Challenges in reproducing results from publicly available data: an example of sexual orientation and cardiovascular disease risk

Nichole Austin[1], Sam Harper[1], Jay S. Kaufman[1], Ghassan B. Hamra[2]

[1] Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Canada, [2]Department of Environmental and Occupational Health, School of Public Health, Drexel University, Philadelphia, PA, USA

Background

Reproducibility

- The production of reproducible research findings is a hallmark of the scientific method, but a number of high profile studies suggest that many results are not replicable. (1, 2)
- Various factors have been cited as barriers to replication, including publication bias, selective outcome reporting, and genuine heterogeneity.
- There have been numerous calls to increase the emphasis on reproducibility, but this is not always possible given inadequate and selective reporting practices. (2, 3)
- Data sharing facilitates replication (4), but the difficulty of replication is well documented even when data and protocols are available. (2)

Sexual orientation & cardiovascular disease

- Farmer et al (5) used the US National Health and Nutrition Examination Survey (NHANES) to explore sexual orientation and risk of cardiovascular disease (CVD) in men.
- They reported that bisexual men were at increased risk for CVD, while homosexually-experienced heterosexuals (HEH) were at decreased risk of CVD.
- The authors concluded that CVD risk differs across subgroups of sexual minority men, and that more attention should be paid to the mechanisms through which risk is conferred.

Objectives

- We aimed to reproduce the findings originally reported by Farmer et al. (5)
- Because the data source was publicly available and the study's methods were generally well-described, we sought to replicate these findings without assistance from the original authors.
- We also extended the original analysis and performed several sensitivity analyses.

Methods

Data

- Source: NHANES, five two-year cycles (2001 to 2010)
- Inclusion criteria: men with informative responses on sexual orientation question and no personal history of CVD
- Exposure categories: gay, bisexual, heterosexual, and heterosexually-identified with at least one same-sex partner in their lifetime (homosexually-experienced heterosexual/HEH)
- Outcome: CVD risk, operationalized as vascular age divided by chronological age and calculated using the Framingham Risk Score (FRS)
- Covariates: Age, race, education, income, smoking, diabetes, alcohol/drug use, cholesterol, systolic blood pressure, BMI

Replication analysis

- We estimated crude and adjusted associations between sexual orientation and vascular age using linear regression (the same approach employed by Farmer et al. (5)).
- We accounted for the survey design and weighting structure described in the NHANES analytic guidelines.
- The CVD risk score can be calculated with a point system or parametric formula (6); we relied on point-based calculation in the interest of exact replication.

Sensitivity analysis & extensions

Age restriction

- The FRS is designed for adults aged 30 and over, but young men (18-29) were included in the original analysis
- We re-estimated the authors' original models restricting to individuals aged 30 and over

Missing subjects

- A number of men provided a non-informative response to the sexual orientation question
- We used a simulation strategy to randomly reassign these men to the four exposure categories

Results

Replication attempt

Vascular age ratio comparison by sexual orientation category (point-based, all ages)

	Farmer et al.			Replication		
	Ratio	Diff	95% CI for difference	Ratio	Diff	95% CI for difference
Unadjusted						
Heterosexual	1.20	Ref	Ref	1.18	Ref	Ref
Gay	1.11	-0.09	(-0.14 to -0.04)	1.10	-0.08	(-0.12, -0.03)
Bisexual	1.29	0.08	(0.01 to 0.15)	1.27	0.09	(0.01, 0.17)
HEH[a]	1.14	-0.07	(-0.12 to -0.02)	1.13	-0.05	(-0.09, -0.00)
Adjusted[b]						
Heterosexual	1.09	Ref	Ref	1.07	Ref	Ref
Gay	1.05	-0.04	(-0.09 to 0.003)	1.04	-0.03	(-0.07, 0.02)
Bisexual	1.16	0.07	(0.00 to 0.13)	1.14	0.08	(0.00, 0.15)
HEH[a]	1.02	-0.07	(-0.12 to -0.02)	1.02	-0.04	(-0.08, -0.00)

[a] *Homosexually-experienced heterosexuals*

[b] *Adjusted for history of hard drug use and education*

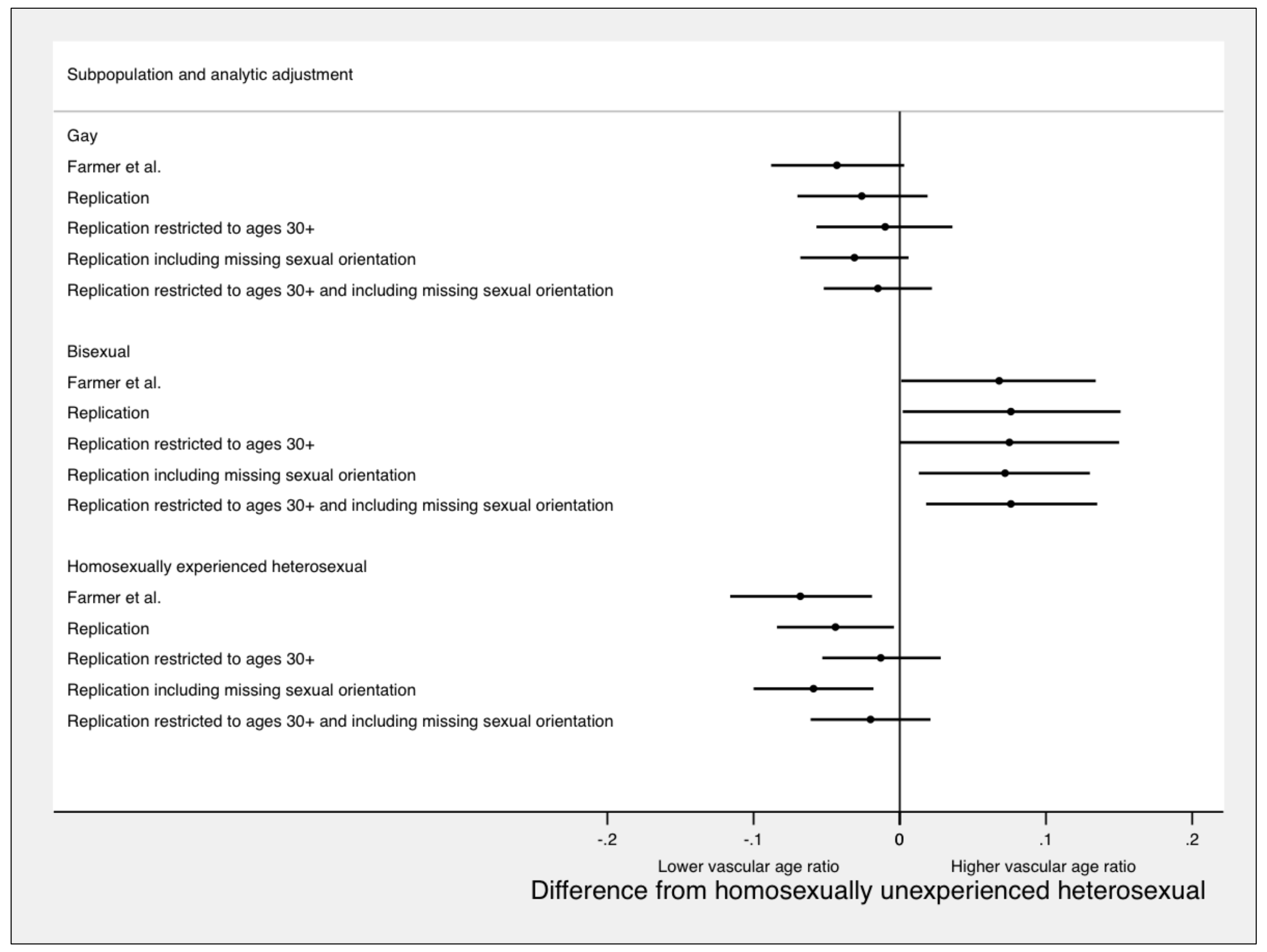
- Our covariate distribution was very close to the original findings with a few exceptions (BMI, family history of CVD, and alcohol use).
- None of the discrepant covariates were components of the Framingham CVD risk algorithm.
- Our results suggested that the average heterosexual subject's vascular age, adjusting for education and drug use, was 1.07 times higher than his chronological age (slightly lower than Farmer's estimate of 1.09).

Sensitivity analyses

- Most age ratios decreased following age restriction: the adjusted heterosexual age ratio decreased considerably to 1.04, suggesting that the average heterosexual subject's vascular age in the modified sample was actually 4%, rather than 7%, higher than their chronological age.
- Our simulation strategy increased the precision of the bisexual estimate and pushed the lower bound away from the null.
- The bisexual point estimate was relatively robust to model specification; the HEH estimate was less robust.

Age ratio calculation comparison by SMM category: Age restriction vs. original

	Ages 30-69			Ages 18-69		
	Ratio	Diff	95% CI for diff	Ratio	Diff	95% CI for diff
Unadjusted						
Heterosexual	1.12	Ref	Ref	1.18	Ref	Ref
Gay	1.07	-0.05	(-0.10, 0.00)	1.10	-0.08	(-0.12, -0.03)
Bisexual	1.20	0.08	(0.01, 0.16)	1.27	0.09	(0.01, 0.17)
HEH[a]	1.10	-0.01	(-0.06, 0.03)	1.13	-0.05	(-0.09, -0.00)
Adjusted[b]						
Heterosexual	1.04	Ref	Ref	1.07	Ref	Ref
Gay	1.03	-0.01	(-0.06, 0.04)	1.04	-0.03	(-0.07, 0.02)
Bisexual	1.12	0.08	(0.00, 0.15)	1.14	0.08	(0.00, 0.15)
HEH[a]	1.03	-0.01	(-0.05, 0.03)	1.02	-0.04	(-0.08, -0.00)



Conclusions

- We were able to identify the trends reported by Farmer et al., but not the exact effect estimates.
- The original findings should have been reproducible given the publicly available data source: the fact that they were not supports the recent calls for greater transparency and improved reporting in research.
- Sensitivity analyses revealed a potentially inappropriate application of the FRS; correcting for this yielded different conclusions about CVD risk in sexual minority men.
- This work elucidates the utility and importance of replication, and the need for rigorously testing assumptions, particularly when data are readily available for reanalysis.

References

1. Baggerly KA, Berry DA. Reproducible research. *Amstat News*. 2011;403(1):16–17.
2. Ioannidis JP, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*. 2014;383(9912):166-175.
3. Chan AW, Song F, Vickers A, et al. Increasing value and reducing waste: addressing inaccessible research. *Lancet*. 2014;383(9913):257-266.
4. Hernán MA, Wilcox AJ. Epidemiology, data sharing, and the challenge of scientific replication. *Epidemiology*. 2009;20(2):167-168.
5. Farmer GW, Bucholz KK, Flick LH, et al. CVD risk among men participating in the National Health and Nutrition Examination Survey (NHANES) from 2001 to 2010: differences by sexual minority status. *Journal of epidemiology and community health*. 2013;67(9):772-778.
6. D'Agostino RB, Sr., Vasan RS, Pencina MJ, et al. General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation*. 2008;117(6):743-753.