# Causal Learning with Large-Scale Propensity Scores to Predict Treatment Outcomes: A Study of Bipolar disorder in Adults with Attention-deficit/hyperactivity disorder

**Junhyuk Chang[1,2,3], Dong Yun Lee[4], Rae Woong Park[1,2,3,4]**
**[1]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea**
**[2]Center for Biomedical Informatics Research, Ajou University Medical Center, Suwon, Korea**
**[3]BK21 R&E Initiative for Advanced Precision Medicine, Suwon, Korea**
**[4]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea**

## Background

Methylphenidate (MPH) is widely used as a first-line treatment for patients with attention-deficit/hyperactivity disorder (ADHD) with depression. However, several case reports and observational studies have suggested that MPH may induce manic episodes in certain individuals[1,2]. In patients with depression, the emergence of a manic episode may lead to a transition to bipolar disorder (BD), that requires different therapeutic approaches and is generally associated with a less favorable prognosis. While research on the potential risk of MPH-induced mania is ongoing, there remains a lack of studies focusing on specific high-risk subpopulations or on individual-level treatment response.

The causal machine learning method is able to estimate treatment effects on individual patients by calculating average treatment effects and non-parametrically adjusting for potential confounding predictors inherent in observational data.[3,4] These techniques predict outcomes based on complex variable interactions and identify heterogeneity in treatment effects, enhancing the robustness and reliability of our findings. We aim to analyze the treatment effect of administrating MPH on bipolar disorder occurrence in ADHD with depression patients with a causal forest model using common data model (CDM) and large-scale propensity score method which is CDM methodology.

## Methods

In this study, we used the Health Insurance Review and Assessment Service - Attention Deficit/Hyperactivity Disorder (HIRA-ADHD) database, which was collected from January 1, 2016, to December 31, 2020, and contained ADHD patient data from nationwide claims data. The HIRA-ADHD database was converted to Observational Medical Outcomes Partnership-CDM. We defined MPH-used patients with an ADHD and depression diagnosis based on the following criteria: 1) aged greater or equal to 18; 2) patients with an ADHD and depression record; 3) patients without other anti-ADHD agents.

The target outcome in this study is the occurrence of bipolar disorder. For developing and validating the causal forest model, we divided the dataset into 70% for training and 30% for validation, ensuring the same outcome prevalence in both sets. We extracted patient baseline covariates to employ a large-scale propensity score utilizing the FeatureExtraction, OHDSI open-source package. Initial screening was conducted to exclude rare covariates by 10-fold cross-validation in the training sample using a logistic regression model generated from the binomial outcome.

A large-scale propensity score approach was used to adjust for differences in baseline covariates between MPH-treated and non-treated patients. Propensity scores (PS) were generated using random forests (RF) with screened covariates, and each patient was weighted using inverse-propensity weighting. With calculated PS and weights, the constructed causal forest model estimated the average treatment effect

(ATE) via a doubly robust method. This method adjusts for baseline covariate differences by combining weights based on MPH treatment probability and outcome regression using a random forest.[3,5]

We estimated treatment heterogeneity using the conditional ATE (CATE). Initially, CATEs were estimated in the 30% test sample by adopting a causal forest model trained from the training dataset. The validation dataset was divided into quintiles based on calculated CATEs, and ATEs were estimated within each quintile. To ensure statistical validity, ATEs were estimated using a doubly robust approach based on the targeted minimum loss estimation (TMLE) framework, which combines propensity score and outcome regression models. This approach aimed to verify the estimated ATE in each CATE quantile group has heterogeneity.

To identify characteristics of high and low CATE groups, we compared additional characteristics by analyzing the distribution of the top 15 variables based on variable importance from the causal forest model. Using the mean CATE of the validation dataset, we divided the groups into high and low CATE groups and analyzed whether there were statistically significant differences in the distribution of the top 15 variables between these groups.

**Results**

Among the total of 28,939 patients, 19,939 patients were prescribed MPH and 1,881 patients had occurrences of BD. After dividing the outcome prevalence equally into training and validation datasets, the training dataset included a total of 20,256 patients, of whom 13,939 were treated with MPH, and the validation dataset included a total of 8,683 patients, of whom 6,000 were treated with MPH. From these patients, we initially extracted a total of 4,608 baseline covariates, which were then reduced to 4,477 covariates after 10-fold cross-validation.

The ATE for the validation dataset is 0.018 [0.011-0.053]. Based on the CATE calculated in the validation dataset, the cut-off values for the quantile groups Q1 to Q5 were as follows: 0.014, 0.017, 0.018, 0.019, and 0.020. The validation dataset was divided equally into 5 quantile groups, which contained 1,736-1,737 patients.

Figure 1 represents the estimated ATE of the quantile groups. In order of increasing quantile, the calculated ATEs were 0.044, 0.036, 0.029, -0.002, and 0.015. Among ATE of quantile groups, the ATE of the Q1 to Q3 group is statistically significant. The results in the Q1 to Q3 group suggest that MPH may significantly affect bipolar disorder in this subgroups. The estimated heterogeneity among the Q1 to Q5

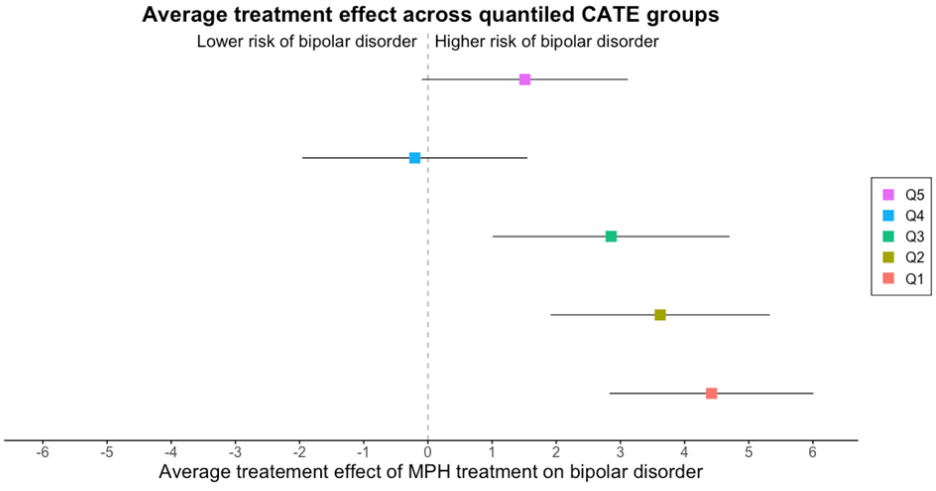


**Figure 1. Average treatment effect of quantile groups**

groups was statistically significant ($\chi^2_4$ = 17.98, p = 0.0012), indicating meaningful variation in treatment effects across quantiles.

We extracted top 15 covariates from the variance importance of the constructed causal forest model, including antipsychotics (measured on the index date), zolpidem, and lorazepam (both measured within 1 year prior to the index date). Figure 2 illustrates the density distributions of the top 15 baseline covariates for the high and low CATE groups, and Table 1 summarizes the distribution of each covariate across the two groups. With the exception of lorazepam, all covariates showed statistically significant differences in distribution between the high and low CATE groups.

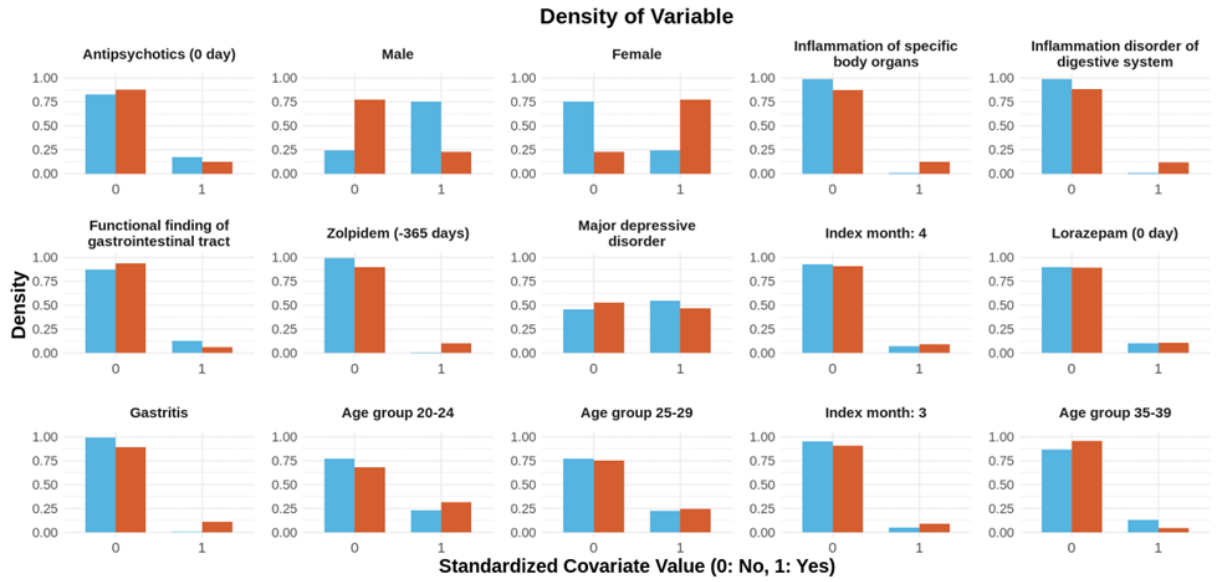


**Figure 2. Density of top 15 covariates**

**Table 1. Comparison of Top 15 Variables Between Low and High CATE Groups**

| Top 15 variables | Low CATE group (n=4,338) | High CATE group (n=4,345) | *P* |
|---|---|---|---|
| Antipsychotics[a], n (%) | 745 (17.2%) | 525 (12.1%) | < 0.001* |
| Male, n (%) | 3281 (75.6%) | 983 (22.6%) | < 0.001* |
| Female, n (%) | 1057 (24.4%) | 3362 (77.4%) | < 0.001* |
| Inflammation of specific body organs, n (%) | 43 (1.0%) | 540 (12.4%) | < 0.001* |
| Inflammation disorder of digestive system, n (%) | 43 (1.0%) | 505 (11.6%) | < 0.001* |
| Functional finding of gastrointestinal tract, n (%) | 265 (6.1%) | 562 (12.9%) | < 0.001* |
| Zolpidem[b], n (%) | 436 (10.1%) | 26 (0.6%) | < 0.001* |
| Major depressive disorder, n (%) | 2042 (47.1%) | 2368 (54.5%) | < 0.001* |
| Index month: 4, n (%) | 391 (9.0%) | 312 (7.2%) | < 0.001* |
| Lorazepam[a], n (%) | 466 (10.7%) | 450 (10.4%) | 0.582 |
| Gastritis, n (%) | 35 (0.8%) | 472 (10.9%) | < 0.001* |

| | | | |
|---|---|---|---|
| Age group 20-24, n (%) | 996 (23.0%) | 1376 (31.7%) | < 0.001* |
| Age group 25-29, n (%) | 982 (22.6%) | 1076 (24.8%) | 0.021 |
| Index month: 3, n (%) | 214 (4.9%) | 399 (9.2%) | < 0.001* |
| Age group 35-39, n (%) | 567 (13.1%) | 189 (4.3%) | < 0.001* |

a: the variable was measured after 0 day from the index date; b: the variable was measured between 365 days before and 0 day after the index date; *: statistically significant

**Conclusion**

Our findings suggest that MPH treatment may be associated with increased risk of BD in specific populations, with substantial heterogeneity observed across CATE-based subgroups. Applying this heterogeneity into individualized treatment strategies may help refine guidelines for MPH use.

**Acknowledgment**

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MOE) (No. 2120240615426, BK21 R&E Initiative for Advanced Precision Medicine); a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (No. RS-2024-00335936); and a grant (No. 25212MFDS002) from the Ministry of Food and Drug Safety in 2025.

**References**

1. Jefsen OH, Østergaard SD, Rohde C. Risk of mania after methylphenidate in patients with bipolar disorder. J Clin Psychopharmacol. 2023;43(1):28–34. doi:10.1097/JCP.0000000000001631
2. Chakraborty K, Grover S. Methylphenidate-induced mania-like symptoms. Indian J Pharmacol. 2011;43(1):80–81. doi:10.4103/0253-7613.75678
3. Ross EL, Bossarte RM, Dobscha SK, et al. Estimated Average Treatment Effect of Psychiatric Hospitalization in Patients With Suicidal Behaviors: A Precision Treatment Analysis. JAMA Psychiatry. 2024;81(2):135–143. doi:10.1001/jamapsychiatry.2023.3994
4. Feuerriegel, S., Frauen, D., Melnychuk, V. et al. Causal machine learning for predicting treatment outcomes. Nat Med 30, 958–968 (2024). https://doi.org/10.1038/s41591-024-02902-1
1. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. Am J Epidemiol. 2011 Apr 1;173(7):761-7. doi: 10.1093/aje/kwq439. Epub 2011 Mar 8. PMID: 21385832; PMCID: PMC3070495.