Learning Causally Predictable Outcomes from Psychiatric Longitudinal Data

Eric V. Strobl

Departments of Biomedical Informatics & Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania 15206, United States of America

Causal inference in longitudinal biomedical data remains a central challenge, especially in psychiatry, where symptom heterogeneity and latent confounding frequently undermine classical estimators. Most existing methods for treatment effect estimation presuppose a fixed outcome variable and address confounding through observed covariate adjustment. However, the assumption of unconfoundedness may not hold for a fixed outcome in practice. To address this foundational limitation, we directly optimize the outcome definition to maximize causal identifiability. Our DEBIAS (Durable Effects with Backdoor-Invariant Aggregated Symptoms) algorithm learns non-negative, clinically interpretable weights for outcome aggregation, maximizing durable treatment effects and empirically minimizing both observed and latent confounding by leveraging the time-limited direct effects of prior treatments in psychiatric longitudinal data. The algorithm also furnishes an empirically verifiable test for outcome unconfoundedness. DEBIAS consistently outperforms state-of-the-art methods in recovering causal effects for clinically interpretable composite outcomes across comprehensive experiments in depression and schizophrenia. R code is available at `github.com/ericstrobl/DEBIAS`.

Keywords: causal inference, psychiatry, longitudinal data, outcome learning, latent confounding, empirical unconfoundedness, composite outcomes

1. Introduction

Causal inference seeks to identify cause-and-effect relationships from data, enabling scientific discovery and effective intervention design.[1,2] In psychiatry, most causal knowledge has come from randomized clinical trials (RCTs), but RCTs are often infeasible or unethical to perform. For example, we cannot randomize individuals to different income levels to study the effects of poverty on mental health. Meanwhile, rich longitudinal observational datasets that provide far greater temporal and phenotypic resolution than cross-sectional studies[3] or electronic health records[4] are increasingly accessible through protected repositories in psychiatry. However, inferring causality from such data remains difficult in the absence of randomized assignment.

One of the primary obstacles to causal inference from observational data is latent confounding,[1] where unmeasured variables $\boldsymbol{C}$ may influence both treatment T_2 and outcomes $(\boldsymbol{Y}_3, \ldots, \boldsymbol{Y}_m)$, biasing causal effect estimates (Figure 1 (a)). Here, subscripts denote time steps in the longitudinal data, so T_1 and T_2 refer to treatments at different times, and $\boldsymbol{Y}_i$ represents the vector of *multiple* clinical rating scale items measured at time i. While standard practice adjusts for observed confounders $\boldsymbol{X}$ in regression models, this ignores clinical

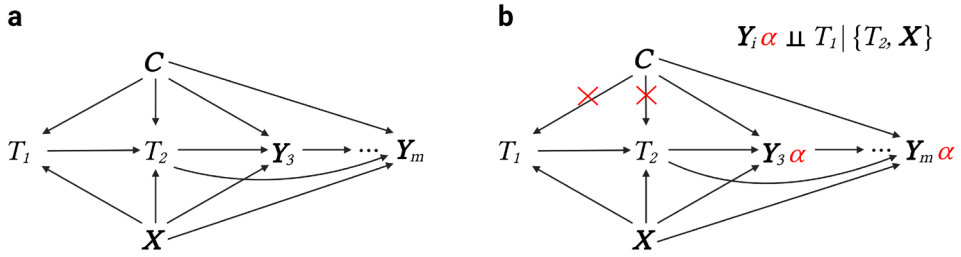

Fig. 1. **Main idea.** (a) Causal diagram where T_2 denotes current treatment assignment, T_1 denotes historical treatment assignment (not necessarily representing the same treatments as in T_2), each $\boldsymbol{Y}_i$ represents multiple items of a clinical rating scale, $\boldsymbol{X}$ indicates observed confounders available for adjustment, and $\boldsymbol{C}$ denotes latent confounders. In this structure, T_1 has time-limited direct effects: its influence on $\boldsymbol{Y}_3, \ldots, \boldsymbol{Y}_m$ is entirely indirect – mediated through T_2 – with no direct arrows from T_1 to the later outcomes. Moreover, $\boldsymbol{Y}_i$ is either unweighted or combined in a non-optimized manner, such as using total severity scores. (b) We learn non-negative weights $\alpha \geq 0$ to form the weighted severity scores $\boldsymbol{Y}_3\alpha, \ldots, \boldsymbol{Y}_m\alpha$ such that $\boldsymbol{Y}_i\alpha \perp\!\!\!\perp T_1 \mid \{T_2, \boldsymbol{X}\}$ for each time step. This outcome projection removes the statistical influence of spurious backdoor associations between T_2 (as well as T_1) and the outcomes, as indicated by the red crosses.

knowledge that many historical treatments T_1 have only *time-limited direct causal effects* and thus do not directly impact later outcomes[a]. Instead, any effect of T_1 on outcomes $\boldsymbol{Y}_3, \ldots, \boldsymbol{Y}_m$ occurs indirectly – mediated through T_2 – with no direct causal arrows from T_1 to the later outcomes (Figure 1 (a)). For instance, antidepressants do not directly affect symptom severity a year after discontinuation in active major depression.[5]

We propose to incorporate the prior knowledge of short-term treatment effects by searching for a non-negative weight vector α such that the aggregated outcome $\boldsymbol{Y}_i\alpha$ is conditionally independent of past treatment T_1, given current treatment T_2 and $\boldsymbol{X}$, for all time steps. The non-negativity constraint on α ensures that $\boldsymbol{Y}_i\alpha$ remains a clinically interpretable *severity score*, preserving the additive contribution of each symptom without introducing cancellation effects. Moreover, in this longitudinal setting, conditioning on the current treatment T_2 activates the collider on the path $T_1 \rightarrow T_2 \leftarrow \boldsymbol{C} \rightarrow \boldsymbol{Y}_i$ for any latent confounders $\boldsymbol{C}$ that affect T_2 and $\boldsymbol{Y}_i$. As a result, achieving conditional independence between T_1 and $\boldsymbol{Y}_i\alpha$ (after adjusting for T_2 and $\boldsymbol{X}$) eliminates the statistical influence transmitted along $\boldsymbol{C}$-mediated non-causal paths linking T_2 (and T_1) to each $\boldsymbol{Y}_i\alpha$, thereby enabling unbiased inference of the causal effect of T_2 on each $\boldsymbol{Y}_i\alpha$ (e.g., Figure 1 (b)). Applying the same α across all time points ultimately yields a single, durable, and empirically unconfounded severity score, enabling credible causal inference for a structured combination of items – even when such inference is infeasible for other combinations.

We specifically make the following contributions in this work:

[a]We relax this assumption in Supplementary Materials 8.1, where we also provide more technical discussion. Supplementary Materials link: https://github.com/ericstrobl/DEBIAS/tree/main/PDF

(1) We introduce a principled approach for learning outcome scores as non-negative combinations of clinical rating scale items, maximizing the correlation between current treatment and subsequent outcomes across time.
(2) We propose a novel regularization criterion that minimizes latent confounding bias by leveraging the time-limited direct effects of past treatments.
(3) We generalize the framework to extract not just a single score, but all unconfounded severity scores, each defined by a distinct non-negative weight vector and interpretable as a composite severity score.
(4) We instantiate the above concepts into a single algorithm, Durable Effects with Backdoor-Invariant Aggregated Symptoms (DEBIAS), which employs cross-validation to minimize latent confounding and maximize correlation.
(5) We demonstrate that DEBIAS effectively eliminates latent confounding and consistently outperforms existing algorithms in depression and schizophrenia.

Collectively, these advances enable robust causal inference from complex longitudinal datasets by learning *multiple* durable severity scores that are empirically unconfounded, even in the presence of latent sources of bias.

2. Related Work

We introduce the term *outcome learning* to describe algorithmic identification of the optimal set or combination of outcome variables for a given analytical task. Traditionally, investigators predefine a single clinically meaningful outcome for prediction,[6,7] which is effective when the target is clear, but problematic in complex diseases like depression or psychosis where symptoms are heterogeneous and not easily summarized by a single composite measure.

Recent algorithms in precision psychiatry attempt to learn multiple outcome measures tailored to specific analytical goals. For example, the Supervised Varimax algorithm constructs outcome measures that maximally differentiate psychiatric treatments in RCTs,[6,7] and the method has been modified to detect subtle differences between subgroups within patient populations.[8] The Sparse Canonical Outcome Regression (SCORE) algorithm, in contrast, optimizes multiple outcome measures for predictability rather than discrimination between groups.[9] Thus, different outcome learning approaches have emerged depending on whether the goal is to distinguish treatments, differentiate patient subgroups, or maximize predictability. Importantly, none of these existing methods explicitly optimize outcome measures to enable causal inference – a gap the DEBIAS algorithm is designed to fill.

Almost all existing causal inference methods address confounding by transforming or leveraging observed covariates but leave outcome variables fixed. For example, Difference-in-Differences relies on observed covariates to support the parallel trends assumption,[10] and Inverse Probability of Treatment Weighting (IPTW) uses propensity scores based on observed covariates to balance treatment groups.[11] In principle, IPTW can be combined with outcome-learning methods such as Non-Negative Canonical Correlation Analysis (NNCCA),[12] but such strategies still rest on the assumption that adjustment for observed covariates is sufficient to re-

move confounding. More recent approaches – including meta-learners[13,14] and causal forests[15] – extend this logic by using flexible machine learning models to control for high-dimensional covariates and estimate the conditional average treatment effect (CATE). Yet, all these methods implicitly assume that confounding can only be addressed by modifying the set of predictors, rather than the outcomes themselves.

Instrumental variable (IV) methods provide an alternative approach by leveraging external sources of variation to identify causal effects.[16] However, they depend on the strong assumption of no latent confounding linking the instrument (e.g., T_1) to the outcome – an assumption that is rarely plausible in complex observational settings. In practice, observed covariates and putative instruments are frequently inadequate to eliminate confounding for all outcome variables. Moreover, no empirical test can definitively establish that all sources of bias have been addressed by observed covariates.[1] Our approach instead relaxes the requirement by finding projections of the outcomes that minimize confounding, thereby retaining robustness even in the presence of residual latent confounders.

In summary, outcome learning methods in psychiatry have so far focused on RCT settings, subgroup differentiation, or improving predictivity rather than optimizing for causal inference from observational data. Meanwhile, causal inference algorithms have almost exclusively targeted the covariate space, neglecting the potential of outcome transformation. To our knowledge, DEBIAS is the first method to empirically learn outcome definitions specifically for causal inference within a unified, testable, and interpretable framework.

3. Assumptions

We adopt the potential outcomes framework to rigorously formalize causal inference in longitudinal data. While Figure 1 illustrates the case of two treatment time points (T_1 and T_2) for simplicity, our general framework allows the treatment of interest T_p to occur at any time step p, with the possibility of multiple prior treatments T_j for all $j < p$. Moreover, to unambiguously distinguish between temporal indices and hypothetical interventions in the outcomes, we denote by $\boldsymbol{Y}_i(t_p)$ the observed or potential outcome vector at time point i, under the hypothetical scenario in which treatment $T_p = t_p$ was assigned at time $p < i$. This notation enables a clear analysis of intervention effects across time.

Our methodology is predicated upon three standard assumptions from the causal inference literature:[1]

(1) **Consistency**: If a unit receives treatment t_p, then the observed outcome at time i coincides with the potential outcome under t_p; that is, $\boldsymbol{Y}_i = \boldsymbol{Y}_i(t_p)$ when $T_p = t_p$.
(2) **Stable Unit Treatment Value Assumption (SUTVA)**: The potential outcomes for any unit are unaffected by the treatment assignments of other units (i.e., there is no interference), and each treatment condition is consistently and uniquely defined.
(3) **Positivity**: The conditional density of treatment assignment is strictly positive; that is, $p(T_p = t_p \mid T_1, \boldsymbol{X}) > 0$ for all $(T_1, \boldsymbol{X}, t_p)$ within the support of the data. Similarly, $p(T_j = t_j \mid T_p, \boldsymbol{X}) > 0$ for all $j < p$ and all $(T_p, \boldsymbol{X}, t_j)$ in the support.

Most existing studies also assume *unconfoundedness* (ignorability), where $\boldsymbol{Y}_i(t_p)\alpha \perp\!\!\!\perp T_p|\boldsymbol{X}$

for some fixed weight vector $\alpha \geq 0$ and all t_p. For example, $\boldsymbol{Y}_i(t_p)\alpha$ with $\alpha = 1$ corresponds to the total score at time point i that would be observed if $T_p = t_p$. However, in practice, a potential outcome $\boldsymbol{Y}_i(t_p)\alpha$ may remain probabilistically dependent on T_p given $\boldsymbol{X}$ due to confounding from latent variables like $\boldsymbol{C}$ (Figure 1 (a)).[17] The dependence can also extend to earlier treatments T_j for all $j < p$ because the factors influencing earlier treatment assignments frequently continue to affect the assignment of treatment at the time point of interest.

We forego the standard unconfoundedness assumption and instead propose to *learn* α such that the antecedent in the following assumption holds:

(4) **Projected Unconfoundedness**: If there exists $\alpha \geq 0$ such that $\boldsymbol{Y}_i(T_p)\alpha \perp\!\!\!\perp \{T_1, \ldots, T_{p-1}\} | T_p \cup \boldsymbol{X}$ for all $i > p$, then $\boldsymbol{Y}_i(t_p)\alpha \perp\!\!\!\perp T_p | \boldsymbol{X}$ for all $i > p$ and all t_p.

This condition is justified by the temporal and graphical structure of confounding in longitudinal data. Specifically, latent factors $\boldsymbol{C}$ – such as historical severity or health insurance status – often exert influence on historical treatments $T_1, \ldots, T_{p-1}$ and the treatment of interest T_p, as depicted in Figure 1 (a). Moreover, historical treatments frequently have a direct causal effect on T_p but are time-limited without direct effects on $\boldsymbol{Y}_i$ for any $i > p$. This induces a graphical structure where T_p functions as a collider on backdoor paths of the form $T_j \to T_p \leftarrow \boldsymbol{C} \to \boldsymbol{Y}_i\alpha$ for some $j < p$. By choosing α such that $\boldsymbol{Y}_i(T_p)\alpha \perp\!\!\!\perp \{T_1, \ldots, T_{p-1}\} | T_p \cup \boldsymbol{X}$, we eliminate statistical dependence induced by latent confounders, thereby removing bias from unmeasured confounding along non-causal paths (Figure 1 (b)). Consequently, the standard unconfoundedness condition is restored for the specific learned outcome $\boldsymbol{Y}_i\alpha$, enabling unbiased causal effect estimation of T_p on $\boldsymbol{Y}_i\alpha$ for each time point $i > p$. See Supplementary Materials 8.1 for an even more general treatment that requires time-limited direct effects for only a subset of items.

Notice that our outcome-centric strategy departs from conventional approaches that presume no unmeasured confounding for combinations of all outcome items, such as total severity scores or remission rates – a strong and generally untestable assumption. Our approach algorithmically learns a non-negative weight vector α that is typically sparse in practice, so the resulting outcome score depends only on a small subset of items. Accordingly, the projected unconfoundedness condition and resulting causal guarantees apply specifically to this empirically identified subset of outcomes, rather than to composite measures constructed a priori without regard to confounding structure. This yields both improved interpretability and greater robustness to latent confounding, as the learned outcome focuses on those items for which credible causal inference is feasible.

4. Maximizing Correlation to Causally Distinguish Patients

In causal inference, larger values of $\boldsymbol{Y}_i(T_p)$ are conventionally interpreted as improvement, whereas most clinical rating scales assign higher scores to greater symptom severity. To align these conventions, we multiply each rating item by -1 so that higher values indicate improvement. The canonical quantity for binary treatment assignment with an outcome composite is then the average treatment effect (ATE):

$$\text{ATE}_\alpha = \mathbb{E}[\boldsymbol{Y}_i(1)\alpha] - \mathbb{E}[\boldsymbol{Y}_i(0)\alpha],$$

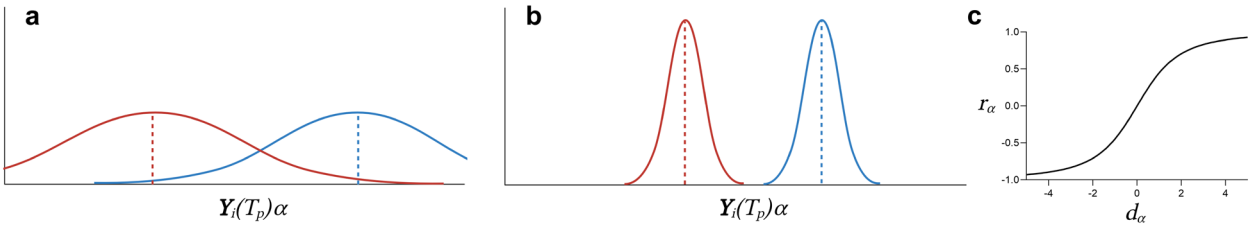

Fig. 2. **Comparison of effect size measures.** (a) The ATE quantifies the raw difference in group means (red vs. blue), without accounting for within-group variability or distributional overlap. (b) Cohen's d and Pearson's correlation coefficient both capture the separability of groups by standardizing the mean difference using the pooled within-group standard deviation; separability increases as within-group variance decreases and group means diverge. (c) Although Cohen's d and correlation are monotonically related under binary treatment, only correlation extends naturally to ordinal (more than two levels) or continuous treatment variables, making it more broadly applicable.

where $\boldsymbol{Y}_i(1)\alpha$ and $\boldsymbol{Y}_i(0)\alpha$ denote the potential outcomes under treatment and control, projected via $\alpha \geq 0$. The goal is to identify α that maximizes the treatment effect along a learned score. However, maximizing the ATE alone does not account for the variance of $\boldsymbol{Y}_i(T_p)\alpha$ within each treatment group. As a result, even if the group means are well separated, the individual outcomes may still overlap substantially due to high within-group variance (Figure 2 (a)), rendering it difficult to predict which patients will benefit from treatment in practice.

The separation between treatment assignments must be meaningful relative to the within-group variability of outcomes. Let $p = P(T = 1)$ denote the proportion of individuals assigned to the treatment group, and $1-p = P(T = 0)$ the proportion in the control group. One classical approach to quantifying such standardized separation is to maximize Cohen's d:

$$d_\alpha = \frac{\text{ATE}_\alpha}{\sqrt{p\text{Var}(\boldsymbol{Y}_i(1)\alpha) + (1-p)\text{Var}(\boldsymbol{Y}_i(0)\alpha)}},$$

where the denominator represents the pooled within-group variance.[18] Cohen's d thus quantifies the magnitude of group separation relative to internal variability. Maximizing d aligns the learned outcome score with maximal distinguishability of treatment groups (Figure 2 (b)).

Cohen's d is unfortunately limited to binary treatment assignment, but the measure admits a monotonic transformation into Pearson's correlation coefficient:[19]

$$r_\alpha = \frac{\sqrt{p(1-p)} \cdot d_\alpha}{\sqrt{1 + p(1-p) \cdot d_\alpha^2}}.$$

This is a bijective mapping that preserves the ranking of outcome scores (Figure 2 (c)), so maximizing d_α necessarily maximizes r_α and vice versa. Importantly, while Cohen's d is limited to binary comparisons, Pearson's correlation generalizes naturally to ordinal and continuous treatment variables (e.g., income levels or medication dosages), thereby offering a more versatile and broadly applicable objective. Accordingly, we adopt the correlation coefficient as our primary optimization criterion, given its capacity to maximally distinguish between treatment conditions regardless of whether the treatment variable is discrete or continuous.

5. Algorithm

5.1. *First Learned Severity Score*

We now describe the DEBIAS algorithm. The algorithm begins by constructing the first non-negatively weighted outcome score that is durably responsive to treatment and robust to confounding. As detailed in Section 4, our objective is to maximize the correlation between T_p and each outcome measure $\boldsymbol{Y}_i\alpha$ for $i > p$. This leads to the following optimization problem:

$$\underset{\alpha\geq 0,\|\alpha\|_1=1}{\arg\max} \sum_{i=p+1}^{m} \Big[\underbrace{\operatorname{cor}(\boldsymbol{Y}_i\alpha, T_p|T_1, \boldsymbol{X})}_{(a)} - \underbrace{\frac{\lambda}{p-1}\sum_{j=1}^{p-1}\operatorname{cor}^2(\boldsymbol{Y}_i\alpha, T_j|T_p, \boldsymbol{X})}_{(b)} \Big], \tag{1}$$

where we enforce $\|\alpha\|_1 = 1$ to ensure uniqueness of the solution.

The **main correlation** term in (a) seeks to maximize the durable partial correlation between the learned outcome and the target treatment across all future time points, adjusting for $\boldsymbol{X}$ to account for observed confounders. We also adjust for the most distant prior treatment T_1 to ensure the algorithm remains robust when treatment assignment is constant over time; without this step, the confounding penalty in (b) – explained below – becomes vacuously zero and cannot control for confounding. Including T_1 as a covariate in (a) ensures the correlation objective is also zero in such cases, preventing spurious associations when there is no temporal variation in treatment.

The **confounding penalty** in (b) targets association arising from latent confounding. Here, $\sum_{j=1}^{p-1}\operatorname{cor}^2(\boldsymbol{Y}_i\alpha, T_j|T_p, \boldsymbol{X}) = 0$ is a necessary condition for $\boldsymbol{Y}_i(T_p)\alpha \perp\!\!\!\perp \{T_1, \ldots, T_{p-1}\}|T_p \cup \boldsymbol{X}$ in the projected unconfoundedness condition (Section 3). Minimizing this penalty thus addresses confounding and also serves as a regularizer in finite samples: if the empirical squared partial correlation is large, it counteracts over-optimistic correlations in (a). Notably, this approach uses a linear approximation to conditional independence, enabling empirical unconfoundedness and statistical hypothesis testing of latent confounding; while not equivalent to full conditional independence, partial decorrelation is often sufficient for practical purposes in high-dimensional observational data.

To balance predictive accuracy and robustness, we optimize λ by cross-validation over a predefined grid and select the value that maximizes the correlation in (a), *provided* that the p-value associated with the squared correlation in (b) exceeds a specified threshold (default 0.05), meaning we cannot reject the null hypothesis of no partial association. In this way, the penalty both mitigates confounding and regularizes the model to avoid overfitting.

5.2. *Sequential Extraction of Severity Scores*

The optimization problem in Expression (1) only produces a single set of non-negative weights, α_1. However, our goal is to identify all α that maximize persistent correlation with treatment while minimizing confounding. We thus introduce an additional **diversity-promoting penalty** term (c) that encourages each new score to be as clinically distinct as possible from

all previously extracted scores:

$$\underset{\alpha\geq 0,\|\alpha\|_1=1}{\arg\max} \sum_{i=p+1}^{m} \Bigg[\mathrm{cor}(\boldsymbol{Y}_i\alpha, T_p|T_1, \boldsymbol{X}) - \frac{\lambda}{p-1}\sum_{j=1}^{p-1} \mathrm{cor}^2(\boldsymbol{Y}_i\alpha, T_j|T_p, \boldsymbol{X}) \\ \underbrace{- \frac{1}{K-1}\sum_{k=1}^{K-1}\left(\frac{\alpha_k^T M_i \alpha}{\sqrt{\alpha_k^T M_i \alpha_k}\sqrt{\alpha^T M_i \alpha}}\right)}_{(c)} \Bigg], \tag{2}$$

where $K \geq 2$ denotes the total number of learned scores including the current score under optimization. The diversity-promoting penalty measures the correlation-weighted cosine similarity between the current score α and each previously discovered score α_k, where M_i is the correlation matrix of the outcome variables $\boldsymbol{Y}_i$. This similarity ranges from -1 to 1 like the main correlation term, so the penalty is naturally scale-matched to the main objective and does not require a separate hyperparameter for balancing its influence.

We intentionally do *not* square the correlation-weighted cosine similarity so that the optimization procedure favors negative similarity values when permitted by the data and the non-negativity constraint. Recall that (1) the individual items in $\boldsymbol{Y}_i$ represent *pathological* symptoms, where higher values indicate worse clinical severity, and (2) each item was sign-flipped in Section 4 so that higher values instead denote clinical *improvement*, aligning with the causal interpretation of treatment effects. Within this transformed space, all weights in α are constrained to be non-negative, ensuring that learned outcome scores reflect additive improvements across symptoms without cancellation. These two conditions impose a directional structure: outcome scores represent strictly non-negative combinations of improvements in underlying pathologies. Consequently, a negative correlation-weighted cosine similarity between scores becomes possible only when the symptom correlation matrix M_i contains sufficiently negative off-diagonal elements – i.e., when some symptoms improve inversely with others. In such cases, the algorithm can extract new scores that are anti-aligned with prior ones within the non-negative feasible region, capturing distinct, clinically interpretable axes of symptom variation.

For example, manic and depressive symptoms in bipolar disorder are clearly clinically distinct yet strongly negatively correlated. In this context, DEBIAS may recover one composite score weighted toward manic symptoms and another toward depressive symptoms, both with non-negative weights and exhibiting negative correlation-weighted cosine similarity. In contrast, minimizing the *squared* correlation-weighted cosine similarity only enforces orthogonality, which can result in a new score that emphasizes peripheral symptoms – such as chronic irritability, distractibility, or somatic complaints – rather than capturing the strong negative correlation between core mood domains. The diversity-promoting penalty thus directly exploits the negative correlation structure to extract outcome scores that are clinically distinct, rather than conflating mathematical orthogonality with clinical distinctiveness.

We ultimately solve the optimization problem in Expressions (1) and (2) by projected gradient ascent using backtracking line search with the Armijo condition[20] (Supplementary Materials 8.2). Computational complexity analysis revealed that DEBIAS scales linearly with

the number of subjects and time points, but quadratically with the number of outcome items, covariates, and summary scores (Supplementary Materials 8.3).

6. Results

6.1. *Comparator Algorithms*

We ran DEBIAS with 5-fold cross-validation evaluating λ over the grid $0, 1, \ldots, 10$. We extracted $s = 3$ severity scores. We compared DEBIAS against the following algorithms:

(1) **Non-Negative Canonical Correlation Analysis (NNCCA) with IPTW**:[12] NNCCA estimates time point-specific outcome weights by independently maximizing the IPTW-weighted correlation at each time point: $\max_{\alpha \geq 0} \mathrm{cor}_w(\boldsymbol{Y}_i\alpha, T_p)$. NNCCA further applies a weighted deflation procedure to iteratively extract approximately orthogonal sets of non-negative outcome weights.

(2) **R-Learner with Extreme Gradient Boosting (RBoost)**:[13] RBoost first estimates the baseline outcome function and the propensity score using XGBoost,[21] and then computes residualized outcomes and residualized treatment assignments. Next, the algorithm fits a model for the CATE by regressing the residualized outcomes onto the residualized treatments using XGBoost. This orthogonalization-based procedure improves robustness to model misspecification and enables efficient use of modern machine learning methods for CATE estimation. We used 3 folds for cross-fitting and cross-validation, 100 trees, 3 search rounds, and 5 early stopping rounds.

(3) **Causal Forests (CF)**:[15] Causal forests estimate the CATE by building an ensemble of honest causal trees, each of which partitions the feature space into locally estimate treatment effects. Honesty is ensured by splitting the data into separate subsamples for tree construction and effect estimation. Averaging over the ensemble yields a flexible, nonparametric estimator with valid asymptotic inference. We learned 2000 trees and otherwise used default parameters.

We further assessed DEBIAS against **three ablated variants**: (i) replacing the correlation objective with mean squared error (MSE), (ii) removing the confounding penalty in Expression (2) ($\lambda = 0$), and (iii) replacing the correlation objective and removing the confounding penalty. Notably, the variant with the confounding penalty removed can be considered a variant of the SCORE algorithm,[9] except that we employ a diversity-promoting penalty (correlation-weighted cosine similarity) instead of deflation to extract multiple scores when treatment is univariate (a setting where deflation fails). All ablated variants used the same cross-validation protocol, λ grid, and number of scores as DEBIAS.

DEBIAS differs from existing comparator methods along several key dimensions. First, while all comparators adjust only for observed confounding, DEBIAS uniquely leverages the temporal structure of longitudinal data and the time-limited nature of treatments to additionally reduce the influence of latent confounding. Second, DEBIAS is the only method capable of extracting multiple interpretable severity scores in the setting of binary treatment assignment, owing to its use of a diversity-promoting penalty rather than deflation. Although NNCCA can also produce multiple scores, these do not correspond to strict severity scores and

lack a clear causal interpretation. Third, DEBIAS estimates non-negative outcome weights by jointly utilizing all future time points, thereby borrowing statistical strength across the outcome trajectory. In contrast, NNCCA, RBoost, and CF estimate effects or representations independently at each time point, precluding information sharing across time. Taken together, DEBIAS is the only approach that integrates temporal consistency and confounding control – both observed and latent – to derive outcome scores that are both longitudinally predictive and causally valid.

6.2. *Evaluation Metrics*

As detailed in Section 4, our primary goal is to assess how effectively each method differentiates patients based on treatment assignment, motivating the use of **partial correlation** $\text{cor}(\widehat{m}_i, T_2 \mid T_1, \boldsymbol{X})$ as the main performance metric for each time point $i > 2$; here, $\widehat{m}_i$ represents the learned outcome at time i – specifically, $\boldsymbol{Y}_i\alpha$ for DEBIAS and NNCCA, or the estimated CATE of the total severity score for RBoost and CF. However, naive correlation is only meaningful in the absence of confounding. To address this limitation, we additionally evaluate all algorithms using the **confounding p-value** derived from the squared partial correlation, $\text{cor}^2(\widehat{m}_i, T_1 \mid T_2, \boldsymbol{X})$. This squared partial correlation quantifies the residual association between the model output and historical exposures, reflecting the extent of any remaining confounding bias. We further evaluated each method by the **sum of confounded coefficients**, or the sum of the coefficients associated with outcome items known to suffer from latent confounding; methods that achieve a low total are preferred, as they minimize reliance on confounded information. Finally, we compare all algorithms in terms of **run-time** to evaluate computational efficiency.

We evaluated all metrics using 1,000 bootstrap replicates, where we sampled with replacement such that approximately 63.2% of the samples acted as the training set and the remaining 36.8% as the held-out test set in each replicate. We computed all performance measures exclusively on the test samples. For all comparisons of relative algorithm performance, we applied two-sided paired t-tests and interpreted significance at a Bonferroni-corrected threshold of $0.05/u$, where u denotes the total number of comparator algorithms.

6.3. *Adolescent Depression*

We evaluated each algorithm's ability to maximize correlation and minimize confounding using data from the Treatment for Adolescents with Depression Study (TADS; ClinicalTrials.gov ID NCT00006286), an RCT assessing the efficacy of cognitive-behavioral therapy (CBT), fluoxetine, their combination, and placebo in adolescent major depressive disorder.[22] We excluded the placebo group since patients received placebo for only 8 weeks before transitioning to standard care, whereas 323 patients in the other groups maintained their assigned treatments for a 36-week duration.

NNCCA, CF, and RBoost are canonically applied to binary treatment variables; therefore, treatment groups were dichotomized as medication (fluoxetine or combination therapy) versus CBT to facilitate a fair comparison across methods. The resulting medication indicator was used as the current treatment variable, $T_p = T_2$. The prior treatment indicator, T_1, was defined as a binary variable indicating whether the patient had received any mental health services

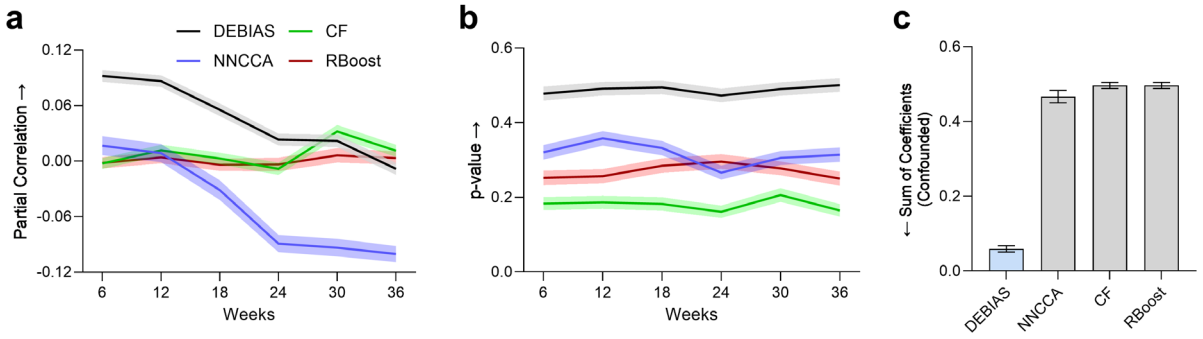

Fig. 3. **Accuracy results for the first score in the TADS dataset.** (a) DEBIAS achieved the highest correlation with outcomes from weeks 6-24, despite incorporating regularization to mitigate latent confounding bias. (b) DEBIAS also attained the largest mean p-value associated with the squared correlation coefficient, reflecting reduced confounding influence relative to competing methods. (c) The sum of the outcome coefficients assigned by DEBIAS to confounded items was also the smallest among all methods. Error bands and bars denote 95% confidence intervals of the mean.

prior to treatment assignment, as measured by the Child and Adolescent Services Assessment (CASA),[23] given its likely correlation with T_2. The outcome vector $\boldsymbol{Y}_i$ comprised the change from baseline in each item of the Children's Depression Rating Scale–Revised (CDRS-R).[24]

Although the original RCT design ensures perfect control of confounding, we introduced an artificial latent confounder, $\boldsymbol{C}$, sampled from a standard normal distribution. We then additively incorporated $\boldsymbol{C}$ into T_1, T_2, and a randomly selected subset of 5 to 12 items within each outcome vector $\boldsymbol{Y}_i$, using time-varying weights for each item and time point sampled uniformly from $[0.2, 1]$. We then re-binarized T_1 and T_2 by thresholding at the value 0.25. Crucially, this experimental setup provides ground-truth knowledge regarding which $\boldsymbol{Y}_i$ items are confounded to compute the sum of confounded coefficients metric. All analyses included age and biological sex as observed covariates in $\boldsymbol{X}$.

Figure 3 presents the primary results. DEBIAS achieved the highest correlation coefficients among all methods across the first 24 weeks (Figure 3 (a)). This superior performance reflects the enforcement of a consistent set of non-negative weights α, which facilitates information sharing across time points. In later weeks, all algorithms exhibited a decline toward zero or negative correlations, reflecting the true convergence of treatment effects among fluoxetine, CBT, and combination therapy.[22] We conclude that DEBIAS maximizes outcome predictability relative to all comparators.

DEBIAS also yielded the largest p-values associated with the squared correlation regularization term, indicating effective mitigation of latent confounding effects (Figure 3 (b)). Inspection of the outcome coefficients revealed that DEBIAS preferentially assigned the smallest weights to the confounded outcomes, providing further evidence of superior confounding control (Figure 3 (c)). Collectively, these results demonstrate that DEBIAS achieves both maximal correlation and minimal confounding in the TADS dataset. Ablation results against three variants in Supplementary Figure 6 confirmed that both correlation maximization and confounding minimization are essential for DEBIAS to achieve optimal performance, even for

the scores beyond the first. Finally, DEBIAS completed within 10 seconds on average in this dataset (Supplementary Figure 7).

6.4. *Chronic Schizophrenia*

We next evaluated each algorithm's ability to learn causally predictable outcomes in schizophrenia. We downloaded data from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE; NCT00014001), which was a large RCT comparing the effectiveness of antipsychotic medications in adults with chronic schizophrenia.[25] We restricted our analysis to 664 subjects treated with olanzapine (coded as $T_2 = T_p = 1$) or quetiapine ($T_2 = 0$), in order to maximize the contrast in antipsychotic efficacy for core psychotic symptoms; olanzapine is widely regarded as one of the most effective antipsychotics, whereas quetiapine is generally considered among the least effective.[26] We set T_1 to the number of prior inpatient psychiatric visits, since this variable influences antipsychotic prescribing in patients with schizophrenia.[27] We introduced time-varying confounding by additively including a standard normal latent confounder in T_1, T_2, and a uniformly selected subset of 15 to 25 items within each outcome vector $\boldsymbol{Y}_i$. Each outcome vector represented the change from baseline for the 30 individual items of the Positive and Negative Syndrome Scale (PANSS) measured across 15 months.[28]

We present the results in Figure 4. DEBIAS once again achieved the highest correlation with treatment assignment among all competing methods (Figure 4 (a)). The algorithm also improved confounding control, as evidenced by a markedly higher p-value for the confounding penalty and a much lower total sum of non-negative coefficients assigned to confounded outcome items. These results indicate that DEBIAS consistently maximized out-of-sample correlation and minimized confounding, paralleling its performance in the TADS dataset. Ablation analyses further demonstrated that both the correlation maximization objective and the confounding penalty were necessary to achieve optimal results across multiple extracted scores

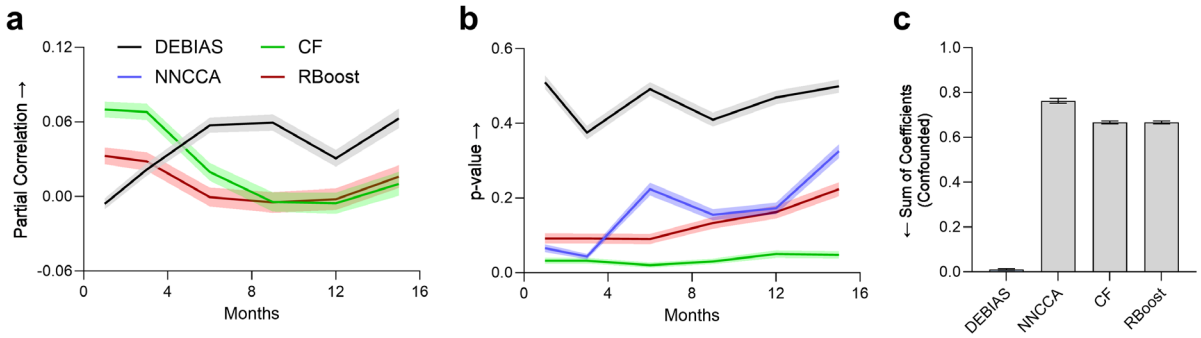

Fig. 4. **Accuracy results for the first score in the CATIE dataset.** (a) DEBIAS attained the highest correlation with treatment assignment as treatment effects diverged over time. The correlation achieved by NNCCA was consistently poor (below -0.20 on average) and therefore omitted from the plot to focus on the competitive algorithms. (b) Compared to all other methods, DEBIAS exhibited substantially superior control of confounding, as reflected by higher p-values for the squared correlation coefficient. (c) DEBIAS also assigned markedly lower non-zero coefficients to confounded outcome items, indicating greater selectivity and robustness to confounding.

(Supplementary Figure 8). Finally, DEBIAS completed its computation within 20 seconds on this dataset (Supplementary Figure 9).

7. Discussion

We introduced the DEBIAS algorithm, which learns combinations of symptom severity items that maximize predictability and minimize latent confounding bias, enabling robust causal inference from psychiatric longitudinal data. Unlike conventional approaches that rely on fixed, user-defined outcomes and assume adequate adjustment for confounding, DEBIAS is the first method to algorithmically identify outcome measures empirically amenable to valid causal inference even with latent confounders. Experiments in adolescent depression and chronic schizophrenia demonstrated that DEBIAS recovers causally predictable severity scores by leveraging the time-limited direct effects of past treatments, a critical advantage when comprehensive covariate collection is infeasible. DEBIAS thus represents a paradigm shift, showing that identifying suitable outcomes for causal inference can be just as important as adjusting for observed confounders in complex real-world settings.

Our work has several limitations. First, DEBIAS is purposefully designed as a linear method to emphasize foundational conceptual advances in causal inference rather than algorithmic complexity. While linear decorrelation is often practically effective for removing confounding, it does not guarantee conditional independence in the strict mathematical sense. Thus, extending DEBIAS to nonlinear settings, where more general forms of dependence can be addressed, remains important for future research. Second, we currently treat T_1 as a historical treatment with time-limited direct effects, an assumption justified for most psychiatric medications. The assumption may be violated for interventions such as CBT, which is believed to exert long-lasting effects by imparting enduring skills.[29] However, in practice, the presence of persistent symptoms among non-remitting patients included in clinical longitudinal datasets suggests that any enduring direct effects of prior therapy are likely minimal in this cohort. Our experimental results further support this claim: DEBIAS consistently achieved the lowest levels of confounding across all methods, despite not explicitly manipulating or blocking potential pathways of the form $T_1 \to \boldsymbol{Y}_3, \ldots, \boldsymbol{Y}_m$. This empirical finding suggests that the time-limited direct effects assumption and corresponding regularization strategy are reasonable and effective. Time-limited direct effects on all outcome items are also not required for our identification strategy, as discussed in Supplementary Materials 8.1. Third, outcome learning opens avenues for novel approaches to confounding adjustment and offers potential applications in domains beyond psychiatry, such as the construction of composite endpoints in endocrinology or the definition of multi-omic phenotypes in oncology. Systematically exploring these alternatives represents a particularly promising avenue for future research.

In summary, DEBIAS represents a substantial advance in outcome learning for causal inference, delivering clinically interpretable measures from data readily available in psychiatric research. This enables more accurate and unbiased causal inference for important real-world applications that may ultimately improve mental healthcare.

References

1. G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press, 2015).
2. P. Spirtes, C. Glymour and R. Scheines, *Causation, Prediction, and Search*, 2nd edn. (MIT press, 2000).
3. D. Hedeker and R. D. Gibbons, *Longitudinal Data Analysis* (John Wiley & Sons, 2006).
4. B. A. Goldstein, A. M. Navar, M. J. Pencina and J. P. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, *Journal of the American Medical Informatics Association: JAMIA* **24**, p. 198 (2016).
5. A. J. Rush, M. H. Trivedi, S. R. Wisniewski, A. A. Nierenberg, J. W. Stewart, D. Warden, G. Niederehe, M. E. Thase, P. W. Lavori, B. D. Lebowitz *et al.*, Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a star* d report, *American Journal of Psychiatry* **163**, 1905 (2006).
6. E. V. Strobl and S. Kim, Learning outcomes that maximally differentiate psychiatric treatments, *medRxiv* , 2024 (2024).
7. E. V. Strobl, Consistent differential effects of bupropion and mirtazapine in major depression, *Journal of Affective Disorders* , p. 119551 (2025).
8. E. V. Strobl, Unique behavior profiles that specify mental distress in autism, *medRxiv* , 2025 (2025).
9. E. V. Strobl, Predicting the predictable in the psychiatric high risk, *Proceedings of the Machine Learning for Healthcare Conference* (2025, accepted).
10. O. Ashenfelter and D. Card, Using the longitudinal structure of earnings to estimate the effect of training programs, *The Review of Economics and Statistics* **67**, 648 (1985).
11. J. M. Robins, M. A. Hernan and B. Brumback, Marginal structural models and causal inference in epidemiology, *Epidemiology* **11**, 550 (2000).
12. C. Sigg, B. Fischer, B. Ommer, V. Roth and J. Buhmann, Nonnegative cca for audiovisual source separation, 253 (2007).
13. X. Nie and S. Wager, Quasi-oracle estimation of heterogeneous treatment effects, *Biometrika* **108**, 299 (2021).
14. S. R. Künzel, J. S. Sekhon, P. J. Bickel and B. Yu, Metalearners for estimating heterogeneous treatment effects using machine learning, *Proceedings of the National Academy of Sciences* **116**, 4156 (2019).
15. S. Wager and S. Athey, Estimation and inference of heterogeneous treatment effects using random forests, *Journal of the American Statistical Association* **113**, 1228 (2018).
16. J. D. Angrist, G. W. Imbens and D. B. Rubin, Identification of causal effects using instrumental variables, *Journal of the American Statistical Association* **91**, 444 (1996).
17. T. S. Richardson and J. M. Robins, Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality, *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper* **128**, p. 2013 (2013).
18. J. Cohen, *Statistical power analysis for the behavioral sciences* (Routledge, 2013).
19. M. W. Lipsey and D. B. Wilson, *Practical meta-analysis.* (SAGE publications, Inc, 2001).
20. L. Armijo, Minimization of functions having lipschitz continuous first partial derivatives, *Pacific Journal of Mathematics* **16**, 1 (1966).
21. T. Chen and C. Guestrin, Xgboost: A scalable tree boosting system, 785 (2016).
22. J. March, S. Silva, S. Petrycki, J. Curry, K. Wells, J. Fairbank, B. Burns, M. Domino, S. McNulty, B. Vitiello *et al.*, Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: Treatment for adolescents with depression study (tads) randomized controlled trial., *JAMA* **292**, 807 (2004).

23. B. H. Ascher, E. M. Z. Farmer, B. J. Burns and A. Angold, The child and adolescent services assessment (casa) description and psychometrics, *Journal of Emotional and Behavioral Disorders* **4**, 12 (1996).
24. E. O. Poznanski, J. A. Grossman, Y. Buchsbaum, M. Banegas, L. Freeman and R. Gibbons, Children's depression rating scale–revised, *Journal of the American Academy of Child Psychiatry* (1995).
25. J. A. Lieberman, T. S. Stroup, J. P. McEvoy, M. S. Swartz, R. A. Rosenheck, D. O. Perkins, R. S. Keefe, S. M. Davis, C. E. Davis, B. D. Lebowitz *et al.*, Effectiveness of antipsychotic drugs in patients with chronic schizophrenia, *New England Journal of Medicine* **353**, 1209 (2005).
26. S. Leucht, A. Cipriani, L. Spineli, D. Mavridis, D. Örey, F. Richter, M. Samara, C. Barbui, R. R. Engel, J. R. Geddes *et al.*, Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis, *The Lancet* **382**, 951 (2013).
27. J. Tiihonen, E. Mittendorfer-Rutz, M. Majak, J. Mehtälä, F. Hoti, E. Jedenius, D. Enkusson, A. Leval, J. Sermon, A. Tanskanen *et al.*, Real-world effectiveness of antipsychotic treatments in a nationwide cohort of 29 823 patients with schizophrenia, *JAMA Psychiatry* **74**, 686 (2017).
28. S. R. Kay, A. Fiszbein and L. A. Opler, The positive and negative syndrome scale (panss) for schizophrenia, *Schizophrenia Bulletin* **13**, 261 (1987).
29. S. D. Hollon, M. O. Stewart and D. Strunk, Enduring effects for cognitive behavior therapy in the treatment of depression and anxiety, *Annu. Rev. Psychol.* **57**, 285 (2006).
30. M. T. Treadway and D. H. Zald, Reconsidering anhedonia in depression: lessons from translational neuroscience, *Neuroscience & Biobehavioral Reviews* **35**, 537 (2011).