

Evaluating data-driven automated discovery of disconnected negative controls using plasmode simulations



Jeong-eun Park, MPH, PhD¹; **Myeonghun Yu**, PhD²; **Rishi Desai**, MS, PhD¹; **Shirley Wang**, PhD¹; **Sarah K. Dutcher**, MS, PhD³; **Fatma M. Shebl**, MD, PhD, MS³; **Fang Tian**, PhD, MPH, MHS³; **Wei Hua**, MD, PhD, MHS, MS³; **Yong Ma**, PhD, MS³; **Motiur Rahman**, PhD, MS, MPharm³; **Richard Wyss**, MSc, PhD^{1*}; **Xu Shi**, PhD^{2*}

¹ Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
² Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA
³ Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA

* Richard Wyss and Xu Shi contributed equally as co-senior authors.

BACKGROUND

- Negative controls (NCs) can help detect unmeasured confounding in observational studies, but valid NCs are difficult to identify in practice.
- Data-driven Automated Negative Control Estimation (DANCE)** is an algorithm designed to identify valid **disconnected NCs**, a special type of NC that is associated with unmeasured confounding but has no direct causal relationship with the primary exposure or outcome.¹
- Its performance in common pharmacoepidemiologic research settings remains untested, particularly in studies involving high-dimensional covariates, binary variables, and time-to-event outcomes.

OBJECTIVES

- To adapt DANCE for high-dimensional real-world data and extend its application to studies with binary variables and time-to-event outcomes.
- To evaluate DANCE performance for identifying valid disconnected NCs using plasmode simulations informed by real-world healthcare data.

METHODS

Plasmode simulation

- Plasmode simulations were informed by an analytic cohort from a real-world study comparing Sodium-Glucose Cotransporter-2 inhibitors vs Dipeptidyl Peptidase-4 inhibitors for genital infection.²
- Observations with 86 measured confounders (X) adjusted in the original analysis were resampled with replacement from the source cohort.
- Within each resampled dataset, we simulated an unmeasured confounder (U), treatment (A), outcome (Y), and candidate NCs (Z).
- We evaluated DANCE under three simulation models:
 - Continuous model:** A, Y, U, and Z were generated from linear models.
 - Binary model:** A, Y, U, and Z were generated from Bernoulli models.
 - Time-to-event outcome model:** Event (T) and censoring times (C) were simulated from Weibull models, and a survival outcome with *event indicator* ($1:T \leq C$, $0:T > C$) and *observed time* ($\min(T, C)$) were derived.
- A total of 330 candidate NCs (Z) were generated with prespecified causal structures, including valid disconnected NCs and invalid/non-disconnected candidates (**Figure 1**).
 - Prevalence of true NC triplets in the current simulation: $11\% = \text{Total \# of possible true NC triplets} (666,760) / \text{Total \# of all possible triplet combinations} (5,935,160)$.

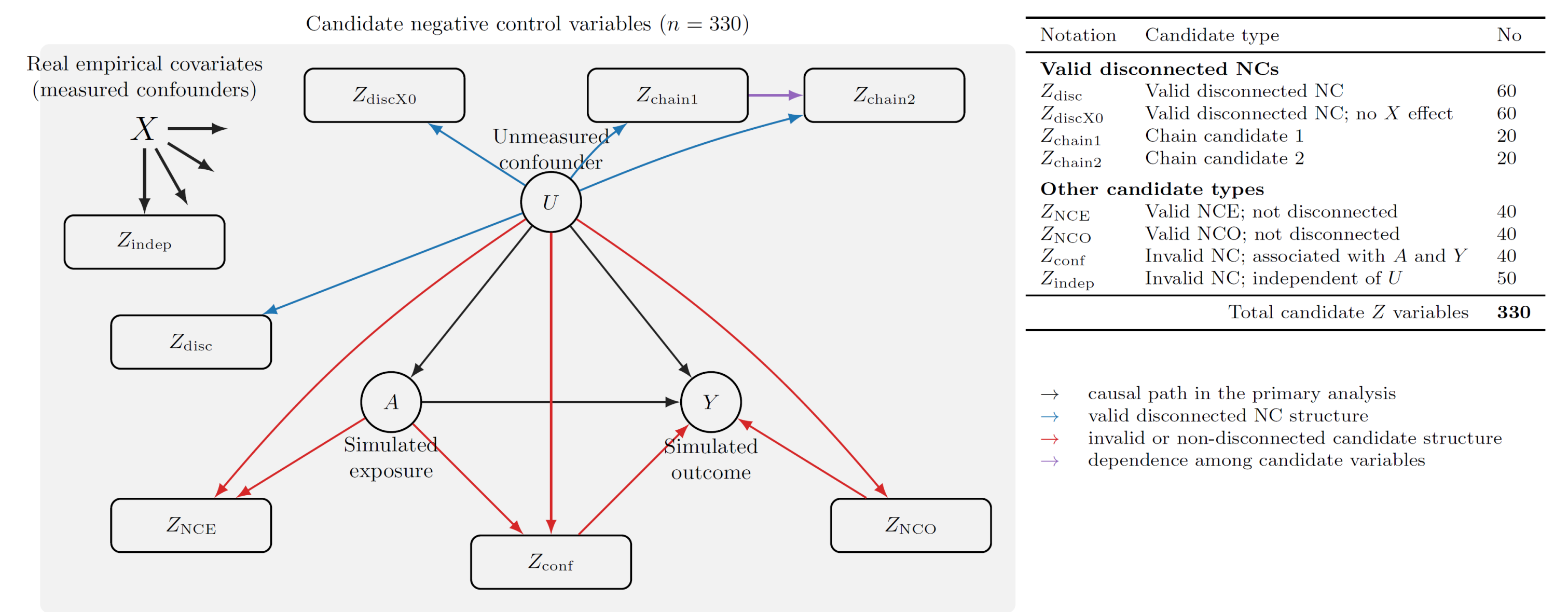


Figure 1. Causal diagram of simulated candidate NC types

Adaptations to DANCE implementation

- Pre-screening:** Candidates were pre-screened based on prespecified levels ($p < 0.2$) of A-Z and Z-Y associations for improved efficiency in screening a large number of triplet combinations.
- Conditional covariance:** We incorporated measured-confounder adjustment into DANCE using conditional covariance, improving performance in settings with many measured confounders.

Key parameters and simulation scenarios

	Parameter	Continuous	Binary	Time-to-event
Data generation scenarios	Sample size (n)	1k, 10k	5k, 10k, 50k	10k
	U-Z association	1, 2	OR=2, OR=3	OR=3
	Prevalence of U or Z	-	15%, 50%	50%
	U-Y structure	U-Y		
DANCE scenarios	Outcome input*	Continuous Y	Binary Y	U-Y, U-Y, C
				Event indicator, observed time, log(observed time)

*For time-to-event outcomes, **the DANCE outcome input was either the event indicator or observed time (or a transformed version)**; the current version of DANCE does not jointly incorporate both components of time-to-event outcomes.

Simulation outcomes

- We evaluated positive predictive value (**PPV**) and the number of triplets discovered (**discoveries**) across the DANCE selection threshold (**$v\alpha$**) over 500 simulation runs.
 - $v\alpha$** is the threshold for the min p-value across all vanishing tetrad tests for a triplet in DANCE; **higher $v\alpha$ values indicate stricter triplet selection**.

RESULTS

DANCE performance across data-generating scenarios

- Sample size:** In the binary model, at $v\alpha=0.5$, average PPV increased from 90% to 97%, as sample size increased from 5k to 10k (**Figure 2A**).
- Strength of U-Z associations:** At $n=10k$ and $v\alpha=0.5$, increasing the OR from 2 to 3 increased PPV from 84% to 93% (**Figure 2B**).
- Prevalence of U and Z:** Increasing both prevalences from 15% to 50% increased PPV from 87% to 93% (**Figure 2C**).
- Continuous model showed similar PPV patterns but returned more discoveries than the binary model (e.g., 30k vs 18k at $n=10k$, $v\alpha=0.5$).

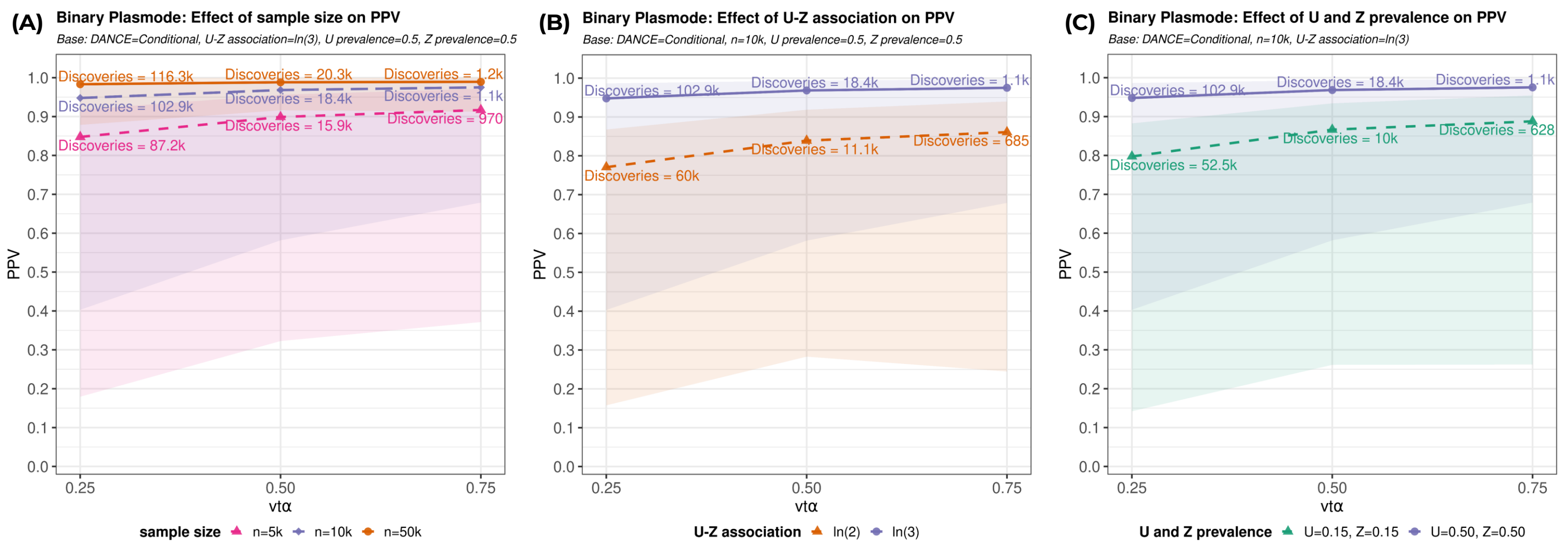


Figure 2. Average PPV and empirical 95% interval across iterations (shaded area) from the binary model by (A) sample size, (B) strength of the U-Z association, and (C) prevalence of U and Z.

DANCE performance for time-to-event outcome model

- Performance varied by how U affected the survival process (U-Y structure).
 - Outcome input: event indicator** – more true positives when U-Y (Figure 3A, green bars) compared with observed time; maintained robust PPV and discoveries when U-Y, C (Figure 3B, green bars).
 - Outcome input: observed time variables** – more true positives when U-Y, C, but also more false positives (Figure 3B, orange and purple bars).
- Overall, the *event indicator* provided more robust and conservative performance across U-Y structures affecting time-to-event outcomes.

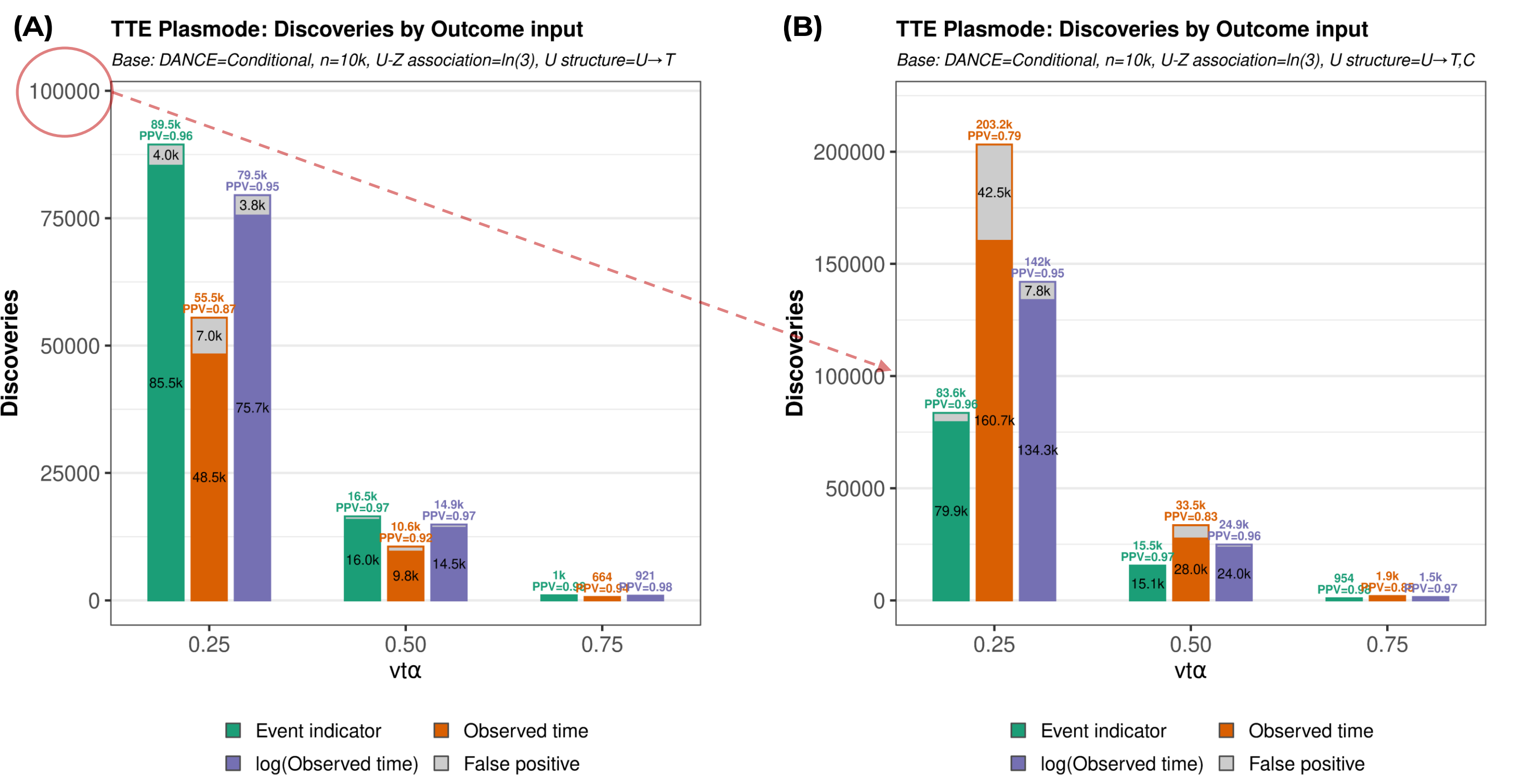


Figure 3. Average number of discoveries (split into true and false positives) from the time-to-event model by outcome input under different U-Y structures: (A) U-Y: U affecting event time (T) only; (B) U-Y, C: U affecting both event (T) + censoring time (C).

CONCLUSION

- Adapted DANCE robustly identified NCs in high-dimensional plasmode simulations, supporting its potential use in pharmacoepidemiologic studies to select NCs for detecting unmeasured confounding.
- DANCE algorithm performance improved with larger samples, stronger U-Z associations, and higher U and Z prevalence.
- For time-to-event outcomes, the *event indicator* was the most robust and conservative input across U-Y structures.

REFERENCES

- Kummerfeld E, Lim J, Shi X. Data-driven Automated Negative Control Estimation (DANCE): Search for, Validation of, and Causal Inference with Negative Controls. Journal of Machine Learning Research. 2024;25(229):1-35.
- Desai RJ, Wang SV, Sreedhara SK, et al. Process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs (PRINCIPLED): considerations from the FDA Sentinel Innovation Center. BMJ. 2024;384:e076460.

ACKNOWLEDGEMENTS/DISCLOSURES

- This project was supported by Task Order 75F40124F19011 under Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA).
- The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS or the U.S. Government.