

UNO: Neural Pruning Enhances AI Reliability of Observational Data-based Findings

Yue Wu^{1,2,†}, Yiwen Lu^{1,3,†}, Huiyuan Wang^{1,2}, Dazheng Zhang^{1,2}, Huilin Tang^{1,2}, Bingyu Zhang^{1,3}, Lu Li^{1,3}, Wenjie Hu^{1,2}, Jiajie Chen^{1,2}, Tingyin Wang^{1,3}, Han Wu^{1,4}, Ting Zhou^{1,2}, Yuqing Lei^{1,2}, Xinyao Jian^{1,3}, Tsai Hor Chan^{1,2}, Siqi Chen^{1,3}, Yong Lin⁵, Demissie Alemayehu⁶, James Zou^{7-9,*}, David Asch^{10,11,*}, Yong Chen^{1-3,11-13,*}

1. Center for Health AI and Synthesis of Evidence (CHASE), Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
2. Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine
3. The Graduate Group in Applied Mathematics and Computational Science, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA
4. The Harker School, San Jose, CA, USA
5. Eli Lilly and Company, Indianapolis, IN, USA
6. Statistical Research and Data Science Center, Pfizer Inc., New York, NY, USA
7. Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
8. Department of Electrical Engineering, Stanford University, Stanford, CA, USA
9. Department of Computer Science, Stanford University, Stanford, CA, USA
10. Division of General Internal Medicine, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
11. Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA, USA
12. Penn Medicine Center for Evidence-based Practice (CEP), Philadelphia, PA, USA
13. Penn Institute for Biomedical Informatics (IBI), Philadelphia, PA, US

†: Co-first authors

*: Senior authors

Background

Observational healthcare data are increasingly used to generate real-world evidence for drug safety, comparative effectiveness, and drug repurposing¹⁻³. These data enable large-scale, reproducible analyses across heterogeneous populations and clinical settings. However, treatment assignment in observational data is not randomized and may reflect disease severity, healthcare-seeking behavior, access to care, and other factors that are incompletely measured in structured records. As a result, systematic error from residual confounding, selection bias, and measurement error can distort treatment effect estimates.

Target trial emulation and empirical calibration have improved the design and interpretation of observational studies, but residual bias remains difficult to eliminate when important confounders are unobserved. Neural networks provide flexible tools for modeling high-dimensional covariates and are increasingly used in propensity score estimation and high-throughput observational pipelines^{4,5}. However, when trained on biased data, they may overfit to spurious correlations and learn patterns driven by hidden confounding rather than true treatment-related signals. We introduce UNO, Unlearning Neural networks using Observational data, a negative-control-guided neural pruning framework designed to improve the reliability of AI-enabled observational analyses.

Methods

UNO is a debiasing module for neural network-based observational analyses. We apply UNO to propensity score estimation. The method has three steps. First, a neural network is trained to estimate treatment assignment using baseline covariates. Second, multiple negative control outcomes (NCO), defined as outcomes not plausibly affected by treatment but susceptible to similar confounding and coding processes, are used to detect residual bias. For each negative control outcome, individual-level conditional average treatment effects are estimated.

Individuals with repeatedly large negative control signals are assigned to a high-NCO-signal subset, while the remaining individuals form a low-NCO-signal subset. Third, UNO scores neurons by comparing their activations between these two subsets. Neurons disproportionately activated by high-NCO-signal individuals are pruned, and the model is fine-tuned on the low-NCO-signal subset. The resulting debiased propensity scores are then applied to the full study population for downstream effect estimation.

We evaluated UNO in two observational drug development use cases. The first was a study of GLP-1 receptor agonists versus DPP-4 inhibitors using TriNetX data. The cohort included 303,580 treatment initiators and evaluated 189 Clinical Classifications Software Refined outcomes across multiple clinical domains. The second was a Penn Medicine EHR-based drug repurposing study for Alzheimer's disease and related dementias among 16,184 patients with mild cognitive impairment, screening 33 candidate therapies. Residual bias was assessed using Expected Absolute Systematic Error across held-out negative control outcomes, and propensity score calibration was assessed using Expected Calibration Error.

Results

Across all 189 outcomes in the GLP-1RA versus DPP-4 inhibitor study, UNO consistently reduced Expected Absolute Systematic Error compared with the base neural network propensity score model, indicating lower residual systematic error (Fig. 1). UNO also improved propensity score calibration, with lower Expected Calibration Error across outcomes, suggesting better alignment between predicted treatment probabilities and observed treatment assignment (Fig. 2).

After debiasing with UNO, GLP-1RAs were associated with reduced risks of several cardiovascular and kidney outcomes, including acute myocardial infarction, cardiac arrest, acute renal failure, and chronic kidney disease. Protective associations were also observed for several respiratory, infectious, and neoplastic outcomes, while elevated risks were observed for selected endocrine, reproductive, otologic, and sleep-related outcomes. These findings demonstrate that UNO can support outcome-wide profiling while reducing bias in high-throughput observational evidence generation (Fig. 3).

In the AD/ADRD repurposing study in Fig. 4, UNO identified multiple candidate therapies associated with lower incident dementia risk. Protective associations were observed for cardiovascular agents such as amlodipine, atorvastatin, losartan, enoxaparin, and rosuvastatin; central nervous system-related agents such as melatonin, mirtazapine, propofol, trazodone, and lidocaine; gastrointestinal agents including omeprazole, pantoprazole, and lansoprazole; and vitamins or minerals including vitamin B12, cholecalciferol, ferrous sulfate, and ascorbic acid. These associations should be interpreted as hypothesis-generating signals requiring further validation.

Injury, poisoning and certain other consequences of external causes. MAL: Congenital malformations, deformations, and chromosomal abnormalities. MBD: Mental, behavioral, and neurodevelopmental disorders. MUS: Diseases of the musculoskeletal system and connective tissue NEO: Neoplasms. NVS: Diseases of the nervous system. RSP: Diseases of the respiratory system. SKN: Diseases of the skin and subcutaneous tissue. SYM: Symptoms, signs and abnormal clinic *D* and laboratory findings, not elsewhere classified.



Figure 2: UNO improves calibration of propensity score models compared to the Base Model. Calibration plots are shown for 28 outcomes in the GLP-1RAs vs. DPP-4is study. Predicted probabilities are grouped into 10 equally spaced bins; within each bin, confidence is the average predicted probability, and accuracy is the fraction of individuals with treatment. Each panel compares predicted treatment probabilities with observed treatment frequencies for the Base Model (top) and UNO (bottom), with lower ECE values indicating better calibration. The top 14 outcomes (upper set) show the largest reductions in ECE under UNO, while the bottom 14 outcomes (lower set) show the smallest reductions. Across both sets, UNO consistently achieved lower ECE than the Base Model, demonstrating improved calibration in high-throughput RWD analyses. ECE: Expected calibration error

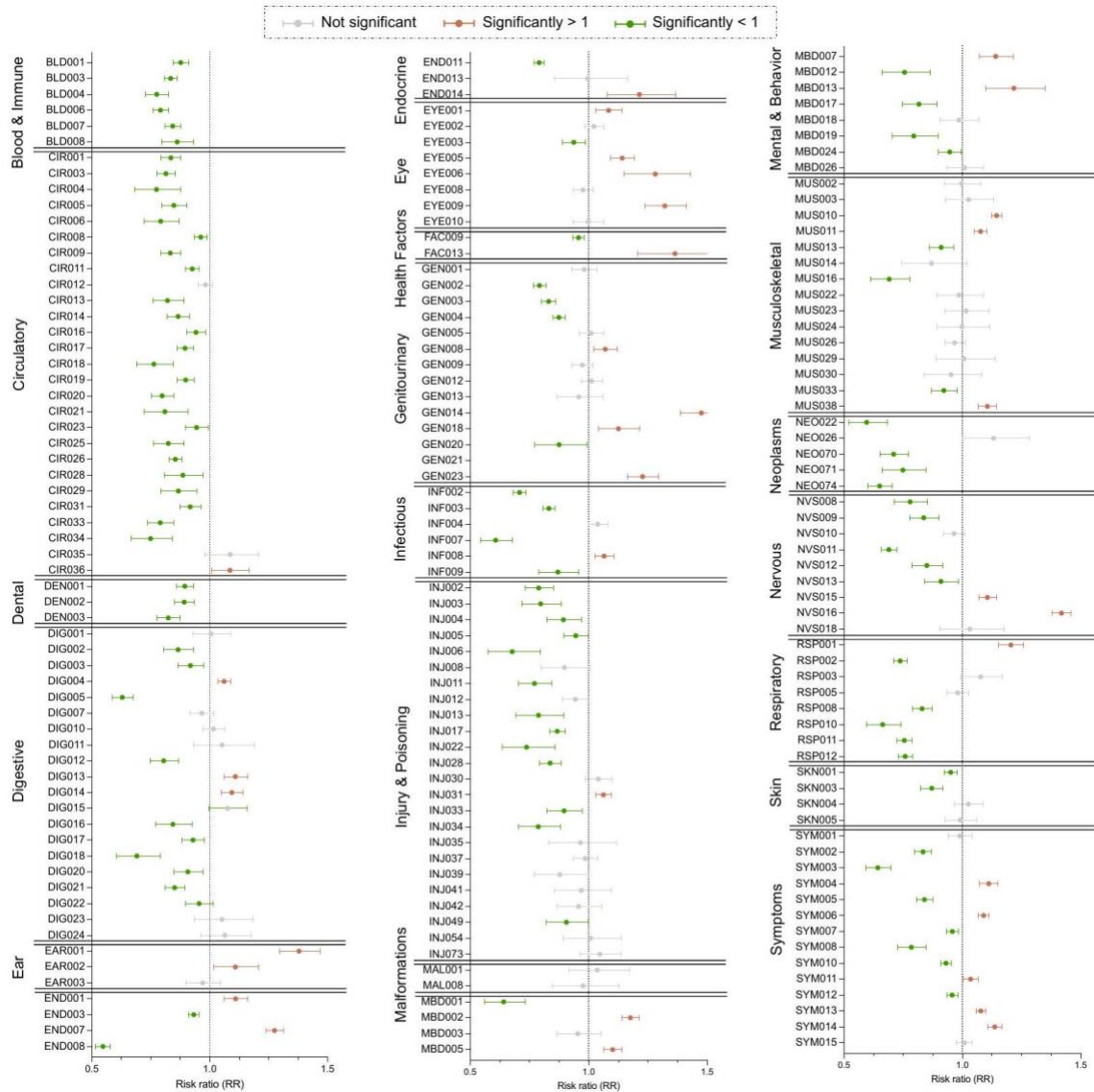


Figure 3: Outcome-wide relative risk estimates from UNO comparing GLP-1RAs versus DPP-4is across 189 CCSR-defined outcomes. Each point represents the estimated risk ratio (RR) with 95% confidence intervals for a given outcome, grouped by ICD-10-CM body system categories. Outcomes with RRs significantly greater than 1 are shown in red, outcomes with RRs significantly less than 1 are shown in green, and non-significant associations are shown in black. Results highlight both protective and elevated risks of GLP-1RAs across diverse clinical domains following adjustment with UNO.

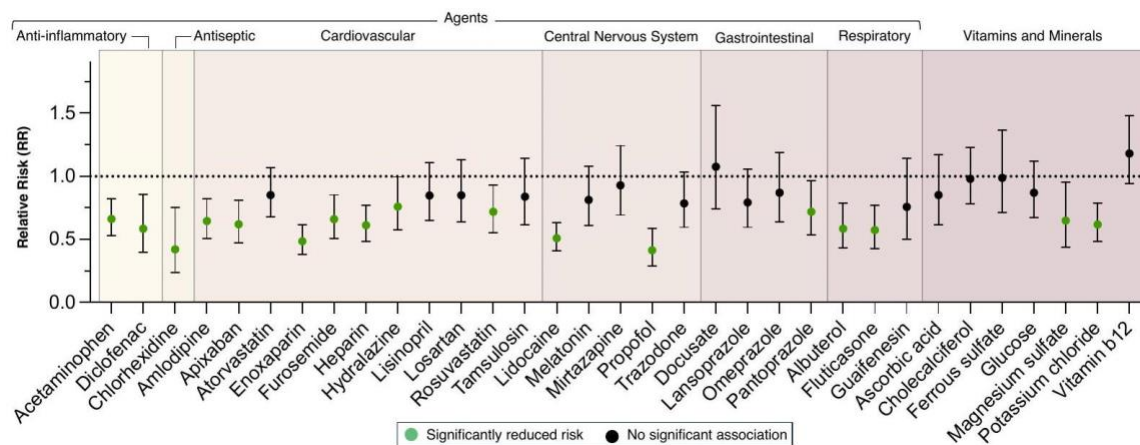


Figure 4: High-throughput drug repurposing for ADRD using UNO. Estimated relative risks of incident ADRD for pre-specified candidate medications using UNO. The medications span multiple therapeutic classes, including anti-inflammatory agents, antiseptics, cardiovascular agents, central nervous system agents, gastrointestinal agents, respiratory agents, and vitamins and minerals. The y-axis shows relative risk estimates, with colors indicating statistical significance levels.

Conclusion

UNO provides a general, architecture-agnostic strategy for improving the reliability of neural network models trained on observational healthcare data. By using negative control outcomes to identify and prune neurons associated with biased signals, UNO reduces residual systematic error without excluding patients from the final analysis or requiring complete model retraining. For the OHDSI community, UNO offers a practical framework for strengthening AI-enabled population-level effect estimation, outcome-wide screening, and drug repurposing studies. Future work will focus on broader validation across databases, expansion to other model architectures, and integration with standardized observational research pipelines.

References

- Booth, C. M., Karim, S. & Mackillop, W. J. Real-world data: towards achieving the achievable in cancer care. *Nat Rev Clin Oncol* **16**, 312–325 (2019).
- Corrigan-Curay, J., Sacks, L. & Woodcock, J. Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. *JAMA* **320**, 867–868 (2018).
- Liu, F. & Demosthenes, P. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med Res Methodol* **22**, (2022).
- Collier, Z. K. & Leite, W. L. A Tutorial on Artificial Neural Networks in Propensity Score Analysis. *J Exp Educ* **90**, 1003–1020 (2022).
- Zang, C. *et al.* High-throughput target trial emulation for Alzheimer’s disease drug repurposing with real-world data. *Nat Commun* **14**, 1–16 (2023).