

TRIM: Unlearning Transformer for Multimodal Trial Emulation

Tsai Hor Chan¹, Xinyao Jian¹, Yiwen Lu¹, Yue Wu¹, Huiyuan Wang¹, Huilin Tang¹, Guosheng Yin², Yong Chen¹

¹ Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

² Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China.

Correspondence: Yong Chen, ychen123@pennmedicine.upenn.edu

Background

Target trial emulation specifies the study design and estimand before analyzing observational data [1,2]. Multimodal cohorts combine clinical assessments, imaging-derived features and genetics, but richer data do not guarantee adequate confounding control. Flexible propensity models can retain treatment-selection signals that destabilize effect estimates. TRIM uses reference outcomes to identify residual-bias signals and guide structured pruning of a multimodal Transformer. The aim is to reduce systematic error before interpreting treatment effects, with a workflow relevant to repeated comparative-effectiveness analyses in OHDSI.

We use the term systematic error reference outcome (SERO) for a post-treatment outcome whose mean potential outcome is unchanged by treatment. For SERO S and binary treatment A , the defining assumption is $E[S(1)] = E[S(0)]$. This mean-null condition is weaker than requiring zero treatment effect for every individual, as for an ideal negative control outcome [3]. SEROs provide benchmarks for systematic error [4]. They are diagnostics and pruning signals, not propensity covariates. Baseline covariates precede treatment.

Methods

We evaluated nine target-trial emulations in the Parkinson's Progression Markers Initiative (PPMI): three medication contrasts among carbidopa/levodopa, rasagiline and selegiline, crossed with 12-month changes in Montreal Cognitive Assessment (MoCA) and Movement Disorder Society Unified Parkinson's Disease Rating Scale Parts II and III (UPDRS-II/III). Inputs were baseline clinical variables, imaging-derived summary features and genetic indicators. Twelve clinically reviewed SEROs from neurologic history, daytime sleepiness and mood were assessed in the post-index diagnostic window and screened for minimal plausible responsiveness to the short treatment exposure.

A shared encoder learned cross-modal representations using Transformer and contrastive-learning components [5-7]. Pair-specific propensity and SERO heads produced SERO-derived bias scores. A training subset of SEROs guided patient ranking into higher- and lower-bias subsets and neuron scoring from differences in activation. TRIM zeroed high-scoring neurons, retained the remaining structure, and refit on the lower-bias subset before stabilized inverse-probability-of-treatment weighting (IPTW) and augmented inverse-probability weighting (AIPW). Figure 1 summarizes the framework.

Held-out SEROs were reserved for evaluation. Expected absolute systematic error (EASE), the primary metric, is the expected absolute value of the fitted empirical-null systematic-error distribution estimated from held-out SERO log-relative-risk estimates. Lower EASE indicates less residual bias. We also assessed propensity-score calibration using expected calibration error (ECE), covariate balance, and variability in average treatment-effect (ATE) estimates across analysis settings.

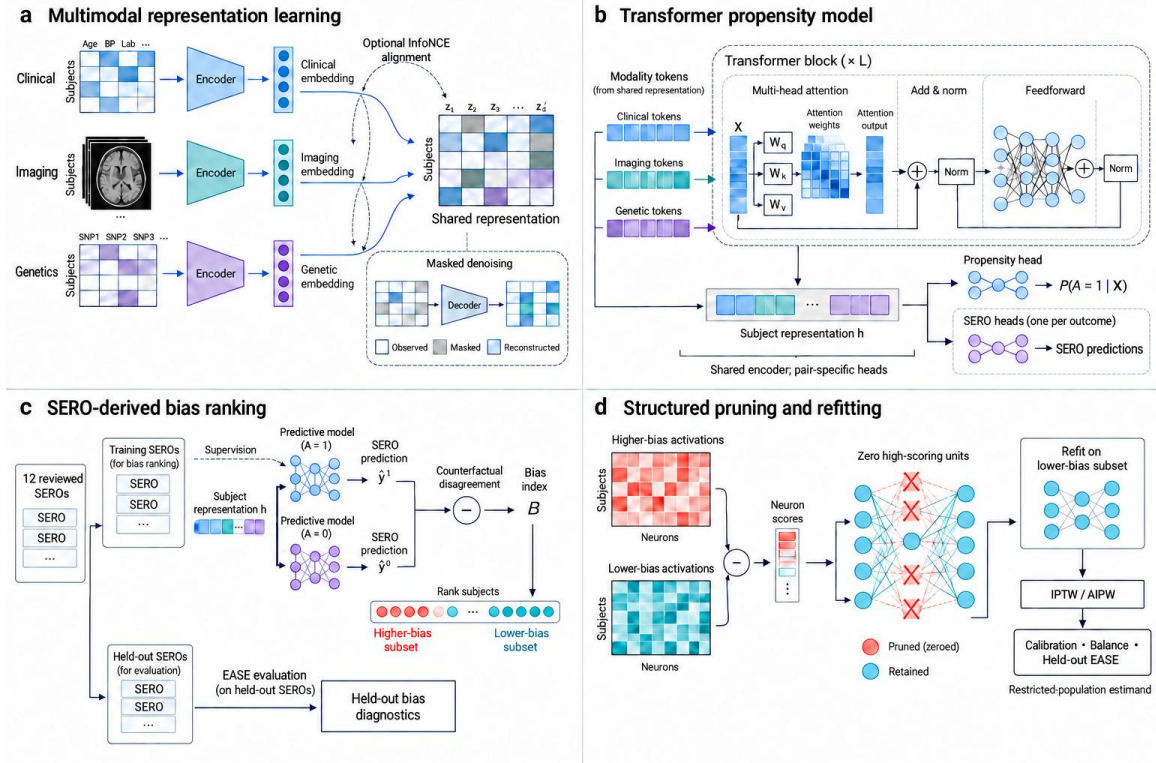


Figure 1. TRIM framework showing shared multimodal representation learning, pair-specific propensity and SERO heads, SERO-derived ranking, and structured neuron pruning. SEROs remain outside propensity inputs. Held-out SEROs evaluate systematic error after refitting.

Results

Across 54 analysis settings and 486 matched unpruned-pruned comparisons, mean EASE fell from 0.294 to 0.257, a 12.6% relative reduction (Figure 2). Eight of nine emulations improved on mean EASE, and EASE decreased in 57.4% of matched settings. Benefits varied by setting, so the aggregate reduction should not be interpreted as uniform improvement.

Mean ECE changed modestly from 0.331 to 0.326. A separate frozen-encoder experiment also showed lower and less variable EASE for the Transformer than for a matched multilayer perceptron. That experiment was distinct from the 54-setting sensitivity sweep.

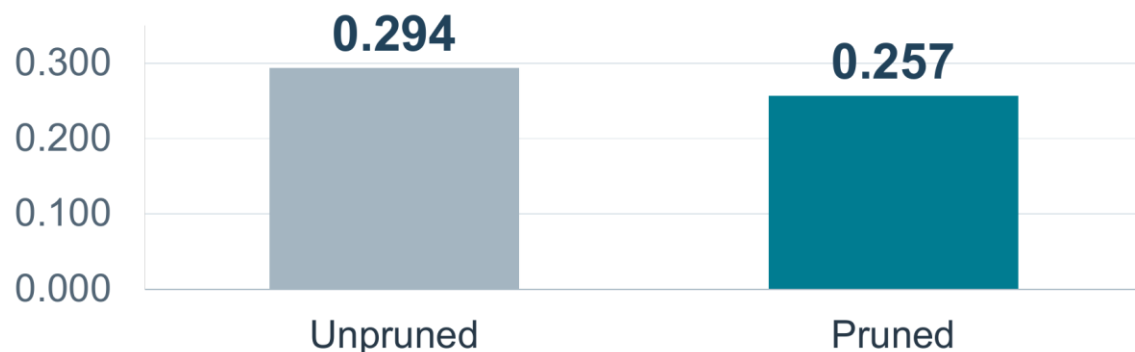


Figure 2. Mean EASE before and after SERO-guided pruning across 486 matched comparisons. The reduction from 0.294 to 0.257 corresponds to 12.6%. Lower EASE indicates less residual systematic error.

Mean within-pair ATE standard deviation across settings decreased from 0.200 to 0.092 for IPTW (54.0%) and from 0.166 to 0.077 for AIPW (53.6%). These results indicate greater stability across settings, not proof that causal effects are unbiased. Weighting largely preserved covariate balance, although a subset of design-related indicators retained imbalance.

Conclusion

SERO-guided pruning reduced mean EASE and stabilized effect estimates in PPMI. For OHDSI, TRIM offers a transferable encoder with lightweight pair-specific heads and an explicit SERO-based diagnostic layer. Lower bias diagnostics do not establish exchangeability or replace careful study design and sensitivity analyses. Refitting on the lower-bias subset changes the population targeted by the estimand. Under the weaker mean-null SERO condition, patient-level prediction contrasts are diagnostic pruning signals rather than proof of individual bias. The method requires valid SEROs and further assessment in larger observational networks.

References

1. Hernan MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183:758-764.
2. Hubbard RA, et al. Target trial emulation for observational studies: potential and pitfalls. *N Engl J Med*. 2024;391:1975-1977.
3. Lipsitch M, Tchetgen Tchetgen EJ, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*. 2010;21:383-388.
4. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc Natl Acad Sci U S A*. 2018;115:2571-2577.
5. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med*. 2022;28:1773-1784.
6. Vaswani A, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:5998-6008.
7. van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. *arXiv*. 2018:1807.03748.