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

Case Example Protocol

Richard Wyss, Msc, PhD[1*]; Xu Shi, PhD[2*]; Rishi J. Desai, MS, PhD[1]; Meighan Rogers Driscoll, MPH[3]; Sarah 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[1]; 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[1]; Myeonghun Yu, PhD[2]

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

1. Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA
2. Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI
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

Case Example 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 round 1 of WG feedback	Sentinel Innovation Center
3.0	03/03/2025	Addressing round 2 of WG feedback	Sentinel Innovation Center
4.0	03/26/2025	Addressing round 3 of WG feedback	Sentinel Innovation Center
5.0	04/10/2025	Addressing round 4 of WG feedback	Sentinel Innovation Center

1 Project Overview

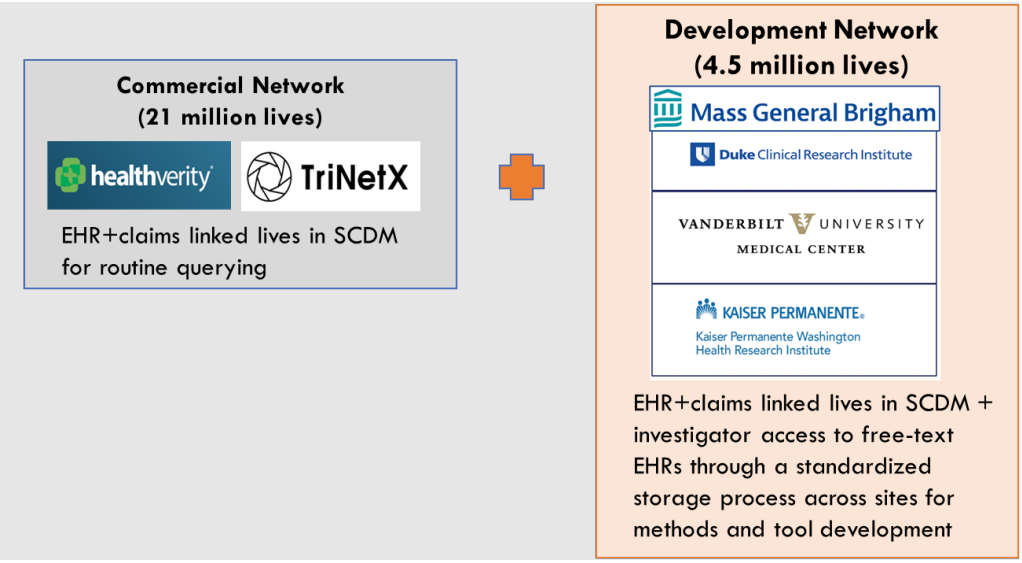
Negative controls are variables associated with unmeasured confounders but not causally related to either the treatment or outcome of interest.[1,2] A negative control exposure is a variable associated with the unmeasured confounder and does not causally impact the outcome, while a negative control outcome is associated with the unmeasured confounder and not causally affected by the treatment. With known null effects, negative control variables can help to identify analyses that are unlikely to fully control for confounding bias. The identification of negative control variables, however, can be time consuming, difficult to scale, and sometimes prone to error. The data-driven automated negative control estimation, or DANCE[3] algorithm, is an automated approach to identify disconnected negative controls. A disconnected negative control is a special type of negative control variable that is associated with unmeasured confounders but not causally related to the treatment nor to the outcome.[3]

This protocol describes the approach for Aim 2, where the objective is to apply the DANCE algorithm to a drug safety question use case in a multisite implementation. Prior to applying the DANCE algorithm to the use case, the DANCE algorithm will be evaluated and tailored to settings relevant to large-scale healthcare database studies using plasmode simulation (Aim 1), 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."[4]

Data Source and Empirical Cohort

The Sentinel Innovation Center (IC) has access to a distributed data network- the Real-World Evidence Data Enterprise (RWE-DE)- containing electronic health records (EHRs) linked to insurance claims for two commercial partners and four academic ("development network") partner sites (**Figure**).[5]



We will use two data assets from the RWE-DE for implementation of DANCE in a multisite fashion i.e., Mass General Brigham (MGB) and HealthVerity (HV). We will use a query that has previously been leveraged from a prior IC project.[6,7] This query, previously implemented at the MGB site, compared users of sodium-glucose cotransporter 2 inhibitors (SGLT2i) vs dipeptidyl peptidase 4 inhibitors (DPP4i) on the risk of genital infections (safety outcome). In this project, we will implement this query at a second site (HV).

The protocol from the prior project, that provides full details for the query to generate the empirical cohort comparing SGLT2i vs DPP4i, is provided in the supplemental material of the publication (e.g., definition of baseline covariates, definition of outcome, etc.).

2 Analytic Approach

Propensity Score (PS) Matching

Prior to implementing DANCE, we will first conduct propensity score matching to balance predefined baseline covariates. Balancing measured confounders is necessary prior to running the DANCE algorithm to select valid disconnected negative controls and estimate the causal effect.

PheWAS Mapping

Prior to implementing DANCE, we will map ICD-9 and ICD-10 codes to clinically meaningful phenotypes called phecodes that are routinely used for Phenome-wide association studies (PheWAS). There are existing mappings and code packages to process and group ICD codes into a manageable number of clinically meaningful phecode categories.[8,9] The clinical concepts defined by phecodes will be used as

input variables in the DANCE algorithm to identify disconnected negative controls.

Implementing the DANCE Algorithm

DANCE is an automated approach that incorporates a statistical test to discover disconnected negative control variables. Details of the DANCE algorithm are provided elsewhere.[3] We will apply DANCE where the candidate negative control variables will include binary features representing the phecodes mapped from ICD codes.[8] Since the outcome is time-to-event, we will first dichotomize the outcome to binary when running DANCE for identification of disconnected negative controls. It is important to note that the binarization of the outcome is only used for identification of disconnected negative controls, and the original time-to-event outcome will be used for bias detection. To ensure that follow-up time between treatment groups is approximately equal, we will censor after a short follow-up of 1-year. If follow-up times are still differential, we will consider censoring at the minimum follow up of the PS-matched set to force follow-up times to be approximately equal.

Including the pre-specified baseline confounders that are used for propensity score matching [5] (see protocol from the prior project) within DANCE introduces additional noise and can decrease power to detect true disconnected negative controls. Therefore, we will not include predefined baseline confounder variables as inputs to DANCE. We will further evaluate the impact of including/excluding phecodes (i.e., candidate disconnected negative controls) that are highly correlated with predefined confounders. We will consider 3 approaches to screening phecodes:

1) Exclude all phecodes that share a common ICD-9 or ICD-10 code with any predefined baseline confounder variable.
2) Exclude all phecodes that are strongly correlated with any predefined baseline confounder variable.. We will consider variables to be strongly correlated if the correlation coefficient is >0.7. We will consider alternative thresholds of 0.5, 0.6, 0.8, and 0.9 in sensitivity analyses.
3) Don't exclude any phecodes regardless of how strongly correlated they are with predefined confounders.

We will consider multi-site implementation of DANCE using various approaches including:

1) Using the union of all negative controls across sites (all)
    a. This will include a list of all phecodes that are selected regardless of site (the union of all selected phecodes across the two sites)
2) Using negative controls that appear at both sites (common denominator)
    a. This will include a list of phecodes that are selected at both sites (the intersection of selected phecodes between the sites).
3) Using site-specific negative controls (tailored).

a. This will include two lists: 1) all codes selected at the first site; 2) all codes selected at the second site. We will consider these two lists separately when performing the evaluation to determine the validity of the selected candidate negative controls. The purpose for separate site-specific negative controls lists is to test the hypothesis that DANCE may provide more reliable selection of candidate negative controls at larger sites.

Each of the generated lists will then be considered for evaluation to determine the accuracy of the selected candidate negative controls to make a determination of which multi-site implementation is preferable.

3 Evaluation

We will assemble a team of clinicians and epidemiologists with expertise in this clinical area to manually review the selected candidate controls in terms of their expected validity as true negative control variables. We will evaluate the proportion of negative controls validated as true negative controls by experts.

Bias Detection:

Each of the disconnected negative controls identified by DANCE can serve as either a negative control exposure or a negative control outcome. We will estimate (1) the hazard ratio between the disconnected negative control and outcome and (2) the risk ratio between treatment and the disconnected negative control within the PS matched cohort. We will plot the distribution of negative control effect estimates and 95% confidence intervals for bias detection. We will visualize the distribution of negative control effect estimates and their 95% confidence intervals. As an exploratory analysis, we will compare these negative control effect estimates with the primary analysis results to investigate possible unmeasured confounding bias.[10]

We will provide recommendations on the most suitable approach for multisite implementation based on learnings in consultation with the FDA and the workgroup. The team will provide a practical process guide to walk investigators through the steps needed to implement DANCE for future Sentinel studies.

Hypothetical Results Table.

Site	Selected Disconnected Negative Control	Valid Negative Control[1]	HR for effect of NC on outcome	95% CI	RR for effect of treatment on NC	95% CI
MGB	Feature A	Yes				
	Feature B	Yes				

	Feature C	No				
⋮	⋮	⋮	⋮			⋮
HealthVerity	Feature A	Yes				
	Feature C	Yes				
	Feature D	No				
	Feature E	No				
⋮	⋮	⋮	⋮		⋮	⋮

[1]Valid negative controls will be determined by team of clinical experts after reviewing list of selected negative controls.

4 References

1. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative Controls. *Epidemiology*. 2010; 21 (3): 383-388. doi: 10.1097/EDE.0b013e3181d61eeb.

2. Shi X, Miao W, Tchetgen ET. A selective review of negative control methods in epidemiology. Current epidemiology reports. 2020 Dec;7:190-202.

3. 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 25 (2024):1-35.

4. 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

5. Desai RJ, Marsolo K, Smith J, et al. The FDA Sentinel Real World Evidence Data Enterprise (RWE-DE). *Pharmacoepidemiology and Drug Safety*. 2024;33(10):e70028. doi:10.1002/pds.70028

6. 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

7. Development and Illustration of a Framework for Conducting Nonrandomized Studies of Medication Safety and Effectiveness Using Healthcare Databases | Sentinel Initiative. Accessed April 7, 2025. https://www.sentinelinitiative.org/methods-data-tools/methods/development-and-illustration-framework-conducting-nonrandomized-studies

8. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenom-wide association studies in the R environment. Bioinformatics. 2014; 30(16):2375-2376. doi:10.1093/bioinformatics/btu197.

9. Wu P, Gifford A, Meng X, et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Medical Informatics*. 2019;7(4):e14325. doi:10.2196/14325

10. Izurieta HS, Wernecke M, Kelman J, Wong S, Forshee R, Pratt D, Lu Y, Sun Q, Jankosky C, Krause P, Worrall C. Effectiveness and duration of protection provided by the live-attenuated herpes zoster vaccine in the Medicare population ages 65 years and older. Clinical infectious diseases. 2017 Mar 15;64(6):785-93.