Evaluate the use of an automated approach to find disconnected negative controls using the data-driven automated negative control estimation (DANCE) algorithm

Simulation Protocol

Xu Shi, PhD[1*]; Richard Wyss, Msc, PhD[2*]; Rishi J. Desai, MS, PhD[2]; Meighan Rogers Driscoll, MPH[3]; Sarah K. Dutcher, MS, PhD[4]; Ryan Hickson, PharmD, MPH, PhD,[5]; Wei Hua, MD, PhD, MHS, MS[4]; Chanelle Jones, MHA, CPhT[4]; Natasha Kasid, MD[5]; Erich Kummerfeld, PhD[6]; Yong Ma, PhD, MS[4]; Haritha S. Pillai, MPH[2]; Motiur Rahman, PhD, MS, MPharm[4]; Fatma M. Shebl, MD, PhD, MS[4]; Eric Tchetgen Tchetgen, PhD[7]; Fang Tian, PhD, MPH, MHS[4]; Darren Toh, ScD[3]; Shirley Wang, PhD[2]; Myeonghun Yu, PhD[1]

*Primary Investigators contact information: rwyss@bwh.harvard.edu; shixu@umich.edu

1. Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI

2. Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

3. Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA

4. Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD

5. Office of New Drugs, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD

6. Institute for Health Informatics, University of Minnesota, Minneapolis, MN

7. Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA

Sentinel Innovation Center

Version 5.0

April 10, 2025

The Sentinel System is sponsored by the U.S. Food and Drug Administration (FDA) to proactively monitor the safety of FDA-regulated medical products and complements other existing FDA safety surveillance capabilities. The Sentinel System is one piece of FDA's Sentinel Initiative, a long-term, multi-faceted effort to develop a national electronic system. Sentinel Collaborators include Data and Academic Partners that provide access to healthcare data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Innovations Center is funded by the FDA through the Department of Health and Human Services (HHS) Master Agreement Number 75F40119D10037.

Evaluate the use of an automated approach to find Disconnected Negative Controls using the DANCE algorithm

Simulation Protocol

Table of Contents

History of Modifications

Version	Date	Modification	Author
1.0	11/29/2024	First draft for WG review	Sentinel Innovation Center
2.0	12/19/2024	Addressing 1st round of FDA feedback	Sentinel Innovation Center
3.0	03/03/2025	Addressing 2nd round of FDA feedback	Sentinel Innovation Center
4.0	03/26/2025	Addressing 3rd round of FDA feedback	Sentinel Innovation Center
5.0	04/10/2025	Addressing 4th round of FDA feedback	Sentinel Innovation Center

1 Project Overview

This protocol describes evaluation of the data-driven automated negative control estimation (DANCE) algorithm[1] in Sentinel settings via simulation studies. The DANCE algorithm[1], is an automated approach to identify disconnected negative controls. Negative controls are variables associated with the unmeasured confounders but not causally related to either the treatment or outcome variables of primary interest.[2,3] A negative control exposure is a variable associated with the unmeasured confounder and does not causally impact the outcome, while a negative control outcome is one that is associated with the unmeasured confounder and not causally affected by the treatment. Such known-null effects form the basis of falsification strategies to test whether adjustment for observed covariates suffices to control for confounding bias.

This protocol describes the approach for Aim 1, where the objective is to utilize a plasmode simulation framework[4] for data generation to mirror the complexity observed in routinely collected healthcare data. This framework involves resampling from an underlying observed dataset with replacement. By leveraging the observed covariate patterns, it allows us to create datasets with simulated variables (unmeasured confounders, treatment, outcome, and negative controls), while maintaining the correlations among observed covariates. Compared to a fully synthetic simulation approach[5,6], plasmode simulations capture the complex covariate patterns observed in healthcare databases.

A parallel component of this project focuses on applying the DANCE algorithm to a drug safety question use case in a multisite implementation (Aim 2), which is described in more detail in a separate protocol.

This project is being conducted as part of FDA's Prescription Drug User Fee Act (PDUFA) VII commitment on "Use of Real-World Evidence – Negative Controls."[7]

Types of Candidate Negative Controls

We will consider different types of candidate negative controls and their potential relationships with treatment A, outcome Y, and unmeasured confounder U, classified into the following categories:

1. **Invalid negative controls:** Independent of A, Y, and U.
2. **Invalid negative controls:** Influenced by U and affecting both A and Y.
3. **Valid negative controls but not disconnected negative controls:**
    - 3a. **Instrumental variables**
    - 3b. **Impacted by U, affecting A, and without a direct causal impact on Y given A** (satisfying negative control exposure assumptions)

    - 3c. **Impacted by U, affecting Y, and not directly influenced by A or affecting A** (satisfying negative control outcome assumptions)
    - 3d. **Impacted by U and another proxy of U**
4. **Valid disconnected negative controls:** Impacted by U, no direct causal relationship with either A or Y. They serve as either negative control exposures or negative control outcomes.

Directed acyclic graphs (DAGs) encoding each of the above scenarios are presented in the appendix. These include-

- Invalid negative controls: independent of A, Y and U (Figure 1)
- Invalid negative controls: influenced by U and affecting both A and Y (Figure 2)
- Valid negative controls but not disconnected negative controls: instrumental variables (Figure 3a)
- Valid negative controls but not disconnected negative controls: impacted by U, affecting A, and without a direct causal impact on Y given A (Figure 3b)
- Valid negative controls but not disconnected negative controls: impacted by U, affecting Y, and not directly influenced by T or affecting T (Figure 3c)
- Valid negative controls but not disconnected negative controls: impacted by U and another proxy of U (Figure 3d)
- Valid disconnected negative controls (Figure 4)

Note:

- Categories 3a-3c have direct arrows pointing from the negative control to either A or Y, violating the disconnected negative control assumption.
- For category 3d, if the parent proxy is excluded from analysis, the remaining proxies are valid disconnected negative controls.
- We simplify the simulation by not including negative controls impacted by U and A but not directly impacting Y given A.

The Observational Dataset for Plasmode Simulation

We will use MGB data which includes Medicare claims linked to EHR over a study period of 01/01/2013 through 12/31/2019 for our plasmode simulation. The dataset was used in a query comparing the risk of genital infections for new users of sodium glucose cotransporter 2 (SGLT2) inhibitors versus dipeptidyl peptidase 4 (DPP4) inhibitors . SGLT2 inhibitors and DPP4 inhibitors are classes of drugs used for the treatment of type 2 diabetes mellitus (T2DM), and SGLT2i have a known and labeled risk of genital infections.[8] Patients with T2DM, age ≥/= 65 years, no prior use of study medications, no prior or concurrent use of glucagon-like peptide-1 (GLP-1) receptor agonists, no history of end stage renal disease (ESRD), no history of human immunodeficiency virus (HIV), no history of genital infections, and with continuous Medicare A, B, D enrollment for six months will be considered eligible. This dataset originated from a prior Sentinel Innovation Center project, with the final results published in the BMJ article titled “A 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". [9]

2 The Stepwise Plan for Conducting the simulation

Simulation Step 1: Selection of Measured Confounders (X)

From the existing plasmode dataset described in the previous paragraph, pre-specified variables will be selected as measured confounders, denoted by X. Selection of the measured confounders will be guided by the analyses and results of the above Sentinel Innovation Center project.[9]

Simulation Step 2: Simulation of Unmeasured Confounders (U)

We will simulate zero or one unmeasured confounder denoted by U of the following three types, which may be predicted by X.

1. No unmeasured confounding.
2. One continuous unmeasured confounder.
3. One binary unmeasured confounder (with a range of prevalence).

We note that U does not have to be simulated and can in fact be drawn from real data along with X in step 1. However, the potential of U -> X -> NC pathway may complicate the simulation of negative controls of types 1 and 3a, as these scenarios require the absence of U -> NC pathway.

Simulation Step 3: Simulation of Treatment (A), Outcome (Y), and Candidate Negative Controls

In each iteration of the plasmode simulation, after resampling the data with replacement, the following procedure will be performed:

1. Simulate negative controls of type 1 and type 3a, predicted by X only.
2. Simulate negative controls of type 2 and types 3b-4, predicted by U and X.
3. Simulate negative controls of type 3d, predicted by U, X, and a proxy of U.
4. Simulate treatment assignment A, predicted by U, X, and negative controls of type 2 and type 3a.
5. Simulate the observed outcome Y, predicted by A, U, X, and negative controls of type 2 and type 3c. This includes both common and rare outcomes relevant to drug safety questions.

Parameters to vary include the following:

1. The number of candidate negative controls: we will select a range of numbers between 10 and 1000 and also consider extreme numbers (e.g., 5k and 10k) to test the limit of DANCE.

2. The distribution of candidate negative controls: we will consider both continuous and binary/categorical variables. For binary candidate negative controls, we do not plan to study the prevalence of negative controls, because we always screen out binary candidate negative controls with a low prevalence before running the DANCE algorithm. Screening rare candidate negative controls aims to ensure that there is sufficient information and power for DANCE algorithm to detect valid negative controls.

3. The distribution of outcome: we will focus on simulating a time-to-event outcome. Our simulation will consider different types of follow-up time (e.g. fixed or not) and censoring (% censoring).

4. The prevalence of treatment assignment (e.g., 2%, 10%, 25%, 50%).

5. The strengths of association between different variables: this is related to the following considerations (1) To ensure that U shows a strong confounding effect that cannot be explained by any other adjustment, treatment assignment and observed outcomes must be strongly and directly predicted by U; (2) Estimation control is a good proxy of U, thus, we will consider a range of values for the strength of association between the candidate negative control and the treatment/outcome. (3) We would like to consider a range of the true causal effect (the A to Y arrow).

3 Implementing the DANCE algorithm

We will first conduct propensity score (PS) matching such that we adjust for X before conducting the negative control identification and causal effect estimation steps. Then, using the matched sample, we will run the DANCE algorithm to select valid disconnected negative controls and estimate the causal effect.

When the negative control variables are categorical or when the outcome is time-to-event, we will first dichotomize them to binary (try to balance sample size within each category) for identification of valid negative controls, and then use the original categorical negative control or time-to-event outcome for effect estimation. With time-to-event outcome, we will check if the follow-up time is similar between the two groups before dichotomizing. If follow-up time is not similar, methods such as matching to ensure similar follow-up and regression adjustment for following up time will be considered. Although the method for identification of negative control variables, the first component of DANCE, is developed for Gaussian variables only[1], we will apply such a method to binary variables which is not theoretically supported and is expected to result in a loss of power at least. The rationale for applying the DANCE algorithm on binary variables is that the method in DANCE for selection of negative controls is based on covariance of pairs of observed variables, which can be computed for both Gaussian and binary variables.

4 Evaluation

To assess the DANCE algorithm's effectiveness in validating candidate negative control variables, we will plot Receiver Operating Characteristic (ROC) curves using varying thresholds for rejecting the null hypothesis in the validation test. Additionally, the accuracy in estimating the causal effect will be evaluated through proportion bias, variance, and coverage probability. DANCE will be compared with three methods:

1. Naive: A naive adjustment method ignoring unmeasured confounding.

2. Random: Randomly selecting pairs of negative controls from the candidate pool to adjust for unmeasured confounding using the double negative control method (which requires a pair of negative controls).

3. All: Using all possible negative controls

5 Appendix

Figures describing different types of negative controls. The node X with arrows pointing out means that X is potentially causing all other nodes, and in our discussion all arguments will implicitly condition on X.

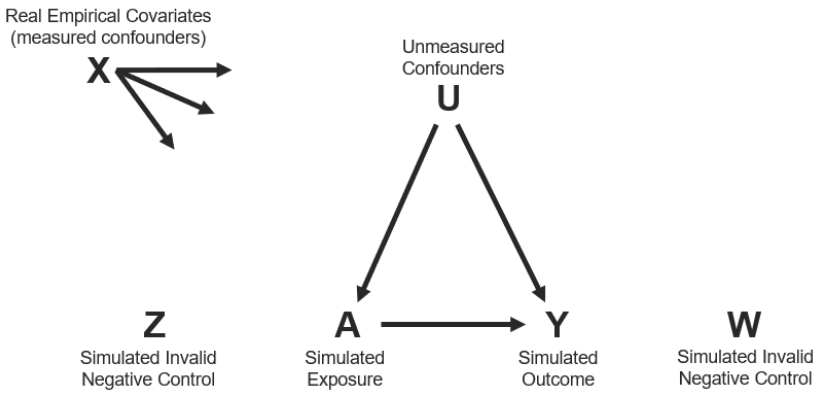

Figure 1. Invalid negative controls: Independent of A, Y, and U

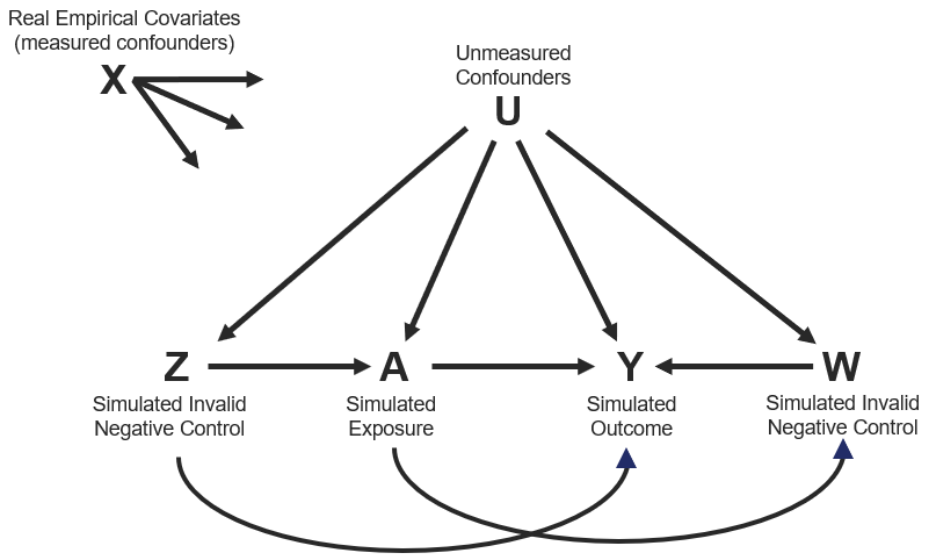

Figure 2. Invalid negative controls: Influenced by U and affecting both A and Y.

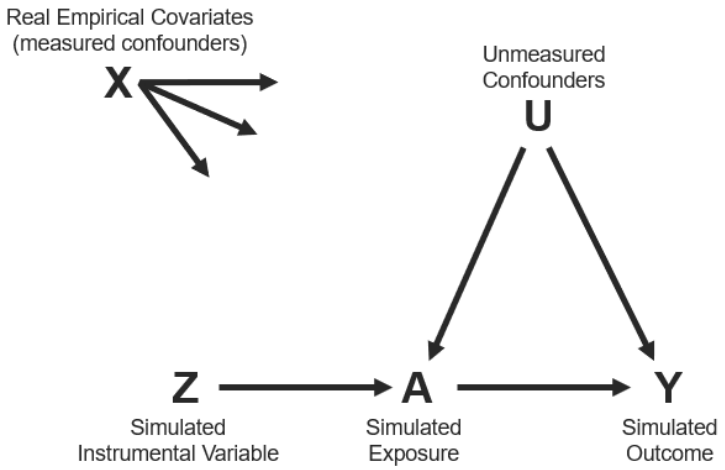

Figure 3a. Valid negative controls but not disconnected negative controls: *instrumental variables*

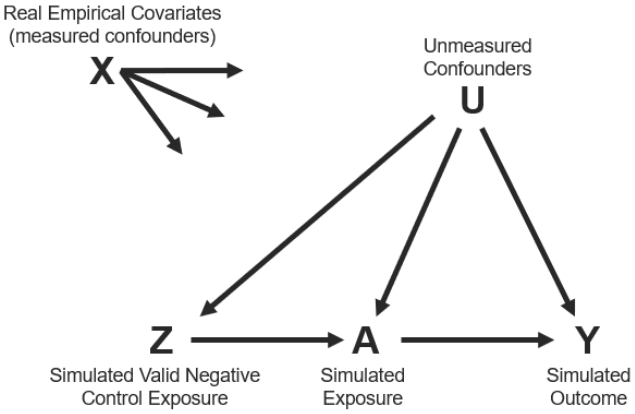

Figure 3b. Valid negative controls but not disconnected negative controls: Impacted by U, affecting A, and without a direct causal impact on Y given A.

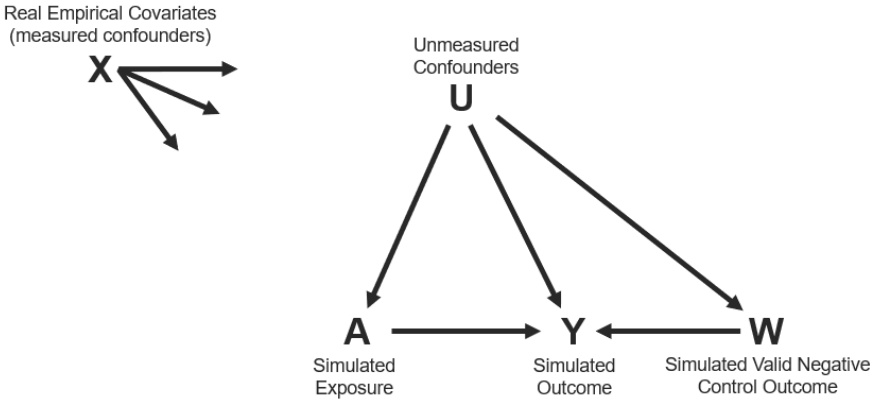

Figure 3c. Valid negative controls but not disconnected negative controls: Impacted by U, affecting Y, and not directly influenced by T or affecting T.

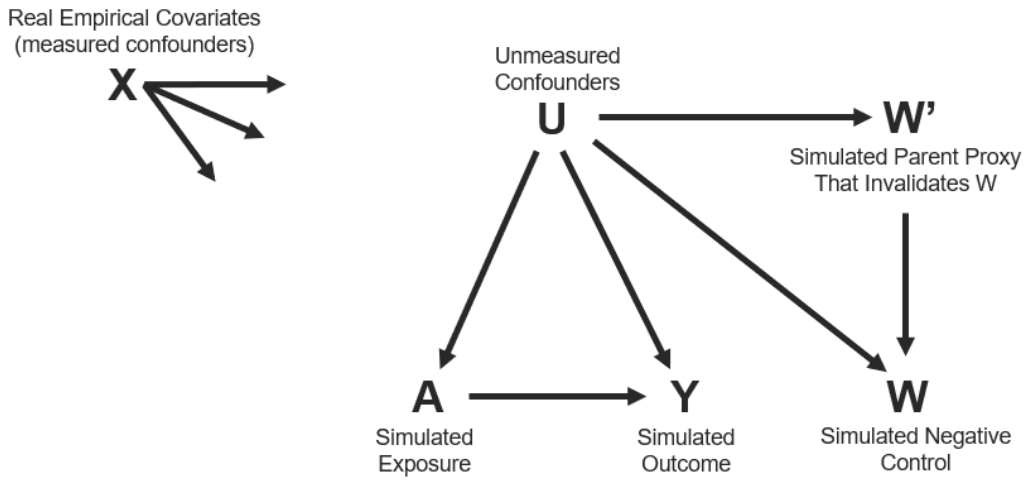

Figure 3d. Valid negative controls but not disconnected negative controls: Impacted by U, and another proxy of U.

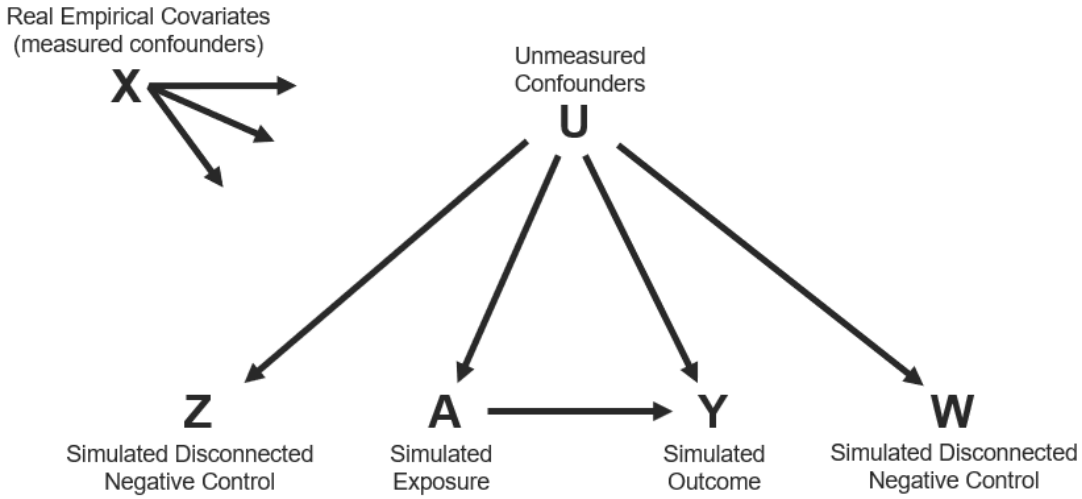

Figure 4. Valid disconnected negative controls

6 References

1. Data-driven Automated Negative Control Estimation (DANCE): Search for, Validation of, and Causal Inference with Negative Controls. ar5iv. Accessed December 16, 2024. https://ar5iv.labs.arxiv.org/html/2210.00528
2. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative Controls. *Epidemiology*. 2010; 21 (3): 383-388. doi: 10.1097/EDE.0b013e3181d61eeb.
3. Shi X, Miao W, Tchetgen ET. A selective review of negative control methods in epidemiology. Current epidemiology reports. 2020 Dec;7:190-202.
4. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*. 2014; **72**: 219-226.
5. Desai RJ, Bradley MC, Lee H, et al. A simulation-based bias analysis to assess the impact of unmeasured confounding when designing non-randomized database studies. American journal of epidemiology. May 31 2024;doi:10.1093/aje/kwae102
6. Williamson BD, Krakauer C, Johnson E, et al. Assessing treatment effects in observational data with missing confounders: A comparative study of practical doubly-robust and traditional missing data methods. *arXiv*. December 2024. Accessed December 23, 2024. https://arxiv.org/abs/2412.15012.
7. PDUFA VII: Fiscal Years 2023 – 2027. *FDA*. Published online April 24, 2023. Accessed April 7, 2025. https://www.fda.gov/industry/prescription-drug-user-fee-amendments/pdufa-vii-fiscal-years-2023-2027
8. U.S. Food and Drug Administration. FDA warns about rare occurrences of serious infection in the genital area with SGLT2 inhibitors for diabetes. U.S. Food and Drug Administration. Published July 2, 2015. Accessed March 3, 2025. https://www.fda.gov/drugs/drug-safety-and-availability/fda-warns-about-rare-occurrences-serious-infection-genital-area-sglt2-inhibitors-diabetes
9. 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. doi:10.1136/bmj-2023-076460