Estimating Individual Treatment Effects with Time-Varying Confounders

Ruoqi Liu*, Changchang Yin*, Ping Zhang*
*The Ohio State University, Columbus, Ohio, USA 43210
Email: {liu.7324, yin.731, zhang.10631}@osu.edu

***Abstract*—Estimating the individual treatment effect (ITE) from observational data is meaningful and practical in healthcare. Existing work mainly relies on the *strong ignorability assumption* that no hidden confounders exist, which may lead to bias in estimating causal effects. Some studies considering the hidden confounders are designed for static environment and not easily adaptable to a dynamic setting. In fact, most observational data (e.g., electronic medical records) is naturally dynamic and consists of sequential information. In this paper, we propose Deep Sequential Weighting (`DSW`) for estimating ITE with time-varying confounders. Specifically, `DSW` infers the hidden confounders by incorporating the current treatment assignments and historical information using a deep recurrent weighting neural network. The learned representations of hidden confounders combined with current observed data are leveraged for potential outcome and treatment predictions. We compute the time-varying inverse probabilities of treatment for re-weighting the population. We conduct comprehensive comparison experiments on fully-synthetic, semi-synthetic and real-world datasets to evaluate the performance of our model and baselines. Results demonstrate that our model can generate unbiased and accurate treatment effect by conditioning both time-varying observed and hidden confounders, paving the way for personalized medicine.**

***Index Terms*—deep learning, electronic medical record, ITE, time-varying confounders**

I. Introduction

Estimating the individual treatment effect (ITE) is a task of evaluating the causal effect of treatment strategies on some important outcomes over individual-level, which is a significant problem in many areas [1]–[5]. For example, in healthcare domain, it is critical to prescribe personalized medicines (treatments) for different patients based on their health conditions. To approach this, randomized controlled trial (RCT) is usually conducted, which is accomplished by randomly allocating patients to two groups, treating them differently (i.e., one group has intervention and the other has a placebo or no intervention) and comparing them in terms of a measured response. However, conducting RCTs in the healthcare domain is extremely expensive and time-consuming, if not impossible, due to the requirement of tremendous expert effort and the consideration of ethical issues.

Observational data contain patient records, including their demographic information, vital signs, lab tests and outcomes but without having complete knowledge of why a specific treatment is applied to a patient. The accumulation of observational data in electronic medical records (EMRs) offers a promising opportunity for ITE estimation when RCTs are expensive or impossible to conduct.

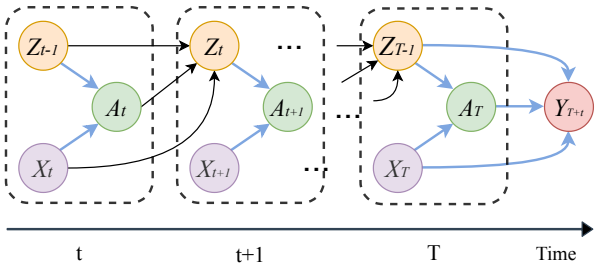

Fig. 1. The illustration of causal graphs for causal estimation from dynamic observational data. During observed window T, we have a trajectory $\{X_t, A_t, Z_{t-1}\}_{t=1}^{T} \cup \{Y_{T+\tau}\}$, where X_t denotes the current observed covariates, A_t denotes the treatment assignments, Z_t denotes the hidden confounders and $Y_{T+\tau}$ denotes the potential outcomes. At any time stamp t, the treatment assignments A_t are affected by both observed covariates X_t and hidden confounders Z_{t-1} (the causal relationship are annotated by blue arrows). The hidden confounders Z_{t-1} are inferred from previous hidden confounders, treatments and covariates (annotated by black arrows). The potential outcomes $Y_{T+\tau}$ are affected by last time stamp covariates X_T, treatment assignments A_T and hidden confounders Z_{T-1}.

There have been lots of existing methods that estimate ITE by leveraging observational data, including propensity score matching method [6], forest based methods [7], [8], and representation learning-based methods [9]–[11]. However, many methods are mainly based on the *strong ignorability assumption* that there are no unobserved confounders and only few studies have considered the influence of hidden confounders [12]. Here, the hidden confounders refer to factors that affect both treatment assignment and outcome, but are not directly measured in the observational data. For example, physicians may prescribe treatments to the patient based on indicators not in the medical records. Ignoring these hidden confounders can lead to bias in estimating causal effects [13]. Besides, these methods are primarily designed for static settings and are hardly extended to longitudinal observation data. However, most real-world observational data is naturally dynamic and consists of sequential information. For example, in EMR, the patient's conditions (e.g., prescribed medicines, lab results and vital signs) and treatment assignments are recorded frequently during their stay in the hospital. Therefore, estimating ITE becomes more challenging when the treatments and covariates change over time, and the potential outcomes are influenced by historical treatments and covariates.

In this paper, we study the problem of *Estimating individual treatment effects with time-varying confounders* (as illustrated by a causal graph in Fig 1). To alleviate the

2374-8486/20/$31.00 ©2020 IEEE
DOI 10.1109/ICDM50108.2020.00047

aforementioned challenges, we propose a novel causal effects estimation framework, Deep Sequential Weighting (DSW), to adjust the time-varying confounders in the longitudinal data. The proposed framework DSW consists of three main components: representation learning module, balancing module and prediction module. To adjust the time-varying hidden confounders, DSW first learns the representations of hidden confounders by leveraging the current observed covariates and all historical information (i.e., previous covariates and treatment assignments) through Gated Recurrent Units (GRU) with an attention mechanism. With the help of the attention mechanism, the model can automatically focus on important historical information. Then, we compute the time-varying inverse probability of treatment for each individual to balance the confounding. The learned representations of hidden confounders and observed covariates are then combined together for both treatment prediction and potential outcome prediction.

To demonstrate the effectiveness of our framework, we conduct comprehensive experiments on synthetic, semi-synthetic and real-world EMR datasets (MIMIC-III [14]). DSW outperforms state-of-the-art baselines in terms of PEHE and ATE. To further illustrate how our method can be used in personalized medicine, we analyze the treatment effects on important outcomes for ICU septic patients. Results demonstrate that our model can generate unbiased and accurate treatment effect by conditioning both time-varying observed confounders and hidden confounders. Our model has the potential to be leveraged as part of clinical decision support systems that assist physicians to determine whether to introduce treatment to a patient or a specific population.

The contributions of this paper are as follows:

- We study the task of estimating ITE with time-varying confounders, on which few attention has been before.
- We propose a novel causal inference framework DSW to solve the task. DSW fully utilize the historical information and current covariates for learning the representations of hidden confounders. A balancing operation is adopted to generate unbiased and accurate ITE estimation.
- We conduct experiments on synthetic, semi-synthetic and real-world datasets to demonstrate the effectiveness of our proposed method. Results show that our method outperforms state-of-the-art causal inference methods and has the potential to be used as part of clinical decision support systems to determine whether a treatment is needed for a specific patient, paving the way for personalized medicine.

II. Method

In this section, we first give a formal definition of the notations used throughout the paper, then present the proposed framework for estimating ITE.

A. Preliminary

Let $X_t \in \mathcal{X}_t$ be the time-dependent covariates of the observational data at time stamp t such that $X_t = \{x_t^{(1)}, x_t^{(2)}, ..., x_t^{(n)}\}$, where $x_t^{(i)}$ denotes the covariates for i-th patient, n denotes the number of patients, and $\mathcal{X}_t$ denotes the time-dependent feature space. The static features (e.g., demographic information), do not change overtime are also considered as observed covariates. We use $C \in \mathcal{C}$ represent the static features for all the patients. At each time stamp t, the treatment assignments are denoted as $A_t = \{a_t^{(1)}, a_t^{(2)}, ..., a_t^{(n)}\}$, where $A_t \in \mathcal{A}$, $a_t^{(i)}$ denotes the treatments assigned to i-th patient. In the case of the binary treatment setting, i.e., $a_t^i = \{0, 1\}$, where 1 is considered as "treated" while 0 as "control", we are interested in estimating the effect of the treatment assigned until time stamp T on the outcomes $Y_{T+\tau} \in \mathcal{Y}$, observed at time stamp $T + \tau$, where τ is the prediction window. Note that in observational data, a patient can only belong to one group (i.e., either treated or control group), thus the outcome from the other group is always missing and referred to counterfactual. To represent the historical sequential data before time stamp t, we use the notation $\overline{X}_t = \{X_1, X_2, ..., X_{t-1}\}$ to denote the history of covariates observed before time stamp t, and $\overline{A}_t$ refers to the history of treatment assignments. Combining all covariates and treatments, we define $\tilde{H}_t^{(i)} = \{\overline{x}_t^{(i)}, \overline{a}_t^{(i)}\} \cup \{c^{(i)}\}$ as all the historical data collected before time stamp t. The observational data for i-th patient can be represented using the notations defined above as: $\mathcal{D}^{(i)} = \{x_t^{(i)}, a_t^{(i)}\}_{t=1}^{T} \cup \{c^{(i)}, y_{a,T+\tau}^{(i)}\}$. We summarize the notations we used in this paper in Table I.

TABLE I
Notations

Notation	Definition
$\mathcal{X}$	The space of time-varying covariates
$\mathcal{C}$	The space of static covariates
$\mathcal{A}$	The set of treatment options of interest
$\mathcal{Y}$	The space of potential outcomes
X_t	The time-varying covariates of all patients at time t
$\mathcal{Z}$	The space of hidden confounders
$x_t^{(i)}$	The time-varying covariates of i-th patient at time t
C	The static covariates of all patients
$c^{(i)}$	The static covariates of i-th patient
A_t	The treatment assigned at time t
$a_t^{(i)}$	The treatment assigned for i-th patient at time t
$Y_{T+\tau}$	The factual (observed) outcomes at time $T + \tau$
$Y_{1,T+\tau}/\hat{Y}_{1,T+\tau}$	The observed/predicted outcome at time $T + \tau$ when receive treatment
$y_{1,T+\tau}^{(i)}/\hat{y}_{1,T+\tau}^{(i)}$	The observed/predicted outcomes of i-th patient at time $T + \tau$ when $a^{(i)} = 1$
$Y_{0,T+\tau}/\hat{Y}_{0,T+\tau}$	The observed/predicted outcome at time $T + \tau$ when not receive the treatment
$y_{0,T+\tau}^{(i)}/\hat{y}_{0,T+\tau}^{(i)}$	The observed/predicted outcomes of i-th patient at time $T + \tau$ when $a^{(i)} = 0$
$e^{(i)}/\hat{e}^{(i)}$	The true/predicted ITE of i-th patient at time t
$\overline{(\cdot)}_t$	The historical covariates collected before time t
Z_t	The learned hidden confounders at time t
$z_t^{(i)}$	The learned hidden confounders for patient i at time t
$\tilde{H}_t^{(i)}$	The historical data consists of $\{\overline{x}_t^{(i)}, \overline{a}_t^{(i)}\} \cup \{c^{(i)}\}$ for i-th patient
T	The number of time stamps (observation window)
τ	The length of prediction window
n	The number of patients in the dataset

We follow the well-adopted potential outcome framework and its variation that considers the time-varying treatment

assignments when estimating the causal effect on the outcomes. The potential outcome $y^{(i)}_{a,T+\tau}$ of i-th patient given the historical treatment can be formulated as $y^{(i)}_{a,T+\tau} = \mathbb{E}[y|x^{(i)}_t, \mathcal{H}^{(i)}_t, a = a^{(i)}]$, where $a^{(i)}$ equals to 1 if the treatment is assigned at time $\{1, 2.., T\}$, otherwise 0. Then the individual treatment effect (ITE) on the temporal observational data is defined as follows:

$$e^{(i)} = \mathbb{E}[y^{(i)}_{1,T+\tau}|x^{(i)}_t, a^{(i)}_t, \tilde{H}^{(i)}_t] - \mathbb{E}[y^{(i)}_{0,T+\tau}|x^{(i)}_t, a^{(i)}_t, \tilde{H}^{(i)}_t] \quad (1)$$

Here, the observed outcome $y^{(i)}_{a,T+\tau}$ under treatment a is called factual outcome, while the unobserved one $y^{(i)}_{1-a,T+\tau}$ is the counterfactual outcome. In observational data, only the factual outcomes are available, while the counterfactual outcomes can never been observed.

B. Assumptions

Our estimation of ITE is based on the the following important assumptions [15], and we further extend the assumptions in our scenario (i.e., time-varying observational data).

Assumption 2.1 (Consistency): The potential outcome under treatment history $\overline{A}$ equals to the observed outcome if the actual treatments history is $\overline{A}$.

Assumption 2.2 (Positivity): For any patient i, if the the probability $\mathbb{P}(\overline{a}^{(i)}_{t-1}, \overline{x}^{(i)}_t, c^{(i)}) \neq 0$, then the probability of receiving treatment 0 or 1 is positive, i.e., $0 < \mathbb{P}(\overline{a}^{(i)}_t, \overline{x}^{(i)}_t, c^{(i)}) < 1$, for all $\overline{a}^{(i)}_t$.

Besides these two assumptions, many existing work are based on *strong ignorability* assumption:

Assumption 2.3 (Strong Ignorability): Given the observed historical covariates $\overline{x}^{(i)}_t$ and static covariates $c^{(i)}$ of i-th patient, the potential outcome variables $y^{(i)}_{1,T+\tau}, y^{(i)}_{0,T+\tau}$ are independent of the treatment assignment, i.e., $(y^{(i)}_{1,T+\tau}, y^{(i)}_{0,T+\tau}) \perp\!\!\!\perp a^{(i)}_t | \overline{x}^{(i)}_t, c^{(i)}$

This assumption holds only if there exist no hidden confounders. However, this condition is hard to guarantee in practice especially in real-world observational data. In this paper, we relax such strict assumption by introducing that there exist potential hidden confounders. Our proposed method can learn the representations of the hidden confounders and eliminate the bias between the treatment assignments and outcomes at each time stamp. The learned representations (denoted by $Z_t \in \mathcal{Z}$) can be leveraged for inferring the unobserved confounders and regarded as the substitutes of hidden confounders. Thus, we extend the *strong ignorability* assumption by considering the existing of hidden confounders Z_t at each time stamp t, which influence the treatment assignment A_t and potential outcomes $Y_{T+\tau}$. Given the hidden confounders Z_t, the potential outcome variables are independent of the treatment assignment at each time stamp.

C. Proposed Method

According to the aforementioned assumptions, we present a novel method that utilizes current variables as well as the historical data to learn the representations of the hidden confounders and estimate the individual treatment effect (ITE) for all patients. The overall framework of the proposed method is illustrated in Fig. 2. We will illustrate the details of each module in the following subsections.

D. Representation Learning Module

As the initial feature vectors are always high-dimensional and sparse in the case of real-world data, we first convert the initial features $x^{(i)}_t \in \mathbb{R}^{d_x}$ of each patient into a lower-dimensional and continuous data representations $u^{(i)}_t \in \mathbb{R}^{d_u}$, where d_u is the dimension of the embedded feature vectors, using a liner embedding layer. That is, we define:

$$u^{(i)}_t = W_{emb} x^{(i)}_t \quad (2)$$

where $W_{emb} \in \mathbb{R}^{d_u \times d_x}$ is the embedding matrix. Simply, the liner embedding layer can be alternatively replaced by other more complex embedding method such as multi-layer perceptron (MLP) [16], which is also widely used in learning representation of EMR data [17].

The representation of hidden confounders $z^{(i)}_t$ is learned through a GRU layer to capture the inherent characteristics of time-varying observational data (i.e., dependency between each time unit and sparsity). The current information $u^{(i)}_t$, treatment assignments $a^{(i)}_t$, last time stamp hidden confounders $z^{(i)}_{t-1}$ and last output hidden state of GRU $h^{(i)}_{t-1}$ are regarded as the input to the GRU. For the convenience of future prediction task, we concatenate the current information $u^{(i)}_t$ and last time stamp hidden confounders $z^{(i)}_{t-1}$ into a new variable $q^{(i)}_t$ as follows,

$$q^{(i)}_t = g([u^{(i)}_t, z^{(i)}_{t-1}]) \quad (3)$$

where $g(\cdot)$ denotes the function for learning the hidden confounders (typically a MLP layer is used for generating the representations), $[\cdot, \cdot]$ denotes the concatenation of two vectors.

We elaborate the the architecture of GRU as follows,

$$\begin{aligned} f_t &= \sigma_g(W_f[q^{(i)}_t, c^{(i)}_t, a^{(i)}_t] + V_f h_t + b_f) \\ r_t &= \sigma_g(W_r[q^{(i)}_t, c^{(i)}_t, a^{(i)}_t] + V_r h_t + b_r) \\ h'_t &= \Phi_h(W_h[q^{(i)}_t, c^{(i)}_t, a^{(i)}_t] + V_f(r_t \odot h_{t-1}) + b_h) \\ h_t &= f_t \odot h_{t-1} + (\mathbf{1} - f_t) \odot h'_t \end{aligned} \quad (4)$$

where σ_g is sigmoid function, Φ_h is hyperbolic tangent function, $\odot$ denotes element-wise product operation, W_f, W_r, $W_h \in \mathbb{R}^{d_h \times (d_q + d_c + 1)}$, V_f, $V_r \in \mathbb{R}^{d_h \times d_h}$, b_f, $b_r \in \mathbb{R}^{d_h}$ are parameters matrices and vectors to learn. f_t denotes the update gate vector, r_t denotes reset gate vector and h_t denotes the hidden output vector. The output vectors are further aggregated via a attention layer for automatically focusing on important historical time stamp. We can use various methods to calculate the attention energies between the each previous hidden state h_s and current state h_t, e.g., dot product $h^\top_t h_s$, linear attention $h^\top_t W_\alpha h_s$. In this paper, we calculate the attention weight $\alpha_{t,s}$ using a method that concatenates each previous hidden state with the current state, and the product of two states. That is,

$$\begin{aligned} \alpha_{t,s} &= \text{score}(h_t, h_s) = \Phi(W_\alpha[h_t, h_s, h_t \odot h_s]) \\ \alpha_t &= \text{softmax}(\alpha_{t,1}, \alpha_{t,2}, \ldots, \alpha_{t,t-1}) \end{aligned} \quad (5)$$

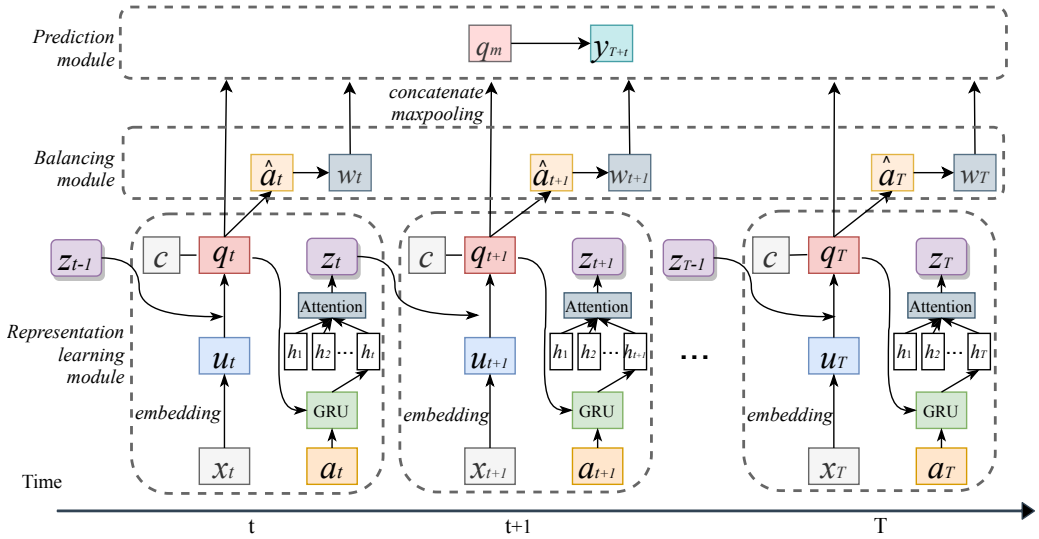

Fig. 2. The framework of DSW. DSW contains three main modules: representation learning module, balancing module and prediction module. At each time stamp t, the model takes the current covariates (X_t, C), treatment assignments A_t, along with the hidden variables Z_{t-1} as input for learning representations of confounder Q_t. The historical information is modeled via a gated recurrent unit (GRU) and aggregated through an attention layer. Then, we use the learned representations of confounding variables for treatment prediction at each time stamp and final outcome prediction after time T.

where Φ is hyperbolic tangent function, $W_\alpha \in \mathbb{R}^{3d_h \times d_h}$ is learnable parameter matrix. Using the generated attention energies, we can calculate the context vector o_t for each patient up to t time stamp as follows,

$$o_t = \sum_{s=1}^{t-1} \alpha_{t,s} h_s \tag{6}$$

We further concatenate the context vector with the current hidden state to generate the current representations z_t up to time t:

$$z_t = \Phi(W_z[h_t, o_t]) \tag{7}$$

where $W_z \in \mathbb{R}^{2d_h}$ is a learnable parameter matrix.

E. Prediction Module

After obtaining the hidden confounding representations, we leverage them for treatment prediction at each time stamp and final outcome prediction during the following prediction window.

1) Global Max Pooling: As the time sequence increases, basic RNN model may forget earlier information due to its long-term dependency. Thus, we adopt a global max-pooling operation over the concatenation of all outputs of $q_t^{(i)}$ vectors. As shown in Fig. 2, the output of max-pooling layer $q_m^{(i)}$ are further used for treatment prediction and potential outcome prediction.

2) Treatment prediction: We predict the treatment assignments for the patient at each time stamp regarding the $q_m^{(i)}$ and static demographic features $c^{(i)}$ as input, and the real treatment $a_t^{(i)}$ as the target label. The predicted treatments $\hat{a}_t^{(i)}$ are obtained through a fully-connected layer with sigmoid function as the last layer,

$$\hat{a}_t = \text{sigmoid}(W_a[q_m, c] + b_a) \tag{8}$$

where $W_a \in \mathbb{R}^{d_q + d_c}$ and $b_a \in \mathbb{R}$ are learnable parameters, $\hat{a}_t^{(i)}$ denotes the the probability of receiving treatment based on the potential confounders of patient i at time t. Typically, the predicted results can also be referred as propensity score [6] that $\hat{a}_t^{(i)} = \mathbb{P}(a_t^{(i)} = 1 | u_t^{(i)}, z_{t-1}^{(i)})$.

As we consider the binary treatment in this paper, we use a cross-entropy loss for the treatment prediction over all patients and all time stamps as follows,

$$\mathcal{L}_a = -\frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T}(a_t^{(i)} \log \hat{a}_t^{(i)} + (1 - a_t^{(i)}) \log (1 - \hat{a}_t^{(i)})) \tag{9}$$

3) Outcome prediction: We finally adopt a potential outcome prediction network to estimate the the outcome $\hat{y}_{t,T+\tau}^{(i)}$ by taking the hidden representations $q_t^{(i)}$ from each time stamp as input. Here, to fully utilize the time series information, we use a max-pooling layer to aggregate all hidden representations. Let $g(\cdot)$ denotes the function learned from outcome prediction network. Then we have,

$$\hat{y}_{t,T+\tau}^{(i)} = g(q_m^{(i)}, a^{(i)} = t) \tag{10}$$

where we estimate the potential outcome for each patient given each treatment assignment situation. Here, we use MLPs to model the function $g(\cdot)$. We minimize the factual loss function as follows,

$$\mathcal{L}_y = \frac{1}{N}\sum_{i=1}^{N} w_a^{(i)} (\hat{y}_{t,T+\tau}^{(i)} - y_{t,T+\tau}^{(i)})^2 \tag{11}$$

where $w_a^{(i)}$ is to re-weight the population for adjusting confounders. We introduce the computation of $w_a^{(i)}$ in the following section.

F. Balancing Module

The predicted probability of receiving treatment is further leveraged for generating weights for each individual to balance the confounding. We compute the weights using inverse probability of treatment weighting (IPTW) and extend to dynamic setting as follows,

$$w_t^{(i)} = \frac{\Pr(A)}{\hat{a}_t^{(i)}} + \frac{(1-\Pr(A))}{(1-\hat{a}_t^{(i)})} \tag{12}$$

where $\Pr(A)$ denotes the probability of being in treated group and $\hat{a}_t^{(i)}$ is the predicted probability of receiving treatment given the current observed data and historical information. We take average of weights computed at each time stamp denoted as $w_a^{(i)}$.

G. Loss Function

The total loss function for the proposed method is defined as,

$$\mathcal{L} = \mathcal{L}_y + \gamma \mathcal{L}_a + \lambda \|W\|_2 \tag{13}$$

where $\mathcal{L}_y$ is the factual prediction loss between estimated and observed factual outcomes, $\mathcal{L}_a$ is the loss from treatment prediction, γ, λ are parameters to balance the loss function. The last term is L_2 regularization on model parameters W. The training process of DSW is presented in Algorithm 1.

Algorithm 1 DSW Model

Input: data of i-th patient: time-varying covariates $x_t^{(i)}$, static covariates c, treatment assignments $a_t^{(i)}$;
Output: potential outcome $y_{T+\tau}^{(i)}$;

Randomly initialize embedding matrix W_{emb} for time-varying covariates, GRU parameters W_f, W_r, W_h, V_f, V_r, b_f, b_r, attention parameters W_α;
repeat
  for covariate v in $x_t^{(i)}$ **do**
    Obtain the embedding of v using Eq. (2);
    Obtain $q_t^{(i)}$ using Eq. (3);
    Input the $q_t^{(i)}$ and $a_t^{(i)}$ into GRU and obtain $h_t^{(i)}$ using Eq. (4);
    Compute the attention weights using Eq. (5); (6)
    Obtain $z_t^{(i)}$ using Eq. (7)
  end for
  Predict the treatment $\hat{a}_t^{(i)}$ using Eq. (8);
  Compute the weights for each patient using Eq. (12)
  Calculate the treatment prediction loss using Eq. (9);
  Predict the potential outcome $\hat{y}_{T+\tau}^{(i)}$ using Eq. (10);
  Calculate the outcome prediction loss using Eq. (11);
  Update parameters according to gradient of mean loss;
until convergence

III. Experimental Setup

To evaluate the performance of the proposed model, we conduct comprehensive comparison experiments on three different datasets: fully-synthetic dataset, semi-synthetic dataset and real-world dataset.

A. Datasets and Simulation

1) Synthetic Dataset: As introduced in the previous section, the treatment assignments $a_t^{(i)}$ at each time stamp are influenced by the confounders $q_t^{(i)}$, which are consist of previous hidden confounders $z_{t-1}^{(i)}$, current time-varying covariates $x_t^{(i)}$ and static features $c^{(i)}$. We first simulate $x_t^{(i)}$ and $z_t^{(i)}$ for each patient at time t following p-order autoregressive process [18] as,

$$\begin{aligned} x_{t,j}^{(i)} &= \frac{1}{p}\sum_{r=1}^{p}(\alpha_{r,j} x_{t-r,j}^{(i)} + \beta_r a_{t-r}^{(i)}) + \eta_t \\ z_{t,j}^{(i)} &= \frac{1}{p}\sum_{r=1}^{p}(\mu_{r,j} z_{t-r,j}^{(i)} + \upsilon_r a_{t-r}^{(i)}) + \epsilon_t \end{aligned} \tag{14}$$

where $x_{t,j}^{(i)}$ and $z_{t,j}^{(i)}$ denote the j-th column of $x_t^{(i)}$ and $z_t^{(i)}$, respectively. For each j, $\alpha_{r,j}, \mu_{r,j} \sim \mathcal{N}(1-(r/p), (1/p)^2)$ control the amount of historical information of last p time stamps incorporated to the current representations. $\beta_r, \upsilon_r \sim \mathcal{N}(0, 0.02^2)$ controls the influence of previous treatment assignments. $\eta_t, \epsilon_t \sim \mathcal{N}(0, 0.01^2)$ are randomly sampled noises.

To simulate the treatment assignments, we generate 1000 treated samples and 3000 control samples. For treated samples, we randomly pick the treatment initial point among all time stamps. The treatments starting from the initial point are all set to 1. For the control samples, the treatments at each time stamp are all set to 0.

The confounders $q_t^{(i)}$ at time stamp t and outcome $y_{T+\tau}^{(i)}$ can be simulated using the hidden confounders and current covariates as follows,

$$\begin{aligned} q_t^{(i)} &= \gamma_h \frac{1}{t}\sum_{r=1}^{t} z_r^{(i)} + (1-\gamma_h) g([x_t^{(i)}, c^{(i)}]) \\ y_{T+\tau}^{(i)} &= w^\top q_T^{(i)} + b \end{aligned} \tag{15}$$

where γ_h is the parameter to control the influence of hidden confounders, $w \sim \mathcal{U}(-1, 1)$ and $b \sim \mathcal{N}(0, 0.1)$. The function $g(\cdot)$ maps the concatenated feature vectors $[x_t^{(i)}, c^{(i)}]$ into the hidden space. In this paper, for each individual, we simulate 100 time-varying covariates, 5 static covariates with 10 time stamps in total. We modify the value of $\gamma_h \in \{0.1, 0.3, 0.5, 0.7\}$ and obtain four variants of the current dataset.

2) Semi-synthetic Dataset based on MIMIC-III: With a similar simulation process, we construct a semi-synthetic dataset based on a real-world dataset: Medical Information Mart for Intensive Care version III (MIMIC-III) [14]. MIMIC-III has more than 61,000 ICU admissions from 2001 to 2012 with recorded patients' demographics and temporal information, including vital signs, lab tests, and treatment decisions. We extracted 11,715 adult sepsis patients fulfilling the sepsis-3 criteria [19] as our studied cohort from MIMIC-III since sepsis contributes to up to half of all hospital deaths and is associated with more than $24 billion in annual costs in the United States [20].

Here, we obtain 27 time-varying covariates (vital signs: temperature, pulse rate, glucose, etc; lab tests: potassium,

sodium, chloride, etc.) and 12 static demographics (i.e., age, gender, race, height, weight, etc.) as potential confounding variables. The full list of covariates is available at Github[1]. We consider vasopressors as treatments since they are commonly used in septic patients. As the MIMIC-III dataset is real world observational data, then it is impossible to obtain the counterfactual outcomes for calculating the ground truth treatment effect. Therefore, we simulate the potential outcomes for each patient using the observed covariates and treatment assignments. The simulation process is similar to the way we generate fully-synthetic dataset, with the exception that we only need to synthesize the potential outcomes (using Eq. 15). By varying the values of $\gamma_h \in \{0.1, 0.3, 0.5, 0.7\}$, we have four variants of the current datasets.

3) Real world Dataset: MIMIC-III: To evaluate the performance of our model in a real-world application, we design a causal inference setting based on the MIMIC-III dataset. (1) **Treatment**. We consider two available treatment assignments: vasopressors (vaso) and mechanical ventilator (mv). For each treatment option, we separately evaluate its causal effect on the important outcome signals. (2) **Outcomes**. To evaluate the treatment effect of vasopressors, we use mean blood pressure (Meanbp) as target outcomes since vasopressors are highly related to Meanbp. For mechanical ventilator, we adopt oxygen saturation (SpO2) as outcome since ventilator is usually assigned to patients with the difficulty of breathing. (3) **Confounders**. We consider the same confounders as in synthetic dataset (27 time-varying covariates and 12 static demographics).

B. Methods for comparison

To evaluate the performance of the proposed framework in estimating the ITE, we conduct comparison experiments on the following state-of-the-art causal inference methods,

- **Linear Regression (LR)**. LR directly regard the treatment as an additional feature for potential outcome prediction. It ignores the confounders and selection bias in observational data.
- **Random Forest (RF)**. The training process is same as LR (using treatment as a feature). We vary the number of trees in range $\{50, 60, .., 150\}$ and select best parameter setting on validation set.
- **K-Nearest Neighbor Matching (KNN)** [21]. KNN is a matching based method that estimates the counterfactual outcomes of treated (control) group from K-nearest neighbors in control (treated) group. We attempts three different distance metrics *euclidean*, *minkowski* and *mahalanobis*, and choose the the metric yields best performance on the validation dataset.
- **Propensity Score Matching (PSM)** [6]. PSM is also a matching based method but instead use propensity score to measure the distance among individuals. Commonly, logistic regression is adopted for propensity score estimation. For logistic regression, we try different solvers: *liblinear*, *lbfgs*, *sag* and *saga* on the training set and select the solver with best performance on the validation set.
- **Counterfactual Regression (CFR)** [9], [10]. CFR is deep representation learning based method. CFR has four different variants according the selection of distribution balancing metrics: Maximum Mean Discrepancy (**CFR MMD**), Wasserstein (**CFR WASS**), **BNN**, **TARNet**.
- **Causal Forest (CF)** [8]. CF is an extension of RF for estimating the ITE, which is designed for causal effect estimation.
- **Bayesian Additive Regression Trees (BART)** [7]. BART is a non-parametric Bayesian regression tree model, which takes the covariates and treatment as inputs and outputs the distribution of outcomes.

C. Performance Measurement

To evaluate the estimated ITE, we adopt mean squared error (MSE) between the ground truth and estimated ITE as follows,

$$\text{PEHE} = \frac{1}{N}\sum_{i=1}^{N}((y_1^{(i)} - y_0^{(i)}) - (\hat{y}_1^{(i)} - \hat{y}_0^{(i)}))^2, \qquad (16)$$

which is also known as Precision in Estimation of Heterogeneous Effect (PEHE). Typically, we report the rooted PEHE $\sqrt{\text{PEHE}}$ in our paper. We are also interested in the causal effect over the whole population to help determine whether a treatment should be assigned to population. Then we calculate the mean absolute error (MAE) between the ground truth and estimated and average treatment effect (ATE):

$$\text{ATE} = |\frac{1}{N}\sum_{i=1}^{N}(y_1^{(i)} - y_0^{(i)}) - \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_1^{(i)} - \hat{y}_0^{(i)})| \qquad (17)$$

Additionally, we adopt rooted mean squared error (RMSE) between the estimated factual outcomes and ground truth outcomes to evaluate the performance on factual prediction task as follows,

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_t^{(i)} - y_t^{(i)})^2} \qquad (18)$$

D. Implement Details

The model is implemented and trained with Python 3.6 and PyTorch 1.4 [2], on a high-performance computing cluster with four NVIDIA TITAN RTX 6000 GPUs. We train our model using the adaptive moment estimation (Adam) algorithm with a batch size of 128 subjects and the learning rate is 0.001. The data is randomly split into training, validation and test sets with a ratio of 70%, 10%, 20%. The information from a given patient is only present in one set. The validation set is used to improve the models and select the best model hyper-parameters. We report the performance of our model and baselines on the test set. We use $\sqrt{\text{PEHE}}$ and ATE error to measure the models' performance on ITE estimation, and RMSE for factual prediction. As all baseline methods are originally designed for static environment, we run these models

[1]https://github.com/ruoqi-liu/DSW

[2]https://pytorch.org/

independently on each time stamp and average the evaluation metrics over all time stamps. The code and more implementation details are available at https://github.com/ruoqi-liu/DSW.

IV. Results

We now report the performance of DSW on synthetic, semi-synthetic and real-world datasets. We focus on answering the following research questions by our experimental results:

- **Q1: How precise is DSW on ITE estimation?**
- **Q2: How accurate is DSW on factual prediction task (i.e., outcome prediction)?**
- **Q3: How can DSW be used for personalized medicine?**

A. How precise is DSW on ITE estimation?

We conduct comprehensive comparison experiments on synthetic and semi-synthetic datasets and report $\sqrt{\text{PEHE}}$ and ATE on each dataset. By varying the parameter γ_h that controls the influence of hidden confounders, we evaluate the how precise is DSW on ITE estimation under different value of γ_h.

Results on Synthetic Dataset Table II shows the performance of our method and baselines on fully-synthetic dataset evaluated by $\sqrt{\text{PEHE}}$ and ATE. The values shown in the table are averaged on 10 realizations. We observe that DSW outperforms all other baselines with different value of γ_h, which confirms that our designed framework can better capture the characteristics of longitudinal data and generate accurate estimation of ITE.

Generally speaking, the representation learning based approaches achieve better performance compared with base methods, matching based methods and tree based methods. Since those linear approaches are not designed for causal effect estimation, they may not able to control the influence of confounding variables. The matching based methods consider the similarity information among treated and control groups to alleviate the selection biases. However, their estimation becomes inaccurate when dealing with high-dimensional and complex data. Tree and forest based method achieve comparable performance with basic random forest method since these two methods are established upon the random forest.

The representation learning based methods use the deep neural network to model the representations of confounding variables. Their methods achieve the best performance among all baselines. CFR MMD, CFR WASS, BNN and TARNet share the similar design of neural network, with exception that they adopt different strategies to minimize the distance between treated and control groups. Specifically, CFR MMD and CFR WASS have the same outcome prediction networks, but the former uses Maximum Mean Discrepancy (MMD) and the latter uses Wasserstein (WASS) to balance the distributions. BNN regards the treatment as additional feature and minimize the distances between treated and control group in latent space. TARNet is a vanilla version without balancing property. Among all these four representation learning based methods, we observe that CFR MMD and CFR WASS generally achieve better performance than other two methods. Although representation learning based models outperform the other baselines, they ignore the time-varying confounders and lose lots of temporal information, which are crucial and common in healthcare data. Our proposed DSW successfully captures the temporal information and thus outperform the baselines.

Note that, four variants of simulated datasets ($\gamma_h \in \{0.1, 0.3, 0.5, 0.7\}$) are not comparable, since the distribution of simulated outcomes are not same. With the increasing of γ_h, less portion of observed confounders are included, which results in smaller values of outcomes. Instead, we focus on the performance of methods within each dataset.

Results on Semi-synthetic Dataset We demonstrate the performance of the proposed model and baselines on a semi-synthetic dataset based on MIMIC-III. As shown in Table III, DSW achieves the best performance among other baseline methods in terms of $\sqrt{\text{PEHE}}$ and ATE. We vary the value of γ_h to control the influence of hidden confounders and conduct comparison experiments under each value of γ_h.

We observe that representation learning based methods in general outperform most causal inference methods which demonstrate that deep learning architecture with distribution balancing is beneficial to the ITE estimation. Base models (LR and Random Forest) cannot perform very well since they ignore the selection bias and confounding factors. As for matching based methods, they consider the similarity information among individuals to alleviate selection bias. Among all the baselines, DSW considers the temporal information in the time-varying data, and thus generate unbiased and accurate treatment effect estimation.

B. How accurate is DSW on factual prediction?

In real-world data, we have no access to counterfactual outcomes for calculating the true treatment effect, and thus we cannot compute the $\sqrt{\text{PEHE}}$ and ATE. Instead, we evaluate the performance of our model through a factual prediction task. We measure the RMSE between observed (factual) outcomes and estimated outcomes on two treatment-outcome pairs: vasopressor-meanBP and ventilator-SpO2.

Table IV shows the estimated RMSE of two selected pairs. Since causal effect methods are not initially designed for factual prediction, we adopt four baselines from each category that can be adapted for factual prediction task: LR, KNN, CFR (WASS) and BART. Matching based methods aim to estimate the counterfactual and cannot be directly used for factual inference. We adapt the KNN matching for factual prediction by combining the treatments as additional features to predict factual outcomes. Among four representation learning based methods, they achieve relative comparative performance in two datasets, so we use CFR WASS as a representative for this kind of method. In two forest based methods, CF directly outputs the estimated ITE without any inference of factual outcomes, so we use BART as our baseline.

As shown in TABLE IV, DSW outperforms all the baselines, which demonstrates that our model yields accurate estimation on factual data. Among the baselines, we find that representation learning based method (CFR WASS) performs better than

TABLE II
Performance comparison on synthetic datasets. We construct four variants of datasets by varying the value of $\gamma_h \in \{0.1, 0.3, 0.5, 0.7\}$. Here, we report the estimated $\sqrt{\text{PEHE}}$ and ATE of each method among four datasets.

		$\gamma_h = 0.1$		$\gamma_h = 0.3$		$\gamma_h = 0.5$		$\gamma_h = 0.7$	
	Method	$\sqrt{\text{PEHE}}$	ATE	$\sqrt{\text{PEHE}}$	ATE	$\sqrt{\text{PEHE}}$	ATE	$\sqrt{\text{PEHE}}$	ATE
Base model	LR	0.640	0.551	0.648	0.558	0.656	0.566	0.663	0.574
	Random Forest	0.656	0.552	0.658	0.553	0.660	0.555	0.663	0.557
Matching based	KNN [21]	0.713	0.604	0.718	0.608	0.724	0.611	0.729	0.615
	PSM [6]	0.699	0.591	0.708	0.602	0.714	0.607	0.720	0.611
Representation learning based	CFR MMD [10]	0.587	0.485	0.594	0.491	0.597	0.492	0.600	0.495
	CFR WASS [10]	0.571	0.470	0.574	0.473	0.576	0.474	0.580	0.476
	BNN [9]	0.586	0.488	0.593	0.495	0.595	0.496	0.597	0.498
	TARNet [10]	0.606	0.512	0.610	0.516	0.615	0.520	0.619	0.523
Forest based	Causal Forest [8]	0.604	0.511	0.612	0.517	0.615	0.521	0.618	0.523
	BART [7]	0.608	0.520	0.614	0.525	0.621	0.530	0.619	0.528
Ours	**DSW**	**0.491**	**0.391**	**0.469**	**0.372**	**0.485**	**0.384**	**0.515**	**0.397**

TABLE III
Performance comparison on semi-synthetic MIMIC-III datasets. We construct four variants of datasets by varying the value of $\gamma_h \in \{0.1, 0.3, 0.5, 0.7\}$. Here, we report the estimated $\sqrt{\text{PEHE}}$ and ATE of each method among four datasets.

		$\gamma_h = 0.1$		$\gamma_h = 0.3$		$\gamma_h = 0.5$		$\gamma_h = 0.7$	
	Method	$\sqrt{\text{PEHE}}$	ATE	$\sqrt{\text{PEHE}}$	ATE	$\sqrt{\text{PEHE}}$	ATE	$\sqrt{\text{PEHE}}$	ATE
Base model	LR	0.823	0.660	1.414	1.134	1.474	1.233	1.542	1.336
	Random Forest	0.837	0.671	1.422	1.143	1.484	1.245	1.553	1.349
Matching based	KNN [21]	0.818	0.650	1.408	1.137	1.472	1.237	1.543	1.337
	PSM [6]	0.767	0.607	1.417	1.143	1.477	1.240	1.548	1.340
Representation learning based	CFR MMD [10]	0.803	0.643	1.257	0.972	1.356	1.099	1.527	1.318
	CFR WASS [10]	0.800	0.641	1.256	0.972	1.370	1.118	1.556	1.352
	BNN [9]	0.802	0.643	1.287	1.006	1.337	1.080	1.515	1.305
	TARNet [10]	0.829	0.664	1.256	0.972	1.373	1.120	1.527	1.318
Forest based	Causal Forest [8]	0.796	0.639	1.413	1.134	1.473	1.233	1.540	1.335
	BART [7]	0.785	0.631	1.413	1.135	1.472	1.234	1.538	1.336
Ours	**DSW**	**0.522**	**0.432**	**0.604**	**0.523**	**0.672**	**0.601**	**0.722**	**0.652**

TABLE IV
Factual Prediction on the MIMIC-III dataset. We select two treatment-outcome pairs: Vasopressor-Meanbp and Ventilator-SpO2, and report RMSE between estimated factual outcome and ground truth.

	MIMIC Dataset (RMSE)	
Method	Vent-SpO2	Vaso-Meanbp
LR	0.909	0.973
KNN [21]	0.901	1.030
CFR WASS [10]	0.870	1.011
BART [7]	0.873	0.966
DSW	**0.814**	**0.814**

the linear regression method (LR) and matching based method (KNN), which is consistent with TABLE II and III.

C. How can DSW be used for personalized medicine?

Based on observational data, we are going to show that our model can adjust time-varying confounders, and generate unbiased and accurate estimation of treatment effect. In this case, our model could potentially help physicians determine whether to apply a specific treatment to a patient. We examine the usages of *Vasopressor* and *Ventilator* in the analysis, which are commonly used treatments for septic patients.

Vasopressor-Meanbp pair Vasopressor (a.k.a., antihypotensive agent) is a group of medications that tend to raise low blood pressure. The patients are expected to have normal blood pressure after receiving a vasopressor. To demonstrate that our model can adjust time-varying confounders, we plot the distribution of ground truth (observed) and predicted Meanbp values after the patients have received vasopressor in Fig. 3. According to the threshold of the normal range of Meanbp (70-100 mmHg), we separate the patients into two groups: one is the patients with observed MeanBP values below 70 mmHg, the other is the patients with observed Meanbp values above 70 mmHg. Figure. 3(a) shows the distribution of patients with Meanbp below 70 mmHg. We observe this group of patients remains low blood pressure even after receiving vasopressor and the average value is far lower than 70 mmHg. If mainly based on the observed data, we may conclude that vasopressor has no effect on raising blood pressure and are unnecessary to be assigned to patients with low blood pressure. However, the predicted values given by our model are higher than the observed values and the average Meanbp belongs to the normal range, which indicates that vasopressor should have a beneficial effect on the blood pressure. Figure 3(b) shows the distribution of patients with Meanbp above 70 mmHg. For these patients whose Meanbp remains in the normal range after receiving vasopressor, the predicted values are still within the normal range, which indicates that vasopressor should be

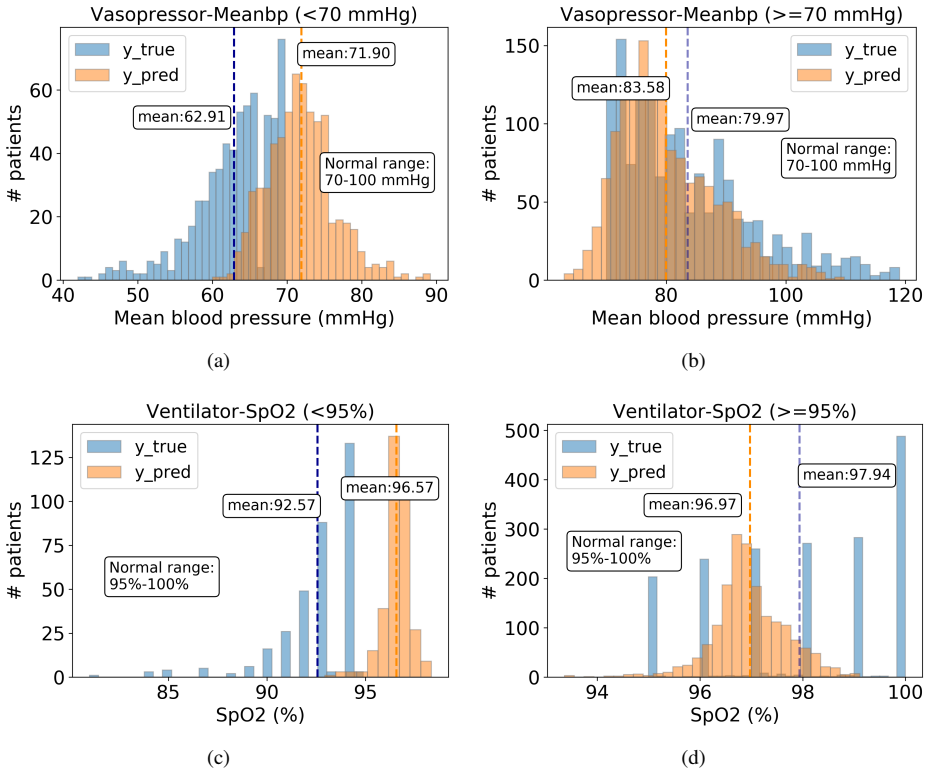

Fig. 3. Distribution of ground truth MeanBP and predicted MeanBP given vasopressor as treatment. Fig. 3(a) displays the distribution of patients with observed MeanBP below than 70 mmHg. Fig. 3(b) displays the distribution of patients with observed MeanBP above 70 mmHg. Fig. 3(c) displays the distribution of patients with observed SpO2 below than 95%. Fig. 3(d) displays the distribution of patients with observed SpO2 above 95%.

assigned to this group of patients to maintain normal blood pressure. If not, their situation may become worse.

Ventilator-SpO2 pair A ventilator is a machine that delivers breaths to a patient who is physically unable to breathe, or breathing insufficiently to maintain blood oxygen. We monitor the value of oxygen saturation (SpO2) to estimate the treatment effect.

Similarly, we show the distribution of ground truth and predicted SpO2 values of patients who have received a ventilator during the observational window in Fig. 3. As the normal range of SpO2 is 95-100%, we separately plot the distribution of patients with observed SpO2 values below 95% in Fig. 3(c), and the distribution of patients with observed SpO2 values above 95% in Fig. 3(b). We observe that, for patients with observed SpO2 lower than normal value, the distribution of predicted SpO2 values lies in normal range with an average of 96.57%. And for patients with observed SpO2 in the normal range, our predicted values still belong to the normal range.

Results show that our model adjusts time-varying confounders and is able to generate unbiased and accurate ITE on important outcome signals (vasopressor's effect on blood pressure and ventilator's effect on blood oxygen in our analysis). Thus, it could potentially assist physicians to determine whether to introduce a treatment to a specific patient, paving the way for personalized medicine.

V. Related Work

In this section, we review the related work for ITE estimation using static and time-varying observational data. We first introduce the causal effect learning framework on static data, and then the framework based on time-varying data.

Learning causal effects with static data According to the way to control the confounders, existing work with static observational data can be divided into four groups: 1) Matching-based methods; 2) Tree-based methods; 3) Reweighting-based methods; 4) Representation-based methods. The matching-based methods are adopted to estimate the counterfactual from the nearest neighbors. The distances among individuals can be measured in several ways (i.e., Euclidean distance, propensity scores). For example, propensity score matching (PSM) [6] is to match a treated (control) sample to a set of control (treated) samples with similar propensity scores. Tree-based methods are also widely adopted in causal effect estimation. Bayesian additive regression trees (BART) [7] is a non-parametric Bayesian regression tree model based on the *strong ignorability assumption*. It is easy to implement and free from parameter adjustment. Causal Forest (CF) [8] is also a tree-based causal effect estimation method, which estimates the treatment effect at the leaf node by mapping the original covariate into tree and forests. The reweighting-based methods attempt to re-weight samples in the population for correcting the bias in observational data. For example, inverse probability of treatment weighting (IPTW) [6] removes the confounding

by assigning a weight to each individual in the population. The weights are calculated based on the propensity score. Recently, representation learning methods are proposed for causal effect estimation via balancing the distribution between treated and control groups in hidden space [9], [10]. Moreover, Yao et al. [11] incorporate the local similarity among individuals with population-level distribution balancing in latent space to better estimate ITE. Yoon et al. propose to use generative adversarial nets (GAN) for inferring the counterfactual outcomes based on factual outcomes. Shi et al. [22] jointly model the propensity prediction and potential outcome prediction as a multi-task learning problem.

Though existing work shows great performance in causal effect estimation, they still have some limitations. First, most of them are built upon *strong ignorability assumption* without considering the influence of hidden confounders. This constrain has been shown to lead to bias in estimating causal effects [13]. Moreover, existing work is initially designed for static data, which is not easy to adapt for ITE estimation under dynamic longitudinal setting.

Learning causal effects with time-varying data As estimating causal effect from observational data is significant and most observational data contains sequential information, some work has been proposed for dealing time-varying confounders. In statistics and epidemiology domains, a group of methods use the inverse probability of treatment and *g-formula* based method to estimate causal effect with sequential data [23], [24]. More recently, Lim et al. [25] propose a recurrent marginal structural network for predicting the patient's potential response to a series of treatments. Bica et al. [26] adopt adversarial training techniques to balance the historical confounding variables. Their method is based on the *strong ignorability assumption*. Later, Bica et al. [27] relax the assumption on strong ignorability and propose to estimate the treatment response with the existence of hidden confounders.

VI. Conclusion

In this paper, we propose Deep Sequential Weighting (DSW), a deep learning based framework for estimating ITE with time-varying confounders. Specifically, DSW infers the hidden confounders by incorporating the current treatment assignments and historical information using a deep recurrent weighting neural network. When combined with current observed data, the learned representations of hidden confounders are leveraged for potential outcome prediction and treatment prediction. We compute the time-varying inverse probabilities of treatment for re-weighting the population. Comprehensive experiments on fully-synthetic, semi-synthetic, and real-world datasets demonstrate the effectiveness of DSW when compared to state-of-the-art baseline methods. Results illustrate that our model can generate unbiased and accurate treatment effect by conditioning on time-varying confounders. Our model has the potential to be used as part of clinical decision support systems to determine whether a treatment is needed for a specific patient, paving the way for personalized medicine.

References

[1] T. A. Glass, S. N. Goodman, M. A. Hernán, and J. M. Samet, "Causal inference in public health," *Annual review of public health*, vol. 34, pp. 61–75, 2013.

[2] N. Baum-Snow and F. Ferreira, "Causal inference in urban and regional economics," in *Handbook of regional and urban economics*. Elsevier, 2015, vol. 5, pp. 3–68.

[3] P. Wang, W. Sun, D. Yin, J. Yang, and Y. Chang, "Robust tree-based causal inference for complex ad effectiveness analysis," in *WSDM'15*, 2015, pp. 67–76.

[4] J. J. Heckman, J. E. Humphries, and G. Veramendi, "Returns to education: The causal effects of education on earnings, health, and smoking," *Journal of Political Economy*, vol. 126, no. S1, pp. S197–S246, 2018.

[5] Z. Song and K. Baicker, "Effect of a workplace wellness program on employee health and economic outcomes: a randomized clinical trial," *JAMA*, vol. 321, no. 15, pp. 1491–1501, 2019.

[6] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[7] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.

[8] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.

[9] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *ICML'16*, 2016, pp. 3020–3029.

[10] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *ICML'17*. JMLR. org, 2017, pp. 3076–3085.

[11] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang, "Representation learning for treatment effect estimation from observational data," in *NeurIPS'18*, 2018, pp. 2633–2643.

[12] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in *NeurIPS'17*, 2017, pp. 6446–6456.

[13] J. Pearl, *Causality*. Cambridge university press, 2009.

[14] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[15] M. A. Hernan and J. M. Robins, "Causal inference," 2010.

[16] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, classifiaction," 1992.

[17] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, "Multi-layer representation learning for medical concepts," in *KDD'16*, 2016, pp. 1495–1504.

[18] T. C. Mills and T. C. Mills, *Time series techniques for economists*. Cambridge University Press, 1991.

[19] M. Singer, C. S. Deutschman, C. W. Seymour *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 801–810, 2016.

[20] V. Liu, G. J. Escobar, J. D. Greene, J. Soule, A. Whippy, D. C. Angus, and T. J. Iwashyna, "Hospital deaths in patients with sepsis from 2 independent cohorts," *JAMA*, vol. 312, no. 1, pp. 90–92, 2014.

[21] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik, "Nonparametric tests for treatment effect heterogeneity," *The Review of Economics and Statistics*, vol. 90, no. 3, pp. 389–405, 2008.

[22] C. Shi, D. Blei, and V. Veitch, "Adapting neural networks for the estimation of treatment effects," in *NeurIPS'19*, 2019, pp. 2503–2513.

[23] J. M. Robins, M. A. Hernan, and B. Brumback, "Marginal structural models and causal inference in epidemiology," 2000.

[24] P. Schulam and S. Saria, "Reliable decision support using counterfactual models," in *NeurIPS'17*, 2017, pp. 1697–1708.

[25] B. Lim, "Forecasting treatment responses over time using recurrent marginal structural networks," in *NeurIPS'18*, 2018, pp. 7483–7493.

[26] I. Bica, A. M. Alaa, J. Jordon, and M. van der Schaar, "Estimating counterfactual treatment outcomes over time through adversarially balanced representations," *arXiv preprint arXiv:2002.04083*, 2020.

[27] I. Bica, A. M. Alaa, and M. van der Schaar, "Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders," *arXiv preprint arXiv:1902.00450*, 2019.