Journal of Machine Learning Research 25 (2024) 1-35 Submitted 9/22; Revised 1/24; Published 4/24

Data-driven Automated Negative Control Estimation (DANCE): Search for, Validation of, and Causal Inference with Negative Controls

Erich Kummerfeld
ERICHK@UMN.EDU
Institute for Health Informatics
University of Minnesota
Minneapolis, MN 55455, USA

Jaewon Lim
IMJAEWON@UW.EDU
Department of Biostatistics
University of Washington
Seattle, WA 98105, USA

Xu Shi
SHIXU@UMICH.EDU
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109, USA

Editor: Joris Mooij

Abstract

Negative control variables are increasingly used to adjust for unmeasured confounding bias in causal inference using observational data. They are typically identified by subject matter knowledge and there is currently a severe lack of data-driven methods to find negative controls. In this paper, we present a statistical test for discovering negative controls of a special type—disconnected negative controls—that can serve as surrogates of the unmeasured confounder, and we incorporate that test into the Data-driven Automated Negative Control Estimation (DANCE) algorithm. DANCE first uses the new validation test to identify subsets of a set of candidate negative control variables that satisfy the assumptions of disconnected negative controls. It then applies a negative control method to each pair of these validated negative control variables, and aggregates the output to produce an unbiased point estimate and confidence interval for a causal effect in the presence of unmeasured confounding. We (1) prove the correctness of this validation test, and thus of DANCE; (2) demonstrate via simulation experiments that DANCE outperforms both naive analysis ignoring unmeasured confounding and negative control method with randomly selected candidate negative controls; and (3) demonstrate the effectiveness of DANCE on a challenging real-world problem.

Keywords: causal discovery; graphical models; negative control; unmeasured confounding; vanishing tetrad

1. Introduction

There are many causal questions in science and medicine that can not be solved with randomized experiments now or in the foreseeable future. For such questions, our best estimates must thus rely on observational data instead. The rich field of causal inference has developed in response to this, providing support for these efforts and developing methods that offer some level of assurance and confidence for learning causal information from observational data (Pearl, 2009; Rubin, 1974). Many causal inference methods assume that there are no unmeasured common causes of treatment and outcome, but it is generally believed that in reality unmeasured confounders are widespread. This is a serious limitation to the methods that make such assumptions. One of the most frequently used approaches to mitigate unmeasured confounding is the instrumental variable (IV) approach (Angrist and Keueger, 1991; Angrist et al., 1996; Hernán and Robins, 2006), which has been previously studied extensively (Greenland, 2000; Baiocchi et al., 2014; Garabedian et al., 2014; Burgess et al., 2017; Swanson et al., 2018).

A more recently developed strategy is negative control (NC) methods (Lipsitch et al., 2010; Shi et al., 2020b; Tchetgen Tchetgen et al., 2020). Negative controls are variables associated with the unmeasured confounders but not causally related to either the treatment or outcome variables of primary interest. Intuitively, such known-null effects form the basis of falsification strategies to test whether adjustment for observed covariates suffices to control for confounding bias: one expects no significant association between the negative control and the treatment or outcome of interest if there is no uncontrolled confounding; on the other hand, the unanticipated presence of an association between the negative control and the treatment or outcome of interest constitutes compelling evidence of residual confounding bias. For example, in a study about the effects of influenza vaccination on influenza hospitalization, injury/trauma hospitalization was considered as a negative control as it was not causally affected by influenza vaccination, but may be subject to the same confounding mechanism mainly driven by health-seeking behavior (Jackson et al., 2006). The authors found that despite efforts to control for confounding, influenza vaccination not only appeared to reduce risk of influenza hospitalization after influenza season (risk ratio 0.82, 95% CI 0.73–0.92), but also appeared to reduce risk of injury/trauma hospitalization (risk ratio 0.83, 95% CI 0.75–0.91). This was interpreted as evidence of bias due to inadequately controlled confounding. NCs have traditionally been used to rule out non-causal explanations of empirical findings (Rosenbaum, 1989; Weiss, 2002; Lipsitch et al., 2010; Glass, 2014). Recently, a sequence of NC methods have been developed to identify causal effects and correct for unmeasured confounding bias (Miao et al., 2018a; Deaner, 2018; Shi et al., 2020a; Singh, 2020; Cui et al., 2023; Ying et al., 2021; Kallus et al., 2021; Dukes et al., 2023; Li et al., 2022).

A key challenge in the use of NC methods is that until now, NC variables have had to be identified laboriously from background knowledge. It also had to be assumed that the identified variables were genuine NCs, as no validation test existed unless one is willing to make additional assumptions. Such situations are common in causal inference, e.g., the assumption of no unmeasured confounding is also untestable. Nevertheless, we will show that under certain conditions, it is possible to leverage certain subcovariance matrix rank

constraints to validate a particular class of NC variables, referred to as disconnected NCs which we formally define in Section 2.1, satisfying a specific causal structural model.

In this paper, we utilize some lesser known theory regarding relationships between sub-covariance matrix rank constraints and the graphical structure of causal models to provide both theory and algorithms for evaluating NC variables. First, we provide a statistical test that can be used to determine whether a triplet of candidate NCs are real disconnected NCs or not. Second, we provide a simple algorithm for searching among a set of candidate NCs, and identifying subsets of those variables that collectively meet the conditions of being disconnected NCs. Third, we combine our proposed method for finding valid NC variables with a recently developed double-NC method for causal inference (Miao et al., 2018b; Shi et al., 2020a; Cui et al., 2023), creating an algorithm that accurately estimates and makes inferences about causal effects from observational data. We refer to the proposed method as the Data-driven Automated Negative Control Estimation (DANCE) algorithm. We prove that our proposed methods are correct under fairly general assumptions, evaluate their finite sample performance with a series of numerical experiments, and demonstrate their usability on a real world data set.

Our novel contributions put forward in this paper include the following: (1) We define a new type of negative control, the disconnected NC, that is a special case of both NC exposure and NC outcome; (2) We provide the first statistical test for validating disconnected NCs and prove correctness of this validation test; (3) We develop a search procedure that uses this test to search from a large number of candidate NCs for sets of three disconnected NCs that pass the validation test; (4) We develop a new method that aggregates a collection of disconnected NCs to produce a point estimate and confidence interval, which improves efficiency compared to the double-NC method that utilizes only a pair of NCs; (5) We develop the DANCE algorithm using the disconnected NC search procedure and the aggregate negative control estimation method, with performance demonstrated via simulations and an applied study.

The rest of the paper is organized as follows. In Section 2 we review the three main topics that the work in this paper builds upon: negative controls, structural models, and rank constraints. We then present a statistical validation test for disconnected NCs in Section 3, and prove its correctness in Section 3.3. Section 4 presents an algorithm that searches a set of candidate NC variables to find sets of disconnected NCs which pass the validation test, and Section 5 presents the DANCE algorithm that combines with the double-NC method to construct an all-in-one method for producing a valid causal effect estimate from a data set containing a collection of candidate NC variables, some of which are not necessarily valid disconnected NCs. Section 6 presents numerical experiments to evaluate our proposed test and algorithms, and compares them to two methods: a simple regression method ignoring unmeasured confounding and a random selection of candidate NCs followed by the double-NC method. An application of DANCE to a real clinical data set is described in Section 7. Section 8 summarizes the strengths and limitations of the methods presented in this paper, and points towards promising directions for future work. R code for implementing our method is available at `https://github.com/imjaewon07/DANCE`.

2. Background

2.1 Unmeasured Confounding and Negative Control Methods

We adopt the potential outcome framework under the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1974, 1980; Cox, 1992) and let $(O(1), O(0))$ denote the pair of potential outcomes under treatment and control conditions, respectively. We are interested in estimating the average treatment effect (ATE), defined as $\Delta = E[O(1) - O(0)]$. It suffices to identify the counterfactual mean $E[O(t)]$ for $t \in \{0, 1\}$. Let O denote the observed outcome and T denote the binary treatment. We suppress measured covariates for simplicity; adjustment for measured covariates is discussed in Section 5.1.

Instead of making the no unmeasured confounding assumption, we allow the presence of an unmeasured confounder U with a latent ignorability assumption that $O(t) \perp\!\!\!\perp T \mid U$. If U was measured, then $E[O(t)]$ is identified under the ignorability assumption (Robins, 1986). However, when U is unobserved and unadjusted, ATE estimation will be biased. In this case, additional information is needed to identify and make inference about the ATE.

An increasingly popular approach to mitigate bias due to unmeasured confounding is to use its proxies. For example, as shown in Figure 1, if U can be measured with error via proxy variables Z and W, then one can leverage Z and W to identify the confounding bias due to U and remove such bias from the estimated causal effect. Such proxy variables have been referred to as negative controls (Lipsitch et al., 2010; Shi et al., 2020b). Formally, a negative control outcome, denoted as W, is a variable known not to be causally affected by the treatment of interest. Likewise, a negative control exposure, denoted as Z, is a variable known not to causally affect the outcome of interest. The negative control exposure and outcome variables should share a confounding mechanism with the treatment and outcome variables of primary interest. In summary, Z and W satisfies

$$(T, Z) \perp\!\!\!\perp (O(t), W) \mid U. \tag{1}$$

There are a number of causal graphs that satisfy the NC assumptions (Shi et al., 2020b). For example, both a valid instrumental variable independent of the unmeasured confounder and an invalid instrumental variable associated with the unmeasured confounder are valid negative control exposures. Alternative directed acyclic graphs encoding the NC assumptions are available in Shi et al. (2020b).

Figure 1 presents a special case where Z and W are causally related to neither the treatment nor the outcome of interest, hence Z and W can serve as either negative control exposure or negative control outcome (Shi et al., 2020a; Tchetgen Tchetgen et al., 2020). We refer to such a special class of NC variables as the *disconnected NCs*. Formally, the disconnected NCs satisfy the following assumption

$$(Z, W) \perp\!\!\!\perp (T, O) \mid U.$$

Compared to the fundamental NC assumption (1), the disconnected NCs satisfy additional assumptions that $Z \perp\!\!\!\perp T \mid U$ and $W \perp\!\!\!\perp O \mid U$.

Miao et al. (2018a) established nonparametric identification of the average treatment effect (ATE) using a pair of negative control exposure and outcome variables, referred to as the double-NC. Intuitively, having additional children of U that are conditionally

independent with T and O allows for identification of the unmeasured confounding bias due to the influence of U on T and O, and subsequently this quantity can be removed from the association between T and O, leaving an unbiased estimate of T's effect on O. Recently, the NC framework has been extended to proximal causal inference, which partitions measured covariates into proxies satisfying NC conditions, acknowledging that covariate measurements are at best proxies of the underlying confounding mechanisms (Tchetgen Tchetgen et al., 2020; Cui et al., 2023; Ying et al., 2021; Dukes et al., 2023).

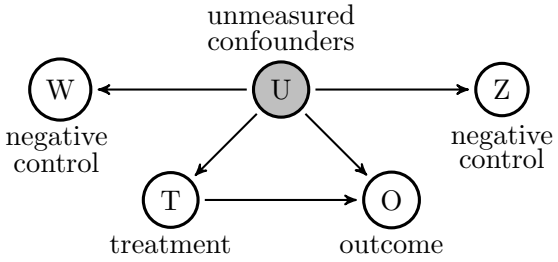


Figure 1: Causal graph of two disconnected NCs, Z and W, suppressing the measured covariates X which is implicitly conditioned on in all arguments.

2.2 Structural Models

The rest of this paper makes heavy use of the theory of structural graphical models, especially structural equation models. This section provides background terminology and definitions from this field, which can be used as an introduction for readers who are unfamiliar with it, or as a reference for readers who are already familiar with it.

A *directed graph* is a pair of sets, $\langle \Phi, \Psi \rangle$, where Φ contains some number of variables, and Ψ contains some number of directed edges, or arrows, pointing from one variable in Φ to another variable in Φ. It is often important to consider not just individual edges, but *paths* in a directed graph, which are ordered lists of edges in a graph such that each edge shares one of its endpoint variables with the edges before and after it. In other words, a path is a connected sequence of edges. A special kind of path is a *trek*, where nowhere along the length of the path are there two consecutive edges with arrows pointing to the variable between them. In other words, in a trek there are no subsections that look like this: $\rightarrow X \leftarrow$. A special type of trek is the *directed path*, where all the edges point in the same direction along the path. Probably the most commonly studied subclass of directed graphs is *directed acyclic graphs* (DAGs), which are directed graphs that contain no directed paths passing through the same variable twice, i.e., directed graphs without any directed cycles.

Directed graphs are often augmented with additional quantitative and statistical information, forming a statistical model. Two commonly used examples are *structural equation models* (SEMs) and *Bayesian networks* (BNs). The specific meanings of these terms often vary from one paper or research group to another. In this paper, we consider SEMs to refer to DAGs which have been augmented with an additive Gaussian noise term ϵ_K for each variable K in the DAG, as well as a linear coefficient β_{JK} corresponding to each edge $J \rightarrow K$. As such, each variable's value is calculable from the value of its *parents* (the nodes with edges pointing to it) and the value of its independent noise term. Let *parents* be a

function that returns all of the parents of the variable in its argument, then the SEM can be written as a collection of equations of the form:

$$K = \sum_{J \in parents(K)} \beta_{JK} J + \epsilon_K.$$

In this paper, we consider BNs to be similar to the SEM class of models, but using categorical variables instead of continuous variables. Each variable in a BN has its probability distribution determined by a conditional probability table based on the values of that variable's parents. As such, like an SEM, it specifies a complete joint distribution that can be decomposed into separate components for each of its variables. As we are defining them here, a BN specifies a multinomial distribution, while an SEM specifies a multivariate Gaussian distribution.

Variables can be either *measured*, or *unmeasured*. Unmeasured variables are also called *hidden* variables or *latent* variables. Whether a variable is measured or unmeasured has no effect on the SEM or BN itself, but rather refers to what variables are available in data collected from that SEM or BN. Unmeasured variables are absent from the data entirely, and so all their values are missing. Notoriously, since unmeasured variables are still part of the data generating model, measured variables in the model can be correlated with each other because an unmeasured variable is the parent of both of them. Such an unmeasured variable is considered an *unmeasured confounder* in many contexts. In causal inference from observational data, unmeasured confounders are a potential source of bias, as it is a source of spurious association that can occur even when neither of the two measured variables influences the other in any way.

2.3 Rank Constraints and the Vanishing Tetrad Test

It is well known that the graphical structure $\mathcal{G}$ of a causal structural model over variables V entails constraints among the partial correlations of the observed (measured) variables $M \subseteq V$. Partial correlation constraints have been leveraged to develop causal discovery algorithms, including algorithms that still operate correctly in the presence of unmeasured common causes, such as Fast Causal Inference (FCI) (Spirtes et al., 2000; Zhang, 2008; Hyttinen et al., 2013). These methods all depend on theoretical work establishing theorems that systematically relate graphical structures to their implied partial correlation constraints.

A lesser known type of constraint that is also implied by graphical structure is rank constraints on the subcovariance matrix. A subcovariance matrix is a covariance matrix between two sets of variables, S_1 and S_2, with $S_1, S_2 \subseteq M$. As first proved by Sullivant et al. (2010), the rank of each subcovariance matrix will have bounds due to features of the graph $\mathcal{G}$. Let $\mathrm{Cov}(X, Y)$ be the covariance between two random variables X and Y, then the subcovariance matrix for $S_1 = \{W, Z\}$ and $S_2 = \{T, O\}$ can be written as:

$$\Sigma_{S_1,S_2} = \Sigma_{\{W,Z\},\{T,O\}} = \begin{pmatrix} \mathrm{Cov}(W,O) & \mathrm{Cov}(W,T) \\ \mathrm{Cov}(Z,O) & \mathrm{Cov}(Z,T) \end{pmatrix}.$$

Sullivant's theorems state that, for the graph in Figure 1, since all paths from W and Z to T and O pass through one variable, U, the rank of matrix $\Sigma_{\{W,Z\},\{T,O\}}$ is less than its dimension. This also implies that the determinant of $\Sigma_{\{W,Z\},\{T,O\}}$ is zero, i.e.,

$\text{Cov}(W,O)\text{Cov}(Z,T) - \text{Cov}(W,T)\text{Cov}(Z,O) = 0$. This is referred to as a *vanishing tetrad*. The same is not true of the subcovariance matrix for $\Sigma_{\{W,T\},\{Z,O\}}$, since there is an additional path from T to O that does not include U. These theorems were then extended by Spirtes (2013) to relax some of their linearity and acyclicity assumptions. A growing number of papers have made use of these theorems to identify the presence of unmeasured common causes, and to even make inferences about the causal relationships among unmeasured common causes (Kummerfeld et al., 2014; Kummerfeld and Ramsey, 2016; Yang et al., 2017).

The theorems relating rank constraints to graphical structure make use of some lesser known graph concepts, in particular the concepts of *trek* and *trek-separation*, typically abbreviated as *t-separation*. We here provide those definitions, including some intermediary definitions necessary for defining trek-separation. In a directed graphical model, a trek is a path with no colliders on it, i.e., no variable within the trek is a direct child of the variables both before and after it in the trek. A variable is said to block a trek if it lies anywhere along that trek. Any trek has exactly one *root* node which has no parents in the trek, which can be any node along the trek including the endpoints. Let t be an arbitrary trek from node A to B in directed graph $\mathcal{G}$, with root node C. The portion of the trek t from C to A is naturally also a trek, as is the portion of t from C to B. Let D be a node along t. We say that D blocks t on the A-side of t if D lies along the trek from C to A, and likewise D blocks t on the B-side of t if it lies along the trek from C to B. An ordered pair of sets of variables $\langle S_A, S_B \rangle$ is said to *t-separate* variables A and B in graph $\mathcal{G}$ if for every trek t from A to B in $\mathcal{G}$, t is either blocked on the A-side by a variable in S_A or on the B-side by a variable in S_B. For example, in the graph presented in Figure 2b, for $A = \{Z_4, Z_5\}$, $B = \{Z_6, Z_7\}$, we have $C = U$, and they can be trek separated in multiple ways, including: $S_A = \{U\}$ and $S_B = \{\}$; $S_A = \{\}$ and $S_B = \{U\}$; $S_A = \{U, T\}$ and $S_B = \{U, Z_1\}$; and so on. By comparison, if $A = \{Z_4, Z_6\}$ and $B = \{Z_5, Z_7\}$, then sets like $S_A = \{U\}$ and $S_B = \{\}$ no longer t-separate A from B, since U does not block Z_4 from Z_5 or Z_6 from Z_7. Additional variables would have to be included in S_A or S_B to block these additional treks.

In this paper we make use of the extended trek separation theorems of Spirtes (2013), which relate statements about t-separation in the data generating model to rank constraints in the covariance matrix among certain sets of nodes. The presence or absence of rank constraints can be determined from empirical data using statistical tests for vanishing determinants. Wishart (1928) created a statistical test for the null hypothesis that the determinant of a subcovariance matrix is zero, assuming that the relevant portions of the structural model are Gaussian. Alternative tests (Bollen, 1990; Bollen and Ting, 1993) that relax this distributional assumption tend to be more computationally intensive and appear to have reduced power. Moreover, in practice, even when the distribution is non-Gaussian, the Wishart test often performs well with a large sample (Spirtes, 2013; Silva and Shimizu, 2017). Therefore in our simulations and applications presented later in this paper, we use the Wishart test (Wishart, 1928).

For sets of variables S_1 and S_2 such that $|S_1| = |S_2|$, let D_{S_1,S_2} be the determinant of the subcovariance matrix Σ_{S_1,S_2} of all the variables in S_1 by all the variables in S_2. The Wishart test calculates the distribution of the empirically observed value of $D_{\{a,b\},\{c,d\}}$, denoted as $\widehat{D}_{\{a,b\},\{c,d\}}$, under the null hypothesis that the true determinant is zero and the corresponding tetrad vanishes, i.e., $H_0 : D_{\{a,b\},\{c,d\}} = 0$. Let N be the sample size of the

observed data. The estimated variance of $\widehat{D}_{\{a,b\},\{c,d\}}$ is $\widehat{\sigma}^2 = \{\widehat{D}_{\{a,b\},\{a,b\}}\widehat{D}_{\{c,d\},\{c,d\}}(N+1)/(N-1) - \widehat{D}_{\{a,b,c,d\},\{a,b,c,d\}}\}/(N-2)$ (Wishart, 1928). The Wishart test then computes

$$W = \widehat{D}_{\{a,b\},\{c,d\}}/\widehat{\sigma},$$

and forms a p-value for a two-sided significance test based on the asymptotic normal distribution with mean zero under the null. This is referred to as the vanishing tetrad test.

While the algorithms we present in this manuscript do not rely on Gaussian distributed data, the specific test of vanishing tetrad we use here—the Wishart test—does assume a Gaussian distribution. The same asymptotic correctness results can be extended to other distributions by using other tetrad tests, such as the distribution-free tetrad test developed by Bollen and Ting (1993). We use the Wishart test because it is simple to compute and appeared to have better power at plausible finite sample sizes during numerical experiments.

3. The Statistical Validation of Negative Controls

3.1 Definitions and Assumptions

Before stating the validation test for NCs and then proving its correctness, it will be useful to present some definitions and assumptions. We first define the following simple NC model that involves an unmeasured confounder U and children of U.

Definition 1 *A* simple NC model *is a structural equation or Bayesian network model, with measured variables M, unmeasured variables U, structural causal relationships S and causal coefficients C, that meets the following criteria:*

1. *$|U| = 1$, i.e., there is only one unmeasured variable. Let this unique unmeasured variable be denoted by U.*
2. *U directly causes all elements of M.*
3. *M has two distinct privileged variables, denoted by T and O. T may cause O, but no other variable in M directly causes or is caused by T or O.*

See Figures 1 and 2 for examples of simple NC models. Let variables in $M \setminus \{T, O\}$ be called candidate NCs which are not necessarily valid disconnected NCs.

We clarify that although the simple NC model may appear to be limited to one unmeasured confounder U, such an unmeasured confounder can be a spectrum of multiple latent variables. For example, suppose U denotes healthcare seeking behavior, a common source of unmeasured confounding bias of concern in vaccine effectiveness studies, then it is likely that $U = \sum_j \alpha_j L_j$, where L_j denotes latent variables such as perceptions of illness and treatment, access to healthcare, and insurance coverage, and α_j denotes the corresponding importance of each latent factor.

In this paper, we will focus on identifying a particular class of NC variables, which we refer to as the disconnected NCs. As presented in Figure 1 and defined in Section 2.1, a disconnected NC is independent of both the treatment T and the outcome O conditional on the unmeasured confounder U. That is, any trek between a disconnected NC and T (or O) must pass through U. Hence a tetrad of three disconnected NCs plus T (or O) may be a

vanishing tetrad as introduced in Section 2.3, while a tetrad of two disconnected NCs plus both T and O will not vanish due to the potential direct path from T to O that does not go through U. Motivated by this observation, our validation test leverages the vanishing tetrad test as detailed in the next section and relies on at least three disconnected NCs.

Definition 2 *A* disconnected negative control triplet *(DNCT) is a set of three candidate NCs in a simple NC model such that all treks from one member of the DNCT to another pass through U.*

Our DNCT definition excludes pathways connecting any two disconnected NCs without passing through U, because such a pathway will lead to a tetrad that does not vanish and will also violate the NC assumption (1) when using such a pair of disconnected NCs as double-NC. We make the following assumptions necessary for detecting DNCT in a graph $\mathcal{G}$. We previously defined $\mathcal{G}$ in Section 2.2.

Assumption 1 *The data is generated by a simple NC model with $|M| \geq 5$.*

This implies that there are at least three candidate NCs.

Assumption 2 Tetrad Faithfulness. *In the data distribution implied by the simple NC model, tetrads vanish only if they are implied to vanish by the structure of the simple NC model.*

In other words, tetrads do not vanish as an "accident" of the model's particular coefficients.

Assumption 3 *The data is generated by a simple NC model which is linear and acyclic among its measured variables.*

While we believe that these results likely extend to a larger class of models, our current results are constrained to the class of simple negative control models (SNCMs), defined above. This is still a large class of models able to represent many real-world situations.

3.2 The DNCT Validation Test

We now introduce our proposed validation test for the DNCT. For a particular treatment and outcome, the DNCT validation test determines whether a set of three candidate NCs is a DNCT, i.e., a triplet of disconnected NC variables. It serves as a validation test for determining whether a proposed set of candidate NCs meets the assumptions necessary for causal inference with NCs. This test takes the following as input:

1. A table of data, $Data$;
2. A variable T in the data identified as the treatment;
3. A variable $O \neq T$ in the data identified as the outcome;
4. Three candidate NCs in the data which cannot include T or O;
5. A vanishing tetrad test, $VanTetrad(S_1, S_2, Data) \rightarrow \{TRUE, FALSE\}$, for two sets S_1 and S_2 each containing two variables;

6. A hyperparameter value between 0 and 1, γ, used by the vanishing tetrad test.

The output of this test is a true/false Boolean value: true indicates that the set of three candidate NCs forms a DNCT for T and O, and false indicates that the set of three candidate NCs is not a DNCT for T and O. Our DNCT validation test requires that six specific vanishing tetrad tests must not reject their corresponding null hypotheses that the determinant of a subcovariance matrix is zero, i.e., the corresponding tetrads vanish. More specifically, let X, Y, and Z, be the three candidate NCs being tested. Our test returns $TRUE$ if and only if the following vanishing tetrad tests all accept their null hypotheses that the corresponding tetrads vanish:

1. $VanTetrad(\{X, Y\}, \{Z, T\}, Data)$;
2. $VanTetrad(\{X, Z\}, \{Y, T\}, Data)$;
3. $VanTetrad(\{Z, Y\}, \{X, T\}, Data)$;
4. $VanTetrad(\{X, Y\}, \{Z, O\}, Data)$;
5. $VanTetrad(\{X, Z\}, \{Y, O\}, Data)$;
6. $VanTetrad(\{Z, Y\}, \{X, O\}, Data)$.

These six tetrad tests cover all combinations where exactly one of T or O is included in the tetrad. Intuitively, this tests whether there is more than one pathway connecting any of the candidate NCs to each other or to T or O.

3.3 Correctness of the DNCT Validation Test

We provide the following theorem which states that the DNCT validation test correctly distinguishes valid DNCTs from a set of candidate NCs.

Theorem 3 *Let the data be generated by a simple NC model, $\mathcal{G}$, and assume Assumptions 1, 2, and 3. The DNCT validation test will return TRUE for any DNCT in $\mathcal{G}$, and FALSE otherwise.*

The proof of Theorem 3 follows directly from the proof of two Lemmas, 4 and 5, which we also present here. The proofs for those lemmas are somewhat complicated to follow, and can be found in Appendix A.

Lemma 4 *Let the data be generated by a SNCM, $\mathcal{G}$. Under Assumptions 1 and 3, the DNCT validation test will return TRUE for any DNCT in $\mathcal{G}$.*

This lemma states that as long as the model is a SNCM that is linear and acyclic amongst its measured variables, and has at least 3 candidate negative controls, then if the model contains any valid DNCTs the test will return TRUE for those triplets. Note that this lemma does not require Assumption 2, so the test recovers true DNCTs even if the data distribution implied by the SNCM is not faithful for vanishing tetrads.

Lemma 5 *Let the data be generated by a SNCM, $\mathcal{G}$. Under Assumptions 1, 2, and 3, the DNCT validation test will return FALSE for any set of three candidate negative controls that is not a DNCT in $\mathcal{G}$.*

This lemma states that as long as the model is a SNCM that is linear and acyclic amongst its measured variables, has at least 3 candidate negative controls, and generates a data distribution such that tetrads vanish only if they are implied to by the model's structure, then if the model contains any invalid DNCTs the test will return FALSE for those triplets. In other words, this means that the test correctly rejects invalid DNCTs so long as tetrad faithfulness holds.

Our proofs make critical use of prior work (Sullivant et al., 2010; Spirtes, 2013). In particular, we rely on two theorems in Spirtes (2013) which we present below. In these theorems, A, B, S_A, and S_B are sets of variables. Note that because these theorems are about populations rather than samples, our own theorems inherit this limitation and do not include additional statistical guarantees about convergence rate (Genin, 2021).

Theorem 6 *Extended Trek Separation Theorem 1 (Spirtes 2013). Suppose $\mathcal{G}$ is a directed graph containing S_A, A, S_B, and B, and $(S_A; S_B)$ t-separates A and B in $\mathcal{G}$. Then for all covariance matrices entailed by a fixed parameter structural equation model S with path diagram $\mathcal{G}$ that is linear acyclic below the sets S_A and S_B for A and B, $rank(\Sigma_{A,B}) \leq |S_A| + |S_B|$.*

Theorem 7 *Extended Trek Separation Theorem 2 (Spirtes 2013). For all directed graphs $\mathcal{G}$, if there does not exist a pair of sets S_A, S_B, such that $(S_A; S_B)$ t-separates A and B and $|S_A| + |S_B| \leq r$, then for any S_A, S_B there is a fixed parameter structural equation model S with path diagram $\mathcal{G}$ that is linear acyclic below the sets $(S_A; S_B)$ for A and B that entails $rank(\Sigma_{A,B}) > r$.*

We here provide a brief rationale for how our proofs make use of theorems 6 and 7. By making Assumption 3, we can directly apply the above theorems and use them in our proof of correctness. In particular, we statistically test whether certain subcovariance matrices are rank deficient, that is, if their rank is less than their dimension. For our application to the simple NC models, we will consider subcovariance matrices of the measured variables with dimension two. The presence of U ensures that there is always one trek between any two sets of variables A and B, and the rank of their subcovariance matrix will be at least one. If any other trek connects A to B and does not pass through U, it will force S_A or S_B to contain an additional variable, increasing their collective size to two. Since two is the dimension of the subcovariance matrix, it would not be rank deficient, and the determinant of the subcovariance matrix would not be 0. In this way, these tetrad tests check to see if there is any trek other than the ones through U that would connect a member on one side of the tetrad to a member on the other side of the tetrad.

4. The Find Negative Controls Algorithm

The Find Negative Controls (FindNC) algorithm, summarized in Algorithm 1, searches through a provided set of candidate NCs to identify triplets of candidate NCs that form a simple NC model along with the provided treatment T and outcome O. It uses the DNCT validation test introduced in Section 3.2, and performs a brute force search through the space of all candidate NC triplets. It outputs a collection of all the candidate NC triplets that passed the DNCT test, for a given T, O, and data set.

The FindNC algorithm takes the same input as the DNCT validation test, except that instead of investigating a set of three candidate NCs, the FindNC algorithm takes the set of all candidate NCs as input to identify DNCTs. The FindNC algorithm has one hyperparameter, γ, which is the threshold used for rejecting the null hypothesis in the six vanishing tetrad tests each time it is applied. In our implementation of the FindNC algorithm, the value for this hyperparameter is optional, and it will default to a value of n^{-1}, where n is the provided data set's sample size. Numerical experiments have indicated that this is a reasonable heuristic for choosing γ, with consistently well-balanced performance at a large range of sample sizes.

The FindNC algorithm is a brute force search of the space of candidate NC triplets, using the DNCT validation test. Its correctness thus depends on the completeness of its search, and the correctness of the DNCT validation test. It is clear from its construction that FindNC checks all possible candidate NC triplets, so its search is complete. The correctness of the DNCT validation test was already proven (see Theorem 3). As such, FindNC is correct under the same conditions as the DNCT validation test.

This brute force search scales approximately $|V|^3$ for data with a set of variables V, since it scans through all possible triplets of variables. $|V|^3$ is computationally feasible for many real world domains where NCs might be used, unlike causal discovery which searches a space that scales super exponentially with $|V|$. This obviates the immediate need for more complex and efficient search procedures, and we leave the development of such procedures for future work.

Algorithm 1: Find Negative Controls (FindNC)

Data: *Data* on a set of Candidate NCs (denoted as $CandidateNCs$), treatment T, and outcome O, the $VanTetrad$ function, the DNCT validation test introduced in Section 3.2 (denoted as $DNCTvalidation$), and the threshold parameter γ

Result: A collection of validated DNCTs

```
Output ← ∅
for X, Y, Z ∈ CandidateNCs, and X ≠ Y ≠ Z do
    if DNCTvalidation(Data,X,Y,Z,T,O,γ,VanTetrad) returns TRUE then
        Output ← Output ∪ {X, Y, Z}
return Output
```

5. Data-driven Automated Negative Control Estimation (DANCE)

In this section we combine the FindNC algorithm with NC estimation of the causal effect of treatment on outcome. We first briefly review the literature on identification and estimation of the ATE with a double-NC pair of Z and W. Then we present methods to aggregate information from multiple NC pairs.

5.1 A Brief Review of Double-Negative Control Methods

Miao et al. (2018a) proved that, under a completeness assumption to ensure that Z is sufficiently informative about U, the ATE is identified by

$$\Delta = \int_{-\infty}^{\infty} \{h(w, t=1) - h(w, t=0)\} f(w) dw$$

where $h(w,t)$ is any solution to

$$E[O \mid T=t, Z=z] = \int_{-\infty}^{\infty} h(w,t) f(w \mid z, t) dw.$$

The function $h(w,t)$ has been referred to as the outcome confounding bridge function (Miao et al., 2018b). The NC framework has been extended to proximal causal inference with an alternative identification via the treatment confounding bridge (Tchetgen Tchetgen et al., 2020; Cui et al., 2023). The outcome and treatment confounding bridge functions are analogous to the outcome regression and propensity score models, respectively, in the classical setting when all confounders are measured. In the following, we focus on the outcome confounding bridge function, as similar estimation strategies can be applied to the treatment confounding bridge function.

In practice, one could specify a parametric model $h(W, T; \alpha)$ and jointly estimate $\theta = (\alpha, \Delta)$ by generalized method of moments (GMM) (Miao et al., 2018b; Tchetgen Tchetgen et al., 2020) with the following moment restrictions

$$g(O, T, Z, W; \theta) = \begin{bmatrix} q(1, Z, T)\{O - h(W, T; \alpha)\} \\ \Delta - \{h(W, T=1) - h(W, T=0)\} \end{bmatrix}, \tag{2}$$

where q is a vector function of the same dimension as α. Specifying an identity weighting matrix, the GMM estimator for θ is

$$\widehat{\theta} = \arg\min_{\theta} \overline{g}^{\top} \overline{g}, \tag{3}$$

where $\overline{g} = n^{-1} \sum_{i=1}^{n} g(O_i, T_i, Z_i, W_i; \theta)$ is the average over a sample of n observations. Asymptotic variance can also be estimated (White, 1980; Hansen, 1982; Newey and West, 1987). The above results and methods are still valid with measured confounders. Specifically, to adjust for measured confounders, denoted as X, we solve for moment condition $E[g(O, T, Z, W, X; \theta)] = 0$ where

$$g(O, T, Z, W, X; \theta) = \begin{bmatrix} q(1, Z, T, X)\{O - h(W, T, X; \alpha)\} \\ \Delta - \{h(W, T=1, X) - h(W, T=0, X)\} \end{bmatrix}. \tag{4}$$

In the special case where each child-parent family in the graph $\mathcal{G}$ represents a linear SEM, we have

$$h(W, T; \theta) = \alpha_0 + \alpha_1 W + \Delta T,$$

where $\theta = (\alpha_0, \alpha_1, \Delta)^{\top}$, and correspondingly,

$$g(O, T, Z, W; \theta) = (1, Z, T)^{\top} \{O - h(W, T; \theta)\}.$$

Therefore the moment condition $E[g(O,T,Z,W;\theta)] = 0$ can be solved by two-stage least squares (Angrist and Imbens, 1995; Wooldridge, 2010; Tchetgen Tchetgen et al., 2020). In fact, under linear SEM, Kuroki and Pearl (2014) showed that the causal effect identified from a pair of disconnected NCs has a closed form solution which is given by

$$\begin{aligned}\Delta &= \frac{\text{Cov}(T,O)\text{Cov}(Z,W) - \text{Cov}(Z,O)\text{Cov}(T,W)}{\text{Cov}(T,T)\text{Cov}(Z,W) - \text{Cov}(T,Z)\text{Cov}(T,W)} \\ &= \frac{\text{Cov}(T,O)\text{Cov}(Z,W) - \text{Cov}(W,O)\text{Cov}(T,Z)}{\text{Cov}(T,T)\text{Cov}(Z,W) - \text{Cov}(T,Z)\text{Cov}(T,W)},\end{aligned}$$

where the second equality is due to the vanishing tetrad $\text{Cov}(Z,O)\text{Cov}(T,W) - \text{Cov}(W,O)\text{Cov}(T,Z) = 0$.

5.2 Aggregate Multiple Double-Negative Controls to Estimate the ATE

The FindNC algorithm outputs one or more validated DNCTs, if it detects any valid DNCTs. For each triplet, one can define six possible double-NC pairs each generating a distinct ATE estimate. Therefore, it is necessary to aggregate all six pairs of double-NCs to estimate the ATE. Aggregation can also improve estimation efficiency compared to using one single pair of NCs. In the presence of multiple triplets, there could be overlapping NC pairs, making aggregation more challenging. In this section, we propose methods to aggregate all possible pairs of double-NCs to estimate the ATE assuming an outcome confounding bridge function $h(W,T,X;\alpha)$ defined in (4).

It is important to note that α in the outcome confounding bridge function $h(W,T,X;\alpha)$ is the same regardless of what Z is, and Δ is the same regardless of what the outcome confounding bridge function is. Therefore, for different NC pairs, α may be shared and Δ must be shared. Jointly estimating all parameters while acknowledging that certain parameters are shared by different NC pairs can improve efficiency. To do so, one could stack all moment restrictions $g(O,T,Z,W;\theta)$ each corresponding to a double-NC pair, then estimate the unique parameters via GMM. We provide details about this approach in Appendix B.

In practice, such a joint estimation method can be computationally challenging due to jointly estimating an excessive number of parameters. We thus propose two practical approaches to aggregate information from multiple DNCTs selected by the FindNC algorithm. The first is a majority vote method: we take the most frequently selected pair of NC variables among all triplets, then follow Section 5.1 to estimate a single ATE. This majority vote method is convenient but does not fully utilize all available information.

The second method is to estimate the ATE based on each double-NC pair, and then take a weighted average, where the weights are proportional to the frequency that each NC pair is selected. Asymptotic variance of the weighted average can be estimated by nonparametric bootstrapping. Alternatively, one could aggregate the moment restrictions (2) from all double-NCs and then compute a sandwich variance estimator that takes into account the correlation between different ATE estimates obtained from different NC pairs.

We summarize this method in Algorithm 2, referred to as the Aggregated Negative Control Estimation.

Algorithm 2: Aggregated Negative Control Estimation

Data: *Data*, DNCTs, treatment T, and outcome O
Result: A point estimate $\widehat{\Delta}$ and confidence interval (CI) for the effect of T on O aggregated from all NC pairs in the DNCTs

$Estimates \leftarrow \varnothing$
for $(NC1, NC2, NC3) \in DNCTs$ **do**
  for *each pair* $(NC, NC') \in (NC1, NC2, NC3)$ **do**
    Let $Z = NC$, $W = NC'$, solve for $\widehat{\Delta}$ from (3)
    $Estimates \leftarrow Estimates \cup \{\widehat{\Delta}\}$
    Let $Z = NC'$, $W = NC$, solve for $\widehat{\Delta}$ from (3)
    $Estimates \leftarrow Estimates \cup \{\widehat{\Delta}\}$
Aggregate point estimates $\widehat{\Delta} \leftarrow$ weighted average of *Estimates*
Compute confidence interval via either nonparametric bootstrapping or GMM with aggregated moment restrictions detailed in Appendix B.2
return $\langle \widehat{\Delta},\ CI \rangle$

5.3 Data-driven Automated Negative Control Estimation

Finally, we combine the FindNC algorithm (Algorithm 1) and the Aggregated Negative Control Estimation algorithm (Algorithm 2) into the DANCE algorithm, summarized in Algorithm 3. DANCE first identifies subsets of a set of candidate NC variables that satisfy the assumptions of disconnected NCs, then applies the double-NC method to each pair of these validated NC variables, and finally aggregates information from the outputs to produce an unbiased point estimate and confidence interval for the causal effect of interest in the presence of an unmeasured confounder.

Algorithm 3: Data-driven Automated Negative Control Estimation (DANCE)

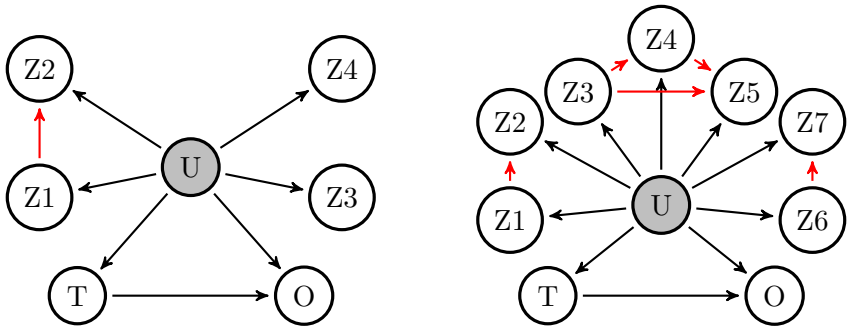
Data: *Data* on a set of Candidate NCs (denoted as $CandidateNCs$), treatment T, and outcome O, the $VanTetrad$ function, the FindNC algorithm, the Aggregated Negative Control Estimation (AggregatedNCE) algorithm and the threshold parameter γ
Result: A point estimate $\widehat{\Delta}$ and confidence interval (CI) for the effect of T on O

DNCTs $\leftarrow FindNC(Data, CandidateNCs, T, O, VanTetrad, \gamma)$
$\langle \widehat{\Delta}, CI \rangle \leftarrow AggregatedNCE(Data, DNCTs, T, O, \theta)$
return $\langle \widehat{\Delta}, CI \rangle$

6. Simulation Study

We perform simulation studies to evaluate the performance of the DANCE algorithm in detecting NCs and estimating the ATE. In these studies, data are generated based on linear

SEMs under two graphical structures: a simple graph and a complex graph, as presented in Figure 2, in which red arrows may lead to violation of the DNCT assumptions for certain NC candidate variables. We first simulate an unmeasured confounder, then all the other variables are generated based on linear SEMs, with edge strengths randomly sampled from uniform distributions. Details about the model parameters and data generating mechanisms are detailed in Appendix C. We consider various sample sizes ranging from 10 to 1,000 for evaluation of NC validation and sample sizes from 300 to 10,000 for evaluation of ATE estimation. We used larger sample sizes for evaluation of ATE estimation to avoid potential computational non-convergence issues. Simulation results are summarized over 200 replications.



(a) Simple Graph (b) Complex Graph

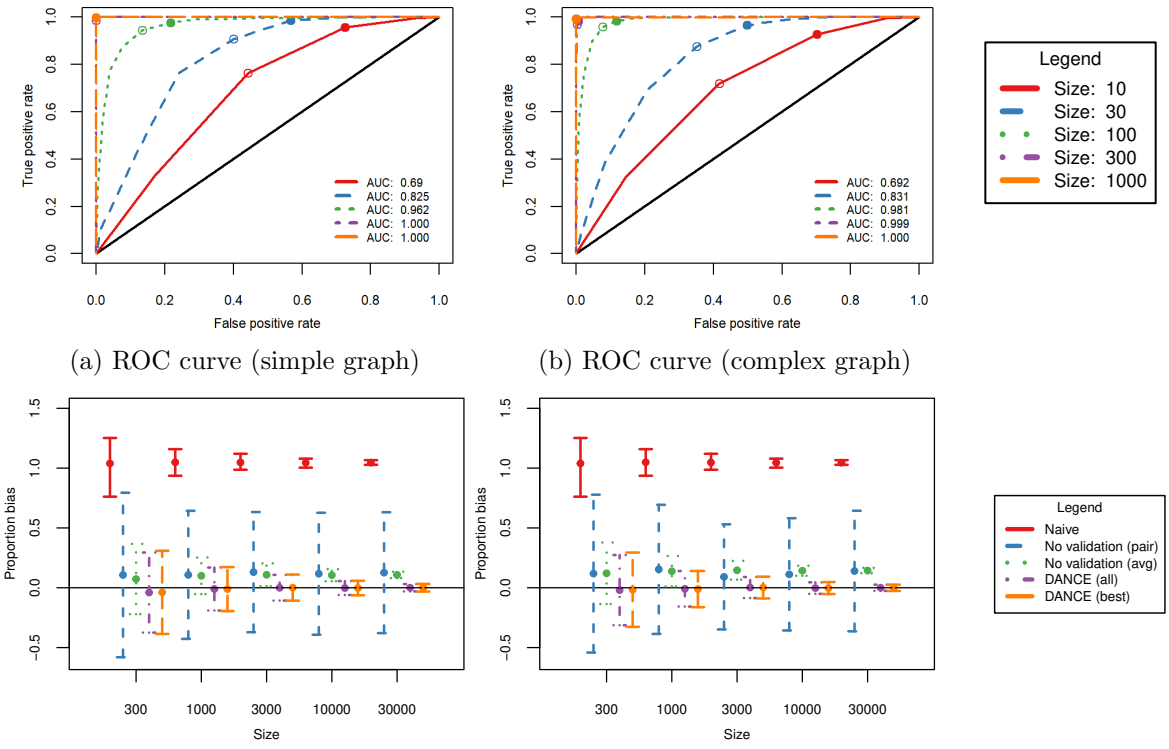
Figure 2: The two different configurations used in the simulation study with a treatment T, an outcome O, an unmeasured confounder U, and multiple candidate NCs $Z_1, \ldots, Z_7$. Due to the existence of red arrows, some candidate NC triplets are not DNCTs.

With data generated from either simple or complex graphical structure, we conduct the following two evaluations. To assess the algorithm's ability to validate candidate NC variables, we plot ROC curves under varying thresholds used for rejecting the null hypothesis in the vanishing tetrad test. To assess the algorithm's accuracy in estimating the causal effect, we compute proportion bias, variance, and coverage probability comparing DANCE algorithm with the following methods:

1. *Naive*: naive regression that ignores unmeasured confounding;
2. *No validation (pair)*: randomly select a pair of NCs without validation to adjust for unmeasured confounding via the double-NC method.
3. *No validation (all)*: use all possible NCs pairs to adjust for unmeasured confounding via Algorithm 2.
4. *DANCE (all)*: use Algorithm 3.
5. *DANCE (best)*: among all DNCTs selected from Algorithm 1, use the one that is most likely to be valid, then apply Algorithm 2

For *DANCE (best)*, we find the DNCT most likely to be valid based on p-values from the vanishing tetrad tests. Recall from Section 3.2 that for each triplet, our DNCT validation test requires that six vanishing tetrad tests must not reject their corresponding null hypotheses that the determinant of a subcovariance matrix is zero, i.e., the corresponding

tetrads vanish. Therefore, we first find the minimum p-value among the six vanishing tetrad tests for each DNCT, then select the DNCT with the hightest minimum p-value.



(a) ROC curve (simple graph) (b) ROC curve (complex graph)

(c) Proportion bias (simple graph) (d) Proportion bias (complex graph)

Figure 3: Simulation results with data generated under simple and complex graphical structures. The first row presents ROC curves for NC validation, and the second row presents proportion bias for each method in estimating the ATE. The solid circle and empty circle on each ROC curve correspond to $\gamma = 1/n$ and $2/n$, respectively.

Graph	Method	Bias (10^{-3})		Proportion Bias (%)		Monte Carlo SE (10^{-3})		Estimated SE (10^{-3})		95% CI Coverage	
	Sample size	1000	3000	1000	3000	1000	3000	1000	3000	1000	3000
Simple	No validation (pair)	67.50	81.90	10.78	13.08	171.10	160.48	68.91	39.27	0.77	0.76
	No validation (all)	62.63	67.96	10.00	10.85	48.95	30.50	50.83	29.10	0.79	0.37
	DANCE (all)	-6.68	0.08	-1.07	0.01	57.38	34.29	57.86	32.95	0.94	0.94
	DANCE (best)	-6.96	0.68	-1.11	1.09	58.77	34.82	58.79	33.46	0.94	0.94
Complex	No validation (pair)	96.31	57.04	15.38	9.10	172.42	140.41	56.24	34.02	0.70	0.79
	No validation (all)	86.55	91.52	13.82	14.61	40.59	25.04	41.95	24.06	0.44	0.05
	DANCE (all)	-5.45	1.11	-0.87	0.18	47.46	28.06	47.99	27.41	0.96	0.93
	DANCE (best)	-6.90	1.03	-1.10	0.17	48.35	29.09	50.51	29.02	0.97	0.95

Table 1: Operation characteristics of different estimators with various approaches for selecting NCs under simple and complex graphs.

Figure 3 presents the ROC curves for validation of NCs and the proportion bias for estimation of the ATE. Each point on each ROC curve is computed by averaging the true and false positive rate over 200 simulated data sets. We further mark two candidate values of the threshold for rejecting the null hypothesis in a vanishing tetrad test, $\gamma = 1/n$ and $2/n$, on the ROC curve with solid circle and empty circle, respectively, to evaluate the impact of choice of γ. The proportion bias is computed by first dividing bias by the true ATE value, then averaging over 200 replications.

Under both simple (Figure 3a) and complex (Figure 3b) graphical structures, we observe improved ROC curve with near perfect separation between valid and invalid NCs when the sample size is greater than or equal to 300. In addition, we found that the default value of γ, $1/n$, tends to lead to high sensitivity but also low specificity, while a larger value such as $2/n$ may improve specificity at the price of a lower sensitivity. The difference between such choices of γ becomes small with a modest sample size of 100. Therefore, in practice, we recommend collecting as much data as possible, and consider the trade-off between sensitivity and specificity when choosing a threshold.

In terms of estimation bias (Figures 3c and 3d and Table 1), DANCE provides an unbiased estimate with improved precision as sample size increases. In particular, DANCE (all) tends to be more efficient than DANCE (best) with smaller Monte Carlo standard error, because it integrates information from all validated NC triplets. In contrast, estimators without validating NCs (the no validation (pair) and no validation (all) methods) are substantially biased. Moreover, DANCE provides valid inference with coverage probability close to the nominal level of 95%. We also observe that, without NC validation, estimation bias tends to be larger under complex graphical structure than simple structure. This is because when data are generated under the simple graphical structure as presented in Figure 2a, there are two DNCTs out of four candidate NC triplets: (Z1, Z3, Z4) and (Z2, Z3, Z4), while under the complex graphical structure as presented in Figure 2b, there are 12 DNCTs out of 35 candidate NC triplets. Therefore, the no validation methods have 50% and 34% chance of correctly selecting valid NCs under the simple and complex graph, respectively.

We further investigate the performance of our proposed method under stronger edge strength in the linear SEMs generating the data. In addition, to assess the sensitivity of the vanishing tetrad test to violations of Gaussian assumptions, we perform simulations in which all variables follow Bernoulli distributions rather than Gaussian distributions. The simulation results are presented in Appendix D. We observe similar results under these more challenging settings.

7. Demonstration on Real World Data

We illustrate our proposed methods with an application to the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) to evaluate the effectiveness of right heart catheterization (RHC) procedure among seriously ill hospitalized adults admitted to the intensive care unit (ICU) (Connors et al., 1996). Many physicians believed that measurements from the RHC procedure can guide therapy and lead to better outcomes for critically ill patients. Due to its popularity and physicians' strong belief, conducting a clinical trial was unethical. In the absence of an RCT, the SUPPORT team

conducted an observational study to evaluate the effectiveness of the RHC procedure. Out of 5,735 critically-ill patients who were considered for the RHC procedure on their admission to an ICU. RHC was performed in 2,184 patients, and the remaining 3,551 patients were managed without RHC. As an observational study, this study does not claim to produce a causal effect estimate. They found that, contrary to expectation, RHC was associated with increased mortality. There was thus concern that such a detected association, with a direction opposite to physicians' belief, is subject to residual confounding. Due to this controversial result, this data set has been further analyzed by many researchers (Lin et al., 1998; Tan, 2006; Li et al., 2018; Mao and Li, 2020; Tchetgen Tchetgen et al., 2020).

A particular concern has been the potential of hidden bias due to confounding. In this application, we aim to find and apply NC variables to estimate the causal effect of RHC on 30-day survival defined as the number of days between admission and death or censoring at 30 days, while accounting for potential unmeasured confounding. The SUPPORT study measured an extensive set of 72 covariates including demographics, comorbidity, vital signs, physiological status, and functional status. We applied our DANCE algorithm to all measured covariates to find valid NCs. Our algorithm identified 43 DNCTs which resulted in 164 unique double-NC pairs out of a total of 258 pairs. The most frequently selected pair of NC variables were two comorbidity variables: one is `dementhx` which stands for dementia, stroke or cerebral infarct, Parkinson's disease, and the other is `gibledhx` which stands for upper gastrointestinal (GI) bleeding. Figure 4 presents the corresponding causal diagram with the two most frequently selected variables. We applied the double-NC method introduced in Section 5.1 using this pair while adjusting for the 42 measured covariates not selected into the DNCTs. Direct adjustment of the measured covariates would lead to a large number of parameters to be estimated for each outcome confounding bridge. With 164 unique double-NC pairs, the total number of parameters to be estimated is extremely large. Therefore, instead of directly adjusting for measured covariates, we adjusted for cubic spline basis of the propensity score obtained by regressing the treatment on the 42 measured covariates via logistic regression. We also applied the method introduced in Section 5.2 to aggregate the estimated RHC effects based on all selected pairs. We further implemented a simple linear regression that regresses the outcome on all measured covariates without any attempt to control for unmeasured confounding. Table 2 presents the estimated effect

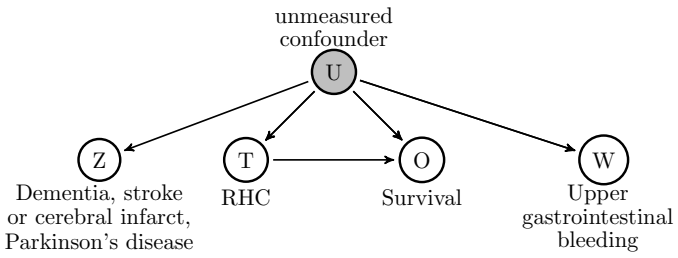


Figure 4: Causal graph for the real-world example with the two most frequently selected disconnected NCs (Z and W), suppressing the measured covariates X which is implicitly conditioned on in all arguments.

of RHC on days alive since admission to ICU up to 30 days. Similar to the simple linear

regression, our methods estimated that RHC has a negative effect on 30-day survival among adults admitted to the ICU. However, different from the regression adjustment approach, our methods provided a 95% confidence interval that covers zero, which is evidence of the uncertainty due to potential unmeasured confounding. An interesting observation is that we obtained very similar results when flipping variables (`dementhx` and `gibledhx`) allocated to W and Z. As discussed in Tchetgen Tchetgen et al. (2020), under the causal graph in Figure 1, the role of W and Z are equivalent, hence causal inference would remain invariant to the choice of W and Z in the double-NC method. Therefore the similarity in our results is strong evidence that the identified pair of NCs satisfy the causal graph in Figure 1. In addition, despite the relatively wide confidence intervals from the majority vote method, we were able to provide more precise inference with a narrow confidence interval by aggregating information from all selected NCs.

Method		RHC effect on days alive (95% CI)
Regression ignoring unmeasured confounding		-1.29 (-1.83, -0.75)
Most frequent NC pair	W=`dementhx`†, Z=`gibledhx`‡	-2.98 (-14.96, 8.99)
	W=`gibledhx`, Z=`dementhx`	-2.80 (-13.42, 7.83)
Aggregate over all NC pairs		-0.19 (-1.50, 1.12)

† Dementia, stroke or cerebral infarct, Parkinson's disease. ‡ Upper GI bleeding.

Table 2: Results from application to the right heart catherization study.

8. Discussion

In this paper we introduced a new test for validating candidate NC variables (i.e., the DNCT validation test), proved its correctness, implemented a correct search procedure using it (i.e., the FindNC algorithm), evaluated the search procedure's performance in simulations, and combined this search procedure with a negative control estimation procedure to produce the DANCE method for causal inference, which we evaluated with simulations and demonstrated on a real world data set. DANCE allows for causal inference in the presence of an unmeasured confounder, but does not assume that the user can identify precisely which variables in the data set meet the strict requirements of NCs. It instead asks only for a collection of candidate NC variables. If no subset of variables passes the DNCT validation test, then DANCE does not produce an estimate, but reports that no variables meet the condition. As such, DANCE does not assume that the provided set of variables includes valid NCs, either.

The methods we provide here still have some limitations. The validation method we provide will throw out some NCs that satisfy a DAG different than Figure 1. Our correctness proof for the NC validation test assumes continuous variables with linear relationships, while many real world scenarios involve binary or categorical variables. DANCE may still work for binary variables, as shown in our simulation study in Appendix D, but this is currently lacking a formal proof. Our correctness proof currently also assumes that the data are generated by a large but still limited set of structures. It is possible to test more general assumptions and more complex structures, and to rule out candidate NCs that do not meet

these structural requirements using methods from the causal discovery literature, but at present there is no proof for the correctness of a more complex method.

In addition, although the proposed Aggregated Negative Control Estimation uses the frequency that each NC pair is selected as weights, one can also consider data-driven methods to select optimal weights that maximize estimation efficiency. Statistical inference should then take into account the selection of weights. Methods for selection of optimal weights warrants future research.

DANCE can be compared to instrumental variable methods (Bowden and Turkington, 1990; Angrist et al., 1996; Greenland, 2000; Martens et al., 2006; Pearl, 2013) and other existing negative control methods for estimating causal effects in the presence of unmeasured confounding. The primary merit of DANCE with respect to instrumental variable methods is that DANCE is bundled with a method for validating many of its own structural assumptions from data. Instrumental variable methods typically require expert opinion to validate assumptions and find instruments. There are some exceptions, such as recent progress on validating instrumental variables and invalid instruments, although these methods (Kang et al., 2016; Silva and Shimizu, 2017; Xie et al., 2022; Burauel, 2023) will not be applicable under circumstances where candidate negative control variables are available but candidate instrumental variables are not. In addition, a valid instrumental variable is a special case of a valid negative control exposure, and an invalid instrumental variable that is associated with the unmeasured confounder is also a valid negative control exposure (Shi et al., 2020b). Thus both valid and invalid instruments may be used in the estimation step of DANCE. Relative to other negative control methods, DANCE can (1) validate candidate negative controls prior to calculating its causal parameter estimate and (2) combine multiple pairs of negative control variables to provide an aggregated estimate, thereby increasing estimation efficiency. To our knowledge DANCE is the first data-driven method to validate negative controls. This does come with some limitations however, as DANCE is more restrictive with regards to the type of negative controls that it can validate. In addition, once validated, there will be at least three negative controls that enter the estimation step, whereas other existing negative control methods (for estimation only) require only one or two negative control variables (Shi et al., 2020b).

Future work can improve the computational efficiency of the validation methods presented here, relax some of the assumptions or limitations, or apply similar techniques to different problems. More complex structural scenarios should also be considered, such as when there is more than one unmeasured confounder.

Acknowledgments

We would like to acknowledge support for this project from the National Institutes of Health (R01GM139926 and NCRR 1UL1TR002494-01).

Appendix A. Technical Lemmas and Proof of Theorem 3

Here we present the proofs for the two technical lemmas that are essential to proving Theorem 3.

Recall that **Lemma 4** states: "Let the data be generated by a SNCM, $\mathcal{G}$. Under Assumptions 1 and 3, the DNCT validation test will return TRUE for any DNCT in $\mathcal{G}$."

Proof Let $\{A, B, C\}$ be an arbitrary DNCT in $\mathcal{G}$. By construction, our DNCT validation test will return TRUE only for candidate negative control triplets in the simple NC model $\mathcal{G}$ for which all 3 vanishing tetrad tests consisting of the members of the triplet and T, and all 3 vanishing tetrad tests consisting of the members of the triplet and O, do not reject the corresponding null hypothesis that the determinant of subcovariance matrix is 0. We proceed by showing that the null hypotheses investigated by these 6 tests will not be rejected.

Since $\{A, B, C\}$ is an arbitrary DNCT in $\mathcal{G}$, by definition of DNCT U must be on all treks from each of $\{A, B, C\}$ to every variable in $\{A, B, C\}$. By the construction of $\mathcal{G}$ as an simple NC model, U is also on all treks from each member of $\{A, B, C\}$ to T and to O. From this point, we can set ourselves up to apply Theorem 6. By the construction of the simple NC model, U is not only on all of the aforementioned treks, but it is also the source of those treks. As such, without loss of generality let $U \in S_A$ and let $S_B = \varnothing$ (as compared to the converse). Then for all partitions of $\{A, B, C, T\}$ and $\{A, B, C, O\}$ into sets S_1 and S_2 with $|S_1| = |S_2|$, $(S_A; S_B)$ t-separates S_1 and S_2 in $\mathcal{G}$, since U is the source of all treks among every pair of variables in $\{A, B, C, T\}$, and the same for $\{A, B, C, O\}$. Finally, with this setup and Assumptions 1 and 3, we can apply Theorem 6, implying that $rank(\Sigma_{S_1,S_2}) \leq |S_A| + |S_B|$. Because $|S_A| + |S_B| = 1$, we have that Σ_{S_1,S_2}, a 2 by 2 matrix, is rank deficient, and has a determinant of 0. As such, $VanTetrad(S_1, S_2)$ will not reject the null hypothesis for all such sets S_1 and S_2. ■

Recall that **Lemma 5** states "Let the data be generated by a SNCM, $\mathcal{G}$. Under Assumptions 1—3, the DNCT validation test will return FALSE for any set of three candidate negative controls that is not a DNCT in $\mathcal{G}$."

Proof Let $\{A, B, C\}$ be an arbitrary triplet of distinct candidate negative controls that is not a DNCT in $\mathcal{G}$. It will suffice to show that for some partition of $\{A, B, C, T\}$ (or $\{A, B, C, O\}$) into sets S_1 and S_2 with $|S_1| = |S_2|$, $VanTetrad(S_1, S_2)$ will reject the null hypothesis that $det(\Sigma_{S_1,S_2}) = 0$.

Assumptions 2 and 3 together imply that if there is no S_A and S_B that t-separate S_1 from S_2 such that $|S_A| + |S_B| \leq 1$, then $rank(\Sigma_{S_1,S_2}) > 1$, and thus $det(\Sigma_{S_1,S_2}) \neq 0$. In the large sample limit, $VanTetrad(S_1, S_2)$ would thus reject the null hypothesis that $det(\Sigma_{S_1,S_2}) = 0$. Therefore, it suffices to show that there is no S_A and S_B that t-separate S_1 from S_2 such that $|S_A| + |S_B| \leq 1$.

Since $\{A, B, C\}$ is not a DNCT, this implies that there is at least one trek from one member of $\{A, B, C\}$ to another that does not pass through U. Without loss of generality, assume this trek goes from A to B. Let $S_1 = \{A, C\}$ and $S_2 = \{B, T\}$. By construction of the simple NC model, U is the source of at least one trek from A to B, with no other variables along that trek.

Let sets S_A and S_B t-separate S_1 from S_2 with S_A on the S_1 side of the t-separating set and S_B on the S_2 side of the t-separating set. By excluded middle, either (I) S_A includes A or S_B includes B, or (II) S_A does not include A and S_B does not include B. We proceed by disjunctive elimination, showing that assuming either (I) or (II) entails that $|S_A| + |S_B| \geq 2$.

First, assume (I). Further, without loss of generality assume that $A \in S_A$ (as compared to $B \in S_B$). This blocks both treks from A to B. By the construction of $\mathcal{G}$ as an simple NC model, there is also a trek from C to T that passes only through U, so either C, T, or U need to be included in either S_A or S_B. Since A is assumed to be in S_A, this entails that $|S_A| + |S_B| \geq 2$.

Second, assume (II). Since S_A and S_B are assumed to t-separate S_1 from S_2, but S_A does not include A, S_B does not include B, and the extra trek between A and B does not include U, there must be another variable, $Y \notin \{A, B, U\}$, that blocks the extra trek between A and B when placed into either S_A or S_B. However Y does not block the trek from A to B that passes only through U, so an additional variable (U) must be added to S_A or S_B in order to block that trek as well, thus $|S_A| + |S_B| \geq 2$.

Therefore, by disjunctive elimination, $|S_A| + |S_B| \geq 2$. This implies that there is no S_A and S_B that t-separate S_1 from S_2 such that $|S_A| + |S_B| \leq 1$, and so the DNCT validation test will return FALSE for $\{A, B, C\}$. ■

Appendix B. Joint Estimation and Inference via Aggregated Moment Restrictions

B.1 Aggregated Estimation With One or More DNCTs

We illustrate our method to aggregate multiple double-negative control pairs across one or more validated DNCTs taking one triplet as an example. For one triplet, $\{A, B, C\}$, one can define six possible double-negative control pairs each generating a distinct ATE, which are $\{W = A, Z = B\}, \{W = B, Z = A\}, \{W = A, Z = C\}, \{W = C, Z = A\}, \{W = B, Z = C\}, \{W = C, Z = B\}$ as listed in Table 3.

W	Z	Freq
A	B	1
B	A	1
A	C	1
C	A	1
B	C	1
C	B	1

Table 3: All possible double-NC pairs from one single triplet, {A,B,C}, and their frequencies.

Correspondingly, we have six moment restrictions and we stack them into the following joint moment restrictions:

$$
\begin{aligned}
&g(O,T,A,B,C,X;\alpha_A,\alpha_B,\alpha_C,\Delta)\\
&=\begin{bmatrix}
q(1,Z=B,T,X)\{O-h(W=A,T,X;\alpha_A)\}\\
q(1,Z=C,T,X)\{O-h(W=A,T,X;\alpha_A)\}\\
q(1,Z=A,T,X)\{O-h(W=B,T,X;\alpha_B)\}\\
q(1,Z=C,T,X)\{O-h(W=B,T,X;\alpha_B)\}\\
q(1,Z=A,T,X)\{O-h(W=C,T,X;\alpha_C)\}\\
q(1,Z=B,T,X)\{O-h(W=C,T,X;\alpha_C)\}
\end{bmatrix},
\end{aligned}
$$

where α_i, $i \in \{A,B,C\}$ is the same regardless of what Z is, and Δ is the same regardless of what the outcome confounding bridge function is. For example, A is an NCO in two negative control pairs: $\{W=A,Z=B\}$ and $\{W=A,Z=C\}$. Consequently, there are two moment functions that share the same parameter α_A in the outcome confounding bridge function. With three disconnected negative controls, there are three unique outcome confounding bridge functions $h(W=i,T=1,X;\alpha_i)$, $i \in \{A,B,C\}$. In addition, all three confounding bridge functions lead to the same ATE Δ, that is, $\Delta = E[h(W=i,T=1,X;\alpha_i) - h(W=i,T=0,X;\alpha_i)]$, for all $i \in \{A,B,C\}$. Therefore, we compute Δ as an average of the three

$$
\Delta = \sum_{i\in\{A,B,C\}} \omega_i \Big\{ E[h(W=i,T=1,X;\alpha_i) - h(W=i,T=0,X;\alpha_i)] \Big\},
$$

where $\omega_i \propto$ {number of all possible moment restrictions for α_i} and $\sum_{i\in\{A,B,C\}} \omega_i = 1$. In fact, the number of all possible moment restrictions for α_i is equal to

$$
\sum_{W=i,Z\in\{A,B,C\}\setminus\{i\}} \text{Freq}_{W,Z},
$$

where $\text{Freq}_{W,Z}$ is the frequency of the NC pair {W,Z} as listed in Table 3. That is, $\omega_i = 2/6 = 1/3$. With the aggregated moment restrictions, we can estimate and make statistical inference on $(\alpha_A,\alpha_B,\alpha_C,\Delta)$ simultaneously using GMM.

Now consider two triplets, {A,B,C} and {A,B,D}. One can define the following ten double-NC pairs in Table 4, where $\{W=A,Z=B\}$ and $\{W=B,Z=A\}$ would appear twice because they are selected in both triplets. In this case, the moment restrictions

W	Z	Freq
A	B	2
B	A	2
A	C	1
C	A	1
B	C	1
C	B	1
A	D	1
D	A	1
B	D	1
D	B	1

Table 4: All possible double-NC pairs from two triplets, {A,B,C} and {A,B,D}, and their frequencies.

become

$$
\begin{aligned}
&g(O,T,A,B,C,X;\alpha_A,\alpha_B,\alpha_C,\Delta)\\
&=\begin{bmatrix}
q(1,Z=B,T,X)\{O-h(W=A,T,X;\alpha_A)\}\\
q(1,Z=C,T,X)\{O-h(W=A,T,X;\alpha_A)\}\\
q(1,Z=D,T,X)\{O-h(W=A,T,X;\alpha_A)\}\\
q(1,Z=A,T,X)\{O-h(W=B,T,X;\alpha_B)\}\\
q(1,Z=C,T,X)\{O-h(W=B,T,X;\alpha_B)\}\\
q(1,Z=D,T,X)\{O-h(W=B,T,X;\alpha_B)\}\\
q(1,Z=A,T,X)\{O-h(W=C,T,X;\alpha_C)\}\\
q(1,Z=B,T,X)\{O-h(W=C,T,X;\alpha_C)\}\\
q(1,Z=A,T,X)\{O-h(W=D,T,X;\alpha_D)\}\\
q(1,Z=B,T,X)\{O-h(W=D,T,X;\alpha_D)\}
\end{bmatrix},
\end{aligned}
$$

Similar to the scenario of one single triplet, all four confounding bridge functions lead to the same ATE Δ. However, because $\{W=A,Z=B\}$ and $\{W=B,Z=A\}$ have a frequency of two, in principle the 1st and the 4th restrictions should in fact both appear twice. This also reflect the fact that more frequently selected double-NC pairs are more likely to be valid NCs, and hence are accounted more during estimation. As such, we have the following weighted average

$$
\Delta=\sum_{i\in\{A,B,C,D\}}\omega_i\Big\{E[h(W=i,T=1,X;\alpha_i)-h(W=i,T=0,X;\alpha_i)]\Big\},
$$

where $\omega_i \propto$ {number of all possible moment restrictions for α_i} and $\sum_{i\in\{A,B,C,D\}}\omega_i=1$. Similarly, the number of all possible moment restrictions for α_i is equal to

$$
\sum_{W=i,Z\in\{A,B,C,D\}\setminus\{i\}}\text{Freq}_{W,Z},
$$

where $\text{Freq}_{W,Z}$ is the frequency of the NC pair {W,Z} as listed in Table 4. That is, $\omega_A=\omega_B=4/12$, while $\omega_C=\omega_D=2/12$.

With multiple DNCTs, there would be more NCOs each corresponding to an outcome confounding bridge function. As such, the number of α parameters as well as the dimension of parameters will grow rapidly, making it computationally challenging to estimate all parameters jointly. The disadvantage of the joint estimation method in terms of lack of computational efficiency may offset the advantage in terms of improved statistical efficiency.

B.2 Variance Estimation

For one triplet, one can define six possible negative control pairs each generating a distinct ATE estimate, while for multiple triplets, there could be overlapping negative control pairs.

For two DCNTs, if there are overlapping negative control pairs selected, then we will have different frequency for different negative control pairs. For example, suppose we identified two DCNTs, {NC1,NC2,NC3} and {NC1,NC2,NC4}, in which NC1 and NC2 appears in both triplets. Then there are ten rather than twelve unique negative control pairs, and the frequency of negative control pairs $\{W = \text{NC1}, Z = \text{NC2}\}$ and $\{W = \text{NC2}, Z = \text{NC1}\}$ is two while the frequency of the rest of the negative control pairs is one. We take an weighted average of the corresponding ten ATE estimates where the weights are proportional to the frequency. Because the aggregated ATE is a linear combination of the unique ATE estimates from individual negative control pairs, it becomes clear that if we can compute the variance-covariance matrix of the ATEs estimated from the unique negative control pairs, then the variance of the aggregated ATE can be computed. We detail our inference method for the aggregated ATE below.

We first introduce notation. Suppose there are K unique negative control pairs denoted by $\{Z^k, W^k : k = 1, \ldots, K\}$. We observe a sample of n observations and for the k-th negative control pair we use $\{T_i, O_i, Z_i^k, W_i^k : i = 1, \ldots, n\}$ to obtain the k-th ATE estimate. Specifically, let $\widehat{\theta}^k = (\widehat{\alpha}^k, \widehat{\Delta}^k) = \arg\min_\theta \overline{g^k}^\top \overline{g^k}$ where

$$\overline{g^k} = \overline{g^k}(O_i, T_i, W_i^k, Z_i^k; \theta) = \frac{1}{n}\sum_{i=1}^{n} g(O_i, T_i, W_i^k, Z_i^k; \theta)$$

and $g()$ is given in eq. (2). Then we define the final aggregated ATE as a weighted average of the ATE estimates from the negative control pairs

$$\widehat{\Delta} = \sum_{k=1}^{K} w_k \widehat{\Delta}^k,$$

where the weight w_k is proportional to the frequency of the k-th negative control pair. Let $\widehat{\theta} = \left\{(\widehat{\theta}^1)^\top, \ldots, (\widehat{\theta}^K)^\top\right\}^\top$ and

$$\begin{aligned}
G_n(\widehat{\theta}) &= \left\{\overline{g^1}(O_i, T_i, W_i^1, Z_i^1; \widehat{\theta}^1)^\top, \ldots, \overline{g^K}(O_i, T_i, W_i^K, Z_i^K; \widehat{\theta}^K)^\top\right\}^\top \\
A_n(\widehat{\theta}) &= \frac{\partial}{\partial\theta} G_n(\widehat{\theta}) \\
B_n(\widehat{\theta}) &= n G_n(\widehat{\theta}) G_n(\widehat{\theta})^\top,
\end{aligned}$$

then we have the following empirical sandwich estimator of the variance-covariance matrix of $\widehat{\theta}$

$$V(\widehat{\theta}) = A_n(\widehat{\theta})^{-1} B_n(\widehat{\theta}) A_n(\widehat{\theta})^{-1}.$$

Note that the aggregated ATE is a linear combination of $\widehat{\theta}$ given by

$$\widehat{\Delta} = \omega^\top \theta,$$

where $\omega = (\overrightarrow{0}_1, w_1, \ldots, \overrightarrow{0}_K, w_K)^\top$ and $\overrightarrow{0}_k$ is a zero vector of the same dimension as α^k, $k = 1, \ldots, K$. Therefore, the variance of $\widehat{\Delta}$ is

$$V(\widehat{\Delta}) = \omega^\top V(\widehat{\theta}) \omega.$$

Appendix C. Data Generating Mechanisms in Simulation Studies

The data generation procedure of the simulation studies is presented below.

In each scenario, data are generated based on linear SEMs under a given graphical structure (Figure 2). Specifically, in the Gaussian graphical model scenario, we first simulate an unmeasured confounder U following a $Normal(0, 2)$ distribution. Then all the other variables are generated based on linear SEMs with exogenous error terms following a $Normal(0, 1)$ distribution. The coefficients in the SEMs were randomly simulated from uniform distributions with certain parameters as detailed in Table 5, and then fixed over all iterations. We generate edges that potentially lead to violation of the disconnected negative control assumptions, i.e., edges between negative controls highlighted in red color in Figure 2, such as the $Z_1 \to Z_2$ in Figures 2a and 2b. We make the strength of such edges stronger by generating the coefficient of negative controls from the uniform distribution with larger parameters than the other coefficients (coefficients of U and T) as detailed in Table 5. For example, under the weak edge strength scenario, coefficients of U and T follow a Uniform distribution between 0.3 and 0.7, and the coefficients of negative controls follow a Uniform distribution between 1.0 and 2.0.

In the binary graphical models scenario, all variables are generated from Bernoulli distributions. We first simulate an unmeasured confounder from a Bernoulli distribution with success probability 0.5. Then we generate all other nodes following Figure 2 from Bernoulli distributions with success probability being the sigmoid function of a linear structural equation. Coefficients of all variables (which are on the log odds ratio scale) are generated from a Uniform distribution between 1.0 and 2.0, and the intercept is fixed at -1.0.

Distribution	Edge Strength	Coefficient of U and T	Coefficient of NCs	Unmeasured Confounder
Gaussian	weak	Unif(0.3, 0.7)	Unif(1.0, 2.0)	Normal(0, 2)
	strong	Unif(0.6, 1.0)	Unif(2.0, 4.0)	Normal(0, 2)
Binary		Unif(1.0, 2.0)	Unif(1.0, 2.0)	Bernoulli(0.5)

Table 5: Parameters in data generating SEMs under Gaussian and binary graphical models.

Appendix D. Additional Simulation Results

To test robustness of the DANCE algorithm, we performed simulation studies under two additional scenarios:

- Stronger edges in the Gaussian graphical models;
- All variables are binary following Bernoulli distributions.

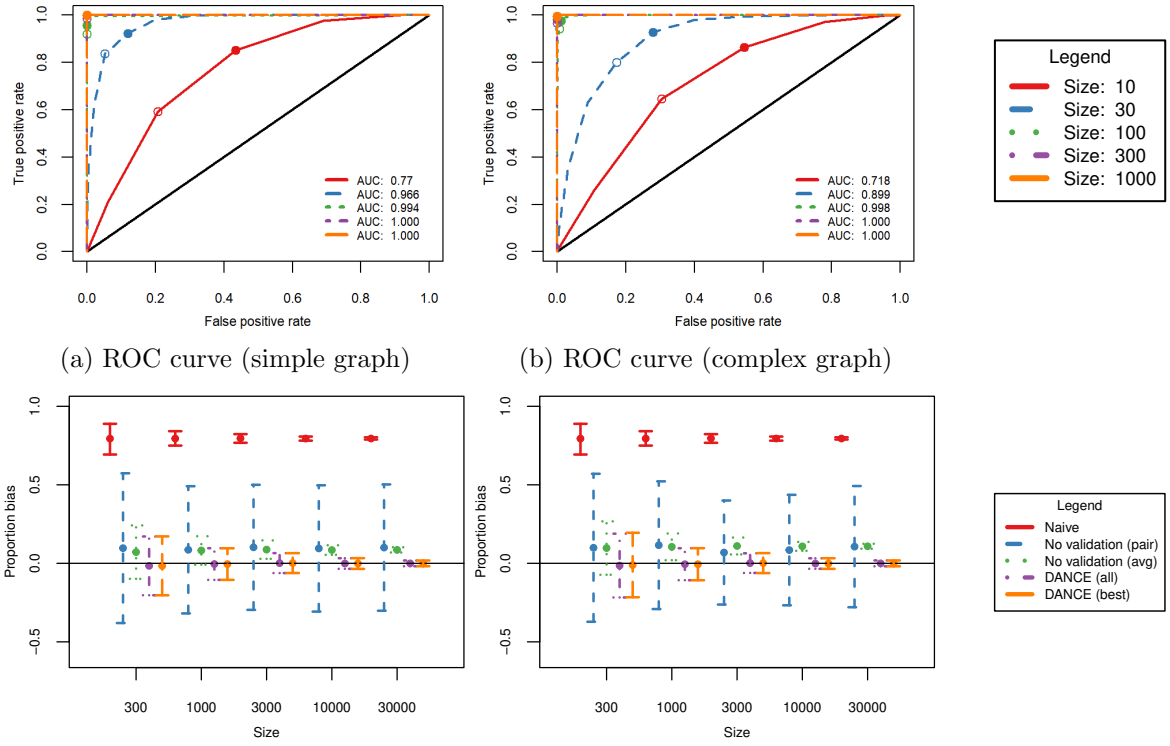
Figure 5 shows the ROC curves for validation of negative controls and the proportion bias for estimation of the ATE when the data were generated under stronger edge strength than Figure 3, and Table 6 presents additional information on the performance of ATE estimation. We observe even more improved ROC curve with near perfect separation between valid and invalid negative controls when the sample size is greater than or equal to 100. In addition, similar to the results under weak edge strength, DANCE provides an unbiased estimate of ATE whereas estimation without NC validation showed statistically significant bias even with large sample size for both simple and complex graphs. DANCE also has near 95% coverage probability.

Figure 6 shows the ROC curves for validation of negative controls and the proportion bias for estimation of the ATE when all random variables generated are binary. As the ROC curve shows, the DANCE algorithm performs well under binary cases under both simple and complex graphical structures. DANCE's estimation of ATE is uniformly better than the Naive and No validation methods.

Notably, in both Figures 5 and 6, we marked on the ROC curve two candidate values of the threshold for rejecting the null hypothesis in a vanishing tetrad test, γ. For continuous data, similar to before, the difference between such choices of γ becomes negligible with a modest sample size of 100; for binary data, due to less information compared to continuous variables, a higher γ level may be preferred to achieve modest level of both sensitivity and specificity when sample size is less than 100.

Graph	Method	Bias (10^{-3})		Proportion Bias (%)		Monte Carlo SE (10^{-3})		Estimated SE (10^{-3})		95% CI Coverage	
Sample size		1000	3000	1000	3000	1000	3000	1000	3000	1000	3000
Simple	No validation (pair)	75.76	89.66	8.61	10.19	181.88	178.83	53.91	29.94	0.78	0.74
	No validation (all)	72.02	76.85	8.18	8.73	40.73	26.34	42.52	24.48	0.58	0.13
	DANCE (all)	-3.99	1.21	-0.45	0.14	45.29	28.53	46.67	26.84	0.95	0.94
	DANCE (best)	-4.25	1.34	-0.48	0.15	45.14	28.43	46.84	26.91	0.95	0.95
Complex	No validation (pair)	101.59	60.47	11.54	6.87	182.58	148.75	51.01	31.20	0.69	0.75
	No validation (all)	92.50	97.10	10.51	11.03	38.20	24.15	39.46	22.61	0.33	0.02
	DANCE (all)	-4.98	1.12	-0.57	0.13	45.72	28.10	46.74	26.70	0.97	0.94
	DANCE (best)	-4.99	1.24	-0.57	0.14	45.83	28.60	47.69	27.46	0.97	0.94

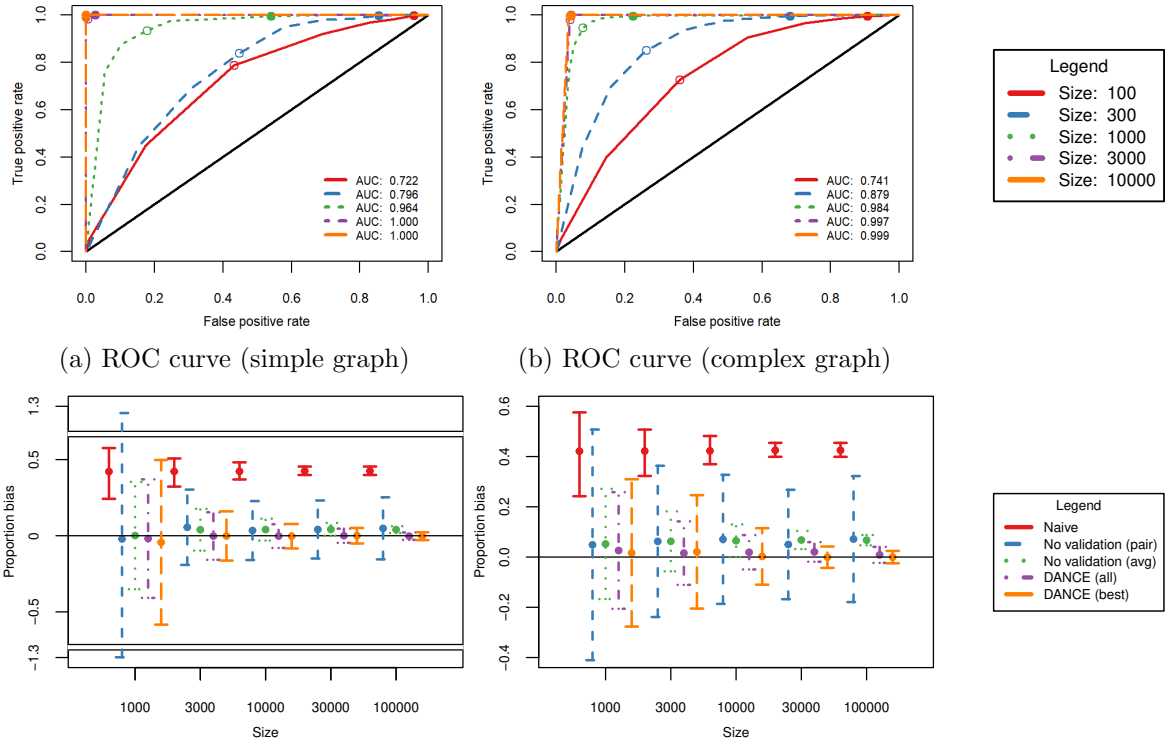
Table 6: Operation characteristics of different estimators with various approaches for selecting NCs under simple and complex graphs with stronger edge strengths.



(a) ROC curve (simple graph) (b) ROC curve (complex graph)

(c) Proportion bias (simple graph) (d) Proportion bias (complex graph)

Figure 5: Simulation results with data generated under simple and complex graphical structures with stronger edge strength. The solid circle and empty circle on each ROC curve correspond to $\gamma = 1/n$ and $2/n$, respectively.



(a) ROC curve (simple graph)

(b) ROC curve (complex graph)

(c) Proportion bias (simple graph)

(d) Proportion bias (complex graph)

Figure 6: Simulation results with data generated under simple and complex graphical structure from binary random variables. The solid circle and empty circle on each ROC curve correspond to $\gamma = 1/n$ and $16/n$, respectively.

References

Joshua D Angrist and Guido W Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995.

Joshua D Angrist and Alan B Keueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996.

Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014.

Kenneth A Bollen. Outlier screening and a distribution-free test for vanishing tetrads. *Sociological Methods & Research*, 19(1):80–92, 1990.

Kenneth A Bollen and Kwok-fai Ting. Confirmatory tetrad analysis. *Sociological Methodology*, pages 147–175, 1993.

Roger J Bowden and Darrell A Turkington. *Instrumental variables*. Number 8. Cambridge university press, 1990.

Patrick F. Burauel. Evaluating instrument validity using the principle of independent mechanisms. *Journal of Machine Learning Research*, 24(176):1–56, 2023.

Stephen Burgess, Dylan S Small, and Simon G Thompson. A review of instrumental variable estimators for mendelian randomization. *Statistical Methods in Medical Research*, 26(5): 2333–2355, 2017.

Alfred F Connors, Theodore Speroff, Neal V Dawson, Charles Thomas, Frank E Harrell, Douglas Wagner, Norman Desbiens, Lee Goldman, Albert W Wu, Robert M Califf, et al. The effectiveness of right heart catheterization in the initial care of critically iii patients. *Journal of the American Medical Association*, 276(11):889–897, 1996.

David Roxbee Cox. Planning of experiments. 1992. Reprint of the 1958 original.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, pages 1–12, 2023.

Ben Deaner. Proxy controls and panel data. *arXiv preprint arXiv:1810.00283*, 2018.

Oliver Dukes, Ilya Shpitser, and Eric Tchetgen Tchetgen. Proximal mediation analysis. *Biometrika*, 110(4):973–987, 2023.

Laura Faden Garabedian, Paula Chu, Sengwee Toh, Alan M Zaslavsky, and Stephen B Soumerai. Potential bias of instrumental variable analyses for observational comparative effectiveness research. *Annals of Internal Medicine*, 161(2):131–138, 2014.

Konstantin Genin. Statistical undecidability in linear, non-gaussian causal models in the presence of latent confounders. *Advances in Neural Information Processing Systems*, 34: 13564–13574, 2021.

David J Glass. *Experimental Design for Biologists*. Cold Spring Harbor Laboratory Press, 2014.

Sander Greenland. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729, 2000.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, pages 1029–1054, 1982.

Miguel A Hernán and James M Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, pages 360–372, 2006.

Antti Hyttinen, Patrik O Hoyer, Frederick Eberhardt, and Matti Järvisalo. Discovering cyclic causal models with latent variables: A general sat-based procedure. In *Uncertainty in Artificial Intelligence*, page 301. Citeseer, 2013.

Lisa A Jackson, Michael L Jackson, Jennifer C Nelson, Kathleen M Neuzil, and Noel S Weiss. Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *International journal of epidemiology*, 35(2):337–344, 2006.

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.

Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

Erich Kummerfeld and Joseph Ramsey. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1655–1664, 2016.

Erich Kummerfeld, Joe Ramsey, Renjie Yang, Peter Spirtes, and Richard Scheines. Causal clustering for 2-factor measurement models. In *Proceedings of the 2014th European Conference on Machine Learning and Knowledge Discovery in Databases-Volume Part II*, pages 34–49, 2014.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

Kendrick Qijun Li, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Doubly robust proximal causal inference under confounded outcome-dependent sampling. *arXiv preprint arXiv:2208.01237*, 2022.

Danyu Y Lin, Bruce M Psaty, and Richard A Kronmal. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, pages 948–963, 1998.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010.

Huzhang Mao and Liang Li. Flexible regression approach to propensity score analysis and its relationship with matching and weighting. *Statistics in Medicine*, 2020.

Edwin P Martens, Wiebe R Pestman, Anthonius de Boer, Svetlana V Belitser, and Olaf H Klungel. Instrumental variables: application and limitations. *Epidemiology*, pages 260–267, 2006.

Wang Miao, Zhi Geng, and Eric Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018a.

Wang Miao, Xu Shi, and Eric Tchetgen Tchetgen. A confounding bridge approach for double negative control inference on causal effects. *arXiv preprint arXiv:1808.04945*, 2018b.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. On the testability of causal models with latent and instrumental variables. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence.*, pages 435–443, 2013.

James M Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

Paul R Rosenbaum. The role of known effects in observational studies. *Biometrics*, 45(2): 557–569, 1989.

Donald Rubin. Discussion of "randomization analysis of experimental data in the fisher randomization test" by D. Basu. *Journal of the American Statistical Association*, 75: 591–593, 1980.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Xu Shi, Wang Miao, Jennifer C Nelson, and Eric Tchetgen Tchetgen. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(2): 521–540, 2020a.

Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. A selective review of negative control methods in epidemiology. *Current Epidemiology Reports*, pages 1–13, 2020b.

Ricardo Silva and Shohei Shimizu. Learning instrumental variables with structural and non-gaussianity assumptions. *Journal of Machine Learning Research*, 18(120):1–49, 2017.

Rahul Singh. Kernel methods for unobserved confounding: Negative controls, proxies, and instruments. *arXiv preprint arXiv:2012.10315*, 2020.

Peter Spirtes. Calculation of entailed rank constraints in partially non-linear and cyclic models. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 606–615, 2013.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT press, 2000.

Seth Sullivant, Kelli Talaska, Jan Draisma, et al. Trek separation for gaussian graphical models. *The Annals of Statistics*, 38(3):1665–1685, 2010.

Sonja A Swanson, Miguel A Hernán, Matthew Miller, James M Robins, and Thomas S Richardson. Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, 113(522):933–947, 2018.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Eric Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Noel S Weiss. Can the "specificity" of an association be rehabilitated as a basis for supporting a causal hypothesis? *Epidemiology*, 13(1):6–8, 2002.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, pages 817–838, 1980.

John Wishart. Sampling errors in the theory of two factors. *British Journal of Psychology*, 19(2):180, 1928.

Jeffrey M Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT press, 2010.

Feng Xie, Yangbo He, Zhi Geng, Zhengming Chen, Ru Hou, and Kun Zhang. Testability of instrumental variables in linear non-gaussian acyclic causal models. *Entropy*, 24(4):512, 2022.

Renjie Yang, Peter Spirtes, Richard Scheines, Steven P Reise, and Maxwell Mansoff. Finding pure submodels for improved differentiation of bifactor and second-order models. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(3):402–413, 2017.

Andrew Ying, Wang Miao, Xu Shi, and Eric Tchetgen Tchetgen. Proximal causal inference for complex longitudinal studies. *arXiv preprint arXiv:2109.07030*, 2021.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.