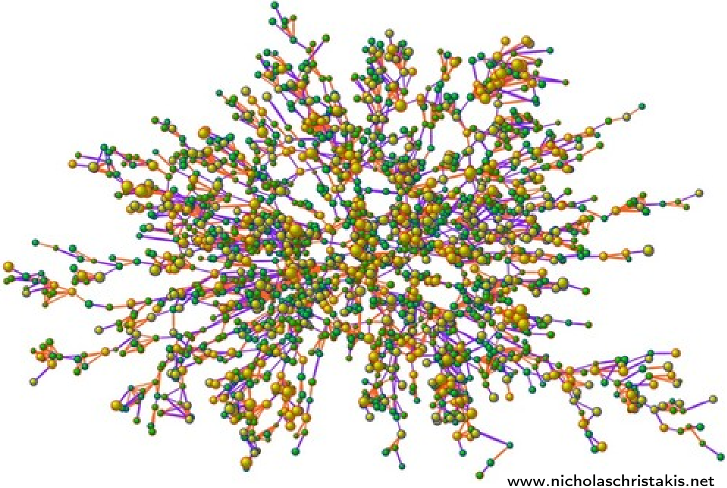
Social network dependence and the replication crisis

Betsy Ogburn

eogburn@jhsph.edu
Department of Biostatistics,
Johns Hopkins University



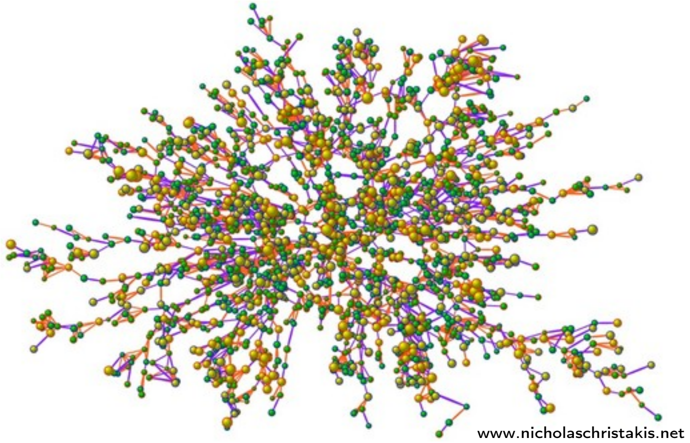
www.nicholaschristakis.net

outline

- Framingham Heart Study
- Ignoring network dependence is dangerous
  - Anticonservative statistical inference
  - Spurious associations due to dependence
- Testing for network dependence...
  - And finding striking evidence of dependence in FHS papers
- Re-analysis of a FHS peer effects model

Framingham Heart Study

- Ongoing cohort study initiated in 1948 to study cardiovascular disease etiology one of the most successful and influential epidemiologic cohort studies in existence
  - arguably the most important source of data on cardiovascular epidemiology
- Thousands of papers published using FHS data, all using i.i.d. statistical methods
- $n \simeq 16,000$, including multiple members of 1538 extended families, representing a sizable portion of the population of Framingham, MA



- FHS is a convenience sample that is comprised of members of an interconnected network rather than independent subjects.
- We expect social network dependence whenever subjects are sampled from one or a small number of schools, communities, hospitals, etc.

Framingham Heart Study

- In the early 2000s, Christakis and Fowler discovered information on social ties that allowed them to reconstruct the (partial) social network underlying the cohort.
- Widely publicized results include significant peer effects for obesity (Christakis and Fowler, 2007), smoking (Christakis and Fowler, 2008), and happiness (Fowler and Christakis, 2008).
- The FHS has since been used to study peer effects by many other researchers (Pachucki et al., 2011; Rosenquist et al., 2010).
- The methods used have come under considerable criticism by statisticians, but little attention has been paid to the fact that i.i.d. methods were used for purportedly non-independent data.

why is (network) dependence a problem?

Lee Y & Ogburn EL (2020)

1. **Anticonservative inference** Failure to adequately account for dependence leads to artificially small p-values, confidence intervals, and standard errors.

2. **Spurious associations** When two variables of exhibit similar types of dependence, association and effect estimates may be biased away from the truth

anticonservative inference

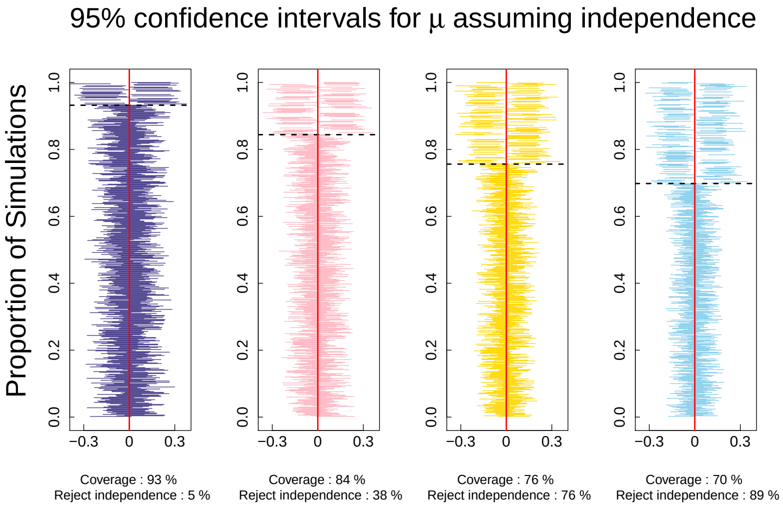
- Suppose we're interested in the average height in the Boston suburbs.
- Let Y be height, and we will estimate $E[Y]$ with the sample average from FHS: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
- If the data are independent, then

$$var(\bar{Y}) = \frac{1}{n^2}\left\{\sum_{i=1}^{n} \sigma^2\right\} = \frac{\sigma^2}{n}$$

- But if there is dependence, then

$$var(\bar{Y}) = \frac{1}{n^2}\left\{\sum_{i=1}^{n} \sigma^2 + \sum_{i \neq j} cov(Y_i, Y_j)\right\}$$

anticonservative inference



spurious associations due to dependence

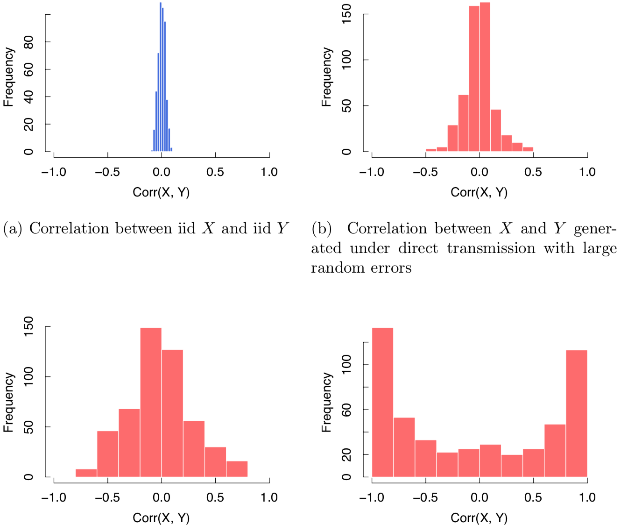
- When an exposure and an outcome both exhibit dependence across units, e.g. due to space, time, genetics, or social network ties, **estimates of associations–and causal effects–may be concentrated away from the truth**.

- Even if the exposure and the outcome are **causally and statistically independent** from one another, tests of independence will tend to reject the null.

- This occurs
  - in the absence of any confounding
  - in a representative sample
  - even if the only interest is in (out-of-sample) prediction

- Well-known in time series and GWAS; I'm not aware of any acknowledgement of this phenomenon outside of those settings

spurious associations due to network dependence



(a) Correlation between iid X and iid Y

(b) Correlation between X and Y generated under direct transmission with large random errors

(c) Correlation between X and Y generated under direct transmission with moderate random errors

(d) Correlation between X and Y generated under direct transmission with small random errors

test for network dependence

- Is it possible that studies based on FHS data report anticonservative s.e.'s (and CIs and p-values) and estimates that are spurious due to network dependence?

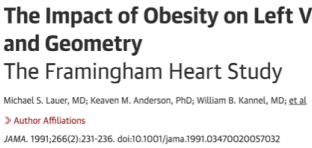
- We adapted Moran's I to test for network dependence, replacing weighted spatial distances with an adjacency matrix.

- We tested:

    1. regression residuals: dependence is (circumstantial) evidence of anticonservative inference

    2. outcome of interest and exposure of interest: dependence in both is (circumstantial) evidence of spurious associations

test for network dependence in FHS papers



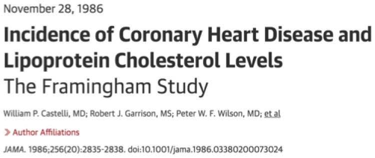
The Impact of Obesity on Left Ventricular Mass and Geometry

The Framingham Heart Study

Michael S. Lauer, MD; Keaven M. Anderson, PhD; William B. Kannel, MD; et al

› Author Affiliations

JAMA. 1991;266(2):231-236. doi:10.1001/jama.1991.03470020057032



November 28, 1986

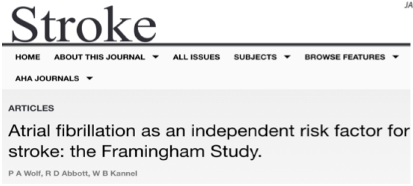
Incidence of Coronary Heart Disease and Lipoprotein Cholesterol Levels

The Framingham Study

William P. Castelli, MD; Robert J. Garrison, MS; Peter W. F. Wilson, MD; et al

› Author Affiliations

JAMA. 1986;256(20):2835-2838. doi:10.1001/jama.1986.03380200073024

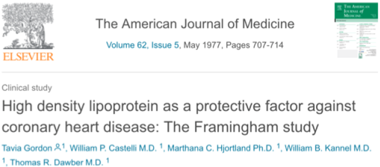


Stroke

ARTICLES

Atrial fibrillation as an independent risk factor for stroke: the Framingham Study.

P A Wolf, R D Abbott, W B Kannel



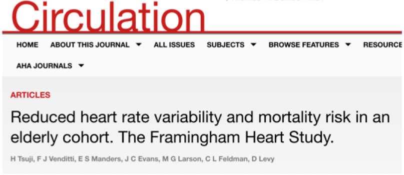
The American Journal of Medicine

Volume 62, Issue 5, May 1977, Pages 707-714

Clinical study

High density lipoprotein as a protective factor against coronary heart disease: The Framingham study

Tavia Gordon [1], William P. Castelli M.D. [1], Marthana C. Hjortland Ph.D. [1], William B. Kannel M.D. [1], Thomas R. Dawber M.D. [1]



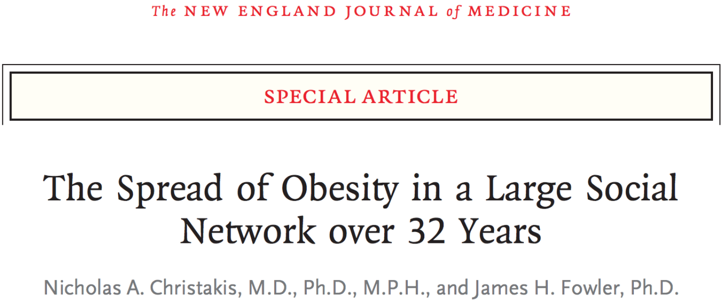
Circulation

ARTICLES

Reduced heart rate variability and mortality risk in an elderly cohort. The Framingham Heart Study.

H Tsuji, F J Venditti, E S Manders, J C Evans, M G Larson, C L Feldman, D Levy

Is there evidence that obesity is "socially contagious" in FHS?



The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL ARTICLE

The Spread of Obesity in a Large Social Network over 32 Years

Nicholas A. Christakis, M.D., Ph.D., M.P.H., and James H. Fowler, Ph.D.

Is there evidence that obesity is “socially contagious” in FHS?

- To assess peer effects of obesity, researchers ran models like this:

$$Y_{ego}^{t} = \alpha + \beta Y_{alter}^{t-1} + \gamma Y_{alter}^{t-2} + \eta Y_{ego}^{t-1} + \lambda X_{alter,ego} + \varepsilon_{ego}^{t}$$

- Y_{ego}^{t} is the ego’s obesity status at time t, Y_{alter}^{t-1} is the alter’s obesity status at time $t-1$, and β is interpreted as the effect of interest.

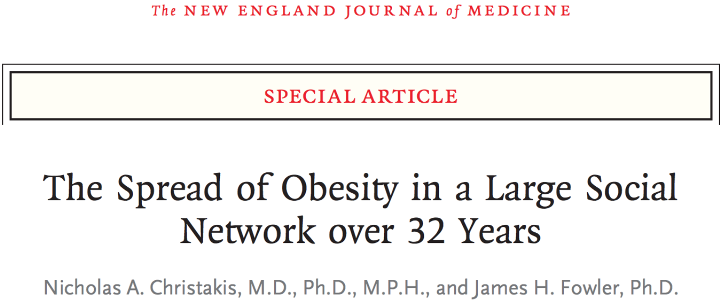
Is there evidence that obesity is "socially contagious" in FHS?

- To assess peer effects of obesity, researchers ran models like this:

$$Y_{ego}^{t} = \alpha + \beta Y_{alter}^{t-1} + \gamma Y_{alter}^{t-2} + \eta Y_{ego}^{t-1} + \lambda X_{alter,ego} + \color{red}{\varepsilon_{ego}^{t}}$$

- These models were estimated assuming that ε_i and ε_j are independent for $i \neq j$ (but ε_i^t and ε_j^s could be dependent).

Is there evidence that obesity is "socially contagious" in FHS?



The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL ARTICLE

The Spread of Obesity in a Large Social Network over 32 Years

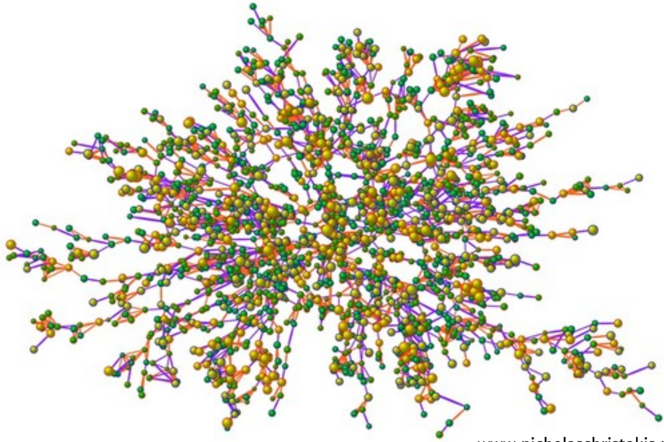
Nicholas A. Christakis, M.D., Ph.D., M.P.H., and James H. Fowler, Ph.D.

- We tested for network dependence in the outcome, the predictor of interest, and the regression residuals.
- $p < 0.01$ for all tests.

Is there evidence that obesity is "socially contagious" in FHS?

- Using a new method to account for network dependence (Ogburn et al. 2020), we re-analyzed the FHS obesity data...
- ... and found no evidence of peer effects.

Is there evidence that obesity is "socially contagious" in FHS?



www.nicholaschristakis.net

- First, we reframed the problem in terms of the entire FHS social network instead of independent pairs.

Is there evidence that obesity is “socially contagious” in FHS?

- We estimated the expected probability of obesity at time t under a hypothetical intervention to increase the number of each node’s obese alters by 1.
- We estimated a causal risk difference of exactly 0, with 95% confidence interval $(-0.01, 0.01)$.

Is there evidence that obesity is “socially contagious” in FHS?

- We also estimated the causal effect of an increase (of half a standard deviation) in the average BMI of each subject’s friends.

- We estimated a causal effect of 0.25, 95% confidence interval $(-0.47, 0.98)$.
- (For context, the empirical mean BMI was 25.51)

- These analyses are consistent with the hypothesis that the strong results in the original paper are spurious, due to dependence and/or model misspecification rather than true associations or causal effects.

conclusion

- Whenever data are dependent, analyses that fail to fully account for dependence can underestimate uncertainty and produce spurious estimates of associations and causal effects.
    - Spurious associations are a problem for out-of-sample prediction, too!
- Data may be dependent more often than you might think.
    - Convenience samples are everywhere in the health and social sciences.
- Statisticians know how to account for Euclidean dependence; non-Euclidean network dependence is a new frontier and lots more research is needed.

Thank you

people

Youjin Lee
Oleg Sofrygin
members of the causal inference groups at JHU and UPenn

funding

ONR Mathematical Data Science Program