Targeted-BEHRT: Deep Learning for Observational Causal Inference on Longitudinal Electronic Health Records

Shishir Rao, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Yikuan Li, Rema Ramakrishnan, Abdelaali Hassaine, Dexter Canoy, and Kazem Rahimi

***Abstract*— Observational causal inference is useful for decision-making in medicine when randomized clinical trials (RCTs) are infeasible or nongeneralizable. However, traditional approaches do not always deliver unconfounded causal conclusions in practice. The rise of "doubly robust" nonparametric tools coupled with the growth of deep learning for capturing rich representations of multimodal data offers a unique opportunity to develop and test such models for causal inference on comprehensive electronic health records (EHRs). In this article, we investigate causal modeling of an RCT-established causal association: the effect of classes of antihypertensive on incident cancer risk. We develop a transformer-based model, targeted bidirectional EHR transformer (T-BEHRT) coupled with doubly robust estimation to estimate average risk ratio (RR). We compare our model to benchmark statistical and deep learning models for causal inference in multiple experiments on semi-synthetic derivations of our dataset with various types and intensities of confounding. In order to further test the reliability of our approach, we test our model on situations of limited data. We find that our model provides more accurate estimates of relative risk [least sum absolute error (SAE) from ground truth] compared with benchmark estimations. Finally, our model provides an estimate of class-wise antihypertensive effect on cancer risk that is consistent with results derived from RCTs.**

***Index Terms*— Causal inference, deep learning, electronic health records (EHRs), machine learning.**

Manuscript received 13 August 2021; revised 7 February 2022 and 5 May 2022; accepted 12 June 2022. Date of publication 23 June 2022; date of current version 5 April 2024. The work of Kazem Rahimi was supported in part by the British Heart Foundation (BHF) under Grant FS/PhD/21/29110 and Grant PG/18/65/33872; in part by the UK Research and Innovation (UKRI) through the Global Challenges Research Fund (GCRF) under Grant ES/P0110551/1; in part by the Oxford National Institute of Health Research (NIHR) Biomedical Research Centre; and in part by the Oxford Martin School (OMS), University of Oxford. The work of Yikuan Li and Dexter Canoy was supported by the British Heart Foundation (BHF) under Grant FS/PhD/21/29110 and Grant PG/18/65/33872. *(Corresponding author: Shishir Rao.)*

Shishir Rao, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Yikuan Li, and Abdelaali Hassaine are with the Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford OX3 9DU, U.K. (e-mail: shishir.rao@wrh.ox.ac.uk).

Rema Ramakrishnan is with the National Perinatal Epidemiology Unit, University of Oxford, Oxford OX3 7LF, U.K.

Dexter Canoy and Kazem Rahimi are with the Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford OX3 9DU, U.K., and also with the NIHR Oxford Biomedical Research Centre, Oxford University Hospitals National Health Service (NHS) Foundation Trust, Oxford OX3 9DU, U.K.

Digital Object Identifier 10.1109/TNNLS.2022.3183864

I. Introduction

ESTIMATING causal effect of a treatment is an important problem in the field of epidemiology [1]. Consider the following example of treatment effect: the investigation of the effect of various antihypertensive drug classes on cancer. The effect of a particular drug class is ideally investigated experimentally with a randomized control trial (RCT). Researchers randomly treat patients with different classes of antihypertensives—e.g., some treated with angiotensin converting enzyme inhibitors (ACEIs) and others and beta-blockers—and then compare incidence of cancer between exposure groups. This comparison is often measured as empirical risk ratio (RR) otherwise known as relative risk, which is the proportion of cancer in one exposure group divided by the same in the other. Since patients are randomized, one can assume that confounders have no meaningful difference across exposure groups. Therefore, trials offer an unconfounded estimate of the association. In the case of association of classes of antihypertensives and cancer, numerous RCTs have indeed pursued this investigation and found it to be null [2]—i.e., any given class of antihypertensives does not cause cancer any more than any other class.

However, where an RCT does not exist, is unfeasible, or fails to generalize, decision makers will have to draw upon evidence from observational data [3]–[5]. In this observational capacity, adjustment of confounding variables (i.e., variables that effect both exposure assignment and outcome) is a necessity, and insufficient adjustment can result in biased conclusions.

Traditionally, conventional statistical models are used for confounding adjustment and estimation of RR in epidemiological studies. However, these models require manual feature engineering and fare poorly in observational settings with issues of selection bias and finite sample estimation bias [6]. To address these issues, statisticians have recently developed "doubly robust" estimators for approximating population-level causal effect [6], [7].

In parallel, the growing availability of routinely collected, administrative clinical records databases with linked electronic health records (EHRs) that capture numerous variables describing large numbers of individuals on a population level provides an encouraging opportunity to conduct observational

studies in the medical domain [8]. These administrative EHR datasets such as Clinical Practice Research Datalink (CPRD) in the U.K. offer a multitude of temporal and static variables in addition to data linkages to other datasets [9], ultimately providing multiple types of routine EHR for modeling.

More recently, developments in the field of deep learning have allowed scalable modeling of such high-dimensional data for various tasks. Across several domains, deep learning modeling has shown superior predictive performance as compared with traditional approaches. In particular, the development of representation learning methods such as multitask unsupervised training strategies has allowed for richer feature extraction and more generalizable performance [10].

Since RR is generally used to estimate effect in most randomized and observational cohort studies, our objective in this work is to develop and evaluate methods that bring together contributions from deep learning and statistics in order to estimate RR more accurately as compared with benchmark models. We address this objective through a combination of three contributions.

Our first contribution is a novel model, targeted bidirectional EHR transformer (T-BEHRT) for more accurate RR estimation. Our method synthesizes the following elements in a novel multitask learning framework: 1) expanded BEHRT architecture (transformer-based feature extractor) for adjustment of both temporal and static variables [11], [12]; 2) novel auxiliary unsupervised learning for richer feature extraction; and 3) doubly robust semi-parametric estimation for mitigating finite-sample estimation biases.

As our second contribution, we develop an environment to objectively evaluate accuracy of RR estimation of models. Focusing on the aforementioned case study: the effect of various classes of antihypertensives on cancer, we form an observational dataset by including patients taking different classes of antihypertensives and investigate risk of cancer. Our reference exposure is ACEIs, one of said classes of antihypertensives. Since counterfactual outcomes (i.e., outcome under a specific exposure status) are missing in our observational dataset, ground truth RR is inaccessible, thereby making model comparisons difficult. Thus, we first construct semi-synthetic derivations of our observational dataset with generated ground truth RR, and then apply our model against statistical and deep learning benchmarks in several experiments to identify the model with best RR estimation. In addition, to test our model in situations of limited data, we demonstrate the utility of T-BEHRT compared with other models in finite-sample estimation experiments.

As our third and final contribution: after validating our model on semi-synthetic derivations of routine clinical observational data, we demonstrate that our model can be applied to the aforementioned observational study: the effect of ACEIs on cancer relative to other drug classes. Where traditional statistical models have demonstrated conflicting results, these associations have been deemed null in numerous RCTs [13], [14] with narrow confidence intervals (CIs), across a wide range of patient groups, for multiple cancer subtypes.

A. Background

Traditionally, semi-parametric and parametric statistical modeling have been predominantly explored in observational longitudinal causal inference in the field of epidemiology. Regression-based models (e.g., logistic regression (LR), log-binomial regression, Poisson regression, and Cox proportional hazards model) utilize the exposure as a covariate and implement regression fitting for estimation of outcomes [15], [16]. In addition to standard regression modeling, recent G-methods for longitudinal data modeling have been developed to relax assumptions of identifiability of causal effect and have been implemented in epidemiological studies [17]–[19]. However, both classes of models have known limitations [20], [21]. First, they require careful manual feature engineering—useful for modeling known confounders but impractical for unknown or interacting variables. Furthermore, these models are susceptible to finite-sample estimation biases (i.e., biased estimation manifesting when the set of adjusted variables is high-dimensional with limited overlap between exposure groups while sample size is limited) [15]. One solution proposed is to adjust solely for the variables that are associated with exposure by propensity score modeling [21]; however, naïve propensity score-based methods require correct specification of the exposure prediction model, often not guaranteed [7]. If misspecified, the errors of the weights rapidly increase rendering highly erratic downstream causal estimates [22].

More recent work in semi-parametric estimation theory—namely "doubly robust" estimation theory—circumvents misspecification issues of these modeling approaches. These doubly robust estimators rely on the consistency of either prediction of propensity score or prediction of outcome to produce unbiased causal effect estimates [23], [24], and examples such as targeted maximum likelihood estimation (TMLE) and derivatives such as the cross-validated TMLE (CV-TMLE) have been prolifically used to explore causal inference problems of average treatment effect (ATE) [25]–[27]. TMLE-based methods have recently also been applied for epidemiological studies on EHR-based causal inference [28].

In contrast, deep learning has been a rapidly evolving field over the past few years. Specifically, for her-based tasks, there have been many convolutional, recurrent, and transformer neural networks models designed for feature extraction (i.e., "representation learning") from raw records and risk prediction [11], [29], [30]. One such transformer model, BEHRT, leverages contextualized embeddings to represent longitudinal clinical encounters and multihead self-attention to achieve the state-of-the-art performances in various EHR-based tasks [31], [32]. Furthermore, there has been much progress in representation learning and generalizability for deep learning. Research has shown that auxiliary unsupervised learning: 1) adds an additional inductive bias ultimately improving generalizability and 2) helps to learn representations shared or beneficial for the main task—in our case, the two tasks being propensity confounding adjustment and causal inference [10], [33], [34].

In the last decade, there have been advances in deep learning for causal inference. Models like treatment agnostic

representation network (TARNET), Dragonnet, causal effect variational autoencoder (CEVAE), and others have been tested on synthetic and semi-synthetic derivations of static tabular data [3], [35]–[37]. The TARNET model has also been applied in epidemiological studies [38]. In particular, Dragonnet exploits the sufficiency of the propensity score to simultaneously model the propensity score and the outcomes [39]. However, these models have not been specifically tested in observational settings involving routine multivariate EHR. And even though deep learning can incorporate multimodal variables, few approaches firstly model both temporal and static variables for causal inference and, secondly, develop environments to objectively test proposed solutions against benchmarks. Lastly, the considerable literature of deep learning for causal inference investigates ATE and conditional ATE/individualized treatment effect (ITE) almost exclusively; methods have rarely been evaluated for accuracy of RR estimation—a metric preferred by clinicians since RR captures risk relative to baseline risk (i.e., risk in the control cohort).

II. Methods

A. Problem Setup

Our objective in this work is to estimate RR in the setup of binary exposure and outcome. Consider the population of patients described by a tuple generated independently and identically: $(X_i, Y_i, T_i) \sim P$. Each patient i is described by medical records, X_i and is assigned exposure status, $T_i \in \{0, 1\}$. The exposures, T_i, in the presented work are two classes of antihypertensives with one of the classes acting as reference group. The variable Y_i corresponds to the observed outcome—cancer—in our proposed investigations. In a fixed amount of "follow-up" time after hypothetical treatment, $T_i = 0$, outcome of cancer is notated as $Y_i^{T=0}$, and similarly for treatment, $T_i = 1$, $Y_i^{T=1}$. These two outcome variables are known as the potential outcomes under the Neyman–Rubin potential outcomes framework [40].

With this, the RR is defined as

$$RR = \frac{\mathbb{E}\left[Y^{T=1}\right]}{\mathbb{E}\left[Y^{T=0}\right]}. \tag{1}$$

As is fundamental to the problem of causal inference, only one of the two outcomes are observed, so equation (1) cannot directly be computed. However, with the following standard assumptions, the exposure effect is identifiable and RR is estimable.

1. *Consistency:* The potential outcome for T is the observed outcome if the given exposure was indeed T.
2. *Positivity:* For all X, there is a nonzero probability of being assigned any exposure status, $T_i \in \{0, 1\}$.
3. *Unconfoundedness or "no Hidden Confounding":* The potential outcomes are independent of the exposure given all confounders are adjusted for. In synthetic data experimental designs, this assumption is more securable. In reality however, this is not measurable, but with richer observational data in the form of comprehensive medical records comprising of various health indicators (e.g., diagnoses, medications, measurements data), confounders can be better adjusted [41]. Under these assumptions, the causal effect is identifiable and naïve RR estimator can be defined as

$$\hat{\psi} = \mathbb{E}\left[\frac{\mathbb{E}[Y|X, T = 1]}{\mathbb{E}[Y|X, T = 0]}\right]. \tag{2}$$

Other more complex estimators utilizing propensity score such as CV-TMLE are also implemented in this work.

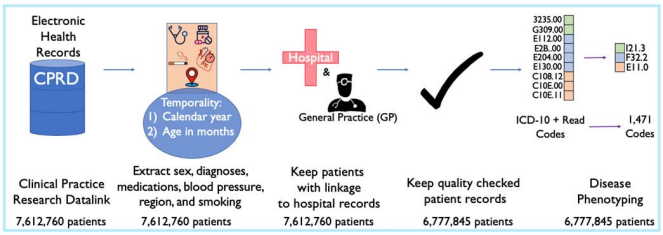

Fig. 1. Representation learning data selection pipeline. We use Clinical Practice Research Datalink (CPRD) and extract diagnoses, medications, blood pressure, smoking, region, and sex records. We homogenize codes from ICD-10 and read to one format. Unmapped read codes were kept for completeness.

B. Dataset and Patient Selection

For our investigations, we used a data cut from CPRD, which has been described previously [31]. The data entail records from 1 January 1985 up to 31 December 2015 and is linked to national administrative databases including hospitalizations (Hospital Episode Statistics, or HES) and death registration (from Office of National Statistics).

The dataset for the investigations was restricted to patients in the database who met the following criteria: 1) registered with the general practice for at least 12 months; 2) aged ≥16 years at registration; 3) registered with the practice considered providing "up-to-standard" data to CPRD; 4) individual data marked by CPRD to be of "acceptable" quality for research purposes (as determined by CPRD); and 5) registered with a practice that provided consent for linking the data with national databases for hospitalizations and death registry.

We extracted diagnoses, medications, blood pressure measurements, sex (male, female), region (ten regions in U.K.), and smoking status (non, previous, or current smoker). We mapped diagnoses and medication codes to a homogenized format for machine readability. This led to a dataset of 6 777 845 patients, which was used for general representation learning (shown in Fig. 1) for deep learning models.

For our causal inference investigation (i.e., investigating the effect of antihypertensive on incident cancer), a dataset containing five subpopulations had to be selected—one for each class of antihypertensives: ACEIs, diuretics, calcium channel blockers (CCBs), beta blockers (BBs), and angiotensin II receptor blockers (ARBs). Patients were selected in one of these groups based on first class of antihypertensive medications recorded before 2009 and if free of cancer report before this first prescription; the year 2009 was chosen conveniently to have sufficient "follow-up" time for the occurrence of potential cancers. The date of this first prescription was defined as "baseline" (a date between 1985 and 31 December 2008). Patients were then followed up from baseline until cancer

diagnosis (including cancer diagnoses as cause of death) or end of five-year follow-up period. The learning period included the entire patients' medical records up to a random point between six and 12 months before baseline; this is to account for any potential inaccuracies in timing of prescription (or decision to prescribe) and to avoid possibility of antihypertensive prescription itself influencing the model training. "CPRD Product codes" are used for identifying classes of antihypertensives and the set of codes were obtained from a dataset published by University of Bristol [42]. Codes for cancer are found in Table II and derived from clinically established publication of codes [43].

C. Semi-Synthetic Data Derivation

Data generation of sequential, temporal variables is a difficult task, and currently, there is no medically validated method of generating realistic EHR medical history. Thus, we utilized the existing medical history in observational data to exclusively simulate binary factual and counterfactual outcomes.

Inspired by other semi-synthetic data simulations [37], [44], intuitively, we first modeled the association between a medical history variable Z_i (e.g., some diagnosis/medication) and exposure T_i with the empirical propensity in the dataset: $\lambda_i = P(T_i = 1 \mid Z_i)$. If associated with an exposure ($\lambda_i \neq 0.5$), we generated the potential outcomes, $Y_i^{T=1}$ and $Y_i^{T=0}$ as a function of λ_i and exposure $T_i = 1$ and $T_i = 0$, respectively. In this way, semi-synthetic outcomes arose from an association between Z_i and exposure and Z_i and the outcome. Thus, the relationship between exposure and outcome is confounded by Z_i. While the empirical RR—the proportion of the outcome in one exposure group divided by the same in the other—would yield confounded causal conclusions, effectively adjusting for the confounder variable, Z_i, would yield identifiable (see Appendix Section A) causal association between exposure and outcome.

In addition, to test model adjustment potential in situations of varying confounding intensity, we weighted the contribution of the confounding with a β factor: the greater the β implies the greater the confounding. More details of the semi-synthetic data generative process and functions modeled are given in Appendix Section A.

In our work, we present investigations in semi-synthetic data utilizing two forms of confounders: persisting and transient confounding. We define persisting confounding as confounders that are assigned at birth and persist through one's life course; e.g., ethnicity, sex, genes, and other variables assigned at birth that associate with variables later in age. We define transient confounding as confounders that manifest at a point or period of one's life effecting events downstream in time; e.g., disease diagnoses, age itself, prescriptions, and other variables not assigned at birth. These two distinctions of confounding are presented in this work because they naturally capture prevalent forms of confounding seen in population health databases [45].

From our observational dataset, we investigated two exposure groups—ACEIs and diuretics and noticed female sex was associated with the diuretics exposure status and thus, chose it to be a persistent confounder and generated conditional outcomes. For another pair of exposures, i.e., ARBs and CCBs, we identified association of incidence of at least one of heart failure, hypertension, ischemic heart disease, and diabetes mellitus to CCBs. Thus, we named occurrence of at least one of these diseases as "cardiometabolic diseases" and utilize it as a transient confounder for the second set of semi-synthetic data experiments. We set low, medium, and high confounding intensity for experiments with sex and cardiometabolic disease as confounder (β values: [1, 5, 10] and [25, 50, 75], respectively) totaling six experiments on semi-synthetic data. In sum, with this confounding generation method, model confounding adjustment ability will be tested with two forms of confounding at various degrees of intensity (β values).

On the semi-synthetic dataset with highest intensity of cardiometabolic disease confounding, we additionally conducted finite-sample causal estimation experiments. Since estimation in limited sample settings is known to be unstable in many cases (e.g., for inverse probability weighted estimators) despite asymptotic guarantees [15], we wished to assess our model for finite-sample estimation ability. And, we specifically set the confounding to the highest intensity level ($\beta = 75$) because we wished to investigate how the model performs in estimation of RR in situations of high confounding. We investigated the finite-sample estimation ability of our proposed model and other deep learning models by applying the models on random subsamples of this dataset: 2.5%, 5%, 10%, 25%, 50%, and finally, the entire dataset.

D. Proposed Model Development

Our model, T-BEHRT, utilizes a modified feature BEHRT extractor to capture both static and temporal medical history variables and captures initial estimates of RR. BEHRT is a state-of-the-art transformer model for EHR data. By using contextualized embeddings to represent longitudinal clinical encounters (e.g., diagnoses/prescriptions) and time of medical visit—both relative in terms of age/visit number and absolute in terms of calendar year—and multihead self-attention for feature extraction, BEHRT has demonstrated the state-of-the-art performances in various EHR-based tasks [5], [16]. After predicting propensity score and conditional outcomes, we use CV-TMLE to correct for bias in initial RR estimate and compute corrected RR (see Fig. 2).

Intuitively, T-BEHRT first extracts latent EHR features from static covariates and fixed subsequences of medical history with BEHRT. Second, the model predicts propensity of exposure and conditional outcome using these learned features. Third, by additionally conducting auxiliary unsupervised learning, the model trains on reconstruction of both static and temporal data with two-part masked EHR modeling (MEM).

The propensity prediction model is modeled as 1-hidden layer multilayer perceptron (MLP) and for each conditional outcome, we use a 2-hidden layer MLP with exponential linear unit (ELU) activation.

With patient data tuple (X_i, Y_i, T_i) as described in Appendix Section A, parameters θ, propensity prediction head $g(X_i)$, and conditional outcome prediction heads, $H(X_i, T_i)$ for input

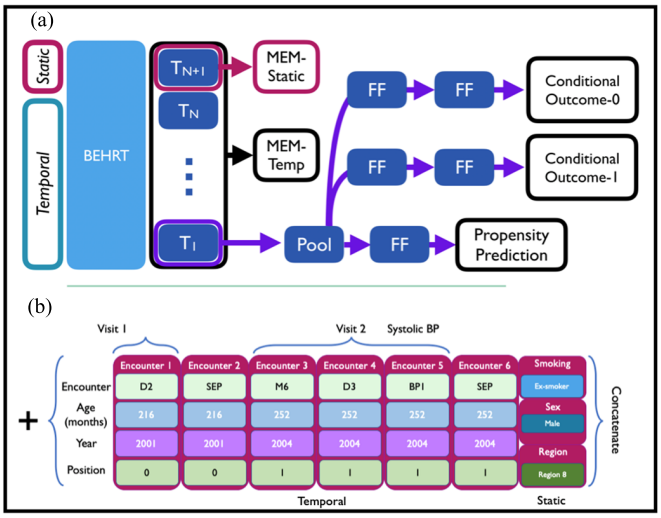

Fig. 2. Targeted BEHRT and embedding structure. (a) Above, the model is shown. Generally, an input x (static and temporal variables) is fed to a feature extractor, which outputs a dense latent state (for EHR modeling, this feature extractor is BEHRT). The output of the final layer of the BEHRT feature extractor is fed to the MEM prediction head to predict any masked encounters. T_{N+1} token state is fed to a variational autoencoder (VAE) neural network to predict masked static variables. The latent state of the first token (T1) is fed to a pooling layer to predict propensity and conditional outcomes with multiple prediction heads with feed forward (FF) neural network layers. The loss consists of the unsupervised loss from two MEM components—temporal (temp) and static (static) unsupervised data training—and the supervised loss of the propensity and factual outcomes. (b) Below, the embedding structure for modeling rich EHR data is shown. Clinical encounters timestamped by age/year/position (visit number) are converted into vector representations and fed to model as temporal variables. Static data variable embeddings: patient sex, region in U.K., and smoking status are concatenated to the temporal variable embeddings.

X_i and exposure T_i for patient i, the loss is

$$\hat{O}(X_i;\theta) = \text{CrossEntropy}(H(X_i,T_i;\theta),Y_i) + \text{CrossEntropy}(g(X_i;\theta),T_i). \quad (3)$$

Next, we conduct MEM for two-part unsupervised learning: 1) temporal variable and 2) static variable modeling. The first part—unsupervised learning on temporal data—functions similar to masked language modeling (MLM) in natural language processing [46]. In MLM, the model receives a combination of masked, replaced, and unperturbed tokens (temporal or textual data) and the task is to predict the masked or replaced encounters. We do the same but additionally enforce another constraint: when replacing encounters, we do not replace encounters with those that define the exposure or outcome—antihypertensives and cancer in the current set of experiments. With encounter j for patient i represented as $E_{i,j} \subset X_i$ (i.e., encounters being a subset of the input X_i), masked/replaced encounters represented as $\tilde{E}_{i,j}$, BEHRT feature extractor B, temporal unsupervised prediction network M, neural network parameters $\phi_{\text{MEM-Temp}}$, we develop objective function

$$\widehat{\mathcal{L}_{\text{MEM-Temp}}}\left(E_{i,j};\phi_{\text{MEM-Temp}}\right) = \sum_{j=1}^{|E_i|} \text{CrossEntropy}\left(M\left(B\left(\tilde{E}_{i,j};\phi_{\text{MEM-Temp}}\right)\right),E_{i,j}\right). \quad (4)$$

For the second part of the MEM, static data modeling, we chose using VAE for unsupervised learning due to cumulative literature empirically demonstrating its strength in representation learning in addition to the utilization of VAE structures in other causal deep learning models such as CEVAE [35]. We model static categorical variables: region, smoking status at baseline, and sex; the three variables are embedded in high-dimensional embeddings (embedding dimensions for each variable are hyperparameters of the T-BEHRT model) and mapped (via 1-layer MLP) to the size of the encounter (temporal) embeddings and, finally, concatenated to the encounter embeddings. Thus, the BEHRT model functions as feature extractor for static/temporal variables and encoder for the VAE (see Fig. 2). The temporal variables interact with the static variables through the multihead self-attention mechanism of the BEHRT architecture [31]. For training the VAE, similar to the temporal modeling, we mask some variables as input and use a variable-specific decoder to decode the variable (if masked). Specifically, for static variable $X_{i,v}$ of a total of V static variables patient i, $q_{\phi_{\text{Enc}}}(Z_i \mid X_i)$ representing the encoder, and $p_{\phi_{\text{Dec}}}(X_{i,v} \mid Z_i)$ representing the multivariate Bernoulli decoder for variable v, and the VAE loss is

$$\hat{\mathcal{L}_{\text{MEM-Static}}}(x_i;\phi_{\text{Enc}},\phi_{\text{Dec}}) = \sum_{v=1}^{V}\sum_{i=1}^{n} \log p_{\phi_{\text{Dec}}}\left(X_{i,v} \mid Z_i\right) - \sum_{i=1}^{n} D_{KL}\left(q_{\phi_{\text{Enc}}}(Z_i \mid X_i)||p_{\phi_{\text{Dec}}}(Z_i)\right). \quad (5)$$

The complete objective function to be minimized is the summation of (3), (4), and (5) as shown in the following equation:

$$\hat{\theta},\hat{\varepsilon},\hat{\phi}_{\text{Enc}},\hat{\phi}_{\text{Dec}},\hat{\phi}_{\text{MEM-Temp}} = \underset{\theta,\varepsilon,\phi_{\text{Enc}},\phi_{\text{Dec}},\phi_U}{\text{argmin}} \sum_{i=1}^{n} \hat{O}(X_i;\theta) + \delta\Big(\widehat{\mathcal{L}_{\text{MEM-Temp}}}\left(E_{i,j};\phi_{\text{MEM-Temp}}\right) + \widehat{\mathcal{L}_{\text{MEM-Static}}}(X_i;\phi_{\text{Enc}},\phi_{\text{Dec}})\Big). \quad (6)$$

With hyperparameter δ for weighting the contribution of the unsupervised MEM loss terms.

E. Feature Selection and Preprocessing

The modalities of CPRD considered for deep learning modeling were sex, region, diagnoses from both primary and secondary care, medications, systolic blood pressure (BP) measurements, and smoking status.

We mapped read codes from primary care and ICD-10 codes from secondary care to 1471 unique ICD-10 diagnostic codes [47], [48] to harmonize disease codes in the dataset; unmapped codes were included for completion. Furthermore, we mapped medication codes to 426 codes in the British National Formulary (BNF) [49] coding format. Since systolic BP is a continuous variable and our feature extractor requires discretized elements (see BEHRT feature extraction

in Appendix Section D), systolic BP measurements (in mm Hg) were grouped into 16 categories based on prespecified boundaries ([90–116], (116,121], (121,126], ..., (181,186], >186). Furthermore, we utilized calendar year, age (months), and relative position (visit number) for the sequential/temporal modalities. Each patient p had n_p encounters, or instances of modalities: diagnoses, medications, and systolic BP measurements. Smoking status at baseline (non, previous, or current smoker), region (ten regions in U.K.), and sex (male, female) were static variables included in modeling.

F. Benchmarks and Causal Estimation

Before pursuing the causal investigations with deep learning modeling, we pretrained contextualized EHR embeddings and network weights through MEM on the pretraining dataset. This MEM task generally trains weights on all patients in CPRD before progressing to causal modeling (6 777 845 patients in Fig. 1).

For semi-synthetic investigations, we implemented statistical and deep learning models to serve as benchmarked comparison models for causal inference. The benchmarks include Bayesian additive regression trees (BART) [50], LR and L1/L2 regularization variants, and LR with TMLE [51]. We chose the covariates for these models to be baseline age, smoking status, sex, region, incidence of 33 curated disease groups, and additionally prescription of four additional medications groups. While inclusion of baseline variables in epidemiological observational studies is standard practice, we specifically include the disease/medication groups to enable a fairer comparison to deep learning modeling. Furthermore, diagnoses and medications are known to be confounders in observational studies, so adjustment of these variables is important for causal estimation. To ensure that the diagnoses and medication groups are medically valid clusters of diseases and medications, respectively, we utilized groups compiled by past medical research [42], [43]. A deeper explication is given in Appendix Section C.

To serve as deep learning benchmarks, we implemented staple deep learning models for average causal effect: TARNET, TARNET + MEM (i.e., with unsupervised MEM component), and Dragonnet with BEHRT feature extractor and the embedding format presented in Fig. 2(a). We initialized these models with pretrained weights. After implementing and evaluating benchmarks, we implemented T-BEHRT with pretrained network weights where applicable and pursue modeling of semi-synthetic data investigations.

For the semi-synthetic data experiments, we did not feed variables denoting cardiometabolic disease and sex, respectively, as input; we wish the statistical and deep learning models to infer confounding from remaining input variables. In routine clinical data, the observational studies would often not have access to all confounding variables—thus, important to test models' ability to adjust for confounding given limited input variables.

For all investigations, we conducted experiments with five-fold cross validation causal estimation. We calculated RR on the test dataset for each fold as advised by Chernozhukov *et al.* [23] and compute 95% CIs over the five folds. We computed RR defined by naïve estimator on a finite sample: $\hat{\psi} = \mathbb{E}[(\mathbb{E}[H(X,1)]/\mathbb{E}[H(X,0)])]$ for TARNET, TARNET-MEM, LR (and L1/L2 regularization variants), and BART. For T-BEHRT, we use the CV-TMLE method for the estimation of RR. For Dragonnet, we implement the model with the CV-TMLE estimator in order to directly compare our model with this benchmark model. In addition, we also implement the Dragonnet model with the naïve estimator (i.e., the original model without post-hoc estimator). For more information on the CV-TMLE method, advantages over TMLE, and implementation, please refer to Appendix Section B. For models that utilized predicted propensity scores, we conducted propensity score trimming and exclude patients with predicted propensity score greater than 0.97 and less than 0.03 [52] before pursuing RR calculation.

We identified the superior model by identifying the model with least sum absolute error (SAE) over the three β values for each confounding experiment. We give the standard error (SE) for the SAE; this was calculated using additive propagation of error [53]. For deep learning models, we also demonstrate change of SAE as modules are removed from our proposed model.

G. Implementation

We developed all statistical and deep learning models on python. The deep learning models were implemented with Pytorch [54]. Hyperparameters for the BEHRT feature extractor are found in Table III. For training all deep learning models, we used the Adam optimizer [55] with exponential decay scheduler (decay rate = 0.95) to ensure training convergence. For TARNET-MEM and T-BEHRT, we pretrained five epochs on exclusively the MEM task before initiating joint MEM-causal task training.

After fitting deep learning and statistical models, in order to derive estimates for RR estimation, we conducted the evaluation of the model on the test fold of the dataset using standard g-computation methods [1]. For all patients in the test set, we first derived risk estimates [e.g., estimation of $P(Y \mid X, T = 0)$] patients as if they were all assigned $T = 0$, and similarly, derived estimates [e.g., estimation of $P(Y \mid X, T = 1)$] as if they were all assigned $T = 1$. In this way, the RR estimate, $\hat{\psi}$, can be derived as a function of these two quantities

$$\hat{\psi} = \mathbb{E}\left[\frac{\mathbb{E}[H(X, T = 1)]}{\mathbb{E}[H(X, T = 0)]}\right]. \tag{7}$$

LR (and regularization variants), BART, TMLE, and CV-TMLE were implemented in python. The code was inspired by past works utilizing TMLE [3]. To fit the nuisance parameter for the TMLE estimate update step, Nelder-mead optimization was utilized [6], [56]. For deep learning models implemented with CV-TMLE, the naïve estimator (7) was not used; rather, the CV-TMLE estimator was implemented utilizing conditional outcome predictions, $H(X, T = 1)$, $H(X, T = 0)$, and propensity score prediction, $g(X)$.

TABLE I
POPULATION STATISTICS

	Classes of antihypertensives				
	ACEIs	**BBs**	**CCBs**	**Diuretics**	**ARBs**
Number (%)	186709 (36)	150098 (29)	128597 (24)	28991 (5)	21970 (4)
Male (%)	101629 (54)	67794 (45)	60395 (46)	8134 (28)	10454 (47)
Smoker (current/ex) (%)	101629 (54)	67794 (45)	60395 (46)	8134 (28)	10454 (47)
YOB (SD)	1938 (15)	1941 (15)	1936 (14)	1934 (16)	1940 (14)
Baseline Age (SD)	63 (14)	59 (14)	64 (13)	63 (15)	63 (13)
Number of visits (SD)	7 (4)	6 (4)	6 (4)	4 (4)	7 (4)
Baseline Year (SD)	2001 (4.2)	1999 (4.3)	2000 (4.9)	1996 (5.2)	2002 (2.8)

YOB: year of birth; baseline: the time of exposure assignment; SD: standard deviation; %: percentage

III. RESULTS

A. Population Statistics

In the dataset for the investigation of antihypertensives on incident cancer, we identified 186 709, 150 098, 128 597, 28 991, and 21 970 patients for ACEIs, BBs, CCBs, diuretics, and ARBs, respectively, totaling 516 365 patients. We demonstrate population statistics in Table I. Cancer incidence counts/percentage of exposure group were 13 728/7%, 9819/7%, 10 232/8%, 1784/6%, and 1709/8% for ACEIs, BBs, CCBs, diuretics, and ARBs, respectively.

B. Semi-Synthetic Data Experiments

In the semi-synthetic experiments on confounders' cardiometabolic diseases and sex, we tested the T-BEHRT models against several statistical and deep learning benchmarks. In Fig. 3(a) and (b), we show SAE with SE measures calculated over all β-specific semi-synthetic data experiments. We include more detailed experimental results in Table IV.

We found that our proposed model, T-BEHRT, outperforms all given deep learning and statistical model solutions in terms of SAE whilst maintaining narrow SE. In addition, across both experiments, we found that deep learning models for EHR benefit from inclusion of CV-TMLE. This is seen by superior performance of both Dragonnet + CV-TMLE and T-BEHRT in comparison with TARNET, which does not handle propensity score modeling. However, by investigating the exclusion of various modules from the chassis of T-BEHRT shown in our ablation analysis [see Fig. 3(c)], we see that exclusion of MEM diminished RR estimation accuracy in a parallel way; the TARNET model with inclusion of MEM (SAE increase of 0.213) did approximately as well as Dragonnet + CV-TMLE (SAE increase of 0.305) averaged over experiments of persistent and transient confounding. Removal of CV-TMLE from Dragonnet + CV-TMLE further deteriorated the performance of the Dragonnet model (SAE increase of 0.077). Ultimately, the improvement in combining both MEM and propensity/CV-TMLE modeling and forming T-BEHRT demonstrated a greatest SAE reduction of 0.676—more so than the sum of its parts: 0.518 (0.231 + 0.305).

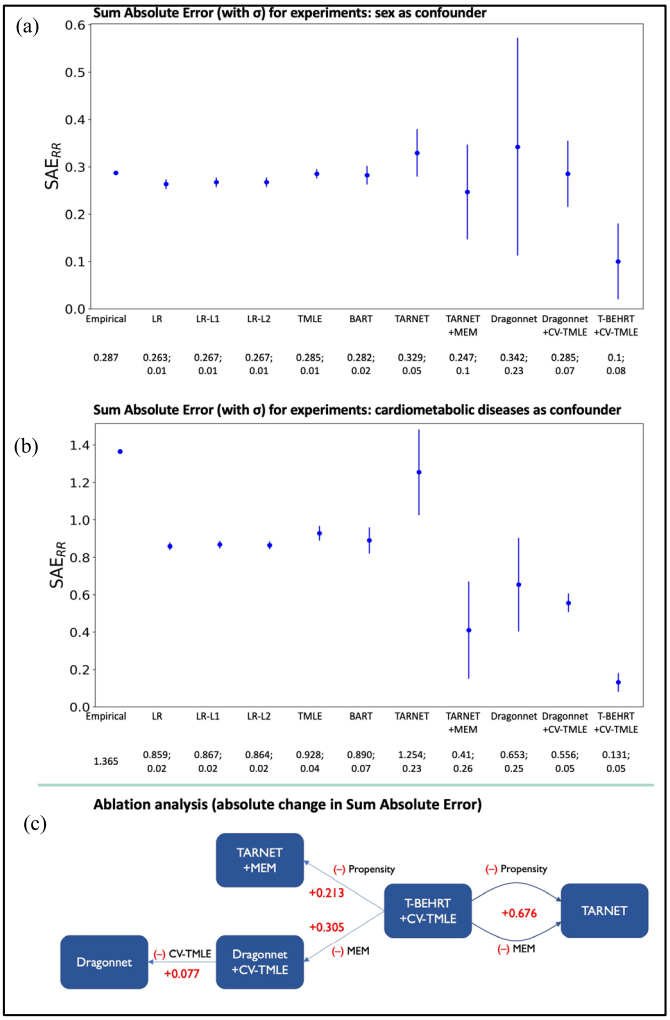

Fig. 3. Experiments on semi-synthetic data with (a) sex and (b) cardiometabolic disease as confounders; (c) module inclusion analysis of causal modules. We show sum absolute error (SAE) between ground truth risk ratio (RR) and estimated RR with standard error measures in both panels. The x-axis is shown by the models implemented on these datasets, and the y-axis is the SAE (lower is better). We present the numerical value and standard error measures underneath the model names. In (c), we present the transformation from T-BEHRT into other deep learning benchmarks. We show increase in average SAE (i.e., increase in error) across experiments of transient and persistent confounding in red as our model strips away components from its architecture indicated by (–).

In the finite-sample estimation experiments shown in Fig. 4, we showed that T-BEHRT outperforms other models in RR estimation in individual and across data subsamples. While improvement of T-BEHRT over Dragonnet + CV-TMLE is less pronounced than over other models, panel B shows that T-BEHRT still demonstrates superior RR estimation performance with respect to the deep learning benchmarks. Furthermore, we found that inclusion of MEM aids more precise estimation of RR; TARNET-MEM and T-BEHRT perform better than TARNET over all finite samples as shown in Fig. 4(b). However, we note that the application of CV-TMLE is more important than MEM in smaller datasets as seen by superior performance of Dragonnet + CV-TMLE as opposed to TARNET + MEM in Fig. 4(b). Furthermore, models equipped with propensity modeling (and CV-TMLE specifically) maintain relatively stable SAE across subsampling

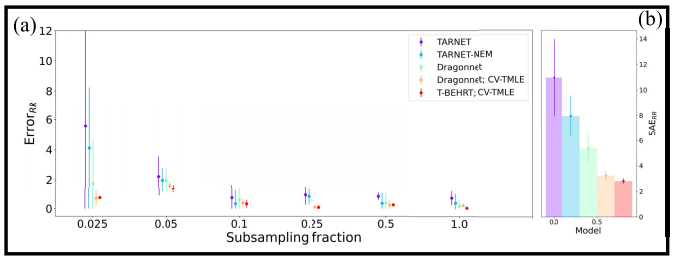

Fig. 4. Finite-sample experiments on semi-synthetic data. (a) We conduct experiments on finite subsamples of the semi-synthetic dataset for cardiometabolic confounding ($\beta = 75$). The subsampling fraction of the dataset is shown on the x-axis. The y-axis shows error from ground truth risk ratio (RR). The models: TARNET (and with MEM), Dragonnet (and with CV-TMLE), and T-BEHRT estimate RR on the fractional samples of the dataset. The point estimate is the mean value on five-fold cross validation and the error bars represent 95% confidence intervals for those point estimates of RR. (b) Sum absolute error (SAE) across the seven subsamples of the dataset is shown for each model (denoted by color). The four models are represented by the four bars with interval defined by standard error (SE) and color scheme is the same as part (a).

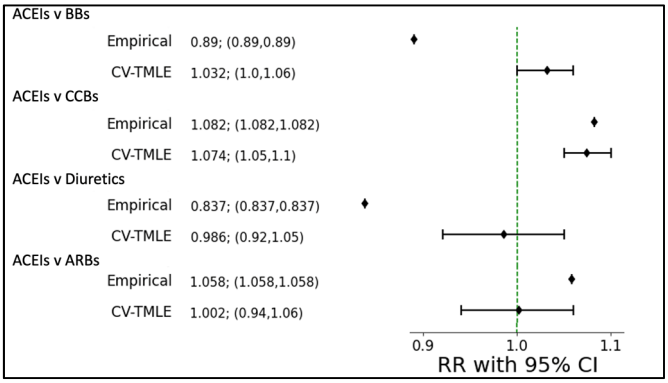

Fig. 5. Application of T-BEHRT on routine clinical data. Effect of ACEI on incident cancer with respect to BBs, CCBs, diuretics, and ARBs. This forest plot has four parts; one for each comparison to other antihypertensive drug classes. We demonstrate CV-TMLE risk ratio (RR) estimates with 95% confidence intervals (CIs) on our T-BEHRT model. In addition, we show empirical RR in the observational cohort selected for these experiments. The ground truth is assumed to be 1.0 (null) for all four associations validated by meta-analysis of RCTs. BBs: beta blockers; CCBs: calcium channel blockers; ACEIs: angiotensin-converting-enzyme inhibitors; ARBs: angiotensin receptor blockers; RR: risk ratio.

fractions, while TARNET and derivatives suffer in RR estimation in smaller datasets. Lastly, across experiments in this work, while Fig. 3 demonstrates that MEM is more important in observational settings with more samples (full dataset), Fig. 4 shows that CV-TMLE provides greater utility in observational settings with limited samples. Implemented simultaneously (i.e., the T-BEHRT model), both components ensure robust estimates across various sample sizes.

As a trend, we saw as dataset size increases, SAE across models began to converge in Fig. 4. Theoretically, as the number of samples increases, we would be slowly mitigating the finite-sample bias, and thus, the performance of TARNET and derivatives should be similar to those of models assisted by propensity modeling also noted by Shi *et al.* [3].

We applied our model on the routine clinical data study of effect of ACEIs on incident cancer with respect to other antihypertensive drug classes and show the results in Fig. 5. Across all four drug class comparisons, while the empirical RR often tended away from null implying a preventive or harmful effect, we showed that our model's 95% CI for RR covered the null hypothesis (1.0 RR) across almost all drug class comparisons with exception of CCBs.

IV. Discussion

In this article, by utilizing large-scale comprehensive EHR and deep learning methods, we have developed a model for observational causal inference. We have validated our model against benchmarks across six semi-synthetic and a finite-sample estimation experiment and found T-BEHRT to demonstrate more accurate RR estimation. Finally, we applied our model to a routine clinical data observational study.

Our work has contributions to the field of EHR-based deep learning research. First, the T-BEHRT model consolidates multiple static and temporal data embeddings into a unified embedding structure, thus allowing adjustment over multiple EHR datatypes. Second, our model conducts novel MEM unsupervised learning using MLM and VAE-based representation learning in tandem with the causal inference objective. We demonstrated the benefits of unsupervised learning in the context of average RR estimation in multiple experiments as well. In our assessment, this is the first work conducting causal inference incorporating unsupervised learning on multiple EHR data types. Third, we utilized CV-TMLE estimation correction for less biased RR estimation for deep learning causal models on EHR data. While the utility of propensity modeling and CV-TMLE is as effective as MEM modeling for RR estimation in larger dataset sizes, we found that in our finite-sample estimation experiments, CV-TMLE is critical for accurate RR estimation. Finally, we show that our model can be easily applied to test a clinical hypothesis regarding treatment effect in an observational setting.

Our work has some limitations and scope for future development. First and most fundamentally, we note that even comprehensive EHR might not completely capture confounding variables and, hence, limit our model to provide a fully unbiased result. A variety of variables affecting outcome may be unadjusted (explicitly or through latent representation modeling) and further modality inclusion is necessary in future work to help mitigate residual confounding. Latent confounding adjustment can be parallelly investigated in future works with latent variable modeling techniques [4] to enrich EHR. Furthermore, we note that we have included the data modalities of diagnoses, medications, smoking, sex, and systolic BP; however, better confounding adjustment might manifest with fuller utilization of the modalities that rich databases like CPRD entail. In addition, in terms of data curation, we have allocated patients into an exposure group based on first prescription of class of antihypertensives. Subgroup investigations stratified by intensity and duration of drug class should be additionally pursued in future studies. In terms of applying our model to a case study, T-BEHRT estimated null in most drug comparisons in the routine clinical data study, but we note that our model finds the comparison to CCBs to deviate from the null (although quite close with $<$1.1 RR). While findings from the RCTs generally demonstrate that antihypertensives

have null effect on cancer, the evidence regarding CCBs is still conflicting and further research is required [14].

In contrast, it must be noted that over-adjustment may also result in biased estimation. Although found to be an uncommon manifestation, M-structure bias variables, a special case of collider variables, might be a source of bias if included in adjustment [57]; although in general, empirical research has shown conditioning on all pretreatment variables is still the optimal course of action [58]. However, more research must be conducted on the effect of these variables specifically in the context of propensity-score modeling and the administrative EHR observational setting.

V. Conclusion

To conclude, we have developed a deep learning model for EHR data for more accurate estimation of RR. T-BEHRT has performed optimally in semi-synthetic data experiments with both persistent and transient confounding and can be applied to an observational study on routine clinical data. Thus, in the future, this model should be further tested and applied to investigate other causal hypotheses questions using routine EHR.

Appendix

A. Supplementary Methods: Semi-Synthetic Data Simulation

Data generation of sequential, temporal variables is a difficult task, and currently, there is no medically validated method of generating realistic EHR medical history, exposure assignment, and outcome. Thus, in gist, we utilize routine clinical data components: 1) medical history with balanced exposure groups and 2) known exposure status data generation of factual/counterfactual outcomes. With generated potential outcomes, we can calculate ground truth RR and compare the deviation of model estimated RR from ground truth.

In order to create this semi-synthetic dataset, we first form the dataset for the investigation: effect of antihypertensives on incident cancer allowing us to access the components (1) and (2). Since confounding often manifests partly due to imbalanced variables between exposure groups, we find an imbalanced variable, Z_i in (1), routine medical history. We then force this imbalanced variable, Z_i, to be a confounder and generate conditional outcome from a sampling function

$$Y_i^{T_i} = \text{Bernoulli}(\sigma(aT_i + m\beta(\lambda_i + c))).$$

λ_i represents $P(T_i \mid Z_i)$, $T_i \in \{0, 1\}$ is the exposure for patient i, $Y_i^{T_i}$ is the outcome for patient i given exposure T_i, σ is the sigmoid function, and β is the intensity of confounding. Variables a, m, and c are coefficients terms to weighting their importance in the function.

Intuitively, we first model the association between a variable Z_i and exposure ($P(T_i \mid Z_i)$) with λ_i. Next, we generate $Y_i^{T_i}$ with two variables: the variable Z_i and T_i. In this way, we form an association between Z_i and exposure and Z_i and the outcome; with association to both exposure and outcome, Z_i becomes a confounder in this data generating process. This process synthesizes controlled confounded observational data; by generating the outcome with this function, we control confounding with a confounder Z_i. Thus, in this way,

TABLE II
ICD-10th Revision Codes for Cancer

Other cancer: C79, D07, D38, D39, C71, D37, C73, C72, C75, C74, C76, C00, C39, C38, C34, D05, C26, C37, C14, C97, C69, R18, C70, C44, C45, C46, C47, C40, C41, C43, D32, C48, C49, D09, D33, D49, D48, D18, D43, D42, D41, D40, D47, D44, C80, B21, E26, C78
Leukaemia: C93, C92, C91, C90, C95, C86, C94, C88, D47, D46, D45
Urinary cancers: C63.9, C67.6, C68.5, C66.4, C66.2, C66.8, C66.0, C68.6, D09.1, C66.5, C68.7, C66.1, C68.1, C66.3, C68.4, C68.3, C66.9, C66.7, C66.6, C68.2, C68.8, C66.X, C68.0, C80.1, C68.9, C67.7
Bladder cancer: C67.9, C67.2, C67.0, C67.1, C67.5, C67.3, C67.8, C67.6, C67.4, D09.0, C67.7
Renal cancer: C65.2, C64.7, C64.X, C65.7, C65.9, C65.4, C65.0, C65.8, C65.X, C64.6, C65.1, C64.3, C64.0, C64.4, C65.5, C65.3, C64.8, C64.5, C64.2, C64.1, C64.9, C65.6
Male reproductive cancers: C63.3, C60.4, C63.2, C60.1, C62.3, C63.9, C63.7, C63.1, C60.2, C62.1, C62.8, C62.0, C63.5, C62.2, C62.6, C62.7, D07.6, C63.8, C63.4, C60.3, C60.6, C60.8, C60.0, C60.5, D07.4, C60.7, C60.9, C62.4, C63.0, C62.9, C63.6, C62.5
Prostate cancer: C61.9, C61.6, C61.2, C61.5, C61.3, C61.7, C61.4, D07.5, C61.8, C61.1, C61.0, C61.X
Female reproductive cancers: C57, D07, C55, C54, C53, C52, C51, D25, C58, D49
Ovarian cancer: C56.1, C57.8, C56.8, C56.9, C56.X, C56.5, C56.3, C56.4, C56.6, C56.7, D63.0, C56.2, C56.0
Cervical cancer: D06.4, D06.2, D06.9, D06.1, C53.8, C53.9, C53.2, D06.7, C53.5, D06.8, C53.1, D06.0, D06.3, C53.6, C53.4, D06.6, C53.3, D06.5, C53.7, C53.0
Breast cancer: D05.6, C50.9, C50.5, D05.7, D05.4, D05.3, D05.2, D05.8, C50.1, C50.4, D05.5, C50.2, D05.1, D05.0, C50.8, C50.3, D05.9, C50.7, C50.0, C43.5, C50.6
Skin cancer: D03.7, D04.4, C44.6, D22.1, D04.3, D22.7, C79.2, C44.0, D04.0, C44.9, C44.5, D04.7, D03.0, D22.5, D03.8, C44.4, D04.8, D22.4, D03.5, C44.3, D22.3, D22.9, D22.6, D49.2, C43.8, C60.0, C44.8, D22.2, D03.2, C44.1, D04.2, D03.4, D04.9, D04.1, D03.6, C46.0, D22.8, D22.0, C44.2, C43.9, C44.7, D03.1, D04.6, D03.3, D03.9, D04.5
Gastrointestinal cancers: C22, C23, C21, C26, D01, C24, D49, C18, C17
Respiratory cancers: C33.7, D02.6, C34.8, C34.0, C76.1, D02.0, D02.2, D02.1, C38.4, D02.8, C45.0, C39.0, C33.8, C33.2, C34.9, D02.9, D02.3, C45.9, C33.9, C33.4, C34.3, D49.1, C33.1, C33.5, C33.6, C33.0, C34.1, C39.8, C34.2, C39.9, C33.X, D02.5, C45.7, C33.3, D02.4, D02.7
Unspecified cancer: D01.2, Z92.8, C46.9, D01.8, D01.9, D09.9, C97.X, Z08.8, D01.0, D49.2, C80.9, D49.9, B20.0, D01.5, Z08.9, D01.1, D01.4, D01.7, D01.6, D01.3, K63.5, D49.5, C80.1, D49.7
Lung cancer: C34.3, C34.4, C34.0, C34.8, C34.9, C34.7, C34.5, C34.1, D02.2, C34.2, C34.6
Head and neck cancers: C79, C39, C13, C12, C11, C10, C31, C30, C32, D00, C14, C46, D02, C43, C49, C08, C09, C00, C01, C02, C03, C04, C05, C06, C07
Pancreatic cancer: C25.3, D49.0, C25.5, C25.4, C25.1, C25.6, C25.9, C25.7, C25.8, C25.0, C25.2
Liver cancer: C22.3, C22.9, C22.6, D01.5, C22.7, C22.4, C22.2, C22.1, C22.5, C22.8, C22.0
Stomach cancer: C16.3, D00.2, C16.4, C16.8, C16.5, C16.7, C16.2, C16.1, C16.0, C16.9, C16.6
Oesophageal cancer: C15.4, C15.7, C15.2, C15.6, C15.8, C15.9, D00.1, C15.0, C15.1, C15.5, C15.3
Rectal cancer: C20.4, C20.8, D01.2, C19.8, C19.9, C20.2, C19.0, C21.8, C20.1, C20.5, C20.3, C19.6, C19.X, C19.5, C20.9, C20.X, D01.1, C19.4, C19.7, C20.6, C19.2, C20.0, C20.7, C19.3, C19.1
Colon cancer: C18.5, C18.7, C18.4, C18.3, C18.6, C18.1, C18.2, C18.0, D01.0, C18.8, C18.9
Lymphatic cancer: C81, C82, C90, C84, C85, C88, C91, B21, C83, D47, C96
Metastatic cancer: C79.3, C79.7, C77.1, G89.3, C79.2, M84.5, C77.8, C78.4, C77.9, C78.0, C78.7, C78.6, C79.1, C78.9, C79.0, C79.5, C78.1, C78.8, C77.5, C77.4, C79.9, C77.3, C79.6, C79.4, C77.7, C77.2, C79.8, C78.5, C43.9, C78.2, C77.6, C78.3, C77.0

The codes considered for incident cancer is shown in the table. The codes are stratified by type of cancer.

TABLE III
HYPERPARAMETERS FOR T-BEHRT

Hyperparameter	Attribute
Hidden BEHRT	150
Intermediate BEHRT Layer size	108
Region embedding size	7
Sex embedding size	1
Smoking status embedding size	2
Hidden dropout probability	0.3
Attention dropout probability	0.4
Number of hidden layers (BEHRT)	4
Hidden activation functions	GeLU
Initialiser range of parameters	0.02
n	200
d	0.1

we can generate factual/counterfactual outcomes and consequently ground truth RR. Lastly, we can modify β value to vary the degree of confounding in the data generation process.

B. Supplementary Methods: CV-TMLE

After using T-BEHRT to compute initial estimates, we use CV-TMLE [6] for the correction of these estimates. We refer readers to the source material for theory behind TMLE and the cross-validated form: CV-TMLE [5], [25]. In brief, the original formulation of the CV-TMLE algorithm requires k targeting steps for each of the k folds for each of the iterations predefined in the iterative version of TMLE. However, Levy forms a simpler construction of the CV-TMLE which is less computationally cumbersome; the advised method is to pool all the initial estimates across folds and compute corrected estimates vis-à-vis a standard TMLE update step [6]. Albeit procedurally different, the original formulation and Levy's more recent formulation of CV-TMLE are identical mathematically and in function. According to our research, this is the first work utilizing CV-TMLE paired with deep learning methods.

CV-TMLE provides a host of benefits to observational causal inference. As recommended by Chernozhukov *et al.* [23], with validation in k-fold framework, CV-TMLE is a form of TMLE which is robust to issues of fold-wise overfitting whilst conducting k-fold cross validation [6]. Furthermore, previous works show that the CV-TMLE estimator provides more robustness than other cross-validated estimators (e.g., CV-AIPTW) in the case of violations of the assumption of overlap [59].

C. Supplementary Methods: Statistical Model Development

In statistics models, we used RR for estimation of causal effect. The covariates adjusted for were: baseline age (continuous variable: [0,1]), sex (male/female), region, smoking status (smoker/nonsmoker), chronic kidney disease (yes/no), human immunodeficiency virus/acquired immune deficiency syndrome (yes/no), ischemic heart disease (yes/no), cardiac arrhythmia (yes/no), stroke (yes/no), heart failure (yes/no), anemia (yes/no), diabetes mellitus (yes/no), hypertension (yes/no), osteoporosis (yes/no), arthritis (yes/no), connective tissue disorder (yes/no), gout (yes/no), rheumatoid arthritis (yes/no), peptic ulcer disease (yes/no), liver disease (yes/no), asthma (yes/no), peripheral arterial disease (yes/no), chronic obstructive pulmonary disorder (yes/no),

TABLE IV
RISK RATIO ESTIMATES ACROSS SEMI-SYNTHETIC DATA EXPERIMENTS

		Risk ratio			Error
Cardio-metabolic disease	**Beta**	**25**	**50**	**75**	
	Modelling				
	Ground Truth	2.207	2.727	3.178	1.555
	Empirical	2.532	3.251	3.883	1.365
Statistical	LR	2.398; (2.37, 2.43)	3.003; (2.97, 3.03)	3.569; (3.5, 3.64)	0.859; 0.02
	LR-L1	2.399; (2.37, 2.43)	3.005; (2.97, 3.04)	3.576; (3.51, 3.64)	0.867; 0.02
	LR-L2	2.399; (2.37, 2.43)	3.004; (2.97, 3.03)	3.574; (3.5, 3.64)	0.864; 0.02
	TMLE	2.411; (2.28, 2.54)	3.005; (2.87, 3.14)	3.622; (3.4, 3.84)	0.928; 0.04
	BART	2.398; (2.37, 2.43)	3.011; (2.98, 3.04)	3.592; (3.53, 3.65)	0.890; 0.07
Deep learning	TARNET	2.283; (2.21, 2.35)	3.183; (2.76, 3.6)	3.899; (3.43, 4.36)	1.254; 0.23
	TARNET + MEM	2.226; (2.14, 2.31)	2.719; (2.35, 3.09)	3.561; (2.94, 4.19)	0.41; 0.26
	Dragonnet	2.308; (2.12,2.50)	3.098; (2.57,3.62)	3.12; (2.55,3.69)	0.529; 0.25
	Dragonnet +CV-TMLE	2.281; (2.26, 2.31)	2.954; (2.91, 3.0)	2.922; (2.85, 2.99)	0.556; 0.05
	T-BEHRT+ CV-TMLE	2.263; (2.24, 2.29)	2.753; (2.71, 2.8)	3.227; (3.14, 3.31)	0.131; 0.05

		Risk Ratio			Error
Sex	**Beta**	**1**	**5**	**10**	
	Modelling				
	Ground Truth	1.465	1.926	2.154	
	Empirical	1.456	1.823	1.979	0.287
Statistical	LR	1.455; (1.45, 1.46)	1.83; (1.81, 1.85)	1.996; (1.95, 2.04)	0.263; 0.01
	LR-L1	1.455; (1.45, 1.46)	1.83; (1.81, 1.85)	1.992; (1.95, 2.04)	0.267; 0.01
	LR-L2	1.455; (1.45, 1.46)	1.829; (1.81, 1.85)	1.993; (1.95, 2.04)	0.267; 0.01
	TMLE	1.453; (1.44, 1.47)	1.824; (1.79, 1.86)	1.982; (1.93, 2.03)	0.285; 0.01
	BART	1.455; (1.45, 1.46)	1.826; (1.8, 1.85)	1.981; (1.94, 2.02)	0.282; 0.02

TABLE IV
(Continued.) RISK RATIO ESTIMATES ACROSS SEMI-SYNTHETIC DATA EXPERIMENTS

	TARNET+ MEM	1.457; (1.44, 1.47)	1.863; (1.77, 1.96)	1.977; (1.72, 2 .24)	0.247; 0.1
	TARNET	1.465; (1.44, 1.49)	1.803; (1.71, 1.89)	1.948; (1.83, 2 .02)	0.329; 0.05
Deep learning	Dragonnet	1.479; (1.45,1 .49)	1.969; (1.84,2 .10)	2.439; (1.82,3. 06)	0.342; 0.23
	Dragonnet +CV-TMLE	1.469; (1.45, 1.49)	1.827; (1.78, 1.87)	1.973; (1.85, 2 .09)	0.285; 0.07
	T-BEHRT+ CV-TMLE	**1.47; (1.45, 1.49)**	**1.854; (1.81, 1.9)**	**2.132; (1.98, 2 .29)**	**0.1; 0.08**

This table shows the risk ratio and standard deviation (five-fold) for statistical and deep learning models over the two semi synthetic experiments with cardiometabolic diseases and sex as confounders (top and bottom respectively). Over various values of Beta, confounding experiments are conducted. Ground truth risk ratio is calculated and displayed for both experiments. Risk ratio and 95% confidence interval for each model is presented in the table. The sum absolute error from ground truth risk ratios for models over all the confounding experiments and standard error is shown in the far-right column. Bolded models are best statistical and deep learning models. LR: Logistic Regression; LR-L1; Logistic Regression with L1 penalty; LR-L2; Logistic Regression with L2 penalty; TMLE: Targeted Maximum Likelihood Estimation; BART: Bayesian Additive Regression Trees; T-BEHRT: Targeted BEHRT

hemiplegia (yes/no), epilepsy (yes/no), dementia (yes/no), learning disorder (yes/no), eating disorder (yes/no), adjustment (yes/no), anxiety (yes/no), affective disorder (yes/no), depression (yes/no), bipolar disorder (yes/no), psychoses (yes/no), schizophrenia (yes/no), hyperlipidemia (yes/no), obesity (yes/no), substance abuse (yes/no), anticholinergics (yes/no), drugs that cause gastrointestinal bleedings (yes/no), statins (yes/no), and drugs for diabetes (yes/no). The exposure variable was antihypertensive medications (class 1/class 2). The outcome was defined as the synthetic outcome/cancer (yes/no). The models were fit and tested using fivefold validation. The naïve RR estimates were calculated on the testing dataset in each fold and mean RR estimate and 95% CIs for estimates were derived.

The TMLE was developed using two LR models—one for outcome prediction and the other for exposure prediction. The outcome prediction model adjusted for covariates and exposure variable listed above, and the exposure prediction model used just the covariates. The TMLE algorithm was fit and tested using five-fold validation. The TMLE RR estimates were calculated on the testing dataset in each fold and mean RR estimate and 95% (CI) for estimates were derived.

D. Supplementary Methods: Codes for Cancer

Provided in Table II, we have ICD-10th revision codes for cancer stratified by type.

E. Supplementary Methods: Hyperparameters for Deep Learning Models

Provided in Table III, we have hyperparameters used for the BEHRT architecture models.

F. Supplementary Results: Results for Semi-Synthetic Data Experiments

Provided in Table IV, we have the raw RR estimates across the semi-synthetic data experiments.

ACKNOWLEDGMENT

The authors would like to thank Prof. Mark van der Laan and Prof. Victor Veitch for constructive comments and discussion during the formative portions of this work.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views expressed are those of the authors and not necessarily those of the OMS, the BHF, the GCRF, the NIHR, or the Department of Health and Social Care.

REFERENCES

[1] M. Hernan and J. Robins, *Causal Inference: What If*. Boca Raton, FL, USA: CRC Press, 2020.
[2] E. Copland *et al.*, "Antihypertensive treatment and risk of cancer: An individual participant data meta-analysis," *Lancet Oncol.*, vol. 22, no. 4, pp. 558–570, Apr. 2021, doi: 10.1016/S1470-2045(21)00033-4.
[3] C. Shi, D. M. Blei, and V. Veitch, "Adapting neural networks for the estimation of treatment effects," 2019, *arXiv:1906.02120*.
[4] L. Zhang *et al.*, "The medical deconfounder: Assessing treatment effects with electronic health records," in *Proc. Mach. Learn. Healthcare Conf.* PMLR, 2019.
[5] A. M. Alaa and M. van der Schaar, "Bayesian inference of individualized treatment effects using multi-task Gaussian processes," Mar. 2017, *arXiv:1704.02801*.
[6] J. Levy, "An easy implementation of CV-TMLE," 2018, *arXiv:1811.04573*.
[7] M. J. van der Laan and D. Rubin, "Targeted maximum likelihood learning," *Int. J. Biostatistics*, vol. 2, no. 1, Jan. 2006, Art. no. 11, doi: 10.2202/1557-4679.1043.
[8] N. Gavrielov-Yusim and M. Friger, "Use of administrative medical databases in population-based researc," *J. Epidemiol. Community Health*, vol. 68, no. 3, pp. 283–287, Mar. 2014, doi: 10.1136/jech-2013-202744.
[9] E. Herrett *et al.*, "Data resource profile: Clinical practice research datalink (CPRD)," *Int. J. Epidemiol.*, vol. 44, no. 3, pp. 827–836, Jun. 2015, doi: 10.1093/ije/dyv098.
[10] L. Le, A. Patterson, and M. White, "Supervised autoencoders: Improving generalization performance with unsupervised regularizers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018.
[11] Y. Li *et al.*, "BEHRT: Transformer for electronic health records," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Dec. 2020.
[12] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–11. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
[13] S. Bangalore *et al.*, "Antihypertensive drugs and risk of cancer: Network meta-analyses and trial sequential analyses of 324 168 participants from randomised trials," *Lancet Oncol.*, vol. 12, no. 1, pp. 65–82, 2011, doi: 10.1016/S1470-2045(10)70260-6.
[14] E. Copland *et al.*, "Antihypertensive treatment and risk of cancer: An individual participant data meta-analysis," *Lancet Oncol.*, vol. 22, no. 4, pp. 558–570, 2021.
[15] M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian, "Doubly robust estimation of causal effects," *Amer. J. Epidemiol.*, vol. 173, no. 7, pp. 761–767, Apr. 2011, doi: 10.1093/aje/kwq439.
[16] M. E. Reichenheim and E. S. Coutinho, "Measures and models for causal inference in cross-sectional studies: Arguments for the appropriateness of the prevalence odds ratio and related logistic regression," *BMC Med. Res. Methodol.*, vol. 10, no. 1, pp. 1–12, Dec. 2010, doi: 10.1186/1471-2288-10-66.
[17] J. M. Robins, M. Á. Hernán, and B. Brumback, "Marginal structural models and causal inference in epidemiology," *Epidemiology*, vol. 11, no. 5, pp. 550–560, Sep. 2000, doi: 10.1097/00001648-200009000-00011.
[18] A. I. Naimi, S. R. Cole, and E. H. Kennedy, "An introduction to G methods," *Int. J. Epidemiol.*, vol. 46, no. 2, Dec. 2016, Art. no. dyw323, doi: 10.1093/ije/dyw323.

[19] J. C. M. Witteman *et al.*, "G-estimation of causal effects: Isolated systolic hypertension and cardiovascular death in the Framingham heart study," *Amer. J. Epidemiol.*, vol. 148, no. 4, pp. 390–401, Aug. 1998, doi: 10.1093/oxfordjournals.aje.a009658.

[20] J. Fan and R. Li, "Statistical challenges with high dimensionality: Feature selection in knowledge discovery," 2006, *arXiv:math/0602133*.

[21] A. N. Glynn and K. M. Quinn, "An introduction to the augmented inverse propensity weighted estimator," *Political Anal.*, vol. 18, no. 1, pp. 36–56, 2010, doi: 10.1093/pan/mpp036.

[22] J. D. Y. Kang and J. L. Schafer, "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Stat. Sci.*, vol. 22, no. 4, pp. 523–539, Nov. 2007, doi: 10.1214/07-STS227.

[23] V. Chernozhukov *et al.*, "Double/debiased machine learning for treatment and structural parameters," *Econometrics J.*, vol. 21, no. 1, pp. C1–C68, Feb. 2018, doi: 10.1111/ectj.12097.

[24] M. J. Van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY, USA: Springer, 2011.

[25] A. Decruyenaere, J. Steen, K. Colpaert, D. D. Benoit, J. Decruyenaere, and S. Vansteelandt, "The obesity paradox in critically ill patients: A causal learning approach to a casual finding," *Crit. Care*, vol. 24, no. 1, pp. 1–11, Aug. 2020, doi: 10.1186/S13054-020-03199-5.

[26] Y. Zhang, L.-A. Lin, L. Starkopf, J. Chen, and W. W. B. Wang, "Estimation of causal effect in integrating randomized clinical trial and observational data—An example application to cardiovascular outcome trial," *Contemp. Clin. Trials*, vol. 107, Aug. 2021, Art. no. 106492, doi: 10.1016/j.cct.2021.106492.

[27] S. Rose and S. Normand, "Double robust estimation for multiple unordered treatments and clustered observations: Evaluating drug-eluting coronary artery stents," *Biometrics*, vol. 75, no. 1, pp. 289–296, Mar. 2019, doi: 10.1111/biom.12927.

[28] O. Sofrygin *et al.*, "Targeted learning with daily EHR data," *Statist. Med.*, vol. 38, no. 16, pp. 3073–3090, Jul. 2019, doi: 10.1002/sim.8164.

[29] B. C. Kwon *et al.*, "RetainVis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 299–309, Jan. 2019, doi: 10.1109/TVCG.2018.2865027.

[30] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, "Deepr: A Convolutional Net for Medical Records," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 22–30, Jan. 2017, doi: 10.1109/JBHI.2016.2633963.

[31] Y. Li *et al.*, "BEHRT: Transformer for electronic health records," *Sci. Rep.*, vol. 10, no. 1, p. 7155, Dec. 2020, doi: 10.1038/s41598-020-62922-y.

[32] S. Rao *et al.*, "An explainable transformer-based deep learning model for the prediction of incident heart failure," Tech. Rep.

[33] O. Melamud, M. Bornea, and K. Barker, "Combining unsupervised pre-training and annotator rationales to improve low-shot text classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3884–3893, doi: 10.18653/v1/d19-1401.

[34] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," 2018, *arXiv:1805.06334*.

[35] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in *Proc. NIPS*, 2017, pp. 1–11.

[36] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2017.

[37] V. Veitch, D. Sridhar, and D. Blei, "Adapting text embeddings for causal inference," in *Proc. Conf. Uncertainty Artif. Intell.* PMLR, 2020.

[38] A. Izdebski *et al.*, "A pragmatic approach to estimating average treatment effects from EHR data: The effect of prone positioning on mechanically ventilated COVID-19 patients," Sep. 2021, *arXiv:2109.06707*.

[39] 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, doi: 10.1093/biomet/70.1.41.

[40] D. B. Rubin, "Causal inference using potential outcomes," *J. Amer. Stat. Assoc.*, vol. 100, no. 469, pp. 322–331, Mar. 2005, doi: 10.1198/016214504000001880.

[41] F. D. Johansson, U. Shalit, N. Kallus, and D. Sontag, "Generalization bounds and representation learning for estimation of potential outcomes and causal effects," 2020, *arXiv:2001.07426*.

[42] R. Payne and R. Denholm, "CPRD product code lists used to define long-term preventative, high-risk, and palliative medication," Univ. Bristol, Bristol, U.K., Tech. Rep., 2018.

[43] J. Tran *et al.*, "Patterns and temporal trends of comorbidity among adult patients with incident cardiovascular disease in the U.K. between 2000 and 2014: A population-based cohort study," *PLOS Med.*, vol. 15, no. 3, Mar. 2018, Art. no. e1002513, doi: 10.1371/journal.pmed.1002513.

[44] L. Yao, M. Huai, S. Li, J. Gao, Y. Li, and A. Zhang, "Representation learning for treatment effect estimation from observational data," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2018, pp. 2633–2643.

[45] M. Nørgaard, V. Ehrenstein, and J. P. Vandenbroucke, "Confounding in observational studies based on large health care databases: Problems and potential solutions & a primer for the clinician," *Clin. Epidemiol.*, vol. 9, pp. 185–193, Mar. 2017, doi: 10.2147/CLEP.S129879.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[47] NHS. *Read Codes*. Accessed: Oct. 19, 2018. [Online]. Available: https://digital.nhs.uk/services/terminology-and-classifications/read-codes

[48] NHS Digital. (2020). *Read Code Map*. [Online]. Available: https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/9

[49] W. J. Trowell, "'British national formulary,"' *BMJ*, vol. 282, no. 6269, p. 1078, Mar. 1981, doi: 10.1136/bmj.282.6269.1078.

[50] H. A. Chipman, E. I. George, and R. E. McCulloch, "BART: Bayesian additive regression trees," *Ann. Appl. Statist.*, vol. 4, no. 1, pp. 266–298, Mar. 2010, doi: 10.1214/09-AOAS285.

[51] M. A. Luque-Fernandez, M. Schomaker, B. Rachet, and M. E. Schnitzer, "Targeted maximum likelihood estimation for a binary treatment: A tutorial," *Statist. Med.*, vol. 37, no. 16, pp. 2530–2546, Jul. 2018, doi: 10.1002/sim.7628.

[52] T. Sturmer, K. J. Rothman, J. Avorn, and R. J. Glynn, "Treatment effects in the presence of unmeasured confounding: Dealing with observations in the tails of the propensity score distribution—A simulation study," *Amer. J. Epidemiol.*, vol. 172, no. 7, pp. 843–854, Oct. 2010, doi: 10.1093/aje/kwq198.

[53] E. W. Weisstein, "Normal sum distribution," MathWorld, Tech. Rep., 2016.

[54] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.

[55] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[56] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Comput. J.*, vol. 7, no. 4, pp. 308–313, Jan. 1965, doi: 10.1093/comjnl/7.4.308.

[57] J. Pearl, "Remarks on the method of propensity score," *Statist. Med.*, vol. 28, no. 9, pp. 1415–1416, Apr. 2009, doi: 10.1002/sim.3521.

[58] P. Ding and L. W. Miratrix, "To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias," *J. Causal Inference*, vol. 3, no. 1, pp. 41–57, Mar. 2015, doi: 10.1515/jci-2013-0021.

[59] I. Díaz and M. J. van der Laan, "Targeted data adaptive estimation of the causal dose–response curve," *J. Causal Inference*, vol. 1, no. 2, pp. 171–192, Dec. 2013, doi: 10.1515/jci-2012-0005.