Biometrics 70, 506–515
September 2014

DOI: 10.1111/biom.12172

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. Christakis[5]

[1]The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire 03766, U.S.A.
[2]Department of Sociology, Center for Demography and Ecology, University of Wisconsin-Madison, Madison, Wisconsin 53706, U.S.A.
[3]Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts 02114, U.S.A.
[4]Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts 02115, U.S.A.
[5]Department 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 *alters*) 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 *mediator*) or $X \leftarrow [M] \rightarrow Y$ (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 $\mathcal{D}$ denote the DAG that represents the assumed causal model, and let $\mathcal{D}_{\text{test}}$ be $\mathcal{D}$ after removing all edges emanating from T ($\mathcal{D}_{\text{test}}$ 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 $\mathcal{D}$.
(2) There is an unblocked path between G and T in $\mathcal{D}_{\text{test}}$ after conditioning on Z.
(3) There is no unblocked path between G and Y in $\mathcal{D}_{\text{test}}$ 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 $\mathcal{D}_{\text{test}}$ be $\mathcal{D}$ 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 $\mathcal{D}$.
(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 $\mathcal{D}_{\text{test}}$ 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 $\mathcal{D}_{\text{test}}$ after conditioning on Z.

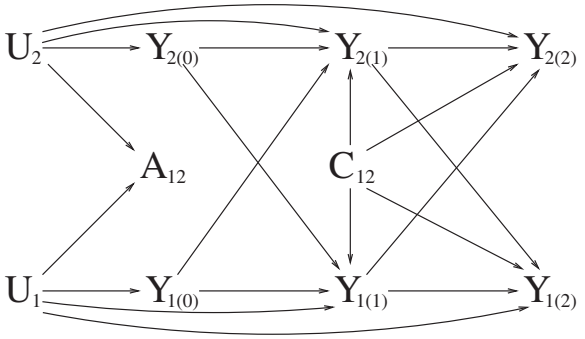

Figure 1. Directed acyclic graph (DAG) representing the common core of causal models for peer effects with observational data. The target of interest is the total causal effect of individual 2's (the alter's) phenotype on individual 1's (the ego's) subsequent phenotype, $Y_{2(1)} \rightarrow Y_{1(2)}$. Latent homophily bias arises from implicit conditioning on the social tie A_{12}, which opens the noncausal path $Y_{2(t)} \leftarrow U_2 \rightarrow [A_{12}] \leftarrow U_1 \rightarrow Y_{1(2)},\ t = 0, 1, 2$. 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_{kt} 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 $T_l \in T$ (Brito, 2010). Note that IV sets identify not only the joint effect of T on Y but also the direct effect of each T_l on Y not mediated by $\{T_k\}_{k \neq l}$, which may coincide with the total causal effect of T_l 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 $Y_{k(t)}$ 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, $Y_{2(t)} \leftarrow C_{12} \rightarrow Y_{1(t)}$ 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, treatment may affect the outcome along the causal path $Y_{2(q-1)} \rightarrow Y_{1(q)}$, the causal effect we aim to identify. Second, they may be associated due to unobserved shared environmental confounding by C_{12} along the unblocked non-causal paths $Y_{2(q-1)} \leftarrow C_{12} \rightarrow Y_{1(q)}$ and $Y_{2(q-1)} \leftarrow C_{12} \rightarrow Y_{1(q-1)} \rightarrow Y_{1(q)}$. 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 generically 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_k (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_k, 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

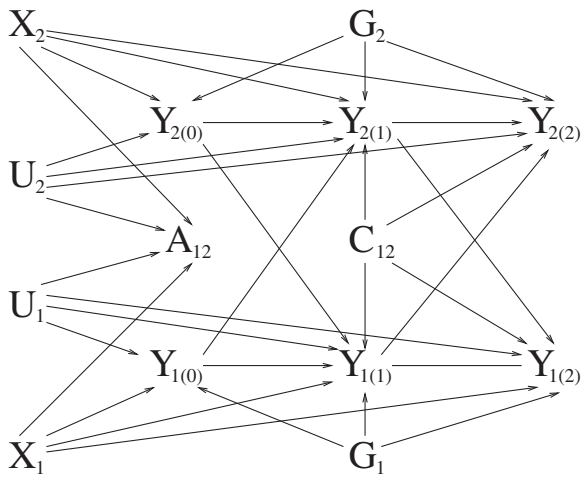

Figure 2. DAG involving time-invariant IV G_2 for causal estimation of $Y_{2(q-1)} \to Y_{1(q)}$ when $t = 0, \ldots, q$. The variables X_k and U_k $(k = 1, 2)$ are observed and unobserved individual predictors of Y_k, respectively, that may also affect tie-formation. While X_k can be conditioned on U_k cannot, necessitating the use of IV-methods. When $q = 1$ (one follow-up period), G_2 instruments $Y_{2(0)}$; when $q = 2$ (the case presented here), G_2 instruments both $Y_{2(1)}$ and $Y_{2(0)}$; and so on until G_2 instruments $Y_{2(q-1)} \ldots Y_{2(0)}$. IV identification is reliant on $Y_{2(0)}, \ldots, Y_{2(q-1)}$ being observed so that they can be instrumented (if $\dim(G_2) \geq q$) and $Y_{2(0)}, \ldots, Y_{2(q)}$ not being causes of A_{12} (i.e., they cannot contribute to homophily).

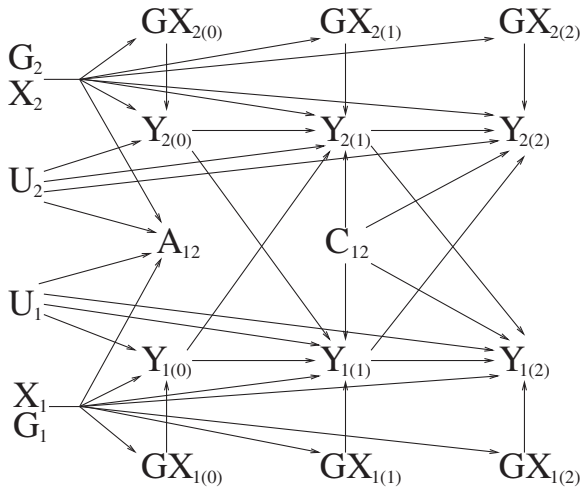

Figure 3. DAG involving time-varying instrumental variable (IV), $GX_{2(t-1)}$, assumed to be a cause of $Y_{2(t-1)}$ through the interaction of G_2 with a time-varying variable (e.g., age) in X_2, $t = 1, \ldots, q$ (presented when $q = 2$). The variables X_k and U_k $(k = 1, 2)$ are observed and unobserved individual predictors of Y_k, respectively, that may also affect tie-formation. While X_k can be conditioned on U_k cannot, necessitating the use of IV-methods. By conditioning on G_2 and X_2, the noncausal pathways from $GX_{2(t-1)}$ to $Y_{1(t)}$ (e.g., $GX_{2(t-1)} \leftarrow G_2 \to [A_{12}] \leftarrow U_1 \to Y_{1(t)}$), $t = 1, 2$, are blocked making $GX_{2(t-1)}$ a valid IV. If $GX_{2(0)} \to GX_{2(1)}$ is added to the DAG, it is necessary to condition on $GX_{2(0)}$.

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 G_2 can serve as an IV for $Y_{2(q-1)} \to Y_{1(q)}$ 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 since tie-formation at $t = 0$ while other cases allow the peer effect to be incrementalized, which is useful if there are time-varying predictors. In this section, we again focus on the peer effect from $t = q - 1$ to $t = q$.

THEOREM 1. *Assume that $q = 1$ in the causal model represented by Figure 2. Then G_2 is an IV for the total causal effect $Y_{2(0)} \to Y_{1(1)}$ conditional on $Z = A_{12}$.*

Proof. Condition (1) of the single-IV criterion is met because A_{12} is not a descendant of $Y_{2(0)}$. Condition (2) is met because the path $G_2 \to Y_{2(0)}$ is a direct effect and hence is unblocked. Condition (3) is met because all paths from G_2 to $Y_{1(1)}$ in $\mathcal{D}_{\text{test}}$ pass through the colliders $Y_{2(0)}$ and $Y_{2(1)}$; since neither $Y_{2(0)}$ nor $Y_{2(1)}$ is conditioned on, and A_{12} is not a descendant of either, all paths from G_2 to $Y_{1(1)}$ in $\mathcal{D}_{\text{test}}$ are blocked. □

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

COROLLARY 1. *In Figure 2 with $q = 1$, any subset of $Z = \{X_2, G_1, X_1, Y_{1(0)}\}$ can be conditioned on in addition to A_{12} without affecting the IV identifiability of $Y_{2(0)} \to Y_{1(1)}$.*

Proof. The single-IV criterion is met because (1) no variable in Z descends from $Y_{2(0)}$; (2) is trivially met; (3) all paths from G_2 to $Y_{1(1)}$ in D_{test} pass through the colliders $Y_{2(0)}$ and $Y_{2(1)}$, which block these paths and are not opened by conditioning on Z since no variable in Z descends from $Y_{2(0)}$ or $Y_{2(1)}$. □

Corollary 1 is useful because all variables in Z are associated with the outcome $Y_{1(1)}$—conditional on A_{12} and the other variables in Z—such that conditioning on them will reduce variance in $Y_{1(1)}$ and lead to more precise estimates.

Gene-alone identification fails when $q \geq 2$ when G_2 is univariate in Figure 2 because no amount of conditioning can remedy several exclusion violations. For example, the open path $G_2 \to Y_{2(q-2)} \to Y_{1(q-1)} \to Y_{1(q)}$ can only be blocked by conditioning on $Y_{2(q-2)}$ or $Y_{1(q-1)}$; but doing so would necessarily induce another exclusion violation by opening the path $G_2 \to [Y_{2(q-2)}] \leftarrow U_2 \to [A_{12}] \leftarrow U_1 \to Y_{1(q)}$ as $Y_{2(q-2)}$ is a collider on this path and $Y_{1(q-1)}$ descends from this collider. However, the total causal effect of $Y_{2(q-1)} \to Y_{1(q)}$ can be identified via the IV-set criterion if G_2 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 $q = 2$, 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) $G_2 \to Y_{2(1)}$ and $G_2 \to Y_{2(0)}$ 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 $\mathcal{D}_{\text{test}}$; 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 $q \geq 2$, $G_2 \to Y_{2(t)}$, $t = 0, \ldots, q$, 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 $Y_{2(t-1)} \to Y_{1(t)}$ 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 $0, \ldots, q-1$ 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 $Y_{2(t)} \to A_{12}$ 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 $G_2 \to [Y_{2(t)}] \leftarrow U_2 \to [A_{12}] \leftarrow U_1 \to Y_{1(q)}$, 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 GX_{kt} denote a variable representing individual k's ($k = 1, 2$) gene-by-age expression at time t (here the notation GX reflects that age is an element of X). The edges $X_k \to GX_k$ and $G_k \to GX_k$ are included at all periods to represent varying gene expression due to age.

THEOREM 3. *In Figure 3 the effect $Y_{2(t-1)} \to Y_{1(t)}$, $t = 1, \ldots, q$ (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_2 an analogous argument as for Theorem 1 completes the proof. □

COROLLARY 4. *Under the DAG in Figure 3, $G_k \to A_{12}$, $GX_{k(t-2)} \to A_{12}$ and $Y_{2(t-2)} \to A_{12}$ may be added for $k = 1, 2$, $t = 2, \ldots, q$ without compromising IV-identification based on $\boldsymbol{GX}_{2(t-1)}$.*

Corollary 4 (proof omitted) illustrates that exploiting time-varying gene expression is advantageous in three ways. First, it allows genetic homophily at (or before) $t-2$, $2 \leq t \leq q$. Second, 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 centrally 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 plausible 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 assumptions (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 (necessitating conditioning on $Z = \{A_{12}, X_2\}$) but not unobserved features directly affecting social-tie formation; that is, $G_2 \to U_2$ (because of the irreparable exclusion violation $G_2 \to U_2 \to [A_{12}] \leftarrow U_1 \to Y_{1(q)}$). By contrast, in Figure 3, adding $G_2 \to X_2$, $G_2 \to U_2$ and even $G_2 \to A_{12}$ are unproblematic, as is $GX_{2(t-2)} \to U_2$ and $GX_{2(t-2)} \to A_{12}$, $t = 2, \ldots, q$, (but not $GX_{2(q-1)} \to U_2$ or $GX_{2(q-1)} \to A_{12}$). Importantly, pleiotropy on unobservables ($G_2 \to U_2$) includes effects of genes on latent 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

Phenomenon	Effect	Change to Z	Applies to figure
Homophily on measured phenotype ($k = 1, 2$)	$Y_{k(0)} \rightarrow A_{12}$ $Y_{k(1)} \rightarrow A_{12}$ $Y_{k(2)} \rightarrow A_{12}$	No implication No remedy No remedy	3 2, 3 2, 3
Homophily on measured genotype ($k = 1, 2$)	$G_k \rightarrow A_{12}$ $GX_{k(0)} \rightarrow A_{12}$ $GX_{k(1)} \rightarrow A_{12}$	No implication No implication No remedy	3 3 3
Pleiotropy on observables	$G_2 \rightarrow X_2$ $G_2 \rightarrow X_2$	Add X_2 No implication	2 3
Pleiotropy on unobservables[a]	$G_2 \rightarrow U_2$ $G_2 \rightarrow U_2$	No remedy No implication	2 3
Population stratification[b]	$\text{PopStrat}_{12} \rightarrow G_k (k = 1, 2)$	Add dyad fixed effects[c]	2, 3
Inter-phenotype Peer effect	$(X_2, U_2) \rightarrow Y_{1(0)}$ $(X_2, U_2) \rightarrow Y_{1(1)}$ $(X_2, U_2) \rightarrow Y_{1(2)}$	No implication No implication No implication	2, 3 2, 3 2, 3
Predictor Associations	$X_2 \rightarrow X_1$ $X_2 \rightarrow C_{12}$ $X_2 \rightarrow U_1, U_2$	No implication Add X_2 Add X_2	2, 3 2, 3 2, 3
Confounding on genotype or gene expression	$C_{12} \rightarrow G_2$ $C_{12} \rightarrow GX_{2(1)}$ $C_{12} \rightarrow GX_{2(0)}$	No remedy No remedy Add $GX_{2(0)}$	2 3 3
Epigenetic Effects	$Y_{2(0)} \rightarrow GX_{2(1)}$ $Y_{2(1)} \rightarrow GX_{2(2)}$	Add $Y_2(0)$ No implication	3 3
Serial dependent gene-expression	$GX_{2(0)} \rightarrow GX_{2(1)}$ $GX_{2(0)} \rightarrow Y_{2(1)}$	Add $GX_{2(0)}$ Add $GX_{2(0)}$	3 3
Relationship status ($k = 1, 2$)	$A_{12} \rightarrow Y_{k(0)}$ $A_{12} \rightarrow Y_{k(1)}$ $A_{12} \rightarrow Y_{k(2)}$	No implication No implication No implication	2, 3 2, 3 2, 3

[a]Including unmeasured prior phenotype, $Y_{k(t)}$ 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 $Y_{1(t)}$, $t = 0, \ldots, q$, the results above hold. Even if 2's unobserved characteristics, U_2, affect $Y_{1(t)}$, 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 on 1's observed characteristics, $X_2 \rightarrow X_1$, have no implications. Sixth, epigenetic confounding on unobserved contextual factors, $C_{12} \rightarrow GX_{2(t-2)}$, $t = 2, \ldots, q$, can be accounted for by conditioning on $GX_{2(t-2)}$ under Figure 3. Even under epigenetic effects due to the phenotype, which imply the addition of $Y_{k(t-1)} \rightarrow GX_{k(t)}$, $t = 1, \ldots, q$, to Figure 3, identifiability is not affected except if $t < q$ then $Y_2(t-1)$ must be added to Z.

Finally, if $GX_{2(t-1)} \rightarrow GX_{2(t)}$, $t = 1, \ldots, q$, (*serial dependence*) is added to Figure 3 it is necessary to condition on $GX_{2(t-2)}$ in addition to G_2 and X_2 for $GX_{2(t-1)}$ (for $t \geq 2$) to be an IV. Therefore, $GX_{2(t-1)}$ must not be fully determined by G_2, X_2, and $GX_{2(t-2)}$. Likewise, if $GX_{2(t-1)} \rightarrow Y_{2(t)}$ is added to Figure 3 then $GX_{2(t-2)}$ 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_k, U_k), $k = 1, 2$, 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, $Y_{1(t)}(\tilde{y}_{2(t-1)}, \boldsymbol{gx}_{2(t-1)})$ 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 $\boldsymbol{gx}_{2(t-1)}$.

Under the DAG in Figure 3, a causal model for the potential outcomes of $Y_{1(t)}$ given the conditioning set $\boldsymbol{Z}_{(t)}$ (which must include $\boldsymbol{G}_2$ and $\boldsymbol{X}_2$) is

$$Y_{1(t)}(\tilde{y}_{2(t-1)}, \boldsymbol{gx}_{2(t-1)}) = \alpha_1 \tilde{y}_{2(t-1)} + \boldsymbol{\beta}^T \boldsymbol{Z}_{(t)} + \boldsymbol{\lambda}_1^T \boldsymbol{U}_{1(t)} + \boldsymbol{\lambda}_2^T \boldsymbol{C}_{12(t)} + \epsilon_{1(t)}, \quad (1)$$

where α_1, $\boldsymbol{\beta}$, $\boldsymbol{\lambda}_1$, and $\boldsymbol{\lambda}_2$ 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 can be relaxed without affecting identification. The involvement of $\boldsymbol{U}_{1(t)}$ and $\boldsymbol{C}_{12(t)}$ in (1) illustrates that causal models make no distinction between observed and unobserved covariates. Due to the exclusion restriction, $\boldsymbol{gx}_{2(t-1)}$ 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 $\alpha_1 = (Y_{1(t)}(\tilde{y}'_{2(t-1)}) - Y_{1(t)}(\tilde{y}_{2(t-1)}))/(\tilde{y}'_{2(t-1)} - \tilde{y}_{2(t-1)})$ 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 $\boldsymbol{Z}_{(t)}$ (including $\boldsymbol{gx}_{2(t-2)}$ and $y_{1(t-1)}$ if conditioned on), yielding the regression

$$y_{2(t-1)} = \boldsymbol{gx}_{2(t-1)}^T \boldsymbol{\theta}_1 + \boldsymbol{z}_{(t)}^T \boldsymbol{\theta}_2 + \delta_{1(t)}, \quad (2)$$

from which the fitted values, $\hat{y}_{2(t-1)}$, are computed. The second-stage applies OLS to

$$y_{1(t)} = \alpha_1 \hat{y}_{2(t-1)} + \boldsymbol{z}_{(t)}^T \boldsymbol{\beta} + \hat{\epsilon}_{1(t)}, \quad (3)$$

where $\hat{\epsilon}_{1(t)} = \epsilon_{1(t)} + \alpha_1(y_{2(t-1)} - \hat{y}_{2(t-1)})$, estimating the peer effect α_1. Because $\boldsymbol{gx}_{2(t-1)}$ is an IV in (2), under OLS estimation $\hat{y}_{2(t-1)}$ is orthogonal to $\hat{\epsilon}_{1(t)}$ and $\boldsymbol{Z}_{(t)}$ 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 ≥ 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 $\boldsymbol{G}_k$, $k = 1, 2$, by four dummy variables for two of the three possible states of each of FTO

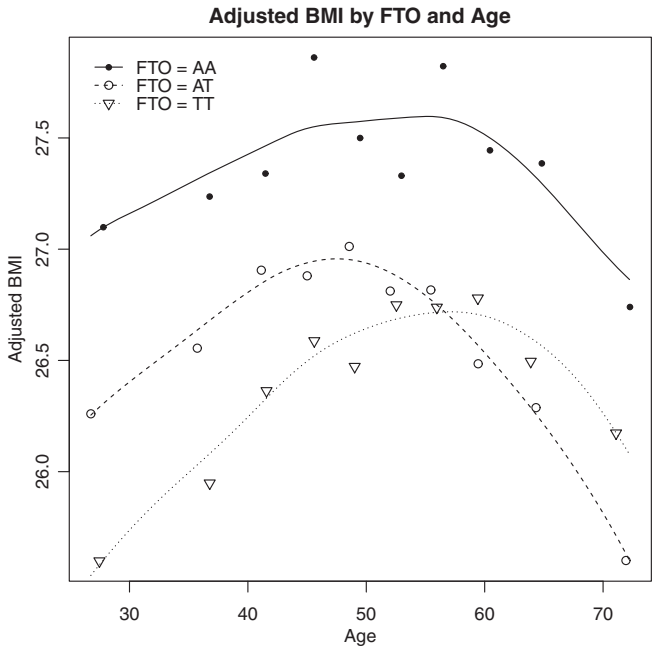

Figure 4. Fitted values of BMI, $\hat{Y}_{i(t)}$, across the $i = 1, \ldots, n$ individuals in the FHS sample are obtained from a regression of BMI on exam (categorical), gender, birth era (categorical), year born, marital status, number of siblings, and smoking status. The smooth curves are computed using a generalized additive spline regression model with smoothing factors judiciously chosen to capture local trends but not overfit the data.

(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 $\boldsymbol{G}_1$, $\boldsymbol{X}_1$, and $\boldsymbol{X}_2$ and are distinguished by whether $\boldsymbol{GX}_{2(t-2)}$ was excluded (as permitted in Figure 3) or conditioned on (to accommodate $\boldsymbol{GX}_{2(t-2)} \rightarrow \boldsymbol{GX}_{2(t-1)}$) 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^2 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 $\boldsymbol{Z}_{(t)} = (\boldsymbol{GX}_{1(t)}, \boldsymbol{X}_{1(t)}, \boldsymbol{X}_{2(t)})$, 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 $\boldsymbol{GX}_{2(t-2)}$ 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 $\boldsymbol{GX}_{2(t-2)}$ to account for possible serial dependence in gene expression (i.e., if $\boldsymbol{GX}_{2(t-2)} \rightarrow \boldsymbol{GX}_{2(t-1)}$ 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 $\boldsymbol{GX}_{2(t-2)}$ as additional covariates in Table 2. Therefore, the absence of $\boldsymbol{GX}_{2(t-2)} \rightarrow \boldsymbol{GX}_{2(t-1)}$ 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
Dyadic peer effect analysis of lag alter BMI using time-varying gene–age expression as an instrument

Discretionary $\boldsymbol{Z}_{(t)}$ terms		IV Regression (2SLS)[a]				Regression (OLS)		
$\boldsymbol{G}\boldsymbol{X}_{2(t-2)}$	$Y_{1(t-1)}$	F_5[b]	Estimate	95% CI		Estimate	95% CI	
			Nominated friend					
Exclude	Exclude	2.150	0.888	0.063	1.713	−0.011	−0.121	0.100
Exclude	Covariate	1.731	0.874	−0.031	1.779	0.009	−0.071	0.089
Covariate	Exclude	1.181	0.133	−0.796	1.062	−0.086	−0.193	0.021
Covariate	Covariate	1.144	−0.003	−0.911	0.906	−0.077	−0.181	0.028
			Spouse					
Exclude	Exclude	4.064	0.099	−0.324	0.522	0.066	0.039	0.094
Exclude	Covariate	4.351	0.101	−0.287	0.488	0.032	0.008	0.055
Covariate	Exclude	0.268	−0.102	−1.855	1.652	0.050	0.017	0.082
Covariate	Covariate	0.181	0.906	−1.832	3.643	0.023	−0.006	0.051

[a] $\boldsymbol{Z}_{(t)} = (\boldsymbol{G}\boldsymbol{X}_{1(t)}, \boldsymbol{X}_{1(t)}, \boldsymbol{X}_{2(t)})$ are exogeneous covariates and $\boldsymbol{G}\boldsymbol{X}_{2(t-1)}$ is an IV in all IV analyses. The elements of $\boldsymbol{X}_{k(t)}$, $k = 1, 2$, 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. $\boldsymbol{G}\boldsymbol{X}_{2(t-2)}$ and $Y_{1(t-1)}$ are added to $\boldsymbol{Z}_{(t)}$ as indicated in the two left-most columns.

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

pairs of friends or spouses. Our analyses, which attempted to account for all sources of confounding, estimated large peer effects but lacked significance in all but one case.

Continued research on the use of genes as IVs for peer effects is motivated by the fact that, if this approach is successful, many important medical, sociological, and economic questions might be more rigorously answered than they have been in the past without having to make strong assumptions about absence of unobserved homophily or unobserved confounding. Conclusive evidence of peer effects would confirm that treatment of traits such as obesity, smoking, alcoholism, and depression could be improved by treating an individual's peers in addition to himself, or by intervening on the composition of his peer group to remove undesirable peer influences.

9. Supplementary Web Appendix

Web Appendices, Tables, and Figures referenced in Sections 6, 6.1, 7, and 7.1 and additional references are available with this paper at the *Biometrics* website on Wiley Online Library. Example code, example data, and associated instructions for running the code are also available as a web supplement (same website).

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

Angrist, J., Imbens, G., and Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**, 444–455.

Brito, C. (2010). Instrument sets. In *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, R. Dechter, H. Geffner, and J. Halpern (eds), 295–307. London: College Publications.

Brito, C. and Pearl, J. (2002). A new identification condition for recursive models with correlated errors. *Structural Equation Modeling* **9**, 459–474.

Carrell, S. E., Fullerton, R. L., and West, J. E. (2009). Does your cohort matter? Estimating peer effects in college achievement. *Journal of Labor Economics* **27**, 439–464.

Centola, D. (2010). The spread of behavior in an online social network. *Science* **329**, 1194–1197.

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine* **357**, 370–379.

Christakis, N. A. and Fowler, J. H. (2008). Dynamics of smoking behavior in a large social network. *New England Journal of Medicine* **358**, 2249–2258.

Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods for Medical Research* **16**, 309–330.

Didelez, V., Meng, S., and Sheehan, N. A. (2010). Assumptions of IV methods for observational epidemiology. *Statistical Science* **25**, 22–40.

Elwert, F. (2013). Graphical causal models. In *Handbook of Causal Analysis for Social Research*, S. Morgan (ed), 245–273. Dodrecht, Netherlands: Springer.

Elwert, F. and Christakis, N. A. (2008). Wives and ex-wives: A new test for homogamy bias in the widowhood effect. *Demography* **45**, 851–873.

Fletcher, J. M. (2008). Social interactions and smoking: Evidence using multiple student cohorts, instrumental variables, and school fixed effects. *Health Economics* **19**, 466–484.

Fowler, J. H. and Christakis, N. A. (2010). Cooperative behavior cascades in human social networks. *PNAS: Proceedings of the National Academy of Sciences* **107**, 5334–5338.

Hernán, M. and Robins, J. (2006). Instruments for causal inference—An epidemiologist's dream? *Epidemiology* **17**, 360–372.

Lasky-Su, J., Lyon, H. N., Emilsson, V., Heid, I. M., Molony, C., Raby, B. A., Lazarus, R., Klanderman, B., Soto-Quiros, M. E., Avila, L., Silverman, E. K., Thorleifsson, G., Thorsteinsdottir, U., Kronenberg, F., Vollmert, C., Illig, T., Fox, C. S., Levy, D., Laird, N., Ding, X., McQueen, M. B., Butler, J., Ardlie, K., Papoutsakis, C., Dedoussis, G., O'Donnell, C. J., Wichmann, H. E., Celedón, J. C., Schadt, E., Hirschhorn, J., Weiss, S. T., Stefansson, K., and Lange, C. (2008). On the replication of genetic associations: Timing can be everything. *The American Journal of Human Genetics* **82**, 849–858.

O'Malley, A. J. and Christakis, N. A. (2011). Longitudinal analysis of large social networks: Estimating the effect of health traits on changes in friendship ties. *Statistics in Medicine* **30**, 950–964.

Palmer, T. M., Lawlor, D. A., Harbord, R. M., Sheehan, N. A., Tobias, J. H., Timpson, N. J., Smith, G. D., and Sterne, J. A. C. (2012). Using multiple genetic variants as instrumental variables for modifiable risk factors. *The International Journal of Biostatistics* **21**, 223–242.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82**, 669–710.

Pearl, J. (2009). *Causality*, 2nd edition. New York: Cambridge University Press.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. Working Paper Number 128, *Center for Statistics and the Social Sciences, University of Washington*, (146 pages), `http://www.csss.washington.edu/Papers/wp128.pdf`.

Rosenquist, J. N., Fowler, J. H., Murabito, J., and Christakis, N. A. (2010). The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine* **152**, 426–433.

Sacerdote, B. (2001). Peer effects with random assignment: results for Dartmouth roommates. *Quarterly Journal of Economics* **116**, 681–704.

Shalizi, C. R. and Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research* **40**, 211–239.

Speliotes, E., Willer, C., Berndt, S., Monda, K., Thorleifsson, G., Jackson, A., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42**, 937–948.

Stock, J., Wright, J., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economics Statistics* **20**, 518–527.

VanderWeele, T. J., Ogburn, E. L., and Tchetgen Tchetgen, E. J. (2012). Why and when "flawed" social network analyses still yield valid tests of no contagion. *Statistics, Politics, and Policy* **3**, doi:10.1515/2151–7509.1050.

Vansteelandt, S., Bowden, J., Babanezhad, M., and Goetghebeur, E. (2011). On instrumental variables estimation of causal odds ratios. *Statistical Science* **26**, 403–422.

Verma, T. and Pearl, J. (1988). Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence*, Mountain View, CA: Association for Uncertainty in Artificial Intelligence, 352–359.

White, H. (1982). Instrumental variables regression with independent observations. *Econometrica* **50**, 483–499.

Wing, R. R. and Jeffery, R. W. (1999). Benefits of recruiting participants with friends and increasing social support for weight loss and maintenance. *Journal of Consulting and Clinical Psychology* **67**, 132–138.

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