Invalid Statistical Inference Due to Social Network Dependence

Youjin Lee[1], and Elizabeth L. Ogburn[1*]

1 Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore MD, USA 21205

* eogburn@jhu.edu

Abstract

Researchers across the health and social sciences generally assume that observations are independent, but when observations are dependent, using statistical methods that assume independence can lead to biased estimates (with bias away from the null) and to artificially small p-values, standard errors, and confidence intervals. This results in inflated false positive rates and may contribute to replication crises. Here, we describe a largely unrecognized but common type of dependence due to social network connections, and explain how such dependence increases variance and engenders confounding that can lead to biased estimates. We describe network dependence and introduce the concept of confounding by network structure. We apply a test for network dependence to several published papers that use the Framingham Heart Study (FHS) data. Results suggest that some of the many decades of research on coronary heart disease, other health outcomes, and peer influence using FHS data may be invalid due to unacknowledged network dependence. The FHS is not unique; these problems could arise whenever human subjects are recruited from one or a small number of communities, schools, hospitals, etc. As researchers in psychology, medicine, and beyond grapple with replication failures, this unacknowledged source of invalid statistical inference should be part of the conversation.

Author summary

Youjin Lee is a doctoral student in the Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health.

Introduction

The replication crises in psychology, medicine, and other fields have drawn attention to many ways that the flawed application of statistics can result in spurious findings. Assuming that data are independent and identically distributed (i.i.d.) is the default for most applications of statistics. Statistical dependence has not received much attention in the discussions around replication crises, but it is well known that when i.i.d. statistical methods are used to analyse data that are in fact dependent, the resulting inference is generally anticonservative: standard errors, p-values, and confidence intervals are artificially small.

In this paper we identify an unacknowledged but potentially pervasive source of dependence, namely social network ties, that can lead to invalid statistical inference and

inflated false positive rates. Whenever human subjects are sampled from one or a small number of communities, schools, hospitals, etc., as is routine in the health and social sciences, they may be connected by social ties, such as friendship or family membership, that could engender statistical dependence. We call this *network dependence*. We show that when an outcome and an exposure of interest both exhibit network dependence, estimates of associations will often be biased away from the null. We call this *confounding by network structure*. The problem is that while many studies in the health and social sciences sample from populations that exhibit some form of network dependence, very few have collected data on social network ties among participants, and the i.i.d. assumption is seldom questioned or tested. It is therefore possible that inflated p-values due to confounding by network structure are pervasive among the health and social science literatures.

Here we show that ignoring social network ties can result in biased and invalid statistical inference. We first define *network dependence* and *confounding by network structure*, then describe tests that can help detect when these might be a problem in real data, and finally apply these tests to real world data from the Framingham Heart Study (FHS), which one of the few studies for which some data on network ties is available. The FHS is a paradigmatic example of an epidemiologic study comprised of individuals from a single tight-knit community, and it has served as a basis for a large literature on phenomena from heart disease to social contagion, all using i.i.d. statistical methods. Our results suggest that the i.i.d. assumption—on which thousands of FHS papers have relied—does not reliably hold, and therefore that confounding by network structure may be widespread at least among studies using FHS data, and likely beyond.

Methods

A network is a collection of nodes and edges [1], where, in a social network, a node represents a person and an edge connecting two nodes represents the existence of some relationship or social tie between them. When the nodes in a network correspond to students in a high school, for example, a tie may indicate that two students are in the same class or that they are members of the same school club; when nodes are patients staying in a hospital, a tie between patients may represent a shared doctor or a shared hospital unit.

In the literature on spatial and temporal dependence, dependence is often implicitly assumed to be the result of latent traits that are more similar for observations that are close than for distant observations. This latent variable dependence [2] is likely to be present in many network contexts as well. Homophily, or the tendency of similar people to form network ties, is a paradigmatic source of latent trait dependence. If the outcome under study in a social network has a genetic component, then we would expect latent variable dependence due the fact that family members, who share latent genetic traits, are more likely to be close in social distance than people who are unrelated. If the outcome is affected by geography or physical environment, latent variable dependence could arise because people who live close to one another are more likely to be friends than those who are geographically distant. In networks, edges often present opportunities to transmit traits or information from one node to another, and such direct transmission will result in dependence that is informed by the underlying network structure [2]. In general, both of these sources of dependence result in positive pairwise correlations that tend to be larger for pairs of observations from nodes that are close in the network and smaller for observations from nodes that are distant in the network.

To illustrate the consequences of treating network observations as if they are i.i.d., consider a hypothetical sample of n nodes in a social network, e.g. students at a U.S. college with ties representing friendship, cohabitation, participation in the same

activities, etc.. Each node provides an outcome Y, e.g. body mass index (BMI). Suppose that, as has been suggested by some researchers [3], BMI exhibits network dependence due to "social contagion." The target of inference is the mean μ of BMI for U.S. college students. The sample average $\bar{Y} = \sum_{i=1}^{n} Y_i/n$ is unbiased for μ as long as the students at this particular college are representative of the overall U.S. college student population. While bias and representability are not necessarily affected by social network connections, the variance of $\bar{Y}$ will be affected by network dependence. For the purposes of this example, suppose that $Y_1, Y_2, ..., Y_n$ are identically but not independently distributed, with common mean μ and variance σ^2. Then

$$\begin{aligned} Var(\bar{Y}) &= Var\left(\sum_{i=1}^{n} Y_i\right)/n^2 \\ &= \frac{1}{n^2}\left\{\sum_{i=1}^{n}\sigma^2 + \sum_{i\neq j} cov(Y_i, Y_j)\right\} \\ &= \frac{\sigma^2}{n/\left(1+\frac{b_n}{\sigma^2}\right)}, \end{aligned} \tag{1}$$

where $b_n = \frac{1}{n}\sum_{i\neq j}^{n} cov(Y_i, Y_j)$. The quantity $n/\left(1+\frac{b_n}{\sigma^2}\right)$ in the denominator is the *effective sample size* of the dependent sample, and under dependence it is generally smaller than the apparent sample size n. But it is the effective rather than the apparent sample size that determines standard errors and rates of convergence for dependent samples. A researcher who failed to question the independence of $Y_1, Y_2, ..., Y_n$ would estimate $Var(\bar{Y})$ with σ^2/n, but whenever b_n is positive (as is expected under network dependence), this underestimates the true variance. Inference using variance estimators based on σ^2/n will be anticonservative: p-values will be artificially low and confidence intervals artificially narrow. With more dependence b_n increases, the effective sample size decreases, and inference that assumes independence is more anticonservative.

Very informally, when subjects are independent, each new observation brings one new "bit" of information about μ; when subjects are dependent, each new observation brings less than one new "bit" of information because some of the information is redundant due to dependence on the previous observations. Therefore, a researcher who falsely assumes independence believes that the data provide more information than they actually do, i.e. the researcher overestimates the strength of evidence provided by the data.

In some settings researchers routinely account for statistical dependence in data analyses: for example, when data are clustered (e.g. clustered randomized trials, batch effects in lab experiments), when studying genetics or heritability in a sample of genetically related organisms, or when data may exhibit spatial or temporal dependence. But outside of these settings it is generally standard practice to use statistical methods that assume independent and identically distributed (i.i.d.) data. Despite increasing interest in and availability of social network data, there is a dearth of valid statistical methods to detect or account for network dependence.

Regression models

Coefficients from regression models suffer from the same problems as sample means in the presence of network dependence. Standard regression models assume independent errors, but when an outcome exhibits network dependence the regression errors generally will, too, rendering inferences drawn from the regression models invalid.

Although researchers have developed regression models for many kinds of dependent data, it is not clear that any of them are generally appropriate for social network data, and certainly none are in wide use for network data.

Confounding by network structure

Bias can result when both an outcome and a covariate of interest exhibit network dependence. In this case, the network structure can act like a confounder, creating a spurious association between the covariate and outcome. Returning to the example above, suppose researchers use data from the college students to ascertain whether choice of academic major is associated with BMI. Students form strong friendships with other students having similar academic interests, engendering network dependence in academic major. An entirely independent process engenders network dependence in BMI: obesity is socially contagious, so students who are friends with one another (regardless of whether the friendship is related to shared academic interests) tend to have similar BMI. Due solely to the underlying network structure, students with the same major are expected to have similar BMI. We would not expect to see this same association in an i.i.d. sample, for example a national sample drawing independent students from many different colleges. Confounding by network structure is analogous to confounding by population stratification and confounding by cryptic relatedness, two well-known sources of bias in population-based genetic association studies when both the outcome and the (in this case genetic or genomic) covariate of interest share a common dependence structure [4].

Testing for network dependence

In a companion technical report [5] we propose statistical methods to test for the presence of network dependence in data with some information about network ties, based on Moran's I, a well-known statistic from the spatial autocorrelation literature. An R package is available [6]. The test takes as inputs a single value associated with each subject, e.g. an outcome, predictor, or regression residual, and a weighted distance matrix with an entry for each pair of subjects. The weight matrix should place higher weights on pairs of subjects who are close in network distance and smaller weights on pairs of subjects who are distant in the network. The choice of weights affects the power, but not the validity, of the test. Similarly, if information is available about some but not all network ties, this will tend to reduce the power of the test but not affect its validity. A robust choice of weight matrix is the *adjacency matrix* for the network, which puts weight 1 on pairs of subjects who share a network tie and weight 0 otherwise; we use this weight matrix throughout. We recommend viewing moderate to large statistics as evidence of possible dependence even if p-values do not meet an arbitrary $\alpha = 0.05$ cut-off, and caution that network dependence may be present even if these statistics are small. If the test statistic calculated from regression residuals is moderate to large, it suggests that standard error estimates from i.i.d. regression models may be underestimated. If both of the test statistics calculated from an outcome and from a covariate of interest are moderate to large, it suggests that confounding by network structure may be present.

Framingham Heart Study

The Framingham Heart Study (FHS), initiated in 1948, is arguably the most important source of data on cardiovascular epidemiology. It is also an influential source of data on network peer effects. FHS is an ongoing cohort study of participants from the town of

Framingham, Massachusetts, that has grown over the years to include five cohorts with a total sample of over 15,000, representing almost 25% of the total population of Framingham. Multiple members (> 3) of more than 1,500 extended families are included in the study population. Study participants are followed through exams every 2 to 8 years. In between exams, participants are regularly monitored through phone calls. Detailed information on data collected in the FHS can be found in [7]. Public versions of FHS data through 2008 are available from the dbGaP database. The FHS data have been analyzed using i.i.d. statistical models (as is standard practice for cohort studies) in over 3,400 peer-reviewed publications since 1950, most of which use multiple regression to explore associations between cardiovascular outcomes and various risk factors. Because the individuals in the FHS are members of a single community, connected by social and familial ties, the outcomes and covariates of interest may be exhibit network dependence. Yet to our knowledge, none of the published studies using FHS data has acknowledged this possibility, including in the literature on peer effects.

Below we demonstrate the potential for bias due to confounding by network structure and show that there is evidence of potentially widespread dependence in the outcomes, predictors, and regression residuals from published papers using FHS data. The problem of network dependence extends to high profile research using FHS data to explicitly study peer effects and social contagion in social networks, but with statistical methods designed for i.i.d. data.

Confounding by network structure

In order to demonstrate the bias that can arise when both a predictor and an outcome share common network structure, we simulated a covariate with dependence structure governed by the FHS social network but otherwise unrelated to any of the variables measured in the FHS. We generated a continuous network dependent covariate, X, conditional on the FHS network, independently 500 times. We regressed a cardiovascular outcome (systolic blood pressure, SBP), a lifestyle outcome (employed or not), a health-seeking behavior outcome (visited a doctor due to illness), and a non-cardiovascular health outcome (diagnosis of corneal arcus) from the FHS data onto X. For each of the four outcomes we fit the same regression model independently 500 times, once for each of the independently generated covariates.

Figure 1 shows the coverage of 95% confidence intervals for β, the coefficient for X in the regression of each outcome onto X plus an intercept. Because the covariate is generated without reference to any of these outcomes, the true value of β for a population-based, rather than network, sample is 0. However, for all four outcomes the confidence intervals are not centered around 0, indicating that estimates of β are biased due to confounding by network structure. For all four outcomes the confidence intervals exhibit undercoverage, ranging from 65% to 85% rather than the nominal rate of 95%. While the bias is due to confounding by network structure; the undercoverage may be due to both confounding and to network dependence in the regression residuals, which could result in underestimated standard errors. Table 1 reports the p-values for tests of dependence in the four outcomes, the predictor X (averaged across 500 replicates), and the residuals from the regression of the outcome on X (averaged across 500 replicates for each outcome). For three of the outcomes (SBP, employment, and corneal arcus) tests based on Moran's I suggested strong evidence of dependence; for visit to doctor the test did not show strong evidence of dependence in the outcome or residuals (though we reiterate that a null test does not imply a lack of dependence). Simulation and analysis details are in the S1 Appendix.

Fig 1. 95% confidence intervals under confounding by network structure Each column contains 95% confidence intervals (CIs) for the coefficient for a random, network dependent covariate. The CIs above the dotted line do not contain the null value $\beta = 0$ (red-line) while the CIs below the dotted line contain 0. Coverage rates of 95% CIs are calculated as the percentages of the CIs covering 0.

Table 1. Results of tests of network dependence for the outcomes, simulated predictor X, and residuals from regressing each outcome onto X. P-values are obtained from permutation tests.

	Systolic blood pressure	Employed	Visited doctor	Corneal arcus
p-value for outcome	0.03	0.00	0.71	0.01
Average p-value for predictor	0.00	0.00	0.00	0.00
Average p-value for residuals	0.04	0.00	0.70	0.02

Cardiovascular disease epidemiology

In order to evaluate whether network dependence and confounding due to network structure may undermine research using FHS data, we chose regression models from five published papers in the epidemiologic and medical literature and applied our tests of dependence to the outcomes, covariates, and regression residuals. We screened for ease of replicability using publicly available data (i.e. models are explicitly defined using variables that are available in the public data), and selected the first five papers that we found on Google Scholar that met the replicability criteria. Because we require social network information for our tests of dependence, and because that information is not available for all individuals and is not straightforward to harmonize across exams, we ran the published regression models on subsets of the data for which network information was readily available. Below we report results from the two papers for which we found the strongest evidence of dependence: the models reported in these two papers show compelling evidence of network dependent outcomes, covariates, and residuals. We also found moderate evidence of dependence in some of the analyses reported in each of the other three papers [8–10]; details are in the S1 Appendix.

Lauer et al. [11] examined the association between obesity and left ventricular mass (LVM); this paper is one of the authors' many highly cited papers on LVM, which is of interest to many researchers due to its relationship with cardiovascular disease [10] and other cardiovascular outcomes. The study assessed the relationship between obesity and LVM using the estimated coefficients for BMI in sex-specific linear regressions adjusted for age and systolic blood pressure, where the outcome was LVM normalized by height. This analysis indicated that obesity is a significant predictor of LVM conditional on age and systolic blood pressure for both men and women.

In order to test whether the assumptions of independence inherently assumed by [11] are valid, we applied Moran's I to normalized LVM and to BMI, separately for males and females, and to the residuals from our replication of the Lauer et al. sex-specific regressions. The results are reported in Table 2. In order for the inference reported in [11] to be valid, the errors from the regressions should be independent, however Moran's I provides evidence of network dependence for the residuals in addition to the marginal LVM variable, for both males and females, undermining the i.i.d. assumption on which the validity of the linear regression model rests. Furthermore, for both sexes there is evidence of network dependence for both LVM and BMI, suggesting that any association may be due to confounding by network structure.

Cox proportional hazards models [12] are commonly applied to the FHS data to assess risk factors for mortality. When the assumptions of the Cox model hold, including i.i.d. observations, Martingale residuals are expected to be approximately

Table 2. Results of tests of network dependence for males and females, for LVM, BMI, and the residuals from regressing LVM onto covariates. P-values are obtained from permutation tests.

Y	I_{std}	p-value
Male		
Normalized LVM	2.26	0.01
BMI	1.36	0.09
Residual from LVM ~ BMI + age + systolic BP	1.34	0.11
Female		
Normalized LVM	2.23	0.02
BMI	1.51	0.06
Residual from LVM ~ BMI + age + systolic BP	2.92	0.00

uncorrelated in finite samples [13,14]. We looked for evidence of residual dependence in a study by Tsuji et al. [15] of the association between eight different heart rate variability (HRV) measures and four-year mortality. We replicated the twenty-four separate Cox models reported in [15]: for each of eight measures of HRV we fit models without adjusting for covariates, adjusting for age and sex, and adjusting for clinical risk factors in addition to age and sex.

Table 3 shows the results of applying tests of independence using Moran's I to the Martingale residuals from the twenty-four different regression models, which suggest that the i.i.d. assumption may be violated in most or all of these regressions. Interestingly, Moran's I statistic is larger with smaller p-values for the covariates that were found to be significant predictors of all cause mortality. This is consistent with a hypothesis that the statistically significant associations are due to confounding by network structure rather than to true population-level associations.

Table 3. Tests of network dependence using Moran's I statistic applied to each HRV measure and to the Martingale residuals from the Cox models for eight different HRV measures. P-values are obtained from permutation tests.

HRV measures:	lnSDNN	lnpNN50	lnr-MSSD	lnVLF	lnLF	lnHF	lnTP	lnLF/HF
Covariate								
I_{std}	0.33	-0.41	-0.12	1.72	1.62	0.83	1.85	-0.03
p-value	0.38	0.59	0.52	0.06	0.08	0.20	0.06	0.47
Residuals from unadjusted model for all-cause mortality								
I_{std}	1.57	1.65	1.64	1.38	1.38	1.54	1.38	1.59
p-value	0.06	0.04	0.04	0.08	0.09	0.06	0.08	0.05
Residuals from model for all-cause mortality adjusted for age and sex								
I_{std}	1.94	2.00	2.05	1.92	1.75	1.95	1.87	1.97
p-value	0.02	0.02	0.02	0.02	0.04	0.02	0.03	0.03
Residuals from model for all-cause mortality adjusted for age, sex, and clinical risk factors								
I_{std}	1.55	1.52	1.56	1.60	1.46	1.53	1.52	1.52
p-value	0.07	0.07	0.07	0.06	0.09	0.07	0.09	0.07

Peer effects

The FHS plays a uniquely influential role in the study of social networks and social contagion. Christakis and Fowler (C&F) discovered an untapped resource buried in the FHS data collection tracking sheets: information on social ties that, combined with existing data on connections among the FHS participants, allowed them to reconstruct the (partial) social network underlying the cohort. They then leveraged this social

network data to study peer effects for obesity [3], smoking [16], and happiness [17]. Researchers have since used the same methods as C&F to study peer effects in the FHS and in many other social networks settings. However, like epidemiologists studying cardiovascular disease, C&F and other researchers using non-experimental data to assess peer effects generally use statistical models that assume independence across subjects [18]; e.g. [17,19,20]. To assess peer influence for obesity, C&F fit longitudinal logistic regression models of each individual's obesity status at exam $k = 2, 3, 4, 5, 6, 7$ onto each of the individual's social contacts' obesity statuses at exam k and $k-1$ (with a separate entry into the model for each contact), controlling for individual covariates and for the node's own obesity status at exam $k-1$. They used generalized estimating equations [21] to account for correlation within individual, but their model assumes independence across individuals. Christakis and Fowler fit this model separately for ten different types of social connections, including siblings, spouses, and immediate neighbors.

We replicated a secondary analysis in which the social contacts' obesity statuses at exams $k-1$ and $k-2$ were used instead of k and $k-1$; we replicated this analysis to avoid the misspecification inherent in the former specification [18]. Although it would be possible to adapt our proposed test of dependence to longitudinal or clustered data, that is beyond the scope of this paper and for simplicity we fit the C&F model at a single time point and selected one social contact for each node in order to have one residual per individual. We chose to use exam 3 for the outcome data because it gave us the largest sample size. We looked at sibling relationships because this gives the largest number of ties in the underlying network compared to the other nine types of relationships considered by Christakis and Fowler and because we had a prior hypothesis that close genetic relationships would evince dependence in obesity status.

We calculated Moran's I for the outcome (obesity status in exam 3), the predictor of interest (sibling's obesity status in exam 2), and the residuals from the logistic regression of each node's exam 3 obesity status onto the node's own obesity status in exam 2, the sibling's obesity status in exam 2, the sibling's obesity status at exam 1, and covariates age, sex, and education. For the outcome $I_{std} = 7.10$ ($p < 0.01$) and for the exposure $I_{std} = 15.91$ ($p < 0.01$) (because BMI is a binary variable I is equivalent to Φ), suggesting that confounding by network structure could contribute to any apparent association between the outcome and the exposure of interest. $I_{std} = 2.76$ ($p < 0.01$) for the regression residuals, providing strong evidence that the i.i.d. assumption on which these analyses rests may be invalid. Details of our analysis can be found in the S1 Appendix.

Discussion

As researchers across many scientific disciplines grapple with replication crises, many sources of artificially small p-values and inflated false positive rates have received attention, but the possible impact of network dependence has been overlooked. In this paper, we used simple tests for independence among observations sampled from a single network to demonstrate that many types of analyses using FHS data may have reported biased point estimates and artificially small p-values, standard errors, and confidence intervals due to unacknowledged network dependence.

Tests for network dependence rely on social network information, which, as we have noted, is not available in most studies that are not explicitly about networks. However, missing data on network ties will generally affect the power but not validity of these tests, so adding information on even just one or two ties per subject to a data collection protocol would enable researchers to test for network dependence. Additionally, when some of the network ties are familial, and when genetic data is available, as is the case

in the FHS, techniques developed to control for confounding due to cryptic relatedness [4] may be helpful for estimating the unknown familial network structure and for controlling for confounding due to that structure.

Existing methods permit testing for network dependence, but do not provide options for data analysis if evidence of dependence is found. Future work is needed on methods to account for dependence if it is detected. Although many statistical methods exist for dealing with dependent data, most of these methods are intended for spatial or temporal data, or, more broadly, for observations with positions in $\mathbb{R}^k$ and dependence that is related to Euclidean distance between pairs of points. The topology of a network is very different from that of Euclidean space, and careful work is needed to justify the use of existing methods for social network data and to develop new methods.

Beyond a call for methods development, our primary recommendation to researchers designing new studies with human subjects is to avoid recruiting from one or a small number of underlying social networks whenever possible, especially if an outcome or exposure of interest could plausibly exhibit network dependence. Careful study design may limit dependence in some settings. For example, for a behavioural outcome restricting attention to within-subject changes during a time period in which subjects have limited interactions with one another could reduce or eliminate network dependence. Researchers working with existing data should be aware of the possibility that social network dependence may undermine the use of i.i.d. models.

Acknowledgements

Youjin Lee and Elizabeth Ogburn were supported by ONR grant N000141512343. 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 and HHSN268201500001I). 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. The authors are grateful to Caroline Epstein, whose M.S. thesis this work builds upon, and to Andrew Gelman, Marshall Joffe, Tom Louis, Eric Tchetgen Tchetgen, Nathan Winkler-Rhoades, and members of the UPenn Causal Inference Reading Group for helpful comments.

Author Contribution

YL and ELO designed the research; YL performed the research, coded the simulations and analyzed the data; YL and ELO wrote the paper.

Supporting information

S1 Appendix. Analysis of the Framingham Heart Study data

References

1. Newman M. Networks: an introduction. Oxford university press; 2010.
2. Ogburn EL. Challenges to estimating contagion effects from observational data. arXiv preprint arXiv:170608440. 2017;.

3. Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. New England journal of medicine. 2007;357(4):370–379.

4. Sillanpää M. Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. Heredity. 2011;106(4):511.

5. Lee Y, Ogburn EL. Testing for Network and Spatial Autocorrelation. arXiv preprint arXiv:171003296. 2018;.

6. Lee Y, Ogburn EL. netdep: Testing for Network Dependence; 2018. Available from: https://CRAN.R-project.org/package=netdep.

7. Tsao CW, Vasan RS. Cohort Profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. International journal of epidemiology. 2015;44(6):1800–1813.

8. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. Stroke. 1991;22(8):983–988.

9. Gordon T, Castelli WP, Hjortland MC, Kannel WB, Dawber TR. High density lipoprotein as a protective factor against coronary heart disease: the Framingham Study. The American journal of medicine. 1977;62(5):707–714.

10. Levy D, Garrison RJ, Savage DD, Kannel WB, Castelli WP. Prognostic implications of echocardiographically determined left ventricular mass in the Framingham Heart Study. New England Journal of Medicine. 1990;322(22):1561–1566.

11. Lauer MS, Anderson KM, Kannel WB, Levy D. The impact of obesity on left ventricular mass and geometry: the Framingham Heart Study. Jama. 1991;266(2):231–236.

12. Cox DR. Regression models and life-tables. In: Breakthroughs in statistics. Springer; 1992. p. 527–541.

13. Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika. 1993;80(3):557–572.

14. Tableman M, Kim JS. Survival analysis using S: analysis of time-to-event data. CRC press; 2003.

15. Tsuji H, Venditti FJ, Manders ES, Evans JC, Larson MG, Feldman CL, et al. Reduced heart rate variability and mortality risk in an elderly cohort. The Framingham Heart Study. Circulation. 1994;90(2):878–883.

16. Christakis NA, Fowler JH. The collective dynamics of smoking in a large social network. New England journal of medicine. 2008;358(21):2249–2258.

17. Fowler JH, Christakis NA. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. Bmj. 2008;337:a2338.

18. Lyons R. The spread of evidence-poor medicine via flawed social-network analysis. Statistics, Politics, and Policy. 2011;2(1).

19. Trogdon JG, Nonnemaker J, Pais J. Peer effects in adolescent overweight. Journal of health economics. 2008;27(5):1388–1399.

20. Rosenquist JN, Murabito J, Fowler JH, Christakis NA. The spread of alcohol consumption behavior in a large social network. Annals of internal medicine. 2010;152(7):426–433.

21. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; p. 13–22.

Appendix to : "Invalid Statistical Inference Due to Social Network Dependence"

Youjin Lee[1], and Elizabeth L. Ogburn[1*]

1 Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore MD, USA 21205

* eogburn@jhu.edu

A1 Analysis of the Framingham Heart Study data

The publicly available data are divided into datasets for individuals with and without non-profit use (NPU) consent, and for each replication we selected the dataset with more eligible individuals or more observed network ties, except for the peer influence analysis, where we merged both consent groups.

A1.1 Confounding by network structure

We used four random outcomes (systolic blood pressure, employed or not, visited a doctor due to illness, diagnosis of corneal arcus) from the Offspring Cohort at Exam 5 with NPU consent. The number of non-missing observations and the number of edges are shown in Table A1.

	Systolic blood pressure	Employed	Visited doctor	Corneal arcus
Sample size (n)	1031	1021	1028	1019
The number of edges (m)	683	670	681	674

Table A1. The number of observations and of undirected edges sampled from the Offspring Cohort at Exam 5.

For each of these four outcomes, we used the $n \times n$ outcome-specific adjacency matrix $\mathbf{A}$ to simulate continuous network dependent covariates $(X_1, X_2, \ldots, X_n)$ conditional on $\mathbf{A}$ as follows:

$$(X_1, X_2, \ldots, X_n) \sim \text{MVN}\left(\mu = (\mu_1, \mu_2, \ldots, \mu_n), \mathbf{\Sigma}_n\right), \tag{A1}$$

where $\mu_i = 1$ if $\sum_{j=1}^{n} A_{ij} > 0$ and $\mu_i = -1$ otherwise $(i = 1, 2, \ldots, n)$. A variance-covariance matrix $\mathbf{\Sigma}_n = [\sigma_{ij}]$ has a diagonal of 0.5, $\sigma_{ij} = \sigma_{ji} = 0.2$ if $A_{ij} = A_{ji} = 1$, and $\sigma_{ij} = \sigma_{ji} = 0.1$ otherwise $(i, j = 1, 2, \ldots, n;\ i \neq j)$.

A1.2 Cardiovascular disease epidemiology

Lauer et al. [1] used data from individuals with echocardiograms between 1979 and 1983, which coincides with the period of the original cohort Exam 16 (1979 - 1982) and the offspring cohort Exam 2 (1979 - 1983). Because we require information on network ties in order to test for dependence, we will consider a subset of the data used in [1], namely the observations from the Original Cohort Exam 16 (1979 - 1982) and the Offspring Cohort Exam 2 (1979 - 1983) without NPU consent. Because the analysis

in [1] is stratified by sex, we constructed sex-specific adjacency matrices using the network ties which were in existence at the start of the cohorts (1979) or were initiated no later than the end of the original cohort Exam 16 (1982), so that any network ties present between 1979 to 1982 are taken account in the adjacency matrices. Figure A1 describes the eligibility criteria we used.

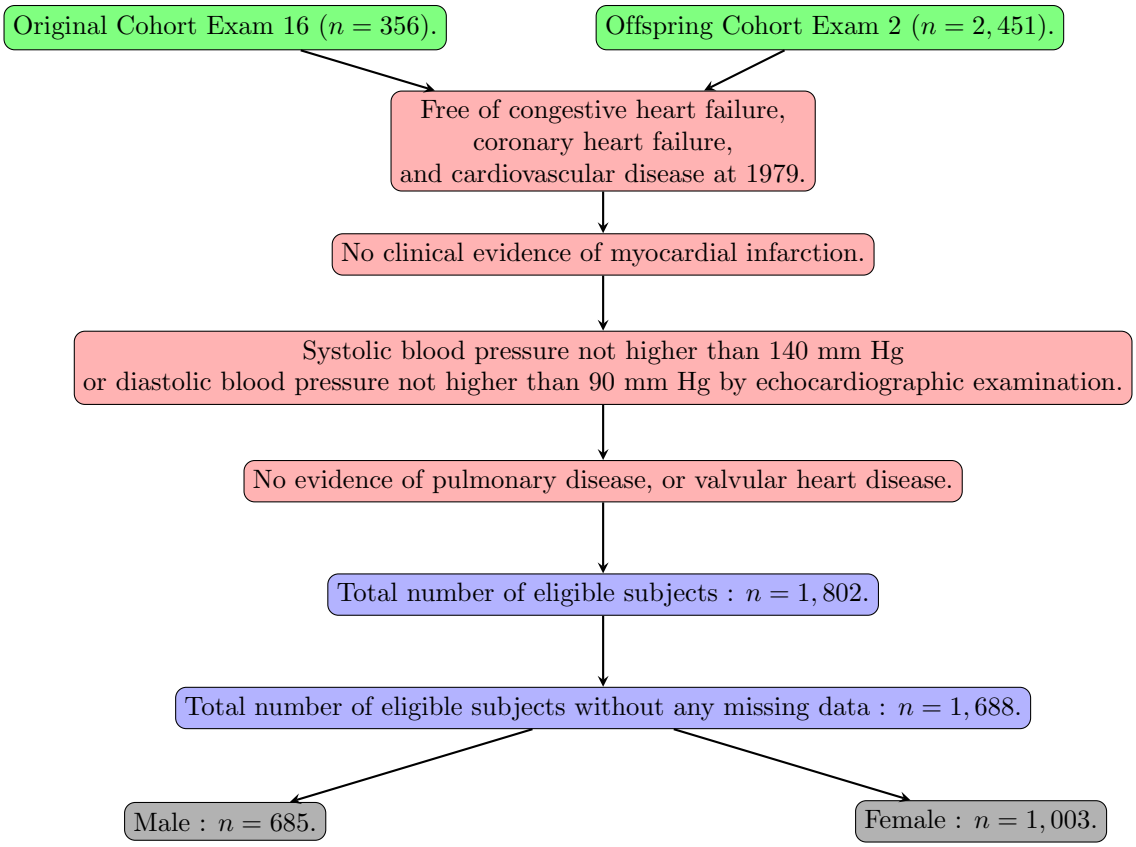
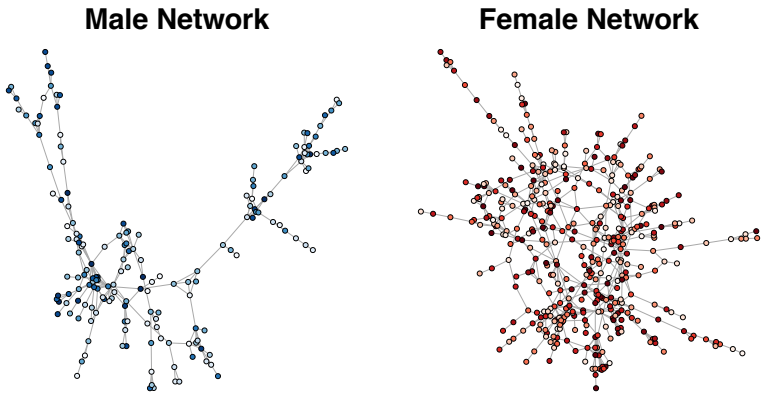


Fig A1. Flowchart for determining eligible healthy subjects from the Original and Offspring Cohorts consent group to replicate the left ventricular mass analysis of Lauer et al.



(a) $I_{std} = 1.34$ (p-value : 0.090) **(b)** $I_{std} = 2.92$ (p-value : 0.002)

Fig A2. The largest connected components of the sex-specific social networks from left ventricular mass (LVM) study [1], displayed using `Fruchterman-Reingold` algorithm. Darker colored nodes represent subjects with higher values of residuals from the regression of normalized LVM onto BMI, age, and systolic blood pressure.

Tables A2 through A5 give summary measures for the variables used in our analyses and report the coefficients from the models that we fit. These can be compared to the published summaries and models in the original papers; we concluded that the summaries and models are sufficiently similar to deem our replications successful.

Table A2. Mean and standard deviations in the parenthesis of characteristics for eligible subjects. This corresponds to Table 1 in the original paper [1] of left ventricular mass study.

	Male (n = 685)	Female (n = 1,003)
Age (year)	40.23 (10.04)	42.04 (10.67)
Weight (kg)	80.61 (10.86)	62.29 (11.10)
Height (cm)	177.2 (7.25)	162.3 (6.32)
BMI (kg/m^2)	2.33 (0.90)	1.75 (0.95)
Systolic BP ($mmHg$)	118.28 (9.41)	112.22 (11.20)
LVM (g)	207.27 (46.45)	135.2 (30.65)
Adjusted LVM (g/m)	116.85 (25.18)	83.3 (18.61)

Table A3. Replication of Lauer et al.'s linear regression of height-corrected left ventricular mass on BMI, age and systolic blood pressure.

	Estimate	Standard error	t-value	Pr(>|t|)
Male ($n = 685$)				
Intercept	70.46	11.35	6.21	0.00
Age	-0.16	0.09	-1.69	0.09
BMI : 23-25.99 kg/m^2	8.87	2.50	3.55	0.00
BMI : 26-29.99	20.43	2.57	7.95	0.00
BMI : $\geq$ 30	27.87	3.57	7.81	0.00
Systolic BP	0.34	0.10	3.40	0.00
Female ($n = 1,003$)				
Intercept	49.19	5.18	9.50	0.00
Age	0.20	0.05	3.89	0.00
BMI : 23-25.99	6.95	1.22	5.70	0.00
BMI : 26-29.99	15.02	1.61	9.34	0.00
BMI $\geq$ 30	27.97	2.02	13.86	0.00
Systolic BP	0.17	0.05	3.45	0.00

Table A4. Standard deviations of eight different heart rate variability measures from [Table 4] of the original paper [2] and from the 516 subjects we used to replicate the original analysis.

	lnSDNN	lnpNN50	lnr-MSSD	lnVLF	lnLF	lnHF	lnTP	lnLF/HF
Original paper	0.33	1.32	0.44	0.76	0.82	0.85	0.73	0.57
Our data	0.33	1.36	0.46	0.74	0.84	0.88	0.73	0.57

Table A5. Replication of twenty-four Cox models from [Table 5] in Tsuji et al. [2].

	Hazard ratio	95% CI	p-value
Unadjusted ($n = 516$)			
lnSDNN	1.31	(1.04, 1.64)	0.0217
lnpNN50	1.03	(0.81, 1.31)	0.8229
lnr-MSSD	1.05	(0.82, 1.34)	0.7092
lnVLF	1.53	(1.23, 1.90)	0.0001
lnLF	1.57	(1.25, 1.98)	0.0001
lnHF	1.27	(0.99, 1.64)	0.0607
lnTP	1.49	(1.20, 1.86)	0.0004
lnLF/HF	1.35	(1.08, 1.68)	0.0095
Age- and sex-adjusted ($n = 516$)			
lnSDNN	1.32	(1.06, 1.65)	0.0146
lnpNN50	1.14	(0.90, 1.45)	0.2781
lnr-MSSD	1.19	(0.94, 1.51)	0.1493
lnVLF	1.53	(1.23, 1.91)	0.0001
lnLF	1.56	(1.24, 1.97)	0.0002
lnHF	1.35	(1.06, 1.72)	0.0150
lnTP	1.51	(1.21, 1.89)	0.0003
lnLF/HF	1.17	(0.93, 1.46)	0.1911
Age, sex, and clinical risk factors adjusted ($n = 512$)			
lnSDNN	1.29	(1.02, 1.62)	0.0312
lnpNN50	1.13	(0.88, 1.44)	0.3480
lnr-MSSD	1.20	(0.94, 1.54)	0.1425
lnVLF	1.55	(1.24, 1.95)	0.0002
lnLF	1.49	(1.16, 1.92)	0.0018
lnHF	1.31	(1.03, 1.68)	0.0304
lnTP	1.51	(1.19, 1.90)	0.0006
lnLF/HF	1.10	(0.86, 1.41)	0.4570

Tables A6 through A8 summarize the results of tests for network dependence applied to the outcomes, primary covariates of interest, and residuals from the models of the three studies [3–5] that we did not include in the main text.

Table A6. Wolf et al. [3] estimated the association between atrial fibrillation (AF) on sex- and age-group-specific two-year incidence of stroke, controlling for coronary heart disease, hypertension, and cardiac failure history. We replicated the analyses using data from the Original Cohort Exam 17 without NPU consent (the original study combined data from 17 exams). Below we report Moran's I statistic and the corresponding permutation-based p-values for the outcome (stroke), the predictor of interest (AF), and the residuals from the full logistic regression model.

	Stroke		**AF**		**Residuals**		n
	I_{std}	p-value	I_{std}	p-value	I_{std}	p-value	
Male							
60-69 yr	-0.19	0.460	0.22	0.152	-0.08	0.408	228
70-79 yr	-0.05	0.304	-0.10	0.438	0.01	0.274	267
80-89 yr	-0.85	0.908	-0.67	0.794	-0.95	0.950	93
Female							
60-69 yr	-1.02	0.984	-0.01	0.690	-0.67	0.942	258
70-79 yr	-0.12	0.544	0.04	0.334	0.15	0.368	398
80-89 yr	1.09	0.120	-0.06	0.410	-0.10	0.476	179

Table A7. Gordon et al. [4] examined the association between HDL cholesterol and four-year incidence of coronary heart disease (CHD) for men and women aged 49 to 82 years old between 1969 and 1971, which coincides with Original Cohort Exam 11. We used the Original Cohort Exam 11 (with NPU consent group) to replicate their univariate logistic regressions of CHD on HDL, and below we report the network dependence test statistics and corresponding permutation-based p-values for the outcome, the predictor, and the residuals. (Due to the large amount of missingness in HDL, the statistics for the residuals are based on smaller sample sizes.)

Y	Sex	n	Moran's I_{std}	p-value
Four-year incidence of CHD	Male	1123	0.32	0.350
	Female	1416	-0.60	0.704
High density lipoproteins (HDL)	Male	552	1.64	0.042
	Female	640	2.05	0.030
Residuals from logistic regression	Male	552	-1.10	0.952
	Female	640	-0.10	0.524

Table A8. Levy at al. [5] investigated the relationship between left ventricular mass (LVM) and cardiovascular disease (CVD) for subjects 40 years old or older. We replicated their analyses of four-year incidence of CVD by running the logistic regression adjusted for age, diastolic blood pressure, pulse pressure, antihypertensive treatment, the number of cigarettes per day, diabetes status, body-mass index, ratio of total to high-density lipoprotein cholesterol, left ventricular hypertrophy on echocardiography, and left ventricular mass; below we report tests of network dependence and corresponding permutation-based p-values for the outcome (CVD), the predictor of interest (LVM), and the regression residuals. As in the original study we used observations ($n = 469$ males and $n = 713$ females) from the Original Cohort Exam 16 and the Offspring Cohort Exam 12, but we restricted our sample to those with NPU consent only.

Sex	Y	Moran's I_{std}	p-value
Male	Incidence of CVD	-0.63	0.744
Female	Incidence of CVD	0.74	0.210
Male	LVM	1.87	0.046
Female	LVM	1.21	0.146
Male	Residuals	-1.10	0.912
Female	Residuals	-0.20	0.450

References

1. Lauer MS, Anderson KM, Kannel WB, Levy D. The impact of obesity on left ventricular mass and geometry: the Framingham Heart Study. Jama. 1991;266(2):231–236.

2. Tsuji H, Venditti FJ, Manders ES, Evans JC, Larson MG, Feldman CL, et al. Reduced heart rate variability and mortality risk in an elderly cohort. The Framingham Heart Study. Circulation. 1994;90(2):878–883.

3. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. Stroke. 1991;22(8):983–988.

4. Gordon T, Castelli WP, Hjortland MC, Kannel WB, Dawber TR. High density lipoprotein as a protective factor against coronary heart disease: the Framingham Study. The American journal of medicine. 1977;62(5):707–714.

5. Levy D, Garrison RJ, Savage DD, Kannel WB, Castelli WP. Prognostic implications of echocardiographically determined left ventricular mass in the Framingham Heart Study. New England Journal of Medicine. 1990;322(22):1561–1566.