ADCB: An Alzheimer's disease simulator for benchmarking observational estimators of causal effects

ADCB: An Alzheimer's disease simulator for benchmarking observational estimators of causal effects

Newton Mwai Kinyanjui MWAI@CHALMERS.SE
Chalmers University of Technology, Sweden

Fredrik D. Johansson FREDRIK.JOHANSSON@CHALMERS.SE
Chalmers University of Technology, Sweden

Abstract

Simulators make unique benchmarks for causal effect estimation as they do not rely on unverifiable assumptions or the ability to intervene on real-world systems. This is especially important for estimators targeting healthcare applications as possibilities for experimentation are limited with good reason. We develop a simulator of clinical variables associated with Alzheimer's disease, aimed to serve as a benchmark for causal effect estimation while modeling intricacies of healthcare data. We fit the system to the Alzheimer's Disease Neuroimaging Initiative (ADNI)[1] dataset and ground hand-crafted components in results from comparative treatment trials and observational treatment patterns. The simulator includes parameters which alter the nature and difficulty of the causal inference tasks, such as latent variables, effect heterogeneity, length of observed subject history, behavior policy and sample size. We use the simulator to compare standard estimators of average and conditional treatment effects.

Data and Code Availability We make use of publicly available longitudinal data, of both Alzheimer's disease (AD) patients and cognitively normal controls, from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). ADNI collects clinical data, neuroimaging data, genetic data, biological markers, and clinical and neuropsychological assessments from participants at different sites in the USA and Canada to study cognitive impairment and and AD. The cohorts used in this work were assembled from ADNI 1, 2, 3 and GO. We use trajectories of 870 unique patients, taking samples in 12-month intervals. An implementation of the simulator can be found at https://github.com/Healthy-AI/ADCB.

1. For the Alzheimer's Disease Neuroimaging Initiative: Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

1. Introduction

Evaluating learned decision-making policies and observational estimators of causal effects is challenging, especially in the healthcare domain. Real-world implementation is often not an option and basing evaluation on observational data must rely on strong assumptions and access to large samples (Rosenbaum et al., 2010). As a result, methods researchers in these areas often turn to simulators for benchmarking (Dorie et al., 2019; Chan et al., 2021).

Simulated data have many advantages but often lack the intricacies observed in reality (Hernán, 2019). For example, two of the most widely used benchmarks in the community studying causal effects, IHDP (Hill, 2011) and ACIC (Dorie et al., 2019), have response surfaces which are hand-crafted from simple mathematical building blocks. To improve on this, researchers have constructed benchmarks from actual samples, simulating a subset (Neal et al., 2020) or all of the observed variables using simulators fit to data (Chan et al., 2021). However, purely data-driven approaches may fail to capture the causal structure of the systems they model. Hernán (2019) argued that, fundamentally, benchmarks must "combine data analysis and subject-matter knowledge".

We propose a new simulator for benchmarking estimators of causal effects, the Alzheimer's Disease Causal estimation Benchmark (ADCB). ADCB combines data-driven simulation with subject-matter

knowledge by fitting a longitudinal causal model of patient variables to real data: i) the simulator is based on a causal structure inferred by Alzheimer's disease experts, ii) average causal effects are based on published results from randomized controlled trials with heterogeneity introduced through an inferred latent variable, iii) overlap and variance in treatment choice is controlled by different behavior policies, and iv) the length of observed subject history is set by the user. This design provides users with tunable parameters which change properties of the system and the difficulty of the benchmark. We use the ADCB simulator to compare standard estimators of causal effects where a) a single time point is used to estimate average and personalized treatment effects, and b) a time series of patient history is used. Based on the results of these experiments, we discuss the benefits and limitations of our approach compared to existing simulators based on experimental data, hand-crafted mechanisms or learned functions.

2. Benchmarks for observational estimation of causal effects

Causal effect estimation studies the outcome $Y(a)$ of intervening with an action (treatment) $a \in \mathcal{A}$ (Rubin, 2005). Here, we define the causal effect of action a as

$$\Delta(a) \coloneqq Y(a) - Y(0),$$

the difference between the potential outcome of $A \leftarrow a$ and that of a baseline action $A \leftarrow 0$. In our setting, $\Delta(a)$ represents the benefit of using treatment a over no treatment. We consider k different actions from a discrete set $\mathcal{A} = \{0, ..., k-1\}$.

Due to the difficulty of trying out different treatments for the same subject under identical conditions, Δ itself is rarely identifiable. Instead, we represent the utility of actions using the *average treatment effect* (ATE), $\tau(a) = \mathbb{E}[\Delta(a)]$ and the *conditional average treatment effect* (CATE),

$$\tau(a \mid x) = \mathbb{E}[\Delta(a) \mid X = x],$$

in a context or stratum $x \in \mathcal{X}$. ATE and CATE measure how well action a performs on average in a population and in a stratum x, respectively. The context X may be a single vector-valued observation or a time-series representing patient history.

Observational estimation refers to estimating τ using samples (a, y, x) of actions, outcomes and context variables without controlling the actions. The following assumptions are sufficient for consistent, unbiased estimation in this setting (Rosenbaum et al., 2010).

Assumption 1 (Identifying assumptions) *Actions $A \in \mathcal{A}$, outcomes $Y \in \mathbb{R}$, a set of context variables X, and an adjustment set of variables $C \subseteq X$ are observed from a distribution $p(X, A, Y)$ such that the following conditions hold for all $a \in \mathcal{A}, c \in \mathcal{C}$,*

Consistency	$Y = Y(A)$
Exchangeability	$Y(a) \perp\!\!\!\perp A \mid C$
Overlap	$p(A = a \mid C = c) > 0$

A wealth of methods have been developed for estimating ATE and CATE under Assumption 1, see e.g., (Dorie et al., 2019; Künzel et al., 2019; Wager and Athey, 2018) for overviews. To assess the qualities of each estimator, various benchmark challenges have been developed (Dorie et al., 2019). See Section 6 for a more in-depth survey.

Fundamentally, the validity of Assumption 1 cannot be verified from data (Pearl, 2009), but must be argued from domain knowledge. Moreover, the assumptions guarantee *identification* of ATE and CATE, but not necessarily good *estimates* when sample sizes are small. Hence, observational data alone are insufficient to determine whether one estimate of a causal effect is more accurate than another. This motivates using simulators for benchmarking, where identifying assumptions can be satisfied by design.

A good benchmark allows users to identify strengths and weaknesses in estimators: Which estimators make efficient use of available data? How does performance scale with dimensionality or sample size? How sensitive are they to (partial) violations of identifying assumptions? Which results are robust to changes in causal structure? Answers to these questions will not be universal, they will depend on the application under study (Hernán, 2019). In this work, we target the healthcare domain, in the context of longitudinal data on clinical variables.

3. The ADCB simulator

Alzheimer's disease is the most common form of dementia, affecting tens of millions of people worldwide (Association, 2019). Despite its toll on public health and vast research investments over several decades, there is currently no cure for AD. Nevertheless, drugs that have disease-modifying effects, alleviating symptoms such as loss of cognitive function,

have shown promise in trials and in practice (Grossberg et al., 2019). For these reasons and more, AD makes an interesting setting for benchmarking causal effect estimators:

- AD is a progressive disorder, deteriorating the health of subjects over time. As a result, data is collected for the same subjects at several time points, allowing for comparing the performance of longitudinal models of causal effects.
- There is evidence that AD is composed of multiple disease subtypes. While the details remain unknown, disease subtypes provide a potential source of heterogeneity in patient outcomes.
- Current treatments are believed to be symptomatic—they affect only symptoms and not the underlying disease cause; their effects disappear once discontinued. This allows for easier attribution of effect to treatment.

The ADCB simulator is based on a longitudinal structural causal model between context, treatment and outcome variables. The remainder of the section describes the components of the simulator, starting with the patient covariates, the assumed causal graph, and a generalization to a longitudinal causal model. Framed boxes are used to indicate readily tunable parameters of the simulator.

3.1. Patient covariates X & outcomes Y

Subjects are represented by covariates $X \in \mathcal{R}^d$ consisting of demographics (sex, age, education level) and various genetic and biomarkers (Aβ plaques, Tau, APOE, FDG, AV45) whose detailed descriptions are provided in Appendix C. The specific variables used to model the time-varying context X_t in this work are presented in Figure 1. The severity of (suspected) Alzheimer's disease is primarily assessed based on cognitive function using tests such as the Alzheimer Disease Assessment Scale (ADAS) (Rosen et al., 1984). We use the ADAS13 variant as our base outcome at time t, $Y_t(0)$, as it has been found to better describe disease progression than the ADAS11 variant (Cho et al., 2021). ADAS13 scores take values between 0-85 where higher scores indicate worse cognitive function. ADNI also contains clinical diagnosis states $DX_t \in \{$Cognitively normal (CN), Mild cognitive impairment (MCI), Alzheimer's disease (AD)$\}$.

3.2. Disease subtype (latent state Z)

It is believed that there are multiple subtypes of Alzheimer's disease (Machado et al., 2020; Satone et al., 2018). One of the signs of this is that in subjects, the level of so-called Amyloid-β (Aβ) plaques form a clearly bimodal distribution, on the ratio of $(\frac{A\beta 42}{A\beta 40})$, see Figure 9 in Appendix. We posit that there are two types of subjects, as indicated by a binary variable $Z \in \{0, 1\}$, which, among other things, give rise to the two modes in the Aβ-ratio. To this end, we infer the subtype Z by fitting a Gaussian mixture model (GMM) with 2 components as in (Dansson et al., 2021) for the Aβ-ratio observations of patients at baseline. We assume that Z is stationary and use the value inferred by the GMM to label all observed trajectories. These values are then used to fit models of downstream variables.

3.3. Baseline Causal Graph

We start by positing a causal graph for the variables of interest at the baseline time point of observation, $t = 0$. A causal graph is a model of the (conditional) dependence structure of variables encoded in a directed acyclic graph, $\mathcal{G} = (\mathrm{V}, \mathrm{E})$ consisting of nodes V and edges, E (Koller and Friedman, 2009). The causal graph, illustrated in Figure 1, was inspired by the structure inferred from data in (Sood et al., 2020) and further verified by a clinically active domain expert in Alzheimer's disease. The graph represents causal relationships among random variables $R \in \{X(1), ..., X(d), A, Y, DX\}$, (where $X(j)$ is a covariate in the set X), each associated with a node in the graph, $V_R \in V$. An edge $(V_R, V_{R'}) \in E$ exists if R is a direct cause of R', and R is therefore a parent of R', $R \in Pa(R')$.

The mechanism for generating each variable is based either on models fit to the ADNI data, on handcrafted functions or on results from the AD literature. The graph is also presented as a table in the Appendix Table 5.

3.4. Longitudinal Model

The longitudinal model is formed by first repeating each variable, except the disease subtype Z, at each time step $t = 1, 2, ..., T$, maintaining the causal structure of the single-time graph in Figure 1. Then, each variable is connected to the previous instance of itself; e.g. Tau_t is assumed to be a direct cause of Tau_{t+1}, and so on. The parents of a variable X at time t is

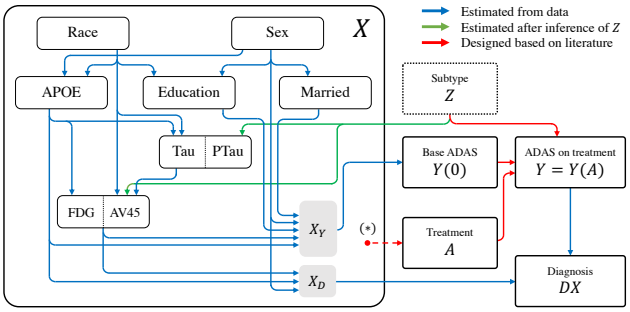


Figure 1: Assumed causal graph for a single time point at baseline. Arrows indicate causal dependencies, with color representing how the mechanism was determined. Blue dependencies were completely estimated from data, green were fit once the subtype Z was inferred, and red were designed based on the Alzheimer's disease literature.

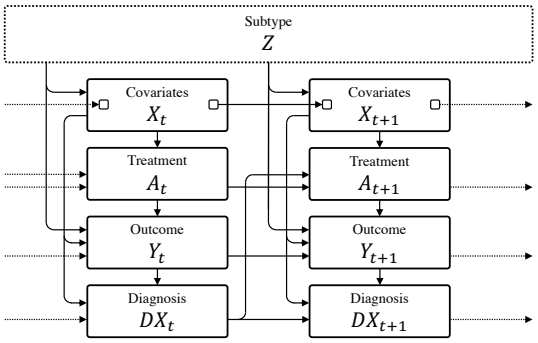


Figure 2: Temporal dependence between variables in the simulator. Each variable obeys the causal dependencies of Figure 1 in addition to depending on the previous value of itself. The small box in the set of covariates X indicates that each variable in the set depends only on the previous value of that specific variable. For example, Tau at time $t+1$ depends only on APOE and Race at time $t+1$, the subtype Z, and Tau at time t. Z is assumed stationary.

therefore the set defined as: $Pa(X_t) = Pa_t(X_0) \cup \{X_{t-1}\}$ where $Pa_t(X) = \{p_t : p_0 \in Pa(X_0)\}$. When used as a benchmark, the user may choose the causal effect of actions at any time point t as their target. The length $H \leq t$ of history used for estimation is a tunable parameter.

History Length, H. History of previous treatment records of a patient is valuable for causal effect estimators that incorporate history, because a longer horizon can increase the capacity to capture heterogeneity in causal effects.

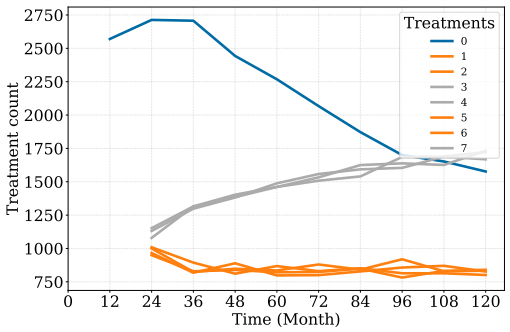
3.5. Treatment assignment A

ADNI does not include significant data on treatments and treatment response, which prevents direct data-driven design of the treatment assignment. Instead, we design policies for treatment assignment and treatment effects based on i) surveys of common treatments and ii) randomized controlled trials (RCT) of their effect. We begin with the former.

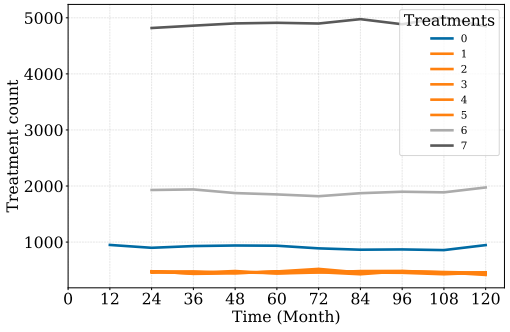
Existing AD drugs have been shown to have at least symptomatic cognitive effects (Livingston et al., 2017; Farlow et al., 2008). In this work, we model a range of such drugs $a = 1, ..., 7$, for which RCT results on treatment effects are available: Donepezil 5mg, Donepezil 10mg, Galantamine 24mg, Galantamine 32mg, Rivastigmine 12mg, Memantine 20mg, Memantine+ChEI, see (Grossberg et al., 2019) for an overview. We assume that the no-treatment option, $a = 0$, corresponds to observations in ADNI. We simulate treatments from two simple policies μ_B, described further below, whose characteristics are shown in Figure 3:

Diagnosis (DX)-based policy With this policy, treatments are assigned based on the diagnosis (DX) observed at the previous time step. We group treatments into 3 classes based on their treatment effect. Patients with mild diagnosis are assigned a randomly chosen treatment from the class with smallest ATE, those with moderate from the class with moderate ATE, and those with the most severe diagnosis from the class with the largest effect.

Hernandez Policy Having access to treatments in ADNI data would have enabled modeling of treatment propensities over the whole covariate set, deriving purely data-driven behavior policies using a much larger subset of covariates. In lieu of this, we draw from Hernandez et al. (2010) who similarly modeled the propensity of the treatments Cholinesterase inhibitors (ChEIs) and Memantine based on clinical variables with a multivariate logistic regression models, with ChEI or Memantine use as the outcome—



(a) DX-based Policy



(b) Hernandez policy

Figure 3: Treatment assignment characteristics of behavior policies over time. Same colour indicates treatments are considered to be in the same treatment assignment class.

we define a policy directly using the coefficients they learned. Treatments are grouped based on drug class $\in$ {ChEIs, Memantine, Combination therapy}. The learned policy depends also on cognitive scores MMSE and CDRSB, which are available in the ADNI database. We generate them in the same way as ADAS13.

Overlap strength ϵ The tuning parameter $\epsilon \in [0, 1]$ interpolates between a random policy ($\epsilon = 1$) and the policies above ($\epsilon = 0$) by assigning a random action with probability ϵ. Note that $\epsilon = 0$ does not always imply a lack of treatment group overlap, depending on the behavior policy.

3.6. Treatment effects Δ

Consistent with the AD literature, we assume that the effects of each existing drug a are primarily symptomatic and temporary, attenuating when treatment is stopped (Grossberg et al., 2019). In addition, we assume that the effect is stationary in time. To this end, we endow each treatment a with an additive effect $\Delta(a, Z)$, depending on the disease subtype Z, and posit that the cognitive function when on drug a is given by $Y_t(a) = \Delta(a, Z) + Y_t(0) + \epsilon_t$. This gives us the response surface on $Y_t = Y_t(A_t)$. $Y_t(0)$ is estimated from observations of the ADAS13 score and is simulated according to the causal graph. We discuss more general forms of treatment effects in Section 6.

To ground our model in domain knowledge, we design $\Delta(a, Z)$ such that the average effect $\tau(a) = \mathbb{E}[\Delta(a, Z)]$ is consistent with real-world effects on cognitive function (in the ADAS-Cog scale) estimated in RCTs (Grossberg et al., 2019). Recall that we define ATE relative to the no-treatment option. For a list of the estimated ATEs $\tau(a)$, for $a = 1, ..., k$, taken from the literature, see Appendix D.

Given the ATE $\tau(a)$ for a treatment a, heterogeneity is introduced through the subtype $z \in Z$. In this work, Z is binary, and we let each subtype-action pair (a, z) have HIGH or LOW effect, with multiplicative margin γ, such that the opposite subtype $(a, 1 - z)$ has HIGH effect, if (a, z) has LOW effect and vice versa.

$$\Delta(a, z) = \begin{cases} \frac{\tau(a)}{p(Z=z)+p(Z\neq z)\gamma}, & \text{if } \Delta(a, z) \text{ LOW} \\ \frac{\gamma\tau(a)}{p(Z=z)\gamma+p(Z\neq z)}, & \text{if } \Delta(a, z) \text{ HIGH} \end{cases}.$$

Whether $\Delta(a, z)$ is HIGH or LOW for a, z is determined by a look-up table that we designed which is presented in the Appendix, Table 3.

Treatment effect heterogeneity γ. The parameter $\gamma \geq 1$ controls heterogeneity in effect such that $\Delta(a, z) = \gamma\Delta(a, 1 - z)$ if $\Delta(a, z)$ is HIGH and vice versa. γ varies heterogeneity without changing the average treatment effect $\tau(a)$. $\gamma = 1$ results in no heterogeneity.

4. Fitting the simulator

Based on the causal the graph presented in Figure 1, we learn a joint distribution of the full set of set of observed variables $X, Y(0), DX$ by fitting each component of the Bayes factorization separately using a variable's parent set, $Pa(X_t)$. For each continuous (or discrete) attribute, a regression (or stochastic classification) model is fit with respect to its parents in the causal graph.

These models are first fit for the baseline timestep ($t = 0$) in patient trajectories for the purpose of i) generating the first time step further downstream in the generation process and ii) data imputation for missing values, as described in Appendix B. The marginal root nodes are sampled from a distribution inferred using the statistics observed in the data. Continuous covariates are further modeled with additive noise ζ sampled from a skewed normal distribution fit to the residuals of the regression, $r_i = y_i - f(x_i)$ where $f(x) \approx \mathbb{E}[Y|X = x]$ is learned from data.

The longitudinal model—the transition models for each variable—is fit similarly. For each covariate at time t, we assume that i) its value is dependent only on its parents in the causal graph at the time t as well as its previous value in the trajectory at time $t-1$. ii) the autoregression is stationary in time. A summary of the different models and their fit characteristics is described in Table 2.

For each time step, classifiers fit $P(X_t|Pa(X_t))$ and generation is done by sampling from this. The regressors fit $f(Pa(X_t)) = E[X_t|Pa(X_t)]$ and samples are generated by $f(Pa(X_t)) + \zeta$. With these models fit, hand-crafted components designed and tunable parameters $\{H, N, \gamma, \epsilon, T, \mu_B\}$ set, we generate N patient trajectories of T time steps with all variables $(Z, X_t, Y_t(0), A_t, Y_t(A_t), DX_t)$ through ancestral sampling.

4.1. ADNI and ADCB cohort statistics

Trajectories of 2254 subjects were downloaded from the ADNI database in December 2020. The full cohort was filtered for availability of measurements of Aβ40 and Aβ42 biomarkers at some point in their trajectory, leaving $n = 870$ subjects for fitting the simulator, 844 of which were observed at baseline. Overview statistics of these subjects at baseline are presented in Table 1. Trajectory lengths varied greatly among subjects, ranging from a single visit at baseline to a total of 8 visits (mean 1.7 visits). The longest trajectory length was 120 months (mean 14 months). Only subjects with observations for all simulator variables (except Z, A and $Y(a)$) were used for fitting baseline and autoregression covariate models. For longitudinal modeling, models were fit based on transitions between pairs of visits (0, 12), (12, 24), (24, 36), (36, 48) for observations present in both time points in the transition in the original data, which was a total of 127 samples.

Table 1: Cohort statistics for the first timestep ($t = 0$) for simulated (ADCB) and observed real-world subjects (ADNI). Continuous variables are described by mean (standard deviation) and categorical variables by count (frequency in %). Complete cohort statistics are provided in the Appendix table 4

	ADCB $t = 0$,	ADNI, $t = 0$
Demographics		
Gender		
Female	4807 (48.1%)	395 (46.8%)
Male	5193 (51.9%)	449 (53.2%)
Biomarkers		
Tau	286.0 (117.3)	279.6 (130.0)
PTau	27.9 (12.7)	26.7 (14.2)
FDG	1.3 (0.2)	1.2 (0.2)
AV45	1.2 (0.2)	1.2 (0.2)
APOE4		
0.0	4196 (42.0%)	460 (54.5%)
1.0	4460 (44.6%)	303 (35.9%)
2.0	1344 (13.4%)	81 (9.6%)
Outcomes		
ADAS13	16.4 (8.4)	15.4 (9.5)

4.2. Model fit for variables in causal graph

We evaluate the model fit on held-out data independently for each variable, as summarized in Table 2. The test split was always 20%. The overall predictability for baseline variables was low, with non-trivial accuracy attained only for a handful of the covariates, including diagnosis and AV45 levels. However, we remind the reader that accurate prediction is not the main goal of this step, but to learn a simulator with similar characteristics as the observed data. In Table 1 and Appendix Table 4 we show the first-order statistics for observed and generated data.

Autoregressors achieved significantly better results due to some variables being more or less static in time or varying very slowly. AV45 had surprisingly poor R^2 fit results for autoregression although the RMSE error was in the range of the standard deviation of the original data. The hyperparameters for the models were obtained by doing a grid search over combinations of parameters over Linear, Random Forest and Gradient Boosting estimators for each variable.

Table 2: Fit statistics for baseline and autoregression models on held-out data. Overall predictability was low at baseline and in autoregression for some continuous variables. This indicates that parents in the causal graph explain only a small amount of variance in the affected variables.. First-order statistics were well matched, see Table 1 and Appendix Table 4.

Target variable	Model	Baseline			Autoregression		
Classifiers		Acc	F1	# Classes	Acc	F1	# Classes
APOE4	KNN	45%	0.42	3	96%	0.94	3
Education (years)	Logistic Regression	21%	0.09	13	100%	1.00	10
Marital status	Logistic Regression	73%	0.62	5	96%	0.94	4
Diagnosis	Logistic Regression	63%	0.63	3	88%	0.87	3
Regressions		R^2	RMSE	σ_Y	R^2	RMSE	σ_Y
Tau	Random Forest	-1.13	105.35	133.43	0.73	47.81	117.95
PTau	Random Forest	-0.55	11.0	14	0.91	3.69	14
FDG	Gradient Boosting	-3.79	0.14	0.15	0.09	0.06	0.09
AV45	Random Forest	0.20	0.15	0.23	-82.03	0.12	0.12
ADAS13	Random Forest	0.21	6.36	9.6	0.55	4.09	6.3
CDRSB	Gradient Boosting	-0.06	1.26	1.5	-0.61	1.20	2.2
MMSE	Gradient Boosting	-0.56	2.03	2.6	-0.26	1.63	1.4

5. Using the benchmark

We run experiments aimed at exploring the utility of the simulator and its generated sequential trajectories in benchmarking causal effect estimators. The experiments compare estimators of the Conditional Average Treatment Effect (CATE) at a given timepoint $0 < t_s \leq T$. We run them in settings with decisions with single-time context X_{t_s} and in settings where context comprises a H-length history of context, treatment and outcome variables, and compare the mean-squared error in estimated CATE (also called precision of estimating heterogeneous effects (PEHE) (Hill, 2011)). Unless otherwise stated, Assumption 1 is satisfied in all experiments by giving estimators access to a valid adjustment set. The adjustment set includes all the covariates in the current demographics, the current biomarkers, the most recent outcome and most recent diagnosis for the DX-policy. A similar adjustment set is used for the Hernandez-based policy, without the most recent diagnosis and with CDRSB and MMSE scores, for validity.

The estimators presented are S-learners (treatment as a covariate) and T-learners (separate regression for each treatment) (Künzel et al., 2019) with Linear Regression, Gradient Boosting or Random Forest base learners, as well as a Sequential T-learner with an RNN base learner to enable incorporation of history. S- and T-learners are trained single-step and the sequential T-learner trained using a history sequence of time points $\{t = t_s - H, ..., t = t_s\}$.

We investigate the effects of sample size, overlap, heterogeneity, history length and confounding as outlined below. Results are from 10 repetitions in each configuration.

Sample size, N: Under Assumption 1, it is expected that the CATE estimation error shall decrease with higher sample sizes as the variance should decrease with more samples, until bias (model misspecification) dominates the error. Estimating CATE with different numbers of samples generated from ADCB is consistent with this across the estimators, as shown in Figure 4 where the base estimator for the T- and S-learners is a Gradient Boosting Regressor. The CATE error with 50,000 samples is comparable with the error using 10,000 samples, so the rest of the experiments have been run with 10,000 samples.

Overlap, ϵ: ADCB enables investigation of overlap with the tunable parameter ϵ, which varies the treatment assignment propensity characteristics of the treatment policies in Figure 3. As ϵ increases and selection bias decreases, the behavior policy approaches a uniform policy and it's expected that the CATE estimation error should decrease. This is ob-

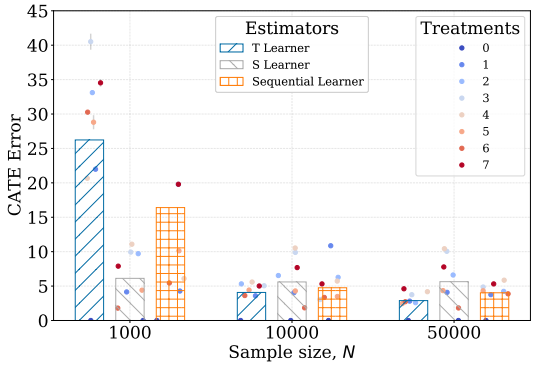
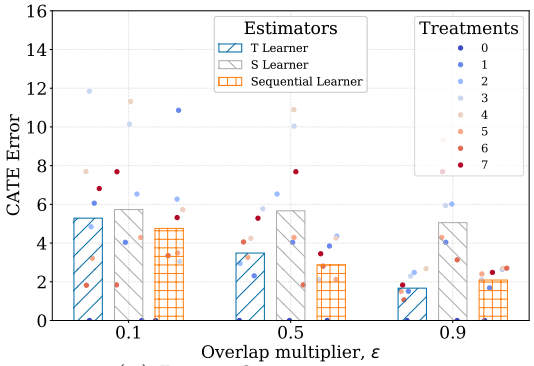
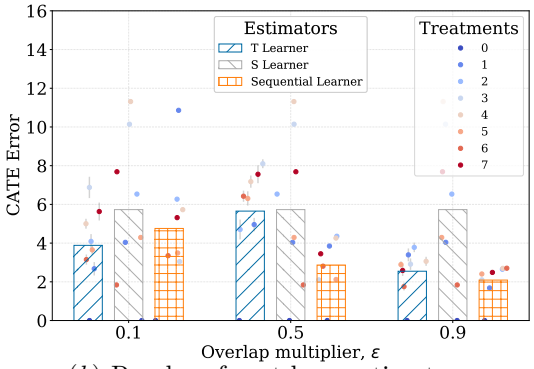


Figure 4: CATE mean squared error varying with sample size, N. ϵ=0.1, γ=2, μ_B=DX-Based, $t_s = 5$, History length, $H = 3$



(a) Linear base estimator



(b) Random forest base estimator

Figure 5: Average (bars) and treatment-specific (dots) mean squared error in estimated CATE, varying with overlap multiplier, ϵ. γ=2, Sample size, $N = 10,000$, μ_B=DX-Based, $t_s = 5$, History length, $H = 3$

served in the T-learner and the sequential learner (RNN), but the S-learner is constant through the three ϵ settings, as shown in Figure 5.

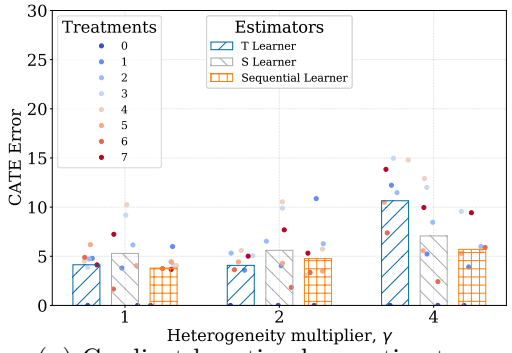
Heterogeneity, γ: With the tunable parameter γ, we can also vary the heterogeneity characteristics of the treatment policies. It is expected that the error should increase as the heterogeneity increases as higher heterogeneity may increase the variance, and the outcomes of actions become harder to predict. Our results in Figure 6 show this across two different base estimators (Gradient boosting and Random forest) in the T- and S- Learners.

History length, H: A key property of the ADCB simulator is access to history. Because physicians usually have access to historical records of a patient, they can use the historical records to personalize their treatment decisions. It is expected that using the history should decrease the error of the estimated CATE for the sequential learner that can incorporate history. This is because access to more history gives the estimator a higher chance of capturing heterogeneity. In Figure 7, the error for the T- and S- learners remains constant because they cannot make use of the history. The error is lowest with a history of length 2, possibly because the DX-based policy uses only the previous diagnosis. It would be interesting to investigate if other sequential estimators are better with longer histories.

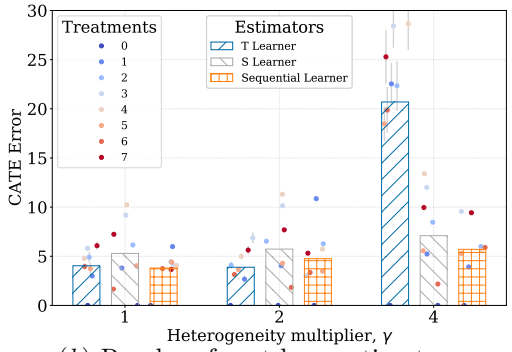
Confounding: Because we know the causal graph of the simulator, we can also investigate confounding effects, e.g. by adding current diagnosis in the adjustment set, which is a post-treatment collider variable, as shown in Figure 8. The estimators are affected differently by this confounding, with the sequential learner showing the highest error increase due to confounding. The T- and S- learners seem to be more robust with the T-learner being slightly more affected.

6. Discussion & Related work

The possibility of producing confounded evaluation metrics prevents using only observational data for benchmarking causal effect estimation, without relying on strong assumptions. There are two main approaches which do not rely on such assumptions: a) making use of data from randomized experiments, and b) simulating all or part the system under investigation, also called the Empirical Monte Carlo Study



(a) Gradient boosting base estimator



(b) Random forest base estimator

Figure 6: CATE error varying with heterogeneity, γ. ϵ=0.1, Sample size, $N = 10,000$, μ_B=DX-Based, $t_s = 5$, History length, $H = 3$

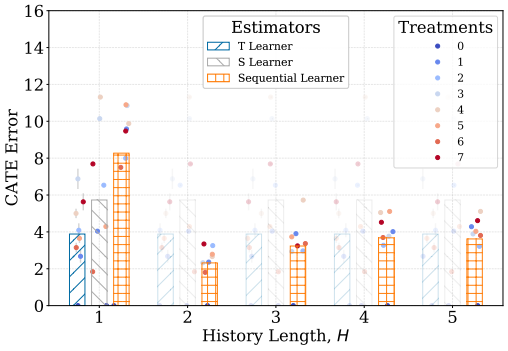


Figure 7: CATE error varying with sequence length, H. ϵ=0.1, γ=2, Sample size, $N = 10,000$, μ_B=DX-Based, $t_s = 5$. Estimators independent of history length are grayed out.

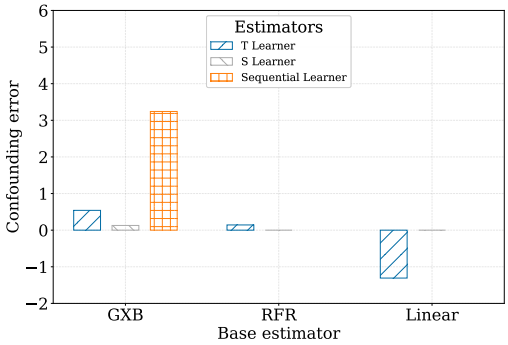


Figure 8: Excess error due to confounding, relative to CATE error of corresponding estimator, when post-treatment covariate DX is added to the adjustment set. ϵ=0.1, γ=2, Sample sizem, $N = 10,000$, μ_B=DX-Based, $t_s = 5$, History length, $H = 3$

(EMCS) approach (Huber et al., 2013; Lechner and Wunsch, 2013). See Gentzel et al. (2019) for a discussion of the pros and cons of each design.

Both methods have limitations that ADCB seeks to remedy. With randomized experiments data, as used in the Jobs dataset (Shalit et al., 2017) or by Neal et al. (2020), the data are guaranteed to be representative of the real world, but it is not possible to vary all characteristics of it, like the sample size or longitudinal horizon length. In contrast, for simulators, it is important to pay attention to the causal structure and mechanisms of the system which most often requires domain knowledge, without which high realism is not easily achievable.

If the goal of a benchmark is to evaluate individual-level or fine conditional treatment effects, access to counterfactual outcomes is required. The only way to reliably achieve this is to simulate the mechanism determining the outcome of interventions, which can be done in isolation or in addition to simulating the treatment assignment, as in the Causal Inference Benchmarking Framework by Shimoni et al. (2018), the Medkit-Learning environment (focused on reinforcement learning) (Chan et al., 2021), and in IHDP (Hill, 2011). Since the outcome mechanisms are often the main target of estimation, these simulations should be as realistic as possible for the domain they aim to represent. To this end, researchers have considered building their simulators on models fit to observational data (Neal et al., 2020; Chan et al., 2021). To incorporate domain knowl-

edge in simulating the outcomes and counterfactual outcomes, ADCB extends these approaches by using treatments and their corresponding effects from Alzheimer's literature, paired with causal generation of a common outcome measurement for cognitive function (ADAS13).

A drawback of simulated data is that, in many cases, simulators "tend to match the assumptions of the researcher" (Gentzel et al., 2019). This is especially problematic in cases where they are introduced to evaluate one particular estimator which may also match those assumptions. As a result, it is important that simulator-based benchmarks contain settings that tweak assumptions to appropriately test the robustness of estimators to these. ADCB enables settings with different configurations for overlap, sample size, patient heterogeneity, behaviour policy and longitudinal history length. Knowledge of the causal graph also enables investigation of estimator performance with confounding. It is also possible to violate consistency by introducing a probability that patients take the assigned treatment.

Limitations

Limitations of ADCB include the following. First, although the treatments, and treatment propensities in the case of the Hernandez policy (Hernandez et al., 2010), are obtained from literature, they are still simulated treatments not originally included in the ADNI data. As such, they may not reflect how subjects in the ADNI cohort would be treated under current practice. Further, behavior policies used only a single time-step context and not patients' entire history. Second, for the treatment effects, we use a simple bi-modal model of heterogeneity and the heterogeneity simulation assumes that heterogeneity is only due to latent covariates Z. A more expressive model would let heterogeneity depend also on X.

As pointed out earlier, several of the autoregressive models (for covariate transitions) had poor accuracy, in three cases with negative R^2. We believe that this could be improved in a future version of the simulator by changing the handling of missing data so that only the target variable for a particular edge in the causal graph is required observed when fitting the model. Currently, transition models are fit to complete cases.

For the presented usage scenario of comparing estimators, we only investigated a handful of simple estimators among a vast array of causal effect estimation methods. Finally, although the assumed causal graph was informed both by conversations with a domain practitioner and by data-driven estimates in (Sood et al., 2020), it would be of interest to test the sensitivity of treatment effect estimates to different adjustment sets or changes to the causal graph such as the addition of new links between covariate nodes.

7. Conclusion

We have introduced the Alzheimer's Disease Causal estimation Benchmark (ADCB), a simulator of clinical variables associated with Alzheimer's disease, aimed to serve as a benchmark for causal effect estimation and policy evaluation. The simulator is fit to covariates and outcomes from the ADNI database and uses models of treatments and treatment effects derived using subject-matter knowledge in the Alzheimer's disease literature. In addition to generating tunable high dimensional observational data with high realism based on a real world Alzheimer's setting, ADCB also generates longitudinal data that includes potential outcomes for all treatments at each step in the longitudinal axis. We also present a method to build semi-synthetic datasets by incorporating results from Alzheimer's literature which is highly effective in attaining realism, and encourages incorporation of inter-disciplinary domain-specific results in building synthetic datasets in machine learning and causal inference.

Usage scenarios for evaluating estimators of causal effects have been presented for varying configurations. Since ADCB generates longitudinal samples of all variables (patient covariates, treatments and outcomes) in the system, it can function as a generator of arbitrarily large observational (batch) data, as an online policy learning environment and for design and evaluation of causally adaptive treatment policies. More complex confounding models based on the AD literature will be explored in future iterations of the simulator, increasing the difficulty of the benchmark. To improve the predictability of the fitted models, the sample sizes will be increased in future iterations by expanding the filtering strategy for the samples included in the training sets.

Institutional Review Board (IRB)

All data collection by ADNI were approved by the Institutional Review Boards of all participating institutions. Written informed consent was obtained from

every research participant according to the Declaration of Helsinki and the Belmont Report.

Acknowledgments

This work was supported in part by WASP (Wallenberg AI, Autonomous Systems and Software Program) funded by the Knut and Alice Wallenberg foundation. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE) partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Collection and sharing of the data used in this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Alzheimer's Association. 2019 alzheimer's disease facts and figures. *Alzheimer's & dementia*, 15(3): 321–387, 2019.

Alex J Chan, Ioana Bica, Alihan Huyuk, Daniel Jarrett, and Mihaela van der Schaar. The medkit-learn (ing) environment: Medical decision modelling through simulation. *arXiv preprint arXiv:2106.04240*, 2021.

Soo Hyun Cho, Sookyoung Woo, Changsoo Kim, Hee Jin Kim, Hyemin Jang, Byeong C Kim, Si Eun Kim, Seung Joo Kim, Jun Pyo Kim, Young Hee Jung, et al. Disease progression modelling from preclinical alzheimer's disease (ad) to ad dementia. *Scientific reports*, 11(1):1–10, 2021.

Hákon Valur Dansson, Lena Stempfle, Hildur Egilsdóttir, Alexander Schliep, Erik Portelius, Kaj Blennow, Henrik Zetterberg, and Fredrik D Johansson. Predicting progression & cognitive decline in amyloid-positive patients with alzheimer's disease. *Alzheimer's Research & Therapy*, 2021.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34 (1):43–68, 2019.

Martin R Farlow, Michael L Miller, and Vojislav Pejovic. Treatment options in alzheimer's disease: maximizing benefit, managing expectations. *Dementia and geriatric cognitive disorders*, 25(5): 408–422, 2008.

Amanda Gentzel, Dan Garant, and David Jensen. The case for evaluating causal models using interventional measures and empirical data. *Advances in Neural Information Processing Systems*, 32:11722–11732, 2019.

George T Grossberg, Gary Tong, Anna D Burke, and Pierre N Tariot. Present algorithms and future treatments for alzheimer's disease. *Journal of Alzheimer's Disease*, 67(4):1157–1171, 2019.

Miguel A Hernán. Comment: Spherical cows in a vacuum: data analysis competitions for causal inference. *Statistical Science*, 34(1):69–71, 2019.

Santiago Hernandez, McKee J McClendon, Xiao-Hua Andrew Zhou, Michael Sachs, and Alan J

Lerner. Pharmacological treatment of alzheimer's disease: effect of race and demographic variables. *Journal of Alzheimer's Disease*, 19(2):665–672, 2010.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Martin Huber, Michael Lechner, and Conny Wunsch. The performance of estimators based on the propensity score. *Journal of Econometrics*, 175(1): 1–21, 2013.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10): 4156–4165, 2019.

Michael Lechner and Conny Wunsch. Sensitivity of matching-based program evaluations to the availability of control variables. *Labour Economics*, 21: 111–121, 2013.

Gill Livingston, A Sommerlad, V Orgeta, SG Costafreda, J Huntley, D Ames, C Ballard, S Banerjee, A Burns, J Cohen-Mansfield, et al. The lancet international commission on dementia prevention and care. *Lancet*, 390(10113): 2673–2734, 2017.

Alejandra Machado, Daniel Ferreira, Michel J Grothe, Helga Eyjolfsdottir, Per M Almqvist, Lena Cavallin, Göran Lind, Bengt Linderoth, Åke Seiger, Stefan Teipel, et al. The cholinergic system in subtypes of alzheimer's disease: an in vivo longitudinal mri study. *Alzheimer's research & therapy*, 12(1):1–11, 2020.

Razvan V Marinescu, Neil P Oxtoby, Alexandra L Young, Esther E Bron, Arthur W Toga, Michael W Weiner, Frederik Barkhof, Nick C Fox, Stefan Klein, Daniel C Alexander, et al. Tadpole challenge: Prediction of longitudinal evolution in alzheimer's disease. *arXiv preprint arXiv:1805.03909*, 2018.

Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. Realcause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*, 2020.

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

Wilma G Rosen, Richard C Mohs, and Kenneth L Davis. A new rating scale for alzheimer's disease. *The American journal of psychiatry*, 1984.

Paul R Rosenbaum, PR Rosenbaum, and Briskman. *Design of observational studies*, volume 10. Springer, 2010.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469): 322–331, 2005.

Vipul Satone, Rachneet Kaur, Faraz Faghri, Mike A Nalls, Andrew B Singleton, and Roy H Campbell. Learning the progression and clinical subtypes of alzheimer's disease from longitudinal clinical data. *arXiv preprint arXiv:1812.00546*, 2018.

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

Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046*, 2018.

Meemansa Sood, Akrishta Sahay, Reagon Karki, Mohammad Asif Emon, Henri Vrooman, Martin Hofmann-Apitius, and Holger Fröhlich. Realistic simulation of virtual multi-scale, multi-modal patient trajectories using bayesian networks and sparse auto-encoders. *Scientific reports*, 10(1):1–14, 2020.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(1): 1–67, 2011.

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

Appendix A. Empirical distribution of the Aβ ratio

The the ratio of $(\frac{A\beta-42}{A\beta-40})$ in subjects showing Amyloid-β (Aβ) plaques form a clearly bimodal distribution;

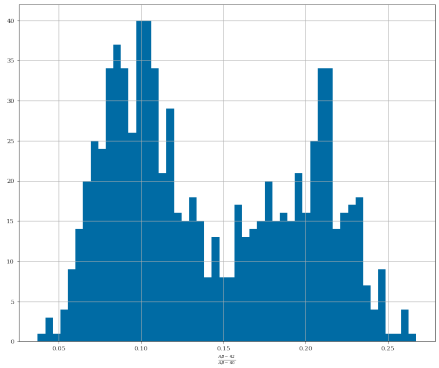


Figure 9: Empirical distribution of the Aβ ratio, used to infer latent disease subtype at baseline.

Appendix B. Imputation of missing data

The patient trajectories have significant missingness along the observation intervals. We impute the missing values based using a method inspired by Multivariate Imputation by Chained Equations(MICE) (Van Buuren and Groothuis-Oudshoorn, 2011), but the chaining is done with respect to a variable's parents in the causal graph. For each attribute with a missing value along the time trajectory, we use the model learned at baseline, from the causal graph, to impute the value for that particular attribute at a given timestep.

Appendix C. Patient covariates description

The subset of covariates used in this work includes the following and their descriptions as outlined in (Marinescu et al., 2018)

1. **FDG PET ROI averages:** Measure cell metabolism, where cells affected by AD show reduced metabolism
2. **AV45 PET ROI averages:** Measure amyloid-beta load in the brain, where amyloid-beta is a protein that mis-folds (i.e. its 3D structure is not properly constructed), which then leads to AD
3. **CSF biomarkers:** Amyloid and TAU levels in the cerebrospinal fluid (CSF)
4. Others:
   - **APOE status:** A gene that is a risk factor for developing AD
   - **Demographic information:** Gender, age, education, race, marital status
   - **Diagnosis:** Either Cognitively Normal (CN), Mild Cognitive Impairment (MCI) or Alzheimer's disease (AD).

Appendix D. Average Treatment Effects from Literature

Table 3: Average treatment effects (ATE), in terms of change in ADAS-Cog compared to no treatment, of various therapies from meta-analyses of clinical trials (Grossberg et al., 2019). Also shown is a look-up table for whether $\Delta(a, z)$ is HIGH or LOW

a	**Treatment**	**ATE** $\tau(a)$	$\Delta(a, z = 0)$
0	No treatment	0	-
1	Donepezil 5 mg	–1.95	L
2	Donepezil 10 mg	–2.48	L
3	Galantamine 24 mg	–3.03	H
4	Galantamine 32 mg	–3.20	H
5	Rivastigmine 12 mg	–2.01	L
6	Memantine 20 mg	–1.29	H
7	Memantine + ChEI	–2.64	L

Appendix E. Cohort statistics

Complete Cohort statistics for synthtetic and real-world cohorts are presented in Table 4.

Appendix F. Causal Graph

The expanded table for the causal graph in Figure 1 is presented in Table 5.

Table 4: Cohort statistics for the first timestep (T=1) for simulated (ADCB) and observed real-world subjects (ADNI). Continuous variables are described by mean (standard deviation) and categorical variables by count (frequency in %).

	ADCB T=1, n=10000	ADNI T=1, n=844
Demographics		
Gender		
Female	4807 (48.1%)	395 (46.8%)
Male	5193 (51.9%)	449 (53.2%)
Marital status		
Divorced	7572 (75.7%)	634 (75.1%)
Married	387 (3.9%)	29 (3.4%)
Never married	1098 (11.0%)	96 (11.4%)
Unknown	889 (8.9%)	80 (9.5%)
Widowed	54 (0.5%)	5 (0.6%)
Ethnicity		
Hisp/Latino	341 (3.4%)	30 (3.6%)
Not Hisp/Latino	9605 (96.0%)	809 (95.9%)
Unknown	54 (0.5%)	5 (0.6%)
Race		
Am Indian/Alaskan	9269 (92.7%)	783 (92.8%)
Asian	384 (3.8%)	31 (3.7%)
Black	148 (1.5%)	13 (1.5%)
Hawaiian/Other PI	17 (0.2%)	1 (0.1%)
More than one	137 (1.4%)	12 (1.4%)
Unknown	18 (0.2%)	2 (0.2%)
White	27 (0.3%)	2 (0.2%)
Education	13.2 (2.7)	13.3 (2.6)
Biomarkers		
Tau	286.0 (117.3)	279.6 (130.0)
PTau	27.9 (12.7)	26.7 (14.2)
FDG	1.3 (0.2)	1.2 (0.2)
AV45	1.2 (0.2)	1.2 (0.2)
APOE4		
0.0	4196 (42.0%)	460 (54.5%)
1.0	4460 (44.6%)	303 (35.9%)
2.0	1344 (13.4%)	81 (9.6%)
Outcomes		
ADAS13	16.4 (8.4)	15.4 (9.5)
MMSE	27.5 (2.0)	27.6 (2.5)
CDRSB	2.0 (1.3)	1.5 (1.7)
Diagnosis		
CN	2700 (27.0%)	275 (32.6%)
Dementia	5817 (58.2%)	438 (51.9%)
MCI	1483 (14.8%)	131 (15.5%)
Subtype, Z		
Subtype, Z	4282 (42.8%)	- (-)

Table 5: Expanded table for the causal graph in Figure 1 at baseline (t=0). For each time step, classifiers fit $P(X_t|Pa(X_t))$ and generation is done by sampling from this. The regressors fit $f(Pa(X_t)) = E[X_t|Pa(X_t)]$ and samples are generated by $f(Pa(X_t)) + \epsilon$. The models are the best performers after a grid search over hyperparameters.

Target variable (X)	Model	Direct causes at baseline ($Pa(X_{t=0})$)
Classifiers		
APOE4	KNN	Ethnicity, Race, Gender
Education (years)	Logistic Regression	Ethnicity, Race, Gender
Marital status	Logistic Regression	Gender
Diagnosis	Logistic Regression	Ethnicity, Race, Gender, Z, Tau, PTau, APOE4, FDG, AV45, ADAS13
Regressions		
Tau	Random Forest	Ethnicity, Race, Gender, Z, APOE4
PTau	Random Forest	Ethnicity, Race, Gender, Z, APOE4
FDG	Gradient Boosting	Ethnicity, Race, Z, Tau, PTau, APOE4
AV45	Random Forest	Ethnicity, Race, Z, Tau, PTau, APOE4
ADAS13	Random Forest	Ethnicity, Race, Education, Gender, Marital status, Z, Tau, PTau, APOE4, FDG, AV45, ADAS13
CDRSB	Gradient Boosting	Ethnicity, Race, Education, Gender, Marital status, Z, Tau, PTau, APOE4, FDG, AV45, ADAS13
MMSE	Gradient Boosting	Ethnicity, Race, Education, Gender, Marital status, Z, Tau, PTau, APOE4, FDG, AV45, ADAS13