

Reducing False Rejections in Covariate Balance Assessment for Target Trial Emulation Using a Joint Wald Diagnostic

Xinyao Jian¹, Jingyue Huang¹, Huiyuan Wang¹, Yuqing Lei¹, Jie Hu¹, Yumou Qiu², Yong Chen¹

1. Department of Biostatistics, Epidemiology, and Informatics Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
2. School of Mathematical Sciences and Center for Statistical Science, Peking University

Background

Target trial emulation is increasingly used to generate real-world evidence from observational health data when randomized trials are unavailable or infeasible.¹ For such evidence to be credible, treated and comparator groups should be comparable with respect to baseline covariates, making covariate balance assessment a critical diagnostic step after confounding adjustment. In a typical target trial emulation workflow, analysts estimate a propensity score and apply matching, stratification, or weighting, followed by balance diagnosis using standardized mean differences (SMDs).² The SMD summarizes the difference in covariate means between treatment groups, scaled by a standard deviation, and is usually assessed separately for each covariate against a fixed threshold such as 0.1.² However, this one-covariate-at-a-time screening rule can falsely flag imbalance due to random variation, especially in finite-sample studies where multiple baseline covariates are examined. Hripcsak et al. recently showed that this chance imbalance problem can be substantial; for example, with a total sample size of 250 and only five covariates, conventional SMD screening may reject approximately 90% of otherwise balanced studies by chance alone.³ To reduce excessive false rejection from chance imbalance, they proposed testing whether each SMD statistically significantly exceeds the prespecified threshold, with multiplicity correction across covariates.³ Motivated by the same concern, we propose a calibration-based joint Wald diagnostic that evaluates selected balance restrictions collectively, rather than screening covariates one at a time.

Methods

We developed a calibration-based joint diagnostic for evaluating covariate balance after propensity-score weighting in target trial emulation. Let T_i denote treatment assignment, R_i the covariate vector used in the propensity-score model, and x_i the prespecified covariates used for balance diagnosis. We assumed a logistic regression model for the propensity score, $e_i(\alpha) = P(T_i = 1 \mid R_i) = \text{expit}(R_i^\top \alpha)$, and estimated the parameter vector α , and constructed inverse-probability weights for the treated and control groups. These propensity-score weights were then treated as initial weights in a generalized entropy calibration problem, with separate balance constraints for the treated and control groups. The key idea is that, if the initial propensity-score weights already balance the selected covariates, then no additional calibration adjustment should be needed in the covariate-balance direction. In the calibration formulation, this additional adjustment is represented by Lagrange multipliers. Therefore, we tested the null hypothesis $H_0: \lambda_{\{T,x\}} = 0, \lambda_{\{C,x\}} = 0$, where $\lambda_{\{T,x\}}$ and $\lambda_{\{C,x\}}$ are the multiplier components corresponding to the treated- and control-group covariate balance constraints. To account for uncertainty from estimating the propensity score, we jointly estimated the propensity-score parameter α and the calibration multipliers (λ_T, λ_C) using stacked estimating equations. We then constructed a Wald statistic $W = \hat{\theta}_x^\top \widehat{\text{Var}}(\hat{\theta}_x)^{-1} \hat{\theta}_x$, where $\hat{\theta}_x = (\hat{\lambda}_{\{T,x\}}^\top, \hat{\lambda}_{\{C,x\}}^\top)^\top$, and the covariance matrix was obtained from the corresponding block of the sandwich variance estimator for the full stacked system. Under the null

hypothesis, W is compared with a χ^2 distribution with degrees of freedom equal to the number of tested balance restrictions. If the joint Wald test does not reject, analysts may proceed to downstream target trial emulation analyses; if the test rejects, the analysis may stop before outcome.

TTE framework with proposed diagnostic test

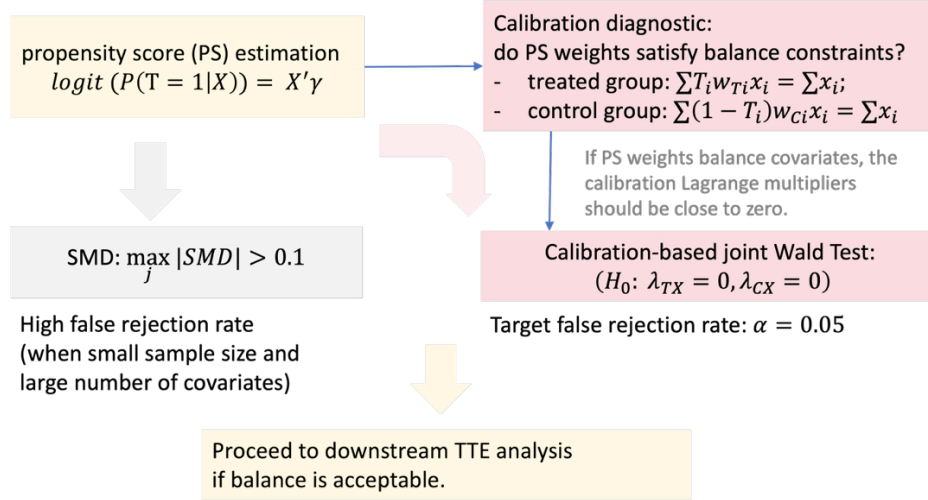


Figure 1. Workflow for covariate balance diagnosis after propensity-score weighting in target trial emulation.

Results

We evaluated the proposed diagnostic in a preliminary simulation study under a null setting where propensity-score weighting should achieve covariate balance in population. We generated $N = 250$ individuals with $J = 50$ independent binary covariates, $X_{ij} \sim \text{Bernoulli}(0.5)$. Treatment assignment was generated from a correctly specified logistic propensity-score model, $\text{logit}\{P(T_i = 1 | X_i)\} = \alpha_0 + \sum^J \alpha_j X_{ij}$, where only the first five covariates affected treatment assignment ($\alpha_j = 0.5, j = 1, \dots, 5$), and the remaining 45 covariates had coefficient zero but were still included in balance assessment. The intercept was set to ($\alpha_0 = -1.25$), giving an approximate marginal treatment probability of 0.5. Because the propensity-score model was correctly specified, rejection of balance in this setting represents false rejection due to finite-sample variability rather than systematic imbalance. Among 766 valid simulation runs, the conventional diagnostic based on inverse-probability weighting followed by SMD rejected balance in 89.30% of runs. In contrast, the proposed calibration-based joint Wald diagnostic rejected in 1.88% of runs at nominal $\alpha = 0.05$. These results align with our intuition: screening many covariates one at a time using a fixed SMD threshold can produce a high false rejection rate, whereas the proposed joint diagnostic evaluates the selected balance restrictions collectively and substantially reduces false alarms under the null.

Conclusion

In target trial emulation, covariate balance diagnosis is an important step before outcome analysis, but conventional SMD screening can falsely flag imbalance when multiple covariates are assessed in finite samples. We proposed a calibration-based joint Wald diagnostic that evaluates whether estimated

propensity-score weights already satisfy selected covariate balance restrictions, rather than testing covariates one at a time. In preliminary simulations under a correctly specified propensity-score model, the proposed diagnostic substantially reduced false rejection compared with the conventional maximum-SMD rule. This approach may help analysts distinguish systematic imbalance from chance imbalance, avoid unnecessary stopping of otherwise acceptable analyses, and improve the reliability of balance assessment in target trial emulation.

References

1. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol.* 2016;183(8):758-764.
2. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med.* 2009;28(25):3083-3107.
3. Hripcsak G, Zhang L, Chen Y, Li K, Suchard MA, Ryan PB, Schuemie MJ. Assessing covariate balance with small sample sizes. *Stat Med.* 2025;44(18-19).