in healthcare (e.g., to accelerate in silico clinical trials by simulating the control counterfactual for individuals in an asthma randomized trial and generating the calcitriol intervention counterfactual for FA patients). As big data and artificial intelligence continue to revolutionize healthcare, our work can be adapted for use beyond clinical trials and clinical decision making and create measurable improvements in public health.
Fig. 14. Tailored predicted scores for each patient under different ADL scale metrics. The vertical black line indicates intervention instance. A higher score is associated with a greater degree of disability. SCouT predicts considerable improvements after using calcitriol across all metrics and all patients.
Fig. 15. Tailored counterfactuals for each patient under 9-HPT metrics. The vertical black line indicates the intervention instance. The scores measure the time it takes for the patient to place the pegs into the hole; a higher score indicates a greater level of disability. SCouT predicts improvement in test times for all patients across both the dominant and non-dominant hands.

APPENDICES

A MODEL DETAILS

Our model is implemented using the Hugging Face library [44]. Our choice of the encoder-decoder model is a Bert2Bert model. The hyperparameters we use for various experiments are outlined in Table 2.

B PROPOSITION 99

Five covariates are used for generating the control:

- Per-capita cigarette consumption (in packs)
- Per-capita state personal income (logged)
- Per-capita beer consumption
- Percentage of the state population aged 15 to 24 years
- Average retail price per pack of cigarettes (in cents)

C ASTHMA

The CAMP dataset uses the FEV (Forced Expiratory Volume) and FVC (Forced Vital Capacity) as the lung functions. A total of 16 continuous covariates from the dataset are used for generating the control:

- Pre-Bronchodilator FEV/FVC ratio % (PREFF)
- Age in years at randomization (age rz)
- Hemoglobin (hemog)
- Pre-Bronchodilator FEV (PREFEV)
- Pre-Bronchodilator FVC (PREFVC)
- Pre-Bronchodilator peak flow (PREPF)
- Post-Bronchodilator FEV (POSFEV)
- Post-Bronchodilator FVC (POSFVC)
- Post-Bronchodilator FEV/FVC ratio % (POSFF)
Table 2. Model and Training Hyperparameters

| Hyperparameter                  | Value                                                                 |
|--------------------------------|----------------------------------------------------------------------|
| Number of layers                | Synthetic: 3                                                          |
|                                | Prop. 99: 3                                                           |
|                                | Asthma: 4                                                             |
|                                | FA: 4                                                                 |
| Number of attention heads       | 1                                                                    |
| Hidden dimension                | Synthetic: 32                                                         |
|                                | Prop. 99: 128                                                          |
|                                | Asthma: 128                                                           |
|                                | FA: 128                                                               |
| Non-linearity                   | GELU                                                                  |
| dropout                         | 0.1                                                                  |
| Pre-intervention length         | 10 (N = 5,10,15)                                                      |
|                                | Synthetic: 20                                                         |
|                                | Prop. 99: 5                                                           |
|                                | Asthma: 2                                                             |
|                                | FA: 5                                                                 |
| Post-intervention length        | 5 (N = 25,50,75,100)                                                  |
|                                | Synthetic: 10                                                         |
|                                | Prop. 99: 2                                                           |
|                                | Asthma: 2                                                             |
|                                | FA: 2                                                                 |
| #Singular values retained       | Synthetic: 5                                                          |
|                                | Prop. 99: 3                                                           |
|                                | Asthma: 55                                                            |
|                                | FA: 5                                                                 |
| Learning rate                   | $1 \times 10^{-4}$                                                    |
| weight decay                    | $1 \times 10^{-4}$                                                    |
| Warmup steps (pre-training)     | Synthetic: $1 \times 10^4$                                            |
|                                | Prop. 99: $1 \times 10^3$                                             |
|                                | Asthma: $1 \times 10^4$                                               |
|                                | FA: $1 \times 10^4$                                                   |
| Pre-training iterations         | Synthetic: $5 \times 10^4$                                            |
|                                | Prop. 99: $5 \times 10^3$                                             |
|                                | Asthma: $5 \times 10^4$                                               |
|                                | FA: $5 \times 10^4$                                                   |
| Fine-tuning iterations          | Synthetic: $1 \times 10^4$                                            |
|                                | Prop. 99: $1 \times 10^3$                                             |
|                                | Asthma: $5 \times 10^3$                                               |
|                                | FA: $5 \times 10^4$                                                   |
| Batch size                      | Synthetic: 128                                                        |
|                                | Prop. 99: 64                                                           |
|                                | Asthma: 16                                                            |
|                                | FA: 16                                                                |
• Post-Bronchodilator peak flow (POSPF)
• Pre-Bronchodilator FEV % prediction (PREFEVPP)
• Pre-Bronchodilator FVC % prediction (PREFVCPP)
• Post-Bronchodilator FEV % prediction (POSFEVPP)
• Post-Bronchodilator FVC % prediction (POSFVCPP)
• White blood cell count (wbc)
• Age of current home (agehome)

D FRIEDREICH’S ATAXIA

D.1 FA Metrics

The following metrics were included in the clinical dataset for FA:

• Age of onset
• Length of shorter GAA Repeat
• Presence of point mutation
• Death Flag
• Ethnicity
• Sex
• Race
• Assistive Device Type
• Level of education
• Living circumstances
• The FARSn examination subscores
• The Timed 25-Foot Walk
• The 9-HPT: Dominant Hand and Non-Dominant Hand
• ADL Scale
• Bladder Control Scale
• Bowel Control Scale
• Impact of Visual Impairment Scale
• Modified Fatigue Impact Scale
• MOS Pain Effects Scale
• SF-10 and SF-36

D.2 Data Formatting for FA

Since the FA dataset we use comes from a natural history study that does not introduce any interventions itself, we have to reformat the dataset so it can be used for an SI. The FA data includes longitudinal information regarding the medications a patient is taking. Out of all the patients, 72 indicated that they were taking a calcitriol supplement at some point during the study. We manually searched for patients who reported taking calcitriol for an extended period of time (typically 3 or more years) and had sufficient pre-intervention data for training. The resulting dataset tailored for the calcitriol SI analysis included $N = 21$ and was aligned using the methodology from Figure 16. Thus, every donor patient began taking the supplement at $T_0 = 8$. Four patients who had not taken the supplement during the study period were then set as the target unit. Furthermore, the narrowed scope of the question we investigate in this section also gives us a use case to test our model on real-world FA data with a much smaller number of donor units.
REFERENCES

[1] Alberto Abadie. 2021. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature* 59 (2021), 391–425.

[2] Alberto Abadie and Matias D. Cattaneo. 2018. Econometric methods for program evaluation. *Annual Review of Economics* 10, 1 (2018), 465–503.

[3] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. 2010. Synthetic control methods for comparative case studies: Estimating the effect of California’s tobacco control program. *Journal of the American Statistical Association* 105, 490 (2010), 493–505.

[4] Alberto Abadie and Javier Gardeazabal. 2003. The economic costs of conflict: A case study of the Basque country. *American Economic Review* 93, 1 (March 2003), 113–132.

[5] Anish Agarwal, Devavrat Shah, and Dennis Shen. 2020. Synthetic interventions. CoRR abs/2006.07691 (2020).

[6] Muhammad Amjad, Vishal Misra, Devavrat Shah, and Dennis Shen. 2019. mRSC: Multi-dimensional robust synthetic control. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 3, 2 (June 2019), Article 37, 27 pages.

[7] Muhammad Amjad, Devavrat Shah, and Dennis Shen. 2018. Robust synthetic control. *Journal of Machine Learning Research* 19, 22 (2018), 1–51.

[8] Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2017. Matrix completion methods for causal panel data models. CoRR abs/1710.10251 (2017).

[9] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. CoRR abs/1607.06450 (2016).

[10] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* 3 (2003), 1137–1155.

[11] Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. 2020. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. arXiv:2002.04083 (2020).

[12] Sarah Bohn, Magnus Lofstrom, and Steven Raphael. 2014. Did the 2007 legal Arizona workers act reduce the state’s unauthorized immigrant population? *Review of Economics and Statistics* 96, 2 (2014), 258–269.

[13] Elena Britti, Fabien Delaspre, A. Sanz-Alcázar, Marta Medina-Carbonero, Marta Llovera, Rosa Purroy, Stefka Mincheva-Tasheva, Jordi Tamarit, and Joaquim Ros. 2021. Calcitriol increases frataxin levels and restores mitochondrial function in cell models of Friedreich ataxia. *Biochemical Journal* 478, 1 (2021), 1–20.

[14] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. CoRR abs/2005.14165 (2020).

[15] David Card. 1990. The impact of the Mariel Boatlift on the Miami labor market. *ILR Review* 43, 2 (1990), 245–257.

[16] David Card and Alan B. Krueger. 1993. *Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania*. Working Paper 4509. National Bureau of Economic Research.
[17] Scott Cunningham and Manisha Shah. 2017. Decriminalizing Indoor prostitution: Implications for sexual violence and public health. Review of Economic Studies 85, 3 (2017), 1683–1715.
[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1. 4171–4186.
[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations.
[20] Nikolay Doudchenko and Guido W. Imbens. 2017. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. CoRR abs/1610.07748 (2017).
[21] M. C. Fahey, L. Corben, V. Collins, A. J. Churchyard, and M. B. Delatycki. 2007. How is disease progress in Friedreich’s ataxia best measured? A study of four rating scales. Journal of Neurology, Neurosurgery and Psychiatry 78, 4 (2007), 411–413.
[22] Gail G. Shapiro et al. 1999. The Childhood Asthma Management Program (CAMP): Design, rationale, and methods. Controlled Clinical Trials 20, 1 (Feb. 1999), 91–120.
[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
[24] Cheng Hsiao, H. S. Ching, and S. K. Wan. 2012. A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong Kong with Mainland China. Journal of Applied Econometrics 27, 5 (2012), 705–740.
[25] Guido W. Imbens and Jeffrey M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. Journal of Economic Literature 47, 1 (March 2009), 5–86.
[26] Henrik Jacobsen Kleven, Camille Landais, and Emmanuel Saez. 2013. Taxation and international migration of superstar: Evidence from the European football market. American Economic Review 103, 5 (2013), 1892–1924.
[27] Inke R. König, Oliver Fuchs, Gesine Hansen, Erika von Mutius, and Matthias V. Kopp. 2017. What is precision medicine? European Respiratory Journal 50, 4 (2017), 1700391.
[28] Christopher D. Manning, Kevin Clark, John Hewit, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. Proceedings of the National Academy of Sciences 117, 48 (2020), 30046–30054.
[29] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. 2022. Causal transformer for estimating counterfactual outcomes. CoRR abs/2204.07258 (2022).
[30] Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Honza Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In Proceedings of the Conference of the International Speech Communication Association.
[31] Huy Hoang Nguyen, Simo Saarikka, Matthew B. Blashko, and Aleksi Tiulpin. 2021. CLIMAT: Clinically-inspired multi-agent transformers for knee osteoarthritis trajectory forecasting. CoRR abs/2104.03642 (2021).
[32] Paolo Pinotti. 2015. The economic costs of organised crime: Evidence from Southern Italy. Economic Journal 125, 586 (2015), F203–F232.
[33] Zhaozhi Qian, Yao Zhang, Ioana Bica, Angela Wood, and Mihaela van der Schaar. 2021. SyncTwin: Treatment effect estimation with longitudinal outcomes. Advances in Neural Information Processing Systems 34 (2021), 3178–3190.
[34] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. Technical Report. Open AI.
[35] Sean R. Regner, Nicholas S. Wilcox, Lisa S. Friedman, Lauren A. Syser, Kim A. Schadt, Karlla W. Braggit, Susan Perlmam, Martin Delatycki, George R. Wilmot, Christopher M. Gomez, Khalaf O. Bushara, Katherine D. Mathews, S. H. Subramony, Tetsuo Ashizawa, Bernard Ravina, Alicia Brocht, Jennifer M. Farmer, and David R. Lynch. 2012. Friedreich ataxia clinical outcome measures: Natural history evaluation in 410 participants. Journal of Child Neurology 27, 9 (Sept. 2012), 1152–1158.
[36] Paul R. Rosenbaum and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 1 (1983), 41–55.
[37] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2016. Estimating individual treatment effect: Generalization bounds and algorithms. CoRR abs/1607.03976 (2016).
[38] Bonnie Sibbald and Martin Roland. 1998. Understanding controlled trials: Why are randomised controlled trials important? British Medical Journal 316, 7126 (1998), 201.
[39] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. CoRR abs/1409.3215 (2014).
[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in Neural Information Processing Systems 30 (2017), 5998–6008.
[41] Xiang Wang, David Sontag, and Fei Wang. 2014. Unsupervised learning of disease progression models. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 85–94.
[42] Ronald J. Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. Neural Computation 1, 2 (1989), 270–280.
[43] Coady Wing, Kosali Simon, and Ricardo A. Bello-Gomez. 2018. Designing difference in difference studies: Best practices for public health policy research. Annual Review of Public Health 39, 1 (2018), 453–469.
[44] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quintin Lhoest, and Alexander M. Rush. 2020. HuggingFace’s transformers: State-of-the-art natural language processing. CoRR abs/1910.03771 (2020).

[45] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In Proceedings of the International Conference on Learning Representations.

Received 21 November 2022; accepted 9 August 2023