Estimating Peer Effects in Longitudinal Dyadic Data Using Instrumental Variables

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Published Version
doi:10.1111/biom.12172

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:13890758

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Estimating Peer Effects in Longitudinal Dyadic Data Using Instrumental Variables

A. James O’Malley,¹,* Felix Elwert,² J. Niels Rosenquist,³ Alan M. Zaslavsky,⁴ and Nicholas A. Christakis⁵

¹The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire 03766, U.S.A.
²Department of Sociology, Center for Demography and Ecology, University of Wisconsin-Madison, Madison, Wisconsin 53706, U.S.A.
³Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts 02114, U.S.A.
⁴Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts 02115, U.S.A.
⁵Department of Sociology, Yale Institute for Network Science, Yale University, New Haven, Connecticut 06520, U.S.A.

*email: james.omalley@dartmouth.edu

Summary. The identification of causal peer effects (also known as social contagion or induction) from observational data in social networks is challenged by two distinct sources of bias: latent homophily and unobserved confounding. In this paper, we investigate how causal peer effects of traits and behaviors can be identified using genes (or other structurally isomorphic variables) as instrumental variables (IV) in a large set of data generating models with homophily and confounding. We use directed acyclic graphs to represent these models and employ multiple IV strategies and report three main identification results. First, using a single fixed gene (or allele) as an IV will generally fail to identify peer effects if the gene affects past values of the treatment. Second, multiple fixed genes/alleles, or, more promisingly, time-varying gene expression, can identify peer effects if we instrument exclusion violations as well as the focal treatment. Third, we show that IV identification of peer effects remains possible even under multiple complications often regarded as lethal for IV identification of intra-individual effects, such as pleiotropy on observables and unobservables, homophily on past phenotype, past and ongoing homophily on genotype, inter-phenotype peer effects, population stratification, gene expression that is endogenous to past phenotype and past gene expression, and others. We apply our identification results to estimating peer effects of body mass index (BMI) among friends and spouses in the Framingham Heart Study. Results suggest a positive causal peer effect of BMI between friends.

Key words: Body-mass index; Causality; Directed acyclic graphs; Dyad; Genes; Homophily; Instrumental variable; Longitudinal; Mendelian randomization; Peer effect; Social network; Two-stage least squares.

1. Introduction

We develop instrumental variable (IV) methods for the estimation of causal peer effects using longitudinal dyadic data from a social network. A peer effect (social contagion, induction) occurs when a behavior, trait, or characteristic of an individual’s peers (those to whom she is connected, or altera) affects her own (the ego’s) health behavior. While evidence exists of associations of observed traits (phenotypes and behaviors) among groups of individuals (such as obesity (Christakis and Fowler, 2007), smoking (Christakis and Fowler, 2008), and alcohol use (Rosenquist et al., 2010)), experiments to prove that such associations are causal are often difficult or impossible due to practical or ethical limitations on randomization, albeit with a few exceptions (Wing and Jeffery, 1999; Centola, 2010; Fowler and Christakis, 2010).

Observational analyses may suffer from selection bias due to non-random assignment of treatment. The challenges are magnified in network contexts as confounding takes several structurally different forms. In addition to the spread of health traits because of peer influence, clusters of similar individuals may form due to both homophily (“birds of a feather flock together”) and unmeasured common causes affecting socially connected individuals (confounding). Because each of these phenomena may lead to correlations between the phenotypes of connected individuals (Christakis and Fowler, 2007; Shalizi and Thomas, 2011), methods to parse these associations apart are required.

One approach to causal inference with observational data emulates randomized trials by using an instrumental variable (IV), a variable that influences exposure but, conditional on the exposure, has no influence on the outcome (Angrist, Imbens, and Rubin, 1996). However, the literature on the use of IVs to estimate peer effects is limited. Randomized dorm-room assignments have been used to estimate peer effects among college students (Sacerdote, 2001) and military recruits (Carrell, Fullerton, and West, 2009). In other settings, covariates averaged over neighboring observations (contextual variables) have been used as IVs for peer effects (Fletcher, 2008).

Directed acyclic graphs (DAGs) can clarify the identification problems of IV analysis for peer effects by focusing attention on the causal relationships among variables to
better align the identification strategy with scientific judgments (Pearl, 2009). We use DAGs to (1) identify subtle dependencies that complicate estimation of peer effects, (2) succinctly notate causal data generating models, and (3) prove theorems about identifiability conditions for causal peer effects. We illustrate our methods using networks with a simple structure consisting of disjoint pairs of individuals (dyads), with no influence (interference) between dyads.

Our motivating application concerns peer effects in the Framingham Heart Study (FHS) (Christakis and Fowler, 2007), specifically the utility of using recently sequenced genetic data to develop IVs for peer effects on body mass index among friends and spouses. The appeal of genes as IVs is that they are inherently randomized by a naturally occurring process, are assigned at conception, and are not directly visible and hence, unlikely to directly influence other individuals. Several recent methodological papers discuss Mendelian randomization as IVs (Didelez, Meng, and Sheehan, 2010; Vansteelandt et al., 2011; Palmer et al., 2012) but none consider peer effects. Our paper explores promises as well as pitfalls facing the use of Mendelian randomization as IVs in the study of peer effects.

In Sections 2–4, we introduce DAGs to develop several increasingly general causal models for peer effects involving IVs to account for latent homophily and unmeasured confounding. Our models accommodate several other features often considered obstacles to identifying peer effects, including pleiotropy (genes affecting multiple individual characteristics), population stratification, and gene-based homophily. Section 5 outlines the potential outcomes representation of our preferred causal model. Estimation of these models of peer effects using longitudinal dyadic network data is described in Section 6. Section 7 describes the FHS network of friend and spouse ties and evaluates the linked genetic alleles as potential IVs for peer effects. Section 8 concludes with a discussion.

2. Directed Acyclic Graphs (DAGs)

We use DAGs to encode the structural (i.e., causal) assumptions of our causal models and prove their identifiability. DAGs represent variables as nodes and the direct causal effects between them as edges. Missing edges denote sharp null hypotheses of no direct causal effect. All DAGs considered in this paper are so-called causal DAGs (Pearl, 2009), which are assumed to contain all observed and unobserved common causes in the process. Paths are non-intersecting sequences of adjacent edges, regardless of the direction of the arrows. Causal paths between a treatment and an outcome contain only edges that point away from treatment and toward the outcome. All other paths are noncausal, or spurious, paths. Variable \( M \) is a collider on a path if the path contains the formation \( X \rightarrow M \leftarrow Y \) (i.e., both edges point to \( M \)). All variables directly or indirectly caused by a given variable are called its descendants. Brackets around a variable indicate that the variable has been conditioned on; for example, \([M]\).

The d-separation rule (Pearl, 1988) translates between the causal assumptions encoded in the DAG and the associations observable in data. A path is said to be d-separated or blocked if (1) it contains a non-collider variable that has been conditioned on, such as \( M \) in \( X \rightarrow [M] \rightarrow Y \) (where \( M \) is a common cause or confounder), or if (2) it contains a collider variable, \( X \rightarrow M \leftarrow Y \), and neither the collider nor any of its descendants has been conditioned on. Paths that are not d-separated are said to be d-connected, unblocked, or open. In causal DAGs, variables that are d-separated along all paths are statistically independent; and variables that are d-connected along at least one path may be associated (Verma and Pearl, 1988). The crucial point is that conditioning on a non-collider blocks the flow of association along a path, whereas conditioning on a collider or one of its descendants may induce an association.

Under conventional axioms (Pearl, 2009; Richardson and Robins, 2013), causal DAGs and potential outcomes are equivalent notational systems for predicting statistical associations and identifying the causal effects of an intervention. Since IV is principally an identification strategy for linear models, we henceforth assume that the DAG represents a linear model, making no assumptions about the distribution of the variables (e.g., joint normality).

3. Graphical IV Criteria

We apply versions of the graphical criteria for detecting IVs for the total causal effect of treatment (variable) \( T \) on outcome (variable) \( Y \) in linear models developed by Brito and Pearl (2002).

Single-IV Criterion: Let \( D \) denote the DAG that represents the assumed causal model, and let \( D_{\text{test}} = D \) after removing all edges emanating from \( T \) \((D_{\text{test}} \) represents the null hypothesis of no treatment effect). Then \( G \) is an IV for the total causal effect of \( T \) on \( Y \) conditional on a set of variables \( Z \) (the so-called conditioning set, which may be empty) if:

1. \( Z \) contains no descendant of \( T \) in \( D \).
2. There is an unblocked path between \( G \) and \( T \) in \( D_{\text{test}} \) after conditioning on \( Z \).
3. There is no unblocked path between \( G \) and \( Y \) in \( D_{\text{test}} \) after conditioning on \( Z \).

The first and third conditions give the exclusion restriction: except for the causal effect of \( T \) on \( Y \), the IV \( G \) must be independent of \( Y \) given \( Z \). (However, these conditions do not imply that \( G \) is independent of \( Y \) conditional on \( (T, Z) \)—in the presence of an unmeasured cause of \( T \) and \( Y \), conditioning on \( T \) opens a path from \( G \) to \( Y \) (Hernán and Robins, 2006).) The IV criterion generalizes to multiple treatments and multiple IVs (IV-sets).

IV-Set Criterion: For multivariate \( T = (T_1, \ldots, T_L) \), let \( D_{\text{test}} = D \) after removing all edges emanating from \( T \). Then a multivariate \( G = (G_1, \ldots, G_K) \) is an IV-set for the joint causal effect of \( T \) on \( Y \) conditional on a set of variables \( Z \) if:

1. \( Z \) contains no descendant of \( T \) in \( D \).
2. For every \( l \in \{1, \ldots, L\} \) there exists, for some \( k \), an unblocked path, called path\(_l\), between \( G_k \in G \) and \( T_l \in T \) in \( D_{\text{test}} \) after conditioning on \( Z \), such that \{path\(_1\), \ldots, path\(_L\)\} have no nodes in common.
3. For \( k \in \{1, \ldots, K\} \) there are no unblocked paths between \( G_k \in G \) and \( Y \) in \( D_{\text{test}} \) after conditioning on \( Z \).
Although presented for the case when $\text{q}k$ individually is a valid IV for any single variable $Y_{k}$, the Two Identification Problems: Confounding and homophily bias arises from implicit conditioning on the social tie $A_{12}$, which opens the noncausal path $Y_{2(i)} \leftarrow U_{2} \rightarrow [A_{12}] \leftarrow U_{1} \rightarrow Y_{1(2)}$, $t = 0, 1, 2$. Confounding bias arises from unobserved common causes, $C_{12}$, satisfying $Y_{2(1)} \leftarrow C_{12} \rightarrow Y_{1(2)}$. Although presented for the case when $q = 2$, other cases are represented by dropping (when $q = 1$) or adding (when $q > 2$) $Y_{k}$ and the analogous edges to those involving $Y_{k}(0), k = 1, 2$. Variables $U$ and $C$ are unobserved, all others are observed.

It follows from condition 2 that $K \geq L$ for an IV-set $G$. Importantly, an IV-set $G$ may exist for $T$ even if no variable $G_{k} \in G$ individually is a valid IV for any single variable $T_{i} \in T$ (Brito, 2010). Note that IV sets identify not only the joint effect of $T$ on $Y$ but also the direct effect of each $T_{i}$ on $Y$ not mediated by $(T_{i})_{k \neq i}$, which may coincide with the total causal effect of $T_{i}$ on $Y$.

4. Causal Models for Peer Effects in Dyads

We first present the common core of our causal models for peer effects (on BMI for illustration) to explicate the two central identification challenges: common cause confounding and homophily bias. We then discuss a series of more realistic models for peer effects and evaluate conditions under which each model can be identified via IV analysis.

4.1. The Two Identification Problems: Confounding and Homophily Bias

Figure 1 gives the core of our causal models for a longitudinally observed population of independent dyads including individuals 1 and 2. Let $Y_{k(i)}$ denote BMI, the phenotype of interest for individual $k = 1, 2$ at time $t$ and let $q$ denote the number of periods before the present that the tie was formed (Figure 1 depicts the case when $q = 2$). Current BMI may affect the same individual’s subsequent BMI: $Y_{k(t-1)} \rightarrow Y_{k(t)}$, $k = 1, 2, t = 1, \ldots, q$. Additionally, each individual’s present BMI may affect the other’s subsequent BMI (peer effect): $Y_{2(t-1)} \rightarrow Y_{1(t)}$, and $Y_{1(t-1)} \rightarrow Y_{2(t)}$, $t = 1, \ldots, q$. We assume there were no effects of 1 and 2 on each other prior to tie-formation.

BMI is affected by two more types of variables, each assumed to be at least partially unobserved. The first is a vector of individual-specific unobserved variables $U_{k}, U_{k} \rightarrow Y_{k(t)}$, $(k = 1, 2, t = 0, \ldots, q)$ such as metabolic functioning, food preferences, etc. Second, each individual’s BMI is potentially affected by shared environmental exposures, $C_{12}$, such as local food sources, restaurant commercials, food fads, etc. Thus, $Y_{2(0)} \leftarrow C_{12} \rightarrow Y_{1(0)}$ for some or all of $t = 0, \ldots, q$. Figure 1 depicts a case where $C_{12}$ corresponds to an event at $t = 1$. Finally, $A_{12}$ represents the existence of a social tie between individuals 1 and 2.

Taking the perspective of individual 1, the goal is to identify the total causal effect of $Y_{2(t-1)}$ on $Y_{1(t)}$; that is, the effect of 2’s BMI at time $t - 1$ on 1’s subsequent BMI at time $t = 1, \ldots, q$. Without loss of generality, we focus on the peer effect from $t = q - 1$ to $t = q$. In the causal model of Figure 1, presented with $q = 2$, treatment $Y_{2(q-1)}$ and outcome $Y_{1(q)}$ share three sources of association—one causal and two spurious. First, they may be associated due to unobserved shared environmental confounding by $C_{12}$ along the unblocked non-causal paths $Y_{2(q-1)} \leftarrow C_{12} \rightarrow Y_{2(q-1)}$ and $Y_{2(q-1)} \leftarrow C_{12} \rightarrow Y_{1(q)} \rightarrow Y_{1(q)}$. Third, and centrally for this investigation, treatment and outcome may be associated due to the preferential (nonrandom) formation of social ties. The status of $A_{12}$ may be affected by $(U_{1}, U_{2})$, because, for example, people bond preferentially with others holding similar tastes in food (homophily—“birds of a feather flock together”) or with opposite tastes (heterophily—“opposites attract”). This preferential formation turns $A_{12}$ into a collider variable. Investigating peer effects among individuals linked by a social tie necessarily implies conditioning on the social tie. Since $A_{12}$ is a collider, conditioning on it opens the noncausal path $Y_{2(q-1)} \leftarrow U_{2} \rightarrow [A_{12}] \leftarrow U_{1} \rightarrow Y_{1(q)}$, and hence induces a noncausal association between treatment and outcome. Bias due to falsely considering this association as causal is generally known as homophily bias (Shalizi and Thomas, 2011) and constitutes a type of selection bias (Elwert and Christakis, 2008; Elwert, 2013). This spurious association cannot be eliminated by conditioning on any set of observed variables if the sources of tie formation are at least partially unobserved, and it will exist even if the causal effect of $Y_{2(q-1)}$ on $Y_{1(q)}$ is zero. In fact, using Pearl (1995), it can be shown that common cause confounding in $C_{12}$ and homophily in $A_{12}$ prevent non-parametric identification of the causal effect of $Y_{2(q-1)}$ on $Y_{1(q)}$ under the causal model of Figure 1.

4.2. IV Identification for Various Causal Models of Peer Effects

We now investigate the identification of peer effects despite confounding and homophily bias in several more realistic causal models. Figures 2 and 3 elaborate on the model in Figure 1 in two ways: first, by explicitly adding the observed exogeneous covariates $X_{1}$ (such as gender, age, education, and the geographic distance between ego’s and alter’s residences) and, second, by adding $G_{k}$ (such as genes or other isomorphic variables) affecting BMI but not tie-formation for $k = 1, 2$. We do not index $(X_{1}, U_{k})$ by $t$ but note that these variables may contain time-varying elements.

Figures 2 and 3 differ in only one, albeit crucial, respect. The model in Figure 2 provides for a scenario where the time-invariant (assigned at conception) gene $G$ alone is the
in instrument, whereas Figure 3 supposes that gene expression varies over time due to an interaction with a time-varying covariate in X, GX. We shall refer to these as gene-alone and gene-interaction identification, respectively.

4.2.1. Gene-alone identification. We now evaluate whether G2 can serve as an IV for Y2(q−1) → Y1(q) under various conditioning strategies, where Z denotes the variables conditioned on. Figure 2 includes several different cases based on q. We first suppose the number of periods since tie-formation is q = 1 and then q = 2, and finally draw conclusions for general q. The case when q = 1 can be thought of as estimating a single peer effect over the entire follow-up period, whereas Figure 3 supposes that gene expression can be conditioned on Z = {X2,G1,X1,Y1(0)}, assumed to be a cause of Y2(q−1) through the interaction of G2 with a time-varying variable (e.g., age) in X2, t = 1,...,q (presented when q = 2). The variables X1 and U1 (k = 1, 2) are observed and unobserved individual predictors of Y1, respectively, that may also affect tie-formation. While X1 can be conditioned on U1 cannot, necessitating the use of IV-methods. By conditioning on G2 instruments Y2(q−1) and Y2(0), IV identification is reliant on Y2(q−1) being observed so that they can be instrumental (if dim(G2) ≥ q) and Y2(0),...,Y2(q−1) being not causes of A12 (i.e., they cannot contribute to homophily).

The model in Figure 2 permits conditioning on certain additional variables.

**Corollary 1.** In Figure 2 with q = 1, any subset of Z = {X2,G1,X1,Y1(0)} can be conditioned on in addition to A12 without affecting the IV identifiability of Y2(0) → Y1(1).

**Proof.** The single-IV criterion is met because (1) no variable in Z descends from Y2(0); (2) trivially met; (3) all paths from G2 to Y1(1) in Dtest pass through the colliders Y2(0) and Y2(1), which block these paths and are not opened by conditioning on Z since no variable in Z descends from Y2(0) or Y2(1).

Corollary 1 is useful because all variables in Z are associated with the outcome Y1(1)—conditional on A12 and the other variables in Z—such that conditioning on them will reduce variance in Y1(1) and lead to more precise estimates.

Gene-alone identification fails when q ≥ 2 when G2 is univariate in Figure 2 because no amount of conditioning can remedy several exclusion violations. For example, the open path G2 → Y2(q−2) → Y1(q−1) → Y1(q) can only be blocked by conditioning on Y2(q−2) or Y1(q−1); but doing so would necessarily induce another exclusion violation by opening the path G2 → Y2(q−2) ← U2 → [A1] ← U1 → Y1(q) as Y2(q−2) is a collider on this path and Y1(q−1) descends from this collider. However, the total causal effect of Y2(q−1) → Y1(q) can be identified via the IV-set criterion if G2 is multivariate (e.g.,
representing multiple genes, or multiple alleles of the same
gene, that each affect $Y_2(t)$ over $t = 0, \ldots, q - 1$; dim($G_2) \geq q$.

**Theorem 2.** In the causal model represented by Figure 2
with $q \geq 2$, if dim($G_2) \geq 2$, then $G_2$ is an IV set for the
total causal effect of $Y_2(1)$ on $Y_1(2)$ after conditioning on $A_{12}$.

**Proof.** The IV-set criterion for the joint causal effect of $Y_2(1)$ and $Y_2(0)$ on $Y_1(2)$ is met because (1) $A_{12}$ does not descend from $Y_2(1)$ or $Y_2(0)$; (2) $G_2 \rightarrow Y_2(1)$ and $G_2 \rightarrow Y_2(0)$ are open
and share no nodes (since $G_2$ is multivariate); (3) all paths from $G_2$ to $Y_1(2)$ must pass through $Y_2(0), Y_2(1)$, or $Y_2(2)$, which
are colliders in $D_{\text{test}}$; since neither $Y_2(0), Y_2(1), Y_2(2)$, nor any
of their descendants are conditioned on, all paths from $G_2$ to $Y_1(2)$ are blocked. Finally, since the total causal effect of $Y_2(1)$ on $Y_1(2)$ is not-mediated by $Y_2(0)$, IV set identification
of the joint causal effect of $Y_2(1)$ and $Y_2(0)$ on $Y_1(2)$ implies identification of the total causal effect of $Y_2(1)$ on $Y_1(2)$.

**Corollary 2.** Theorem 2 generalizes to arbitrary $q \geq 2$,
dim($G_2) \geq q$, where $G_2$ instruments $Y_2(0), \ldots, Y_2(q-1)$ with any
subset of $Z = \{X_2, G_1, X_1, Y_1(0)\}$ together with $A_{12}$ as the conditioning set.

**Proof.** Directly extend the proof of Theorem 2 and Corollary 1.

The solution to the identification problem in Figure 2 when
$q \geq 2$, $G_2 \rightarrow Y_2(t)$, $t = 0, \ldots, q$, and dim($G_2) \geq q$ involves an
unusual use of IV. Whereas typically IVs are used to identify
treatment effects, here, $G_2$ both identifies the treatment effect
and remedies the exclusion violation that would occur if the
paths $Y_2(t-1) \rightarrow Y_1(0)$ were not accounted for by instrumenting
$Y_2(t-1)$ for $t = 1, \ldots, q - 1$.

Corollary 2 illustrates that $G_2$ faces an increasing challenge
with the duration of the social tie as all values of the alter
phenotype over 0, …, $q - 1$ must be instrumented. Because
$G_2$ has limited dimension this will eventually be impossible.
The central limitation of gene-alone identification, however,
is that it breaks down under homophily on phenotype.

**Corollary 3.** If $Y_2(t) \rightarrow A_{12}$ for any $t \in \{0, \ldots, q - 1\}$ is
added to Figure 2 then $G_2$ of any dimension is not a valid IV
to identify the total causal effect of $Y_2(q-1)$ on $Y_1(q)$, conditional
on $A_{12}$.

**Proof.** Because $A_{12}$ is a descendant of $Y_2(t)$, conditioning
on $A_{12}$ is equivalent to conditioning on $Y_2(t)$, which
opens the unblockable noncausal path $G_2 \rightarrow [Y_2(t)] \leftrightarrow U_2 \rightarrow [A_{12}] \leftrightarrow U_1 \rightarrow Y_1(q)$, among others, $t = 0, \ldots, q - 1$, representing
an exclusion violation.

Therefore, we next look beyond using genes alone as IVs.

**4.2.2 Gene-interaction identification.** Even though
genes themselves are not time-varying, their expression often
is. The causal model analogous to that of Figure 2 but
with time-varying gene expression is shown in Figure 3. Let
$G_Xk$ denote a variable representing individual $k$’s ($k = 1, 2$) gene-by-age expression at time $t$ (here the notation $G_X$
reflects that age is an element of $X$). The edges $X_t \rightarrow GX_t$
and $G_k \rightarrow GX_k$ are included at all periods to represent
varying gene expression due to age.

**Theorem 3.** In Figure 3 the effect $Y_2(t-1) \rightarrow Y_1(t)$, $t = 1, \ldots, q$ (the case $q = 2$ is presented), is identified by using
$GX_{2(t-1)}$ to instrument $Y_2(t-1)$ conditional on $G_2$, $X_2$, and $A_{12}$.

**Proof.** Because $GX_{2(t-1)}$ only affects $Y_2(t-1)$ the single-IV
criterion applies. Therefore, after conditioning on $A_{12}$, $G_2$,
and $X_3$ an analogous argument as for Theorem 1 completes the proof.

**Corollary 4.** Under the DAG in Figure 3, $G_k \rightarrow A_{12}$, $GX_{k(t-2)} \rightarrow A_{12}$ and $Y_2(t-2) \rightarrow A_{12}$ may be added for $k = 1, 2$,
t = 2, …, q without compromising IV-identification based on
$GX_{2(t-1)}$.

Corollary 4 (proof omitted) illustrates that exploiting time-
varying gene expression is advantageous in three ways. First,
it allows genetic homophily at (or before) $t - 2, 2 \leq t \leq q$. Sec-
ond, it allows homophily on the phenotype of interest up to
but not including $t - 1$. This restriction appears reasonable
given prior work suggesting that changes in physical appearance
(e.g., BMI) have minimal impact on tie-dissolution even
if initial similar appearance led to tie-formation (O’Malley
and Christakis, 2011). Third, the requirements for identification
do not get more onerous with $q$. These flexibilities cen-
trally motivate our adoption of Figure 3 as the primary causal
model in our empirical analysis.

**4.2.3 Relaxing further assumptions.** In observational
data settings, it is important to evaluate the extent to which a
given identification strategy is consistent with multiple plau-
sible causal models. Table 1 summarizes several substantively
important elaborations of the causal models in Figures 2 and 3,
all of which consist of adding edges; that is, relaxing assump-
tions (proofs omitted).

First, as noted previously, homophily on the phenotype at
any time is lethal for gene-alone identification with a single
IV under the model of Figure 2, but homophily on phenotype
prior to $t - 1$ is not lethal for identifying the peer effect from
t - 1 to $t$ under Figure 3.

Second, $G_2$ may be pleiotropic; that is, affect not only BMI,
but also other characteristics of the individual. In Figure 2,
$G_2$ may additionally affect observed covariates $X_2$ (neces-
sitizing conditioning on $Z = \{A_{12}, X_2\}$) but not unobserved fea-
tures directly affecting social-tie formation; that is, $G_2 \rightarrow U_2$
(because of the irreparable exclusion violation $G_2 \rightarrow U_2 \rightarrow
[A_{12}] \leftarrow U_1 \rightarrow Y_1(q)$). By contrast, in Figure 3, adding $G_2 \rightarrow
X_2$, $G_2 \rightarrow U_2$ and even $G_2 \rightarrow A_{12}$ are unproblematic, as is
$GX_{2(t-2)} \rightarrow U_2$ and $GX_{2(t-2)} \rightarrow A_{12}$, $t = 2, \ldots, q$, (but not
$GX_{2(t-1)} \rightarrow U_2$ or $GX_{2(t-1)} \rightarrow A_{12}$). Importantly, pleiotropy
on unobservables ($G_2 \rightarrow U_2$) includes effects of genes on lat-
et pre-tie formation phenotype (which by virtue of being
unobserved is an element of $U_2$). Pleiotropy on latent pre-
tie formation phenotype thus ruins IV identification only in
the case of Figure 2, but it does not ruin IV identification in
Figure 3.
## Table 1

**Extensions to DAGs and their consequence when \( q = 2 \) and individual 1 is the ego**

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Effect</th>
<th>Change to ( Z )</th>
<th>Applies to figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homophily on measured phenotype ((k = 1, 2))</td>
<td>( Y_{t(0)} \rightarrow A_{12} )</td>
<td>No implication</td>
<td>3</td>
</tr>
<tr>
<td>Homophily on measured genotype ((k = 1, 2))</td>
<td>( G_X \rightarrow A_{12} )</td>
<td>No remedy</td>
<td>2, 3</td>
</tr>
<tr>
<td>Pleiotropy on observables</td>
<td>( G_2 \rightarrow X_2 )</td>
<td>Add ( X_2 )</td>
<td>2</td>
</tr>
<tr>
<td>Homophily on unobservables</td>
<td>( G_2 \rightarrow U_2 )</td>
<td>No remedy</td>
<td>2</td>
</tr>
<tr>
<td>Pleiotropy on unobservables(^a)</td>
<td>( G_2 \rightarrow U_2 )</td>
<td>No implication</td>
<td>3</td>
</tr>
<tr>
<td>Population stratification(^b)</td>
<td>( PopStrat_{12} \rightarrow G_2(k = 1, 2) )</td>
<td>Add dyad fixed effects(^c)</td>
<td>2, 3</td>
</tr>
<tr>
<td>Inter-phenotype</td>
<td>((X_2, U_2) \rightarrow Y_{t(0)} )</td>
<td>No implication</td>
<td>2, 3</td>
</tr>
<tr>
<td>Peer effect</td>
<td>((X_2, U_2) \rightarrow Y_{t(1)} )</td>
<td>No implication</td>
<td>2, 3</td>
</tr>
<tr>
<td>Predictor</td>
<td>( X_2 \rightarrow C_{12} )</td>
<td>Add ( X_2 )</td>
<td>2, 3</td>
</tr>
<tr>
<td>Associations</td>
<td>( X_2 \rightarrow U_1, U_2 )</td>
<td>Add ( X_2 )</td>
<td>2, 3</td>
</tr>
<tr>
<td>Confounding on genotype or gene expression</td>
<td>( C_{12} \rightarrow G_2 )</td>
<td>No remedy</td>
<td>2</td>
</tr>
<tr>
<td>Epigenetic</td>
<td>( Y_{2(0)} \rightarrow G_{X2(1)} )</td>
<td>Add ( G_{X2(0)} )</td>
<td>3</td>
</tr>
<tr>
<td>Effects</td>
<td>( Y_{2(1)} \rightarrow G_{X2(2)} )</td>
<td>No implication</td>
<td>3</td>
</tr>
<tr>
<td>Serial dependent gene-expression</td>
<td>( G_{X2(0)} \rightarrow Y_{2(1)} )</td>
<td>Add ( G_{X2(0)} )</td>
<td>3</td>
</tr>
<tr>
<td>Relationship status ((k = 1, 2))</td>
<td>( A_{12} \rightarrow Y_{t(0)} )</td>
<td>No implication</td>
<td>2, 3</td>
</tr>
<tr>
<td></td>
<td>( A_{12} \rightarrow Y_{t(1)} )</td>
<td>No implication</td>
<td>2, 3</td>
</tr>
<tr>
<td></td>
<td>( A_{12} \rightarrow Y_{t(2)} )</td>
<td>No implication</td>
<td>2, 3</td>
</tr>
</tbody>
</table>

\(^a\) Including unmeasured prior phenotype, \( Y_{t(k)} \) for \( t < 0 \) and \( k = 1, 2 \).

\(^b\) Shared ancestry of individuals 1 and 2.

\(^c\) Add indicator variables for each dyad to \( Z \).

Third, population stratification describes an association between \( G_2 \) and \( G_1 \) based on sharing attributes due to common ancestry (Didelez and Sheehan, 2007). To protect the exclusion restriction, one should control for race and ethnicity and ensure (to the extent possible) that members of the dyad are not directly related (e.g., using the method in Price et al. (2006)). However, because ethnic origin (e.g., Irish, German, Greek) is seldom available within general racial groups, including dyad fixed-effects is a more rigorous strategy of accounting for population stratification.

Fourth, our results also accommodate inter-phenotype peer effects; if \( X_2 \) affects \( Y_{t(k)} \), \( t = 0, \ldots, q \), the results above hold. Even if 2’s unobserved characteristics, \( U_2 \), affect \( Y_{t(k)} \), our results continue to hold. Fifth, effects of 2’s observed characteristics on unobserved shared environmental exposures (e.g., via residential choice), \( X_2 \rightarrow C_{12} \), or 1’s observed characteristics, \( X_2 \rightarrow X_1 \), have no implications. Sixth, epigenetic confounding on unobserved contextual factors, \( C_{12} \rightarrow G_{X2(t-2)} \), \( t = 2, \ldots, q \), can be accounted for by conditioning on \( G_{X2(t-2)} \) under Figure 3. Even under epigenetic effects due to the phenotype, which imply the addition of \( Y_{t(t-1)} \rightarrow G_{X_k(t)} \), \( t = 1, \ldots, q \), to Figure 3, identifiability is not affected except if \( t < q \) then \( Y_{t(t-1)} \) must be added to \( Z \).

Finally, if \( G_{X2(t-1)} \rightarrow G_{X2(t)} \), \( t = 1, \ldots, q \), (serial dependence) is added to Figure 3 it is necessary to condition on \( G_{X2(t-2)} \) in addition to \( G_2 \) and \( X_2 \) for \( G_{X2(t-1)} \) (for \( t \geq 2 \)) to be an IV. Therefore, \( G_{X2(t-1)} \) must not be fully determined by \( G_2, X_2, \) and \( G_{X2(t-2)} \). Likewise, if \( G_{X2(t-1)} \rightarrow Y_{2(t)} \) is added to Figure 3 then \( G_{X2(t-2)} \) must be added to \( Z \). In summary, the IV and IV-set criteria permit identification of peer effects in a surprisingly large class of causal models with latent homophily and confounding.

## 5. Potential Outcomes Representation

From hereon, we assume the causal model of Figure 3 and its extensions, which gives IV point identification under linearity and homogeneity (Brito and Pearl, 2002). We now exhibit model form assumptions using the potential outcomes representation of the DAG in Figure 3. We explicitly allow for time-varying elements of \((X_t, U_t)\), \( k = 1, 2 \), and \( C_{12} \) by adding...
the subscript \((t)\), use bold-face font to denote vectors, and use lower-case letters to denote observed and counterfactual values of random variables.

A potential outcome \(Y(\tilde{v})\) is the value of an outcome \(Y\) that would be observed if a variable \(V\) were set by intervention to \(\tilde{v}\). An observed value of \(V\) is denoted \(v\), distinguishing it from the counterfactual \(\tilde{v}\). Therefore, \(Y_1(t)(\tilde{y}_2(t-1), \tilde{g}_x(t-1))\) denotes the potential outcome that would result for individual 1 if individual 2’s phenotype at \(t - 1\) were set to \(\tilde{y}_2(t-1)\) and her gene-expression were set to \(\tilde{g}_x(t-1)\).

Under the DAG in Figure 3, a causal model for the potential outcomes of \(Y_1(t)\) given the conditioning set \(Z(t)\) (which must include \(G_2\) and \(X_2\)) is

\[
Y_1(t)(\tilde{y}_2(t-1), \tilde{g}_x(t-1)) = \alpha_1 \tilde{y}_2(t-1) + \beta^T Z(t) + \lambda_1^T U_1(t) + \lambda_2^T C_{12}(t) + \epsilon_1(t),
\]

where \(\alpha_1, \beta, \lambda_1, \) and \(\lambda_2\) are coefficients and \(\epsilon_1(t)\) is a random error. We assume \(\epsilon_1(t)\) has constant variance, which simplifies estimation, but note that the assumption estimation cannot be relaxed without affecting identification. The involvement of \(U_1(t)\) and \(C_{12}(t)\) in (1) illustrates that causal models make no distinction between observed and unobserved covariates. Due to the exclusion restriction, \(\tilde{g}_x(t-1)\) is absent from the right-hand-side of (1). Therefore, the left-hand-side of (1) may be denoted \(Y_1(t)(\tilde{y}_2(t-1))\). Then the peer effect we seek to estimate satisfies \(\alpha_1 = (Y_1(t)(\tilde{y}_2(t-1)) - Y_1(t)(\tilde{y}_2(t-1))/\tilde{y}_2(t-1))\) for \(\tilde{y}_2(t-1) \neq \tilde{y}_2(t-1)\).

6. Dyadic Instrumental Variables Analysis

To implement IV analysis of (1), we use a two-stage least squares (2SLS) procedure. The “first-stage” of 2SLS regresses the endogeneous variable \(Y_2(t-1)\), \(t = 1, \ldots, q\), on the IV and the exogeneous variables in \(Z(t)\) (including \(g_x(t-2)\) and \(y_1(t-1)\) if conditioned on), yielding the regression

\[
y_2(t-1) = g_x^T y_2(t-1) \theta_1 + z_{t} \theta_2 + \delta_1(t),
\]

from which the fitted values, \(\hat{y}_2(t-1)\), are computed. The second-stage applies OLS to

\[
y_1(t) = \alpha_1 \hat{y}_2(t-1) + z_{t} \beta + \hat{\epsilon}_1(t),
\]

where \(\hat{\epsilon}_1(t) = \epsilon_1(t) + \alpha_1 (y_2(t-1) - \tilde{y}_2(t-1))\), estimating the peer effect \(\alpha_1\). Because \(g_x(t-1)\) is an IV in (2), under OLS estimation \(\hat{y}_2(t-1)\) is orthogonal to \(\epsilon_1(t)\) and \(Z(t)\) in (3), ensuring unbiased and statistically efficient IV-based estimates. The procedure generalizes to accommodate multiple heterogeneous effects such as two-period dependence (i.e., if \(Y_2(t-2) \rightarrow Y_1(t)\)) and effect heterogeneity in observed effect modifiers (see Web Appendix).

6.1. Variance Estimation

Standard errors are computed using results from the general theory for 2SLS. Because the peer effects are of alter’s lagged as opposed to contemporaneous phenotypes, the complications posed by the simultaneous involvement of the same observation as a predictor and an outcome (VanderWeele, Ogburn, and Tchetgen Tchetgen, 2012) are avoided. To account for repeated observations made on dyads over time, as outlined in the Web Appendix, we compute robust standard errors based on sandwich estimators (White, 1982).

7. Friend and Spouse Peer Effect Analysis of the FHS Network

We illustrate our methods using a novel social network dataset constructed from the first seven health exams of the Offspring Cohort of the Framingham Heart Study (FHS), encompassing 32 years of follow-up. The Offspring Cohort includes 5124 individuals. Genetic data was available for 3462 distinct individuals, appearing in 22,361 exams (see Web Appendix).

The network ties considered here arise from participants naming friends and spouses at their health exams. Participants typically only named a single friend at each exam, which is likely to be the one with the most influence. Given the stability of the Framingham population from 1971 to 2003, approximately 50% of the nominated friend contacts were themselves also participants in the FHS and thus provided the same information, including BMI. Most spouses of FHS participants were also FHS participants. We estimate our model with a sample of 9270 unique dyads comprising spousal and nearly disjoint friendship dyads (ignoring occasional overlap of dyads when the same ego is named by multiple alters).

Because the fat mass and obesity gene (FTO) and the melanocortin-4 receptor gene (MC4R) have been confirmed through original and replication studies to be strongly associated with obesity (Speliotes et al., 2010), we consider them as IVs for peer effects of BMI. There is also evidence suggesting that genetic effects may be moderated by a person’s age (Lasky-Su et al., 2008), justifying consideration of age-dependent gene expression as an IV.

Linearity is assumed for the data analysis and, moreover, we are interested in the linear peer effect of BMI itself. However, we note that in certain applications one might instead be interested in peer effects of obesity (BMI \(\geq 30\)), the effect of some other nonlinear transformation of BMI, or in the extent to which the peer effect of BMI is modified by age or some other individual characteristic of the alter (or the ego). While many interesting specifications could be considered, for illustration, we have chosen to focus on a linear specification.

We adjust for ego’s gender, age, gender–age interaction, birth era, birth year, smoking status, number of siblings, geographic distance between residential locations of ego and alter at tie-formation, and gene–age interactions. Birth era accounts for whether an individual was born before 1942, between 1942 and 1948, or 1948 or later to capture possible cohort effects due to America’s involvement in World War II. Because the offspring cohort is nearly 100% white, we do not adjust for race.

In addition, we adjust for wave number dummies to account for secular trends in BMI. Therefore, one can think of gene-age expression as random with respect to exam timing. Inclusion of alter’s smoking status provides assurance against a possible pleiotropic effect between FTO and smoking and MC4R and smoking.

7.1. Representation of Genes

Genetic alleles are represented in \(G_k\), \(k = 1, 2\), by four dummy variables for two of the three possible states of FTO
(states AA, AT, TT) and MC4R (states CC, CT, TT). The A and C alleles have been recognized by geneticists as the risk-alleles of FTO and MC4R, respectively. Having two copies of the risk-allele is the riskiest state followed by the one-copy heterozygous state. Therefore, we also include a fifth dummy variable corresponding to FTO = AA and MC4R = CC. While we could instrument 5 waves of phenotypes using gene-alone IV identification (Figure 2 and Corollary 2), we can relax more assumptions under gene–age interaction IV identification (Figure 3, Theorem 3, and Table 1). The age-dependent association of the FTO gene with BMI is clearly evident in Figure 4 (see Web Appendix for the same for MC4R).

7.2. Dyadic Peer Effect Analyses

We estimate several statistical models, starting with one that is consistent with the causal model of Figure 3, as well as statistical models obtained by adding several of the Exclusions in Table 1. The four reported here condition on $G_1$, $X_1$, and $X_2$ and are distinguished by whether $G_X_{2(t-2)}$ was excluded (as permitted in Figure 3) or conditioned on (to accommodate $G_X_{2(t-2)} \rightarrow G_X_{2(t-1)}$) and by whether $Y_{1(t-1)}$ was excluded or conditioned on (only allowed under Figure 3) to possibly improve precision. Because population stratification is a major concern in analyses involving genes and phenotypes, we include dyad fixed effects in all analyses. Thus, the five gene–age interaction variables of the alter (individual 2) are the IVs for $Y_{2(t-1)}$. We also performed analyses with the analogous five gene–age interaction variables as additional IVs; results remained essentially unchanged (not shown). We perform separate analyses for friends and spouses and use robust variance estimators to account for repeated observations over time (Section 6.1).

7.3. Estimated Peer Effects

The IV estimates are consistent with positive BMI peer effects among friends and spouses (Table 2). Under the causal model of Figure 3 with $Z_{(i)} = (G_X_{1(i)}, X_{1(i)}, X_{2(i)})$, the estimated BMI peer effect among friends (row 1) is positive and statistically significant ($\hat{\alpha}_1 = 0.888$, 95% CI (0.063, 1.713)), whereas the BMI peer effect among spouses (row 5) is positive but not statistically significant ($\hat{\alpha}_1 = 0.099$, 95% CI (–0.324, 0.522)). In all other specifications (i.e., relaxations of Figure 3), the estimated BMI peer effects among friends and spouses are not statistically significant, although point estimates remain in the expected positive direction in most models. For many IV specifications, the corresponding OLS estimates differ appreciably, consistent with the presence of unobserved confounding and homophily bias in the OLS specifications.

The imprecision (and resulting lack of significance) of many of our IV estimates is owed to relatively weak first stages. F-statistics indicate that only the causal models of Figure 3 (see $G_X_{2(t-2)}$ excluded rows of Table 2) have first stages at which IV strength is modest at best by conventional standards (e.g., under row 1, $F_5 = 2.150$ for friends $F_5 = 4.064$ for spouses) (Stock, Wright, and Yogo, 2002). Note, specifically that conditioning on $G_X_{2(t-2)}$ to account for possible serial dependence in gene expression (i.e., if $G_X_{2(t-2)} \rightarrow G_X_{2(t-1)}$ is added to Figure 3) results in a very weak first stage (e.g., $F_5 \leq 0.268$ for spouses). This explains the noisy estimates of all rows with $G_X_{2(t-2)}$ as additional covariates in Table 2. Therefore, the absence of $G_X_{2(t-2)} \rightarrow G_X_{2(t-1)}$ is crucial to IV peer-effect estimation using FHS data. Other specifications (results not shown) yield first stages of similar strength. To improve precision, one might collect more data to increase sample size; or one might (we believe implausibly) assume the absence of unobserved population stratification, which would permit removal of the dyad fixed effects and result in a stronger first stage (results not shown).

8. Conclusion

We derived IV methodology for the estimation of peer effects using longitudinal data. A key methodological distinction of our approach, compared to past observational approaches, is that we account for latent common causes and homophily. An important theoretical finding is that latent homophily places severe demands on IVs. Genes have appeal as IVs due to their inherent randomness, lack of visibility to peers, and ongoing influence on the phenotype. However, ongoing influence on phenotype is problematic to time-invariant IVs such as genetic alleles as all past values of the alter’s phenotype post tie-formation must be instrumented (even if they only have an indirect effect on ego’s BMI). However, if variation in gene expression across age is used as an IV, the dimension of the instrumented variable does not need to increase with the duration of the social tie.

Using two genes widely recognized as having the strongest effects on BMI or obesity, we explored BMI peer effects among
Table 2

<table>
<thead>
<tr>
<th>Discretary Z((i)) terms</th>
<th>IV Regression (2SLS)(^a)</th>
<th>Regression (OLS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(GX_{2(t-2)})</td>
<td>(Y_{1(t-1)})</td>
<td>(F_s^b)</td>
</tr>
<tr>
<td>Nominated friend</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exclude</td>
<td>Exclude</td>
<td>2.150</td>
</tr>
<tr>
<td>Exclude</td>
<td>Covariate</td>
<td>1.731</td>
</tr>
<tr>
<td>Covariate</td>
<td>Exclude</td>
<td>1.181</td>
</tr>
<tr>
<td>Covariate</td>
<td>Covariate</td>
<td>1.144</td>
</tr>
<tr>
<td>Spouse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exclude</td>
<td>Exclude</td>
<td>4.064</td>
</tr>
<tr>
<td>Exclude</td>
<td>Covariate</td>
<td>4.351</td>
</tr>
<tr>
<td>Covariate</td>
<td>Exclude</td>
<td>0.268</td>
</tr>
<tr>
<td>Covariate</td>
<td>Covariate</td>
<td>0.181</td>
</tr>
</tbody>
</table>

\(^a\)\(Z_{(i)} = (GX_{1(i)}, X_{1(i)}, X_{2(i)})\) are exogeneous covariates and \(GX_{2(t-1)}\) is an IV in all IV analyses. The elements of \(X_{k(i)}, k = 1, 2,\) are: gender, age, gender-age interaction, birth era, birth year, smoking status, number of siblings, and (for \(k = 1\) only) the geographic distance between residential locations of ego and alter at tie-formation. All models include dyad fixed effects. \(GX_{2(t-2)}\) and \(Y_{1(t-1)}\) are added to \(Z_{(i)}\) as indicated in the two left-most columns.

\(^b\)The \(F\)-statistic is for the overall effect of the IV, \(GX_{2(t-1)}\), in the first-stage equation. The critical value of the Cragg-Donald \(F\)-statistic, which quantifies the power of an IV, is at the 20% level ranges from 6.71 to 6.77 across the models.

In our study, we compared several techniques for identifying important variables, with a focus on peer effects. We found that the use of instrumental variables (IVs) provided a way to identify peer effects without making strong assumptions about the absence of unobserved homophily or unobserved confounding. We conducted a supplementary analysis of our data, which included an examination of the role of IV methods for observational epidemiology. This analysis was done using a set of IV techniques, including the use of natural or instrumental variables and Mendelian randomization.

Acknowledgements

Research for the paper was supported by NIH grants R01 AG024448 and P01 AG031093 and by a grant from the Pioneer Portfolio of the Robert Wood Johnson Foundation. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for SHARE Affymetrix genotyping was provided by NHLBI Contract N02-HL-64278. Data was downloaded from NIH dbGap, project #780, with accession phs000153.SocialNetwork.v6.p5.c2.NPU.

References


Received January 2012. Revised February 2014. Accepted March 2014.