A Perspective on Individualized Treatment Effects Estimation from Time-series Health Data

Ghadeer O. Ghosheh[1], Moritz Gögl[1], and Tingting Zhu[1]

[1]Department of Engineering Science, University of Oxford

Abstract

The burden of diseases is increasing worldwide, with unequal treatment efficacy for patient populations that are underrepresented in clinical trials. However, healthcare is driven by the average population effect of medical treatments and therefore, operates in a "one-size-fits-all" approach, not necessarily what best fits for each patient. These facts suggest a pressing need for methodologies to study individualized treatment effects (ITE) to drive personalized treatment. Despite the increased interest in machine learning-driven ITE estimation models, the vast majority focus on tabular data with limited review and understanding of methodologies proposed for time-series electronic health records (EHRs). To this end, this work provides an overview of ITE works for time-series data and insights into future research. The work summarizes the latest work in the literature and reviews it in light of theoretical assumptions, types of treatment settings, and computational frameworks. Furthermore, this work discusses the challenges and future research directions for ITEs in a time-series setting. We hope that this work opens new directions and serves as a resource for understanding one of the exciting yet under-studied research areas.

1 Introduction

Medical treatment and drug budgets are the highest burden on governments and medical institutions. Despite these high costs, only about 90% of the drugs work for 30-50% of the population [29]. These statistics suggest a pressing need to identify subgroups of patients where personalized treatments can be prescribed. The study of treatment effects has gained much attention in recent years, where various tools and approaches have been proposed to help mitigate the financial cost and optimize the effectiveness and efficacy of prescribed treatments. One of the fast-growing applications in clinical Machine Learning (ML) is studying Individualized Treatment Effects (ITEs) [6], where ML capabilities, in many cases, have superseded those of state-of-the-art clinical and pharmaceutical advancements. Using the wealth of observational electronic health record (EHR) data, the individual patient response to various treatments can be estimated [6].

Although there has been increased attention to ITE works from static observational EHR data [51, 9], much less attention has been paid to those from EHR-data measured over time, often referred to as time-series EHRs. Despite similarities in the concepts for treatment estimation between static and time-series data, challenges related to the time-varying nature of covariates make many existing works not directly applicable to time-series data. To this end, our aim is to provide an overview of the main concepts and ideas for estimating ITE from time-series data. The works presented in this paper were identified by searching Google Scholar for the keywords "treatment effects" AND "time series", or "individualized treatment effects" and "time-series", and "counterfactual estimation" AND "time-series" up until February 2024.

To provide a comprehensive resource for a diverse audience, we highlight foundational works and tutorials that underpin individualized treatment effect (ITE) estimation and causal inference in

time-series data. Foundational studies such as Rubin's potential outcomes framework [41] and the advances in causal inference for longitudinal data by Robins et al. [39] lay the theoretical groundwork for treatment effect estimation. Hernán and Robins' widely used book, *Causal Inference: What If* [20], serves as an essential tutorial to understand causal inference principles, including applications to longitudinal and time-series data. Tutorials like those by Austin and Stuart [2] on the inverse probability of treatment weighting (IPTW) explain how to control for confounding in observational data. Geng et al. [13] explore bio-mathematical modeling for dynamic treatment regimes, which is usually used to simulate synthetic datasets for validation. For real clinical validation, the MIMIC-III database [24] remains a cornerstone resource, providing extensive critical care data widely used in research on causal inference and treatment effect estimation. These resources offer a comprehensive foundation for newcomers and support experts in exploring advanced methodologies in this rapidly evolving field.

Starting with introducing the concepts of randomized controlled trials (RCTs) and observational data, then discussing topics such as efficacy and effectiveness of treatments and challenges in ITE estimation. This work aims to bridge theoretical assumptions from causal inference and machine learning to real-world implications of EHR. To the best of our knowledge, this is the first comprehensive review of challenges and methods in ITE estimation for time-series data. We summarize the latest work in the literature and group them based on theoretical assumptions, types of treatment settings, and computational frameworks. Lastly, this work discusses challenges and future research directions for ITEs in a time-series setting. We hope that this work opens new directions and serves as a resource for understanding one of the exciting yet under-studied research areas.

1.1 Randomized Controlled Trial Data

RCT data are considered the gold standard for studying treatment effects. This is mainly due to the random assignment of participants to treatment groups, which effectively eliminates confounding bias. The randomness and control measures used in RCTs make them a lucrative option for studying the effect of treatments; an unbiased estimate of the average treatment effects (ATE) can be directly calculated from the data [33]. Although RCTs offer methodological strengths, various challenges hinder their optimal and complete use for studying treatment effects [7]. Firstly, despite the power of "randomness" to eliminate confounding, certain biases may remain in RCTs. This bias does not come from the level of treatment assignment, but from representatives of samples presented in the study, denoted as sample selection bias hereafter. Most RCTs tend to employ stringent exclusion criteria for enrolled participants, which could introduce bias in the results for members of the unrepresented population [32, 26, 44, 7, 47]. This becomes a bigger issue when a significant percentage of treated patients in real-world data belong to the population excluded from the RCT. For example, the aging population is rarely enrolled in diabetes RCTs due to their age and multimorbid health conditions, although it represents a large proportion of diabetic patients [25]. Such factors might introduce bias that makes the generalizability and external validity of RCTs questionable [18, 7].

The generalizability of the results of the RCT is further limited by the fact that current RCTs are typically designed to only measure ATE. In other words, they estimate how the "average patient" will respond to a given treatment. For a more personalized approach, the unique characteristics of the patient would be needed to predict an individual patient's response to a given treatment, and it may differ significantly from the average response of a population. Furthermore, RCTs tend to be financially costly to design and implement [21, 42] and their data tend to be difficult to share due to privacy concerns [43]. Additional limitations exist in terms of their relatively small sample size [18], ethical issues [16], and short follow-up times, which might miss out on the long-term effects of medications [7]. For example, the effects of oral contraceptives were not quantified until the presence of long-term data, which were not captured in RCTs [46].

1.2 Observational Data: From Efficacy to Effectiveness

Despite all the stated challenges, RCTs are still essential for determining the *Efficacy* for treatments [27] but are not necessarily optimal for studying treatment *Effectiveness*. Making a clear distinction between these two terms will help one understand when data-driven and statistical approaches can improve the use of RCT data. Efficacy refers to the effect of interventions under ideal "theoretical" circumstances, while effectiveness means that an effect is detected not under ideal conditions, but under real-world conditions [27]. Observational EHR data presents a good candidate for better testing of treatment effectiveness. Specifically, longitudinal observational data collected in EHRs typically includes a diverse cohort of patients without strict exclusion criteria, making it more representative of the real or targeted population of patients. Observational data are also less expensive to collect compared to RCTs and capture long-term outcomes [7, 36]. Furthermore, with the widespread use of EHR systems worldwide, observational data can allow one to estimate treatment effects from various clinical settings such as low-middle-income countries (LMICs), where performing RCTs would not be feasible due to high costs. To this end, observational data are a promising resource for studying treatment effects with more inclusive estimations for various patient groups while maintaining low cost and learning from real-world evidence.

1.3 Challenges of Treatment Effects Estimation

Treatment effect estimation is a subfield of causal inference, and as such suffers from the fundamental problem that counterfactual outcomes are never observed [40]. A counterfactual outcome refers to the hypothetical outcome that would have been observed if a different treatment had been given than the actual (factual) one [6]. Of course, for treatments that were not administered, it is not possible to extract the ground truths of individual patient outcomes directly, either from observational or clinical trial data. In RCTs, this problem is resolved by estimating the ATE across the entire study population but not the ITE of an individual. Randomization of treatment allocation ensures that the underlying distribution of patient characteristics is similar in both the treatment and control groups. This allows us to compare the average outcomes in both groups, and thus determine the ATE.

Omitting the randomized control measures used in RCTs and relying on observational data precludes the direct computation of treatment effects. This is because treatment assignment is not random, but biased, as treatment selection in observational data is often driven by the patient's characteristics, such as treatment allocation flowcharts in clinical practice guidelines [48]. Therefore, clinical practice recorded in real-world observational data results in systematic differences in the characteristics of treated and untreated patients. To be able to estimate treatment effects from observational data, it is essential to remove the confounding bias, introduced by the non-random treatment assignment. A confounder is a variable that influences both the intervention and the outcome, which can lead to a spurious association between them [1]. For example, in clinical practice, patients with more severe health conditions might be more likely to receive stronger medications while still being expected to have poorer outcomes. However, it would be wrong to conclude that stronger medication leads to poorer patient outcomes. Failing to adjust for the severity of a health condition as a confounding factor in this case would lead to measuring a spurious association that does not reflect the actual treatment effect.

1.4 From Average to Individualized Treatment Effects Estimation

The modern understanding of estimating treatment effects is highly attributed to the work of Neyman-Rubin's "potential outcomes framework" [41]. In the Neyman-Rubin model, the ITE between a treatment A and a treatment B is defined as the difference between the two potential outcomes (e.g., blood pressure) after administering a treatment A or B to a given patient. Subject to certain assumptions [6], the ATE can be calculated directly from the RCT data by computing the difference between the average outcomes in both treatment groups. However, because of the

absence of counterfactuals in real-world data, ITEs cannot be calculated directly, but must be estimated through the use of models. Based on statistical methods, but also driven by recent developments in machine and deep learning approaches, various models have been proposed to estimate treatment effects on a personalized level [51, 14, 5, 34]. In general, most of these models estimate the potential outcome for an individual patient by learning the underlying effects and interactions between patient characteristics, treatment, and outcome. The patient's ITE can then be calculated as the difference between the predicted potential outcomes with and without treatment. Furthermore, various methods were proposed to address the problem of confounding bias introduced in observational studies as a result of unobserved confounders [39, 4, 28], paving the way for ITE estimations for personalized medicine.

2 Estimating Individualized Treatment Effects from Time-series Data

Various works for estimating ITE using observational EHR data have recently been proposed [51, 45, 14]. Despite the plethora of works, most of them estimate the treatment effect using static data, where each patient is represented as a snapshot of covariates at the exposure of the treatment. Although useful, using only static data and disregarding the time component have many limitations in estimating the impact of treatment over time. Furthermore, the ITE estimation from static data limits the opportunities for learning when to stop or change the treatments when the outcome is dynamic and varies over time, a critical clinical application for ITE. Additionally, in the static treatment effect estimation setup, the treatments are assigned at a single time point and often remain static over time. However, using time-series data for treatment effect estimation would allow monitoring and dynamical change of the treatment plan while simultaneously observing the effect of time-varying treatment on patient covariates and outcomes of interest. A typical example of a time-series ITE problem is in the cancer application, where a patient's treatment option (e.g., radiation or chemotherapy) is adjusted according to his or her clinical response (e.g., tumor size) over time. In the static setting, such a problem cannot be addressed due to the absence of the time factor.

Despite the promise and potential of treatment effect estimation from time-series data, the major problem lies in the time-varying or temporal confounders. Similarly to static confounders, a temporal confounder is typically a time-varying variable that affects both the treatment assignment and outcome. For example, consider that angiotensin receptor blockers (ARBs) (treatment 1) are given when the blood pressure (covariate) of a hypertension patient is outside the normal range value. Suppose also that this patient's covariate was affected by the previous administration of an ACE inhibitor (treatment 0), another type of treatment for uncontrolled blood pressure. Estimating the effect of a different sequence of treatments on a patient's outcome would require adjusting for bias in the current step (treatment 1) and bias introduced by the previous application of ACE inhibitors (treatment 0). Adjusting for time-varying confounders remains a major challenge hindering the direct application of methodologies developed for static treatment effects tasks to dynamic problem settings. Here we have reviewed works in the literature for estimating treatment effects from time-series data, including the estimation frameworks, model architectures, and assumptions used. An overview of the existing work on ITE from time-series data is presented in Table 1. They are categorized into two main groups: (i) outcome estimation methods which focus on inferring the ITE by estimating the potential outcomes of different treatments, and (ii) deconfounder methods, which estimate the ITE in the presence of hidden confounders. More details are presented in Sections 2.2 and 2.3, covering ITE outcome estimation and deconfounder methods for time-series data, respectively. In Figure 1, we show an example causal model that underlies a dynamic ITE estimation setting with time-varying treatments and covariates.

Table 1: A summary of ITE works for time-series data.

	Proposed Methods	Estimation Framework	Assumptions	Model Architecture	Validation Data
Outcome Estimation Methods	MSM [39]	MSM	C/P/SSI	LR	NA
	RMSN [31]	MSM	C/P/SSI	LSTM	simulated tumor dynamics data
	CRN [5]	Balanced Representation	C/P/SSI	LSTM	simulated tumor dynamics data, ICU data
	G-Net [30]	G-formula	C/P/SSI	LSTM	simulated tumor dynamics data, simulated cardiovascular data
	Causal Transformer [34]	Balanced Representation	C/P/SSI	Transformer	simulated tumor dynamics data, (semi-synthetic) ICU data
Deconfounder Methods	Time Series Deconfounder [4]	Latent Factor Model	C/P/SSSI	RNN Factor Model	simulated data, ICU data
	Sequential Deconfounder [17]	Latent Factor Model	C/P/TIUC	GPLVM	simulated data, ICU data
	Deconfounding Temporal AutoEncoder [28]	Noisy Proxies	C/P	AutoEncoder	simulated data, ICU data

*The abbreviations in full form. (C): Consistency, (SO): Sequential Overlap, (SSI): Sequential Strong Ignorability, (SSSI): Sequential Single Strong Ignorability, (TIUC): Time-Invariant Unobserved Confounding, (LR): Logistic Regression, (LSTM): Long Short-Term Memory, (RNN): Recurrent Neural Network, (GPLVM): Gaussian Process Latent Variable Model.

2.1 Assumptions

Estimating treatment effects from time-series data relies on the potential outcome framework [41] and its extensions to the time-varying setting [38]. In the potential outcome framework, an ITE is the difference between the potential outcomes for a specific individual given different treatments. Three main assumptions are typically required for treatment effects to be identifiable from time-series data. We explain each in lay terms and provide examples from real-world clinical problems. For interested readers, we provide references to works that explain and include mathematical notations.

1. **Consistency.** This assumption states that the potential outcome of a patient should be consistent with his/her "factual" outcome if the same treatment plan is applied. For example, consider a patient who has been on a diabetes treatment plan A for several years, where blood glucose is considered a monitored outcome over time. If a model is built to estimate this patient's potential outcome given the same treatment plan A, such outcome would be equivalent to the outcome observed for that patient (i.e., blood glucose). The mathematical notation and theoretical basis for this consistency are explained in [5, 34].

2. **Sequential Overlap or Positivity**. This assumption means that each treatment option has a non-zero probability of being given to the patient at each timestep. For example, let us consider a cancer treatment in which options are radiotherapy or chemotherapy. If a patient was given radiotherapy last month, a doctor might give this patient chemotherapy or radiotherapy this month, and both have a non-zero probability of being given to the patient.

3. **Sequential Strong Ignorability** This assumption means that conditioned on the observed patient history, the current treatment assignment is independent of the potential outcome. In some works, this assumption is referred to as sequential exchangeability or "no unobserved confounders". In other words, no unobserved confounders affect both treatment and outcome. While useful, this assumption cannot be tested in practice since various factors may impact the treatment and the outcome, yet they might not be recorded or observed. For this purpose, various works have been proposed to relax this assumption (see the reference works for estimation with hidden confounders described in Section 2.3). Example variants include *Sequential Single Strong Ignorability*, where the assumption is limited to no hidden single cause confounders [4]. Another example is *Time-Invariant Unobserved Confounding*, where confounders exist under the condition that they are the same random variable at each time step [17].

2.2 ITE Outcome Estimation Methods in time-series data

2.2.1 Marginal Structural Models (MSM)

Various epidemiological approaches have been proposed to account for confounders varying over time, one of which is *inverse probability of treatment weighting* (IPTW) [8]. The main idea be-

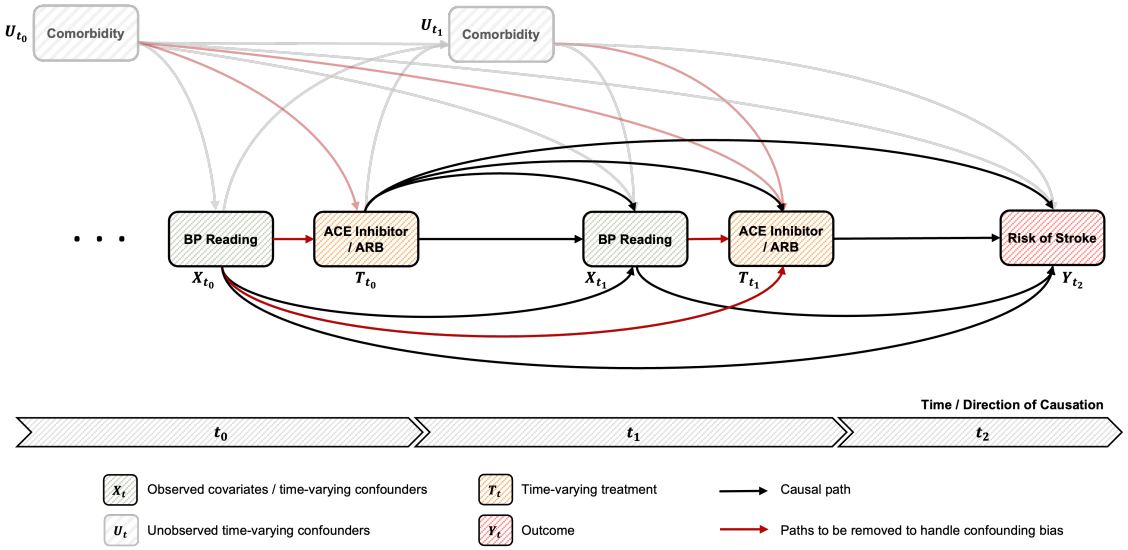

Figure 1: Illustration of a causal model underlying a dynamic ITE estimation setting with time-varying treatments and covariates. The arrows indicate causal dependencies between variables. Here, the observed covariates are blood pressure (BP) readings that act as time-varying confounders affecting all subsequent treatment decisions (ACE inhibitor vs. ARB) and the outcome of interest (risk of stroke). We also showcase an example of an unobserved "hidden" confounder, namely potential comorbidity that is not directly represented in the data. All connections from covariates and unobserved variables to the treatment variables are depicted in red, and would need to be accounted for to remove confounding bias effectively.

hind IPTW involves assigning weights that will redistribute or balance the population such that the effect of time-varying confounding is removed. The weights are derived based on the inverse probability of receiving the patient's treatment at each respective time point conditional on their covariate's history. The IPTW setup creates a pseudo-population set in which receiving treatment assignment is independent of the underlying patient characteristics and previous treatment assignments [3]. One way to implement IPTW in the time-series setting is through *marginal structural models* (MSMs) [39]. MSMs focus on controlling the effects of time-varying confounders affected by previous treatment exposure. The name "Marginal" refers to the approach of estimating the marginal distribution of treatment over time with respect to the outcome [50]. Similarly, the name "structural" refers to the approach of exploring causal relationships inspired by econometrics [50]. An MSM first calculates the weights, most commonly via a regression-based IPTW model, and assigns such weights to each observation. The estimated weights indicate whether each of the observations in confounders is under-represented or over-represented in the sample for a target population [50]. The use of sample reweighing aims to remove the imbalance and bias caused by the uneven distribution of time-varying confounders between treatment groups. Finally, the treatment effect can be estimated using the calculated weights.

Although powerful and useful, MSMs have limitations when dealing with complicated and high-dimensional data dynamics. This is because the treatment effect predictions are calculated using linear or logistic regression models. To address this, [31] proposed *recurrent marginal structural networks* (RMSMs) where recurrent neural networks [10] were used to estimate the inverse probability of treatment weights and counterfactual treatment outcomes. Similarly to the standard two-stage MSM approach, RMSM has two main networks. The first network calculates the treatment probability weights used for the IPTW. The second network, on the other hand, is the prediction used to determine the response to treatment given a sequence of treatments and the calculated weights [31]. Despite the promise shown by the statistical and deep learning approaches, MSMs can be unstable

if the IPTW results in extreme weights, leading to model misspecification [5].

2.2.2 G-formula

Another method for estimating treatment effects in time-varying confounders is *G-formula*. G-formula was first described by Robins et al. [37], where the author proposed a method for generalizing standardization to time-varying treatments and confounders and referred to it as the formula of the G computation algorithm. Most of the epidemiological works use the term G-formula or G-computation to refer to the same method proposed by [37]. The key assumption for the G-computation formula is that the treatment received at each time was assigned conditional on the observed past treatment and the history of the covariates [37]. G-formula works by estimating the conditional distribution of relevant covariates given the covariate and treatment history at each time point, then producing Monte Carlo estimates of counterfactual outcomes by simulating forward patient trajectories under treatment strategies of interest [35]. In most statistical works, the estimation of patient trajectories and outcomes is done using simplistic regression estimators. Although useful, it is important to remember that simple regression models fail to capture complex dependencies over time when dealing with high-dimensional time-varying data. In terms of implementation, there are no well-established G-formula implementations in statistical packages, which limits its applicability when compared to MSMs. Recently, [30] proposed G-NEt, the first deep-learning work that estimates ITEs using an LSTM-based G-formula model. The G-Net results showed improved performance compared to those estimated using a logistic regression estimator and other deep learning-based models [31, 5].

2.2.3 Balanced Representations

Unlike MSM and G-formula-based estimations, a new class of deep learning estimation evolved based on learning representations that balance the distribution of treatment and control groups. Original works for learning balanced representations were first proposed for static settings [22], then several studies used a deep learning architecture to learn treatment-invariant representation for each time step to remove the association between patient history and treatment assignment. For example, the *counterfactual recurrent network* (CRN) [5] is the first work to use a sequence-to-sequence model to learn balanced representations through adversarial training. In their proposed work, CRN aims to learn representations not predictive of treatment assignments yet achieve the highest performance in predicting the outcome. Another related work is that of [34], where the authors proposed *Causal Transformer*, which is a transformer-based model that aims to learn treatment-invariant balanced representations to estimate ITE over time. To do so, the Causal Transformer comprises three transformer sub-networks for processing the time-varying covariates, treatments, and outcomes, all of which are combined via a joint network with cross-attentions.

2.3 ITE Deconfounding methods for time-series data

2.3.1 Latent Factor Models

All the aforementioned studies to estimate ITE in time-series data focus on settings where all confounders are observed, or in other words; they require sequential strong ignorability assumption to hold. Despite the potential of such works, sequential ignorability is not testable in practice. To this end, several studies have proposed approaches in which sequential ignorability is relaxed to account for settings where forms of hidden confounders exist in the data. For example, the first work to propose a deep learning model for deconfounding time-series data was the *Time Series Deconfounder* [4]. In their proposed work, the authors focus on addressing a specific type of hidden confounders which they refer to as multi-cause hidden confounders. The Time Series Deconfounder builds on a factor model to learn the distribution of treatments over time. By leveraging the dependence between multiple treatment options at each given time step, the factor model infers substitutes for unobserved confounders at each time step. The assumption of sequential strong

ignorability is relaxed to sequential single strong ignorability where the assumption is limited to the absence of hidden single-cause confounders [4]. The Times Series Deconfounder can be applied to the datasets before passing the deconfounded data to other outcome estimation models such as RMSM [31] or CRN [5].

While the results show great promise, the Time Series Deconfounder only works when there are multiple treatment options and fails when there is a single treatment option at each time step. This limitation of the Time Series Deconfounder is related to its use on the dependence between multiple treatment options to infer substitutes of hidden confounders. *Sequential Deconfounder*, on the other hand, is a method that deconfounds time-series data for ITE by fitting a Gaussian Process (GP) latent variable model to capture any sequential dependence between the assigned treatments [17], without the limitation of depending on multiple treatments. The GP-based latent variable model aims to control for substitutes and uses them in conjunction with outcome estimation models such as RMSM [31] or CRN [5] to estimate treatment effects over time. Although the sequential deconfounder does not require multiple treatments at each time step, it requires a special case of the ignorability assumption, which they refer to as *Time-Invariant Unobserved confounding* [17]. This assumption requires that the hidden confounder is the same random variable at each time step.

2.3.2 Noisy Proxies

Most aforementioned studies utilize latent models to infer the hidden confounder in time-series ITE by capturing sequential dependencies. Recently, the Deconfounding Temporal AutoEncoder (DTA) is an autorencoder-based model that utilizes noise proxies as an alternative to latent factor models to learn hidden embeddings that resemble the true hidden confounders [28]. The main assumption in DTA builds on the fact that the observed covariates are not necessarily true confounders and assumes that the observed covariates are noisy proxies of the true confounders. DTA aims to learn a hidden embedding for which the ITE is the same when hidden confounders are present and when Sequential Strong Ignorbaility applies. To do so, DTA optimizes over a special loss that is referred to as a cause regularization penalty to yield outcomes and treatment assignments that are conditionally independent for each hidden embedding [28].

3 Datasets and Evaluation

Most studies found in the literature make use of simulated datasets to evaluate their methods for ITE estimation from time-series data. Unlike real-world data, where only the factual outcome is observed, simulations provide ground truths for all potential outcomes. Some simulated datasets used in the literature [4, 17, 28] do not aim to mimic specific medical scenarios; instead, they are based on purely mathematical modeling of time-varying covariates, hidden confounders, treatments, and outcomes. In contrast, models such as the pharmacokinetic-pharmacodynamic (PK-PD) model by Geng et al. [12], which simulates cancer dynamics, strive to provide a realistic perspective on actual medical processes. Simulated "observational" cancer growth data, derived from the PK-PD model, is widely used in the literature [31, 5, 30, 34] for evaluation purposes. The model has been adapted to simulate the change in tumor volume over time under the influence of different treatment options, such as chemotherapy and radiation. Time-dependent confounding can be incorporated into the model by expressing the probabilities of administering chemotherapy and radiation as a function of tumor size [31].

In addition to the PK-PD model, Li et al. [30] evaluate the performance of G-Net on longitudinal data, simulated using Heldt et al. CVSim [19]. CVSim provides a mechanistic model of the human cardiovascular system and enables simulation of the trajectories of outcome parameters such as mean arterial pressure (MAP) or central venous pressure (CVP) in interventions such as different fluid or vasopressor administration strategies.

Moreover, Melnychuk et al. [34] evaluate their Causal Transformer on semi-synthetic and real-world datasets of patient trajectories in the ICU that are based on the MIMIC-III dataset [23]. For

their semi-synthetic data, they combine real-world covariates from the MIMIC-extract by Wang et al. (2020) [49] with simulated trajectories of control outcomes. Treated outcomes are obtained by simulating synthetic binary treatments, incorporating confounding, and applying those treatments to the control outcomes [34]. For their experiments on real-world data, they used the same patient's covariates from MIMIC-III again and considered the effect of vasopressors and mechanical ventilation on blood pressure. However, since counterfactual outcomes are not available for real-world data, they can only report the performance in predicting the factual outcomes.

The fundamental problem of causal inference and the resulting reliance on (semi-) simulated data sets for comprehensive validation poses a major challenge to developing models for ITE estimation. Although models such as CVSim or the PK-PD model can offer valuable insights from a medical standpoint, their data generation process is less complex with simple assumptions, which could result in lower performance when compared to real-world applications.

4 Future Outlook

Discussion and Future Outlook

While this perspective provides an overview of existing methodologies for ITE estimation in time-series EHRs, it also highlights significant gaps and opportunities for future research. One pressing challenge is the integration of multi-modal data sources, such as combining structured EHR data with imaging, genomics, or patient-reported outcomes. This integration could enable a more holistic and accurate modeling of patient responses and treatment effects, especially in complex diseases such as cancer or cardiovascular conditions where treatment pathways and patient outcomes are highly variable. Multi-modal approaches could also enhance the interpretability of ITE models by incorporating richer clinical context. However, these efforts face barriers, including the heterogeneity of disparate data types and computational challenges in training models on large-scale heterogeneous datasets.

Another critical area is the solution to irregular sampling and missing data, as these issues remain prevalent in real-world healthcare datasets. Current imputation techniques, while advanced, often fail to account for the causal structure of the data. Methods like SAITS [11], IGNITE [15] for imputing missing values in time-series data, represent promising advances. However, more research is needed to assess how these methods interact with causal inference frameworks and whether imputation biases might affect downstream ITE estimation. Addressing missingness in ways that align with clinical realities, such as dropout patterns related to patient health, could significantly improve the reliability of these models.

Additionally, the field would benefit from frameworks that explicitly address fairness and bias in ITE models. Healthcare disparities often arise due to the underrepresentation of certain groups in clinical trials or observational datasets, leading to treatment recommendations that may not generalize equitably. Frameworks for evaluating and minimizing bias in causal models are still in their infancy, but are critical to ensuring that ITE-based decision support tools do not inadvertently exacerbate health inequities. For instance, future work could incorporate techniques such as adversarial training or fairness-aware metrics to identify and address disparities in treatment effects across subpopulations.

Model validation remains another pivotal challenge. Although synthetic and semi-synthetic datasets, have been instrumental in developing new methodologies, they often fail to capture the complexity of real-world clinical environments. Validation in large real-world datasets such as MIMIC-III [24] and other clinical datasets from primary care or general wards are essential to understand how models perform in practical settings. Beyond validation, deploying these models in clinical environments requires robust evaluation frameworks that account for clinical workflows and the long-term impact on patient outcomes.

Incorporation of reinforcement learning and adaptive treatment strategies presents an exciting opportunity for the field. Unlike static ITE models, reinforcement learning frameworks can dynam-

ically optimize treatment pathways by continuously updating recommendations based on evolving patient states. For example, in chronic disease management, such approaches could determine not just what treatment to prescribe but when to escalate, de-escalate, or modify interventions. However, this requires significant advances in addressing the exploration-exploitation trade-off in healthcare, where ethical considerations and patient safety limit experimental flexibility.

Another critical area for future research is the need to relax the strong assumptions often required for ITE estimation in time-series data. Many existing models rely on assumptions such as sequential strong ignorability, consistency, and positivity, which may not hold in real-world clinical settings. For instance, sequential strong ignorability assumes that all relevant confounders are observed, but in practice there are often unmeasured variables that influence both treatment assignment and outcomes, such as socioeconomic factors or latent health conditions. Similarly, the positivity assumption—requiring that all patients have a non-zero probability of receiving any treatment at every time step—is frequently violated in clinical practice, where treatments are prescribed based on strict guidelines or patient-specific constraints. Recent advances in deconfounding methods, such as latent factor models [4] and autoencoder-based approaches, such as the Deconfounding Temporal Autoencoder [28], offer promising pathways to address these limitations. More work is needed to develop models that can effectively estimate treatment effects in the presence of hidden confounders or violations of positivity, while maintaining computational efficiency and scalability. By relaxing these assumptions, future methodologies can better align with the complexities of real-world healthcare data, allowing for more robust and clinically relevant predictions.

Finally, the interpretability of ITE models remains a key factor for their adoption in clinical practice. Clinicians must understand and trust the reasoning behind the model predictions to integrate them into decision-making processes. Future research should prioritize developing interpretable models that can communicate treatment recommendations and uncertainties to healthcare providers. Combining ITE with explainable AI techniques, such as counterfactual explanations, could be instrumental in achieving this goal.

In summary, addressing these challenges requires a multi-disciplinary approach that combines advancements in machine learning, causal inference, and domain expertise from clinicians and healthcare stakeholders. By tackling issues of multi-modal data integration, missingness, fairness, validation, interpretability, and dynamic modeling, the field can move closer to bridging the gap between theoretical developments and their real-world impact. The ultimate goal is to create robust, equitable, and clinically meaningful ITE estimation frameworks that transform patient care and improve health outcomes.

References

[1] Sarah C Anoke, Sharon-Lise Normand, and Corwin M Zigler. "Approaches to treatment effect heterogeneity in the presence of confounding". In: *Statistics in medicine* 38.15 (2019), pp. 2797–2815.

[2] Peter C Austin and Elizabeth A Stuart. "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies". In: *Statistics in Medicine* 34.28 (2015), pp. 3661–3679.

[3] Peter C Austin and Elizabeth A Stuart. "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies". In: *Statistics in medicine* 34.28 (2015), pp. 3661–3679.

[4] Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. "Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 884–895.

[5] Ioana Bica et al. “Estimating counterfactual treatment outcomes over time through adversarially balanced representations”. In: *arXiv preprint arXiv:2002.04083* (2020).

[6] Ioana Bica et al. “From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges”. In: *Clinical Pharmacology & Therapeutics* 109.1 (2021), pp. 87–100.

[7] Nick Black. “Why we need observational studies to evaluate the effectiveness of health care”. In: *Bmj* 312.7040 (1996), pp. 1215–1218.

[8] Nicholas C Chesnaye et al. “An introduction to inverse probability of treatment weighting in observational research”. In: *Clinical Kidney Journal* 15.1 (2022), pp. 14–20.

[9] Hugh A Chipman, Edward I George, and Robert E McCulloch. “BART: Bayesian additive regression trees”. In: *The Annals of Applied Statistics* 4.1 (2010), pp. 266–298.

[10] Jerome T Connor, R Douglas Martin, and Les E Atlas. “Recurrent neural networks and robust time series prediction”. In: *IEEE transactions on neural networks* 5.2 (1994), pp. 240–254.

[11] Wenjie Du, David Côté, and Yan Liu. “Saits: Self-attention-based imputation for time series”. In: *Expert Systems with Applications* 219 (2023), p. 119619.

[12] Changran Geng, Harald Paganetti, and Clemens Grassberger. “Prediction of Treatment Response for Combined Chemo- and Radiation Therapy for Non-Small Cell Lung Cancer Patients Using a Bio-Mathematical Model”. In: *Scientific Reports* 7.1 (Oct. 2017). DOI: 10.1038/s41598-017-13646-z. URL: https://doi.org/10.1038/s41598-017-13646-z.

[13] Changran Geng, Harald Paganetti, and Clemens Grassberger. “Prediction of treatment response for combined chemo- and radiation therapy for non-small cell lung cancer patients using a bio-mathematical model”. In: *Scientific Reports* 7.1 (2017), p. 13646.

[14] Shantanu Ghosh et al. “Propensity score synthetic augmentation matching using generative adversarial networks (PSSAM-GAN)”. In: *Computer methods and programs in biomedicine update* 1 (2021), p. 100020.

[15] Ghadeer O Ghosheh, Jin Li, and Tingting Zhu. “IGNITE: Individualized GeNeration of Imputations in Time-series Electronic health records”. In: *arXiv preprint arXiv:2401.04402* (2024).

[16] Cory E Goldstein et al. “Ethical issues in pragmatic randomized controlled trials: a review of the recent literature identifies gaps in ethical argumentation”. In: *BMC medical ethics* 19.1 (2018), pp. 1–10.

[17] Tobias Hatt and Stefan Feuerriegel. “Sequential deconfounding for causal inference with unobserved confounders”. In: *arXiv preprint arXiv:2104.09323* (2021).

[18] RJ Hayes and S Bennett. “Simple sample size calculation for cluster-randomized trials.” In: *International journal of epidemiology* 28.2 (1999), pp. 319–326.

[19] Thomas Heldt et al. “CVSim: An open-source cardiovascular simulator for teaching and research”. en. In: *Open Pacing Electrophysiol. Ther. J.* 3 (2010), pp. 45–54.

[20] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

[21] Susan D Horn et al. “Another look at observational studies in rehabilitation research: going beyond the holy grail of the randomized controlled trial”. In: *Archives of Physical Medicine and Rehabilitation* 86.12 (2005), pp. 8–15.

[22] Fredrik Johansson, Uri Shalit, and David Sontag. “Learning representations for counterfactual inference”. In: *International conference on machine learning*. 2016, pp. 3020–3029.

[23] Alistair E.W. Johnson et al. “MIMIC-III, a freely accessible critical care database”. In: *Scientific Data* 3.1 (May 2016). DOI: 10.1038/sdata.2016.35. URL: https://doi.org/10.1038/sdata.2016.35.

[24] Alistair EW Johnson et al. "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3.1 (2016), pp. 1–9.

[25] Rita R Kalyani, Sherita H Golden, and William T Cefalu. "Diabetes and aging: unique considerations and goals of care". In: *Diabetes Care* 40.4 (2017), pp. 440–443.

[26] Alison M Kim, Candace M Tingen, and Teresa K Woodruff. "Sex bias in trials and treatment must end". In: *Nature* 465.7299 (2010), pp. 688–689.

[27] Helena Chmura Kraemer. "Pitfalls of multisite randomized clinical trials of efficacy and effectiveness". In: *Schizophrenia Bulletin* 26.3 (2000), pp. 533–541.

[28] Milan Kuzmanovic, Tobias Hatt, and Stefan Feuerriegel. "Deconfounding Temporal Autoencoder: estimating treatment effects over time using noisy proxies". In: *Machine Learning for Health*. PMLR. 2021, pp. 143–155.

[29] The Lancet. "Personalised medicine in the UK." In: *Lancet (London, England)* 391.10115 (2018), e1.

[30] Rui Li et al. "G-net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime". In: *Machine Learning for Health*. PMLR. 2021, pp. 282–299.

[31] Bryan Lim. "Forecasting treatment responses over time using recurrent marginal structural networks". In: *Advances in neural information processing systems* 31 (2018).

[32] WA Lindsay et al. "Age, sex, race and ethnicity representativeness of randomised controlled trials in peri-operative medicine". In: *Anaesthesia* 75.6 (2020), pp. 809–815.

[33] Jared K Lunceford and Marie Davidian. "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study". In: *Statistics in medicine* 23.19 (2004), pp. 2937–2960.

[34] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. "Causal transformer for estimating counterfactual outcomes". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 15293–15329.

[35] Ashley I Naimi, Stephen R Cole, and Edward H Kennedy. "An introduction to g methods". In: *International journal of epidemiology* 46.2 (2017), pp. 756–762.

[36] Simon J Newsome, Ruth H Keogh, and Rhian M Daniel. "Estimating long-term treatment effects in observational data: A comparison of the performance of different methods under real-world uncertainty". In: *Statistics in medicine* 37.15 (2018), pp. 2367–2390.

[37] James Robins. "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". In: *Mathematical modelling* 7.9-12 (1986), pp. 1393–1512.

[38] James Robins and Miguel Hernan. "Estimation of the causal effects of time-varying exposures". In: *Chapman & Hall/CRC Handbooks of Modern Statistical Methods* (2008), pp. 553–599.

[39] James M Robins, Miguel Angel Hernan, and Babette Brumback. "Marginal structural models and causal inference in epidemiology". In: *Epidemiology* (2000), pp. 550–560.

[40] Paul R Rosenbaum and Donald B Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (1983), pp. 41–55.

[41] Donald B Rubin. "Causal inference using potential outcomes: Design, modeling, decisions". In: *Journal of the American Statistical Association* 100.469 (2005), pp. 322–331.

[42] Robert William Sanson-Fisher et al. "Limitations of the randomized controlled trial in evaluating population-based health interventions". In: *American journal of preventive medicine* 33.2 (2007), pp. 155–161.

[43] Roosmarijn MC Schelvis et al. “Evaluation of occupational health interventions using a randomized controlled trial: challenges and alternative research designs”. In: *Scandinavian journal of work, environment & health* (2015), pp. 491–503.

[44] Kenneth F Schulz et al. “Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials”. In: *Jama* 273.5 (1995), pp. 408–412.

[45] Uri Shalit, Fredrik D Johansson, and David Sontag. “Estimating individual treatment effect: generalization bounds and algorithms”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 3076–3085.

[46] RW Shaw. “Adverse long-term effects of oral contraceptives: a review”. In: *British journal of obstetrics and gynaecology* 94.8 (1987), pp. 724–730.

[47] Charles A Stiller. “Centralised treatment, entry to trials and survival”. In: *British journal of cancer* 70.2 (1994), pp. 352–362.

[48] Sheldon W Tobe, Diane Hua, and Patrick Twohig. “Clinical practice guidelines”. In: Future Medicine, 2013.

[49] Shirly Wang et al. “MIMIC-Extract”. In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. ACM, Apr. 2020. DOI: 10.1145/3368555.3384469. URL: https://doi.org/10.1145/3368555.3384469.

[50] Tyler Williamson and Pietro Ravani. “Marginal structural models in clinical research: when and how to use them?” In: *Nephrology Dialysis Transplantation* 32.suppl_2 (2017), pp. ii84–ii90.

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