Social Science & Medicine

Available online 14 November 2016

In Press, Corrected Proof — Note to users

Is structural stigma's effect on the mortality of sexual minorities robust? A failure to replicate the results of a published study

Mark Regnerus[a, b,]

Show more

http://dx.doi.org/10.1016/j.socscimed.2016.11.018

Highlights

- Failure to replicate a published study of stigma on mortality of sexual minorities.
- No effect found of structural stigma on premature mortality of sexual minorities.
- Imputation of missing data is sensitive to subjective measurement decisions.
- This study highlights the importance of cooperation and transparency in science.

Abstract

Background

The study of stigma's influence on health has surged in recent years. Hatzenbuehler et al.'s (2014) study of structural stigma's effect on mortality revealed an average of 12 years' shorter life expectancy for sexual minorities who resided in communities thought to exhibit high levels of anti-gay prejudice, using data from the 1988–2002 administrations of the US General Social Survey linked to mortality outcome data in the 2008 National Death Index.

Methods

In the original study, the key predictor variable (structural stigma) led to results suggesting the profound negative influence of structural stigma on the mortality of sexual minorities. Attempts to replicate the study, in order to explore alternative hypotheses, repeatedly failed to generate the original study's key finding on structural stigma. Efforts to discern the source of the disparity in results revealed complications in the multiple imputation process for missing values of the components of structural stigma. This prompted efforts at replication using 10 different imputation approaches.

Results

Efforts to replicate Hatzenbuehler et al.'s (2014) key finding on structural stigma's notable influence on the premature mortality of sexual minorities, including a more refined imputation strategy than described in the original study, failed. No data imputation approach yielded parameters that supported the original study's conclusions. Alternative hypotheses, which originally motivated the present study, revealed little new information.

Conclusion

Ten different approaches to multiple imputation of missing data yielded none in which the effect of structural stigma on the mortality of sexual minorities was statistically significant. Minimally, the original study's structural stigma variable (and hence its key result) is so sensitive to subjective measurement decisions as to be rendered unreliable.

Keywords

United States; Prejudice; Stigma; Sexual orientation; Mortality; Replication; Data imputation; Scientific transparency

1. Introduction

Researchers have successfully documented associations between social stigma toward sexual minorities and the experience of adverse health outcomes among them (Bostwick, 2012 and Hatzenbuehler et al., 2009). Stigma, however, is not simple to define or operationalize, prompting measurement challenges that make it difficult to assess just how influential stigma is on health outcomes. Measurement difficulties, moreover, make it harder to develop broad confidence in conclusions across studies. In his widely disseminated and discussed manuscript on the poor validity of most published research findings, Ioannidis (2005: 698) cites the "flexibility in designs, definitions, outcomes, and analytical modes" as well as the relative popularity of a particular research subject as two key factors apt to weaken confidence in published research findings and elevate the risk of scientific missteps. This, together with the rapid expansion of publication outlets and pressure to publish, has contributed to a surge in scientific overstatements, errors, accusations of fabrications, and the issuing of errata or retractions, as well as a renewed call for greater transparency across the research process (Cumming, 2013, Ioannidis, 2008 and Simmons et al., 2011).

Together with five co-authors, Mark Hatzenbuehler analyzed data from the 1988–2002 survey administrations of the General Social Survey (GSS), linked to mortality outcome data in the 2008 National Death Index (NDI). That study revealed dramatically shorter life expectancy—approximately 12 years—for sexual minorities who resided in communities believed to exhibit high levels of anti-gay prejudice, even after controlling for a variety of demographic and health-related indicators. Their findings were published in this journal in 2014 in a special volume on structural stigma and health that Hatzenbuehler co-edited.

In the present study, the same GSS-NDI linked data is reanalyzed in order first to replicate—and then to assess alternative explanations for—the findings in the original study of structural stigma and all-cause mortality in sexual minority populations. Given that the GSS and NDI are publicly-accessible datasets, this approach seemed reasonable, feasible, and a scientific value, especially when the original study posed such notable findings. However, after initial attempts to replicate the original study's key result about the influence of social stigma on premature mortality failed—and efforts to obtain more information from the first author about their decisions concerning the imputation of missing data on the stigma measures were unsuccessful—a variety of focused attempts at replication were undertaken, with no success. The results of these efforts are reported herein.

1.1. Background

Anti-gay stigma, as the original study's authors and others have pointed out using diverse data sources, is often found to be corrosive to the mental and physical health of sexual minorities (Hatzenbuehler, 2009, Herek and Garnets, 2007 and Meyer, 2003). In the original study under scrutiny here, Hatzenbuehler and his co-authors note that while researchers have believed *structural* stigma to be harmful to individuals' health, few have been able to adequately construct and test a contextual measure of such stigma. Indeed, they note "little or no variation to study" in previous attempts, given "the pervasiveness of structural stigma" in American communities (Hatzenbuehler et al., 2014: 34).

The original study cites scholarly support for the observation that sexual minorities live in social environments that vary widely in their support for gays and lesbians, and notes evidence suggesting that higher rates of contextual stigma, such as state-level amendments prohibiting same-sex marriage, are associated with elevated experience of adverse psychological disorders and attempted suicide (Hatzenbuehler, 2011 and Hatzenbuehler et al., 2009). The key research question they pose in their 2014 publication is whether such stigma contributes to premature death among sexual minorities. The matched GSS-NDI data allow for a unique test of the hypothesis.

The authors found that, after controlling for individual and community-level risk factors, structural stigma was still strongly associated with premature mortality among sexual minorities, displaying a hazard ratio of 3.03 (95% CI: 1.50, 6.13), which translates into a life expectancy difference of 12 years, on average (with a range of 4–20 years). This would indicate that sexual minorities living in communities displaying “high” stigma against homosexuality are apt to die notably sooner than sexual minorities living in communities with lower average stigma. For purposes of comparison, 12 years of reduced life span is greater than that found by the Centers for Disease Control and Prevention (CDC) among regular smokers, among whom life spans are documented to be, on average, 10 years shorter than among nonsmokers (Sakata et al., 2012). The magnitude of this finding—that personal and political attitudes among one's co-residents could be more harmful than the damage self-inflicted by smoking—prompted concern about possible alternative explanations and pathways of influence.

While no research effort is flawless, the original study seemed to overlook several possible confounding variables, including a primary sampling unit (PSU) measure of proportion Black. Given that African Americans are historically both politically liberal and yet cool toward LGBT rights, and communities comprising a higher share of them are more apt to suffer from higher (and earlier) mortality rates, questions about possible omitted variable bias arose. Additionally, the failure to include a measure of personal religiosity in the model seems unusual as well, given the proliferation of a religion-and-health literature in the 1990s that culminated in documenting a seven-year average difference in life expectancy between religious attenders and non-attenders using data from the same source as the Hatzenbuehler et al. study—the NDI (Hummer et al., 1999). It is the original study's process of imputing missing data for its four key social stigma items, however, that appears to bar the way to the successful recreation of the original key predictor variable—structural stigma—and hence hamper the ability to replicate the study's key findings and test alternative pathways of influence.

2. Methods

2.1. The original study's stigma measures

The merged GSS-NDI dataset is publicly available and was prepared for replication, to be followed by the test for possible confounds. The original reported sample of 914 sexual minority respondents out of 21,045 total respondents (4.34 percent) was successfully replicated, as was the 14 percent of respondents who had died by 2008. The individual-level control variables and PSU-level control measures were also replicated, with only tiny differences in a small number of measures.

The effort to replicate the original study was successful in everything except the creation of the PSU-level structural stigma variable. The study's authors constructed this PSU-level structural stigma variable from the following four GSS-NDI items:

1. “If some people in your community suggested that a book in favor of homosexuality should be taken out of your public library, would you favor removing this book, or not?” (GSS variable name: libhomo)
2. “Should a man who admits that he is a homosexual be allowed to teach in a college or university, or not?” (GSS variable name: colhomo)
3.

"Suppose a man who admits that he is a homosexual wanted to make a speech in your community. Should he be allowed to speak, or not?" (GSS variable name: spkhomo)

4. "Do you think that sexual relations between two adults of the same sex is always wrong, almost always wrong, wrong only sometimes, or not wrong at all?" (GSS variable name: homosex)

To construct the PSU-level structural stigma variable, the researchers dichotomized and summed the responses to these four survey items for each case, averaged this value for each PSU, and then constructed a dichotomous (i.e., threshold) measure of "high structural stigma" based on this PSU-level average. That top-quartile cut point, the authors note, was at 1.77, indicating that respondents were considered as living in a PSU with high structural anti-gay stigma if PSU residents responded with an anti-gay answer to (slightly) fewer than two of the four questions.

2.2. Analysis of missing data

The authors noted that, "given the structure of the GSS, not all questions were asked among all respondents each year," and that "(e)ach of these measures had greater than five percent missing due to this planned missing design, meaning that not all respondents were given the chance to respond to all questions." Fig. 1 displays the proportion of missing data for each of the four measures, as well as provides information about the missingness patterns in the 1988–2002 GSS-NDI dataset (N = 21,045). From the left panel of Fig. 1, each of the four measures exhibits around 40 percent missing values, the vast majority of which is intentional given the GSS split-ballot design. The magnitude of missing data was not made plainly evident in the original study, save for a reference to a "sizable portion of our data" on page 35.

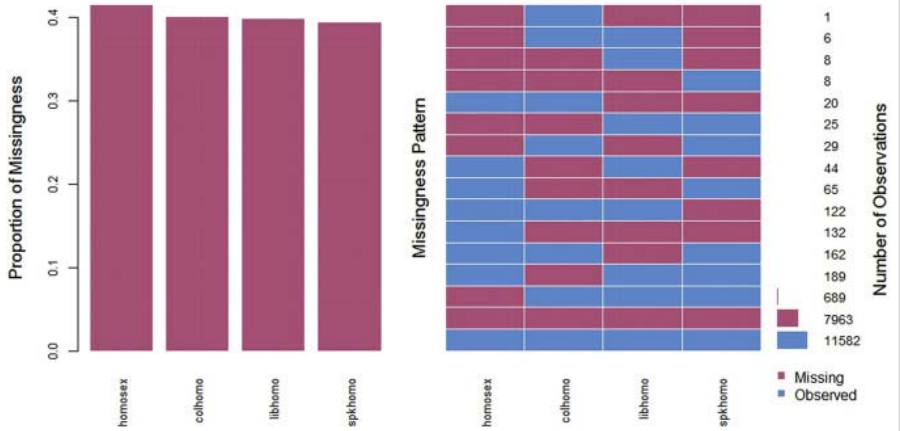


Fig. 1.
Visualization of missingness in the four structural stigma variables (N = 21,045).

Figure options

The right panel of Fig. 1 is an aggregation plot showing the patterns of missingness for the four stigma component measures. It displays all missing and observed variable combinations. Red blocks indicate missingness and blue blocks indicated non-missingness; the numbers to the right are counts (N) for each of the possible missingness patterns. For example, the bottom row contains all blue blocks (i.e., all variables have no missing values) and there are 11,582 cases with this pattern. Just above it is a row of all red blocks, indicating 7963 cases that are missing all four measures. The top-most row contains red, blue, red, and red blocks (i.e., colhomo is not missing, but the other three measures are all missing) and there is only 1 case with this pattern. For 93 percent of cases, then, all measures are either available (55 percent, 11,582 cases) or entirely missing (38 percent, 7963 cases). This high rate of planned missingness suggests that a missing completely at random (MCAR) assumption (i.e., the

probability of data being missing does not depend on the observed or unobserved data) may be appropriate. To assess this MCAR assumption, indicator variables for each structural stigma variable (0 = not missing, 1 = missing) were constructed, correlations with the other variables used in the study were computed, and logistic regression models were estimated to check for associations with the other variables. Logistic regression results revealed that income, education, race/ethnicity, and immigrant status were statistically significant predictors of missingness for the stigma items.

There was also a higher rate of missing values for the survey administration in 2002, which suggests that the year variable should be included in the imputation models. Since these findings do not support an MCAR assumption, the missing at random (MAR) assumption was adopted instead (i.e., missingness is related to the observed variables but not the missing values themselves) and these variables were included in the imputation models to make the assumption more plausible.

2.3. Multiple imputation process

The original study's authors employed Stata 11.2 and its "ice" multiple imputation command in order to estimate population parameters despite missing information, while the present replication attempts to do the same using the software package R 3.3.1 and its "mice" imputation procedure. The two should issue in comparable results, with any differences being miniscule (Berglund, 2015). The authors appeared to abide by the following multiple imputation steps:

1. Generate multiple imputed datasets using multiple imputation by chained equations (MICE).
2. Analyze the imputed datasets using the Cox proportional hazards model.
3. Pool the results to obtain overall estimates, variances, and confidence intervals following Rubin's Rules.

These general procedures are standard for multiple imputation. The most subjective step, however, is the first one, which involves the specification of imputation models. Unfortunately, the authors did not provide sufficient detail—both in the original study and in repeated personal requests—to replicate the exact imputation models they employed. From the original study (p. 35), the following about the authors' multiple imputation strategy can be discerned. It:

- "(U)sed the entire sample"
- Used "all of [their] covariates including the time variable (i.e., year of interview)" in their imputation models. Those covariates are age, race, sex, immigrant status, household income (log), years of education, self-assessed poor health, PSU-level average years of education, PSU-level average household income (log), and PSU-level proportion of conservatives (from the original study's Table 2).

Table 1.
Replication efforts at sample demographics of the sexual minority respondents in the GSS/NDI Study ($N = 914$).

Variable	Weighted mean or proportion in original study	Weighted mean in replication using best imputation
Respondent died by 2008	0.14	0.15
White	0.78	0.78
Black	0.16	0.16
Other race	0.07	0.06
Male	0.51	0.51
Female	0.49	0.49
Age at interview	39.9	40.9
Immigrant	0.12	0.10
Income (ln)	10.27	10.21
Years of education	13.40	13.40

Variable	Weighted mean or proportion in original study	Weighted mean in replication using best imputation
Fair/Poor self rated health	0.18	0.19
Resides in a high prejudice PSU	0.12	0.19
PSU Average education	13.28	13.30
PSU Average income (ln)	10.40	10.73
PSU Proportion conservative	0.34	0.30

Notes. PSU = primary sampling unit. Ln = logarithm transformed. Shading indicates key difference between original study and replication effort.

Table options

Table 2.
Original, replication, and complete-case hazard ratio estimates using Cox proportional hazard models predicting hazards of death for sexual minority individuals (*N* = 914).

	Original Study Model 3	Replication of model 3	Complete case replication of model 3	Original study model 5	Replication of model 5	Complete case replication of model 5
Structural stigma						
Top quartile PSU-level prejudiced score	2.29 (1.40, 3.67)	1.01 (0.68, 1.52)	0.98 (0.62, 1.57)	3.03 (1.50, 6.13)	0.89 (0.52, 1.50)	0.91 (0.38, 2.18)
Demographics						
Age at interview	1.05 (1.04, 1.07)	1.05 (1.04, 1.07)	1.05 (1.04, 1.06)	1.05 (1.04, 1.06)	1.05 (1.04, 1.06)	1.04 (0.96, 1.02)
Black	3.03 (1.90, 4.81)	2.64 (1.71, 4.06)	2.69 (1.65, 4.38)	2.87 (1.76, 4.67)	2.48 (1.54, 3.99)	2.44 (1.28, 4.64)
Other race	2.43 (1.01, 5.84)	3.10 (1.33, 7.22)	2.53 (1.10, 5.83)	2.28 (0.97, 5.37)	2.84 (1.19, 6.82)	1.89 (0.72, 4.98)
Female	0.58 (0.39, 0.87)	0.62 (0.44, 0.87)	0.63 (0.43, 0.91)	0.59 (0.39, 0.88)	0.61 (0.43, 0.87)	0.61 (0.39, 0.95)
Not US born	0.59 (0.26, 1.33)	0.56 (0.26, 1.22)	0.46 (0.19, 1.14)	0.54 (0.25, 1.18)	0.56 (0.26, 1.22)	0.39 (0.14, 1.11)
Socioeconomic factors						
Household income (log transformed)	1.04 (0.86, 1.24)	1.00 (0.84, 1.18)	1.01 (0.83, 1.22)	1.04 (0.86, 1.86)	1.02 (0.86, 1.21)	0.99 (0.79, 1.23)
Years of education	0.99 (0.93, 1.05)	1.00 (0.95, 1.06)	1.01 (0.94, 1.07)	0.99 (0.93, 1.05)	1.01 (0.96, 1.07)	0.99 (0.91, 1.07)
Self-assessed health						
Fair/Poor self-rated health				1.04 (0.61, 1.78)	1.27 (0.87, 1.85)	1.34 (0.75, 2.40)
Whole sample PSU-level covariates						
PSU-Average years of education				1.70 (0.56, 5.17)	0.88 (0.66, 1.15)	0.83 (0.59, 1.17)
PSU-Average income (log)				0.86 (0.61, 1.19)	0.75 (0.23, 2.46)	0.40 (0.10, 1.58)

	Original Study Model 3	Replication of model 3	Complete case replication of model 3	Original study model 5	Replication of model 5	Complete case replication of model 5
PSU-Proportion conservative				0.01 (0.00, 0.32)	0.09 (0.00, 2.26)	0.03 (0.00, 1.44)

Notes. Confidence intervals in parentheses. Shading indicates estimates from best replication effort.

Table options

- Imputed values for the four structural stigma items, as well as the other variables used in the imputation models. This implies that the PSU-level structural stigma average and binary PSU-level structural stigma variable were not imputed directly but instead constructed from that PSU-level average using the imputed structural stigma item scores.
- Adjusted the "imputation command … to ensure proper estimation of missing values on the covariates (i.e., continuous, dichotomous, or ordinal measurement)."
- Imputed 10 datasets using Stata's "ice" command. To check for quality, they examined and found "no statistical differences between the estimates of the means and standard errors of the covariates between imputed datasets."

There are many decisions that remain unknown, however, including whether or how the authors used the raw variables from which Table 2 covariates were constructed, whether they imputed values for the dichotomized version of the homosex variable or the untransformed homosex variable, made any special adjustments for the PSU-level variables, employed passive imputation or summed the individual-level stigma items after imputation, or made any adjustments for dealing with survival data (van Buuren et al., 1999 and White and Royston, 2009). The order in which they imputed the variables, as well as the number of iterations they used to ensure convergence of the MICE algorithm, remain unclear as well.

The lack of information on imputation model specifications presents a special challenge for study replication because the authors used MICE, which is a flexible and commonly-used multiple imputation approach that relies on not one but a set of imputation equations—one for each variable that has missing values—and draws imputations by iterating over them. The MICE algorithm first initializes by filling in missing values arbitrarily from observed values. In each iteration, the algorithm fits a statistical model for the first incomplete variable using the other variables as covariates, draws imputations based on this model, updates the dataset, and moves on to the next incomplete variable.

Since MICE involves specifying an imputation model for each incomplete variable, it can handle different variable types (i.e., nominal, ordinal, interval, ratio), which is a very attractive feature. However, using several imputation models requires making numerous decisions about what predictors and which imputation method (e.g., predictive mean matching, logistic regression) to use for each variable, as well as the order that each variable should be imputed. Because the possible combinations of options are so numerous, the following three limitations were placed on the replication imputation attempts:

1. No interaction terms. This is based on the fact that none of the models in Table 2 of the original study contained an interaction term. The authors did, however, briefly discuss a model (not included in Table 2) that included an interaction between sexual orientation and structural stigma, which they estimated using the whole sample, not just on the 914 cases involving sexual minorities. It is unknown whether they used a separate imputation process—one that includes this interaction term in the imputation models—for the data used for estimating this model.
2.

No multilevel modeling. The original study did not specify any multilevel modeling approach or terminology, other than including PSU-level variables as contextual independent variables in their analytical models.

3. Use of default models for each type of variable. Namely, logit regression was employed for imputing binary variables and predictive mean matching for continuous variables (which were all non-normal). Since the point of the models is to generate plausible values and the actual parameter estimates of those models are not needed, the exact form of those models is generally unimportant (see van Buuren et al., 1999).

Typically, only 5–10 imputations are needed to stabilize the distribution of the regression parameters (Brand, 1999). Rubin (1987) demonstrated in his classic text on multiple imputation that there was little advantage to producing and analyzing any more. The original study employed 10 imputations. However, Graham et al. (2007) demonstrate in a simulation study that more imputations should be performed in order to increase statistical power, and that higher rates of missing information call for more imputations. For a missing information rate of 50–70 percent and a power fall-off over less than one percent, they recommend 40 imputations. But the present replication effort does not allow for a straightforward application of this rule-of-thumb. According to Rubin (1987), the missing information rate and missing data rate are equal in the case of no covariates, but the former is typically less than the latter when there are covariates; later studies approximated the fraction of missing information for any parameter at less than the fraction of incomplete cases (White et al., 2011). This suggests that the missing information rates in this data (which contain many covariates) are smaller than the missing data rates of about 40 percent for the four structural stigma items, and that using Graham et al.'s suggestion of 40 imputations is more than sufficient for good statistical inference. However, it is important to note that these four structural stigma items were not used directly as covariates in the analytical models (in either the original or this replication study); instead, they were combined and transformed into a single binary PSU-level structural stigma variable, which was then used in the analytical models. This means that applying Graham et al.'s rule-of-thumb means basing the appropriate number of imputations on the fraction of missing information (FMI) associated with the parameter of this binary PSU-level structural stigma variable and not on the fraction of missing information associated with the parameters of the four structural stigma items. The exact nature of this relationship between fractions of missing information at different levels of aggregation and its consequences for parameter estimation is beyond the scope of this replication study (see Shin and Raudenbush, 2010 discussion on handling missing data at different levels). The estimated FMI associated with the parameter of the binary PSU-level structural stigma variable is close to zero (2.0e-6), which suggests that even the original authors' use of 10 imputations should be sufficient. Despite these theoretical concerns, the practical consequences are small: in the replication efforts, parameter estimates using 10 imputations were compared with those using 40 imputations, and the parameter estimates were found to be nearly identical.

Another concern is that the original study focuses on Cox regressions of mortality on structural stigma, adjusted for various individual-level and PSU-level characteristics. In multiple imputation, it is necessary to include the outcome variable in the imputation models. For imputation models for incomplete variables that are later used as covariates for survival models, the inclusion of survival variables must be carefully considered. According to White and Royston (2009), incorrect inclusion of survival outcomes in imputation models may subsequently dilute the association between the incomplete covariate and the survival outcome. The authors of the original study did not indicate how they included survival outcomes in their imputation models.

In light of these concerns, three imputation approaches were developed. The first, deemed the "best" approach, involves using MICE to handle the missing data in the four structural stigma items along with the other covariates used in the analytical models in the original study. It is "best" in the sense that it is based on the most generous reading of

the description of the imputation approach in the original study. Since the original study lacks complete information for replication, efforts to do so proceeded as if the best practices discerned from the multiple imputation literature stood in place of the information needed but not provided. Accordingly, a range of imputation was produced (first 5, then 10, 20, 40, and 100 imputed datasets; the results of 40 imputations is reported), using the binary event indicator and the Nelson–Aalen estimate of the cumulative hazard as covariates in the imputation models (White and Royston, 2009). A variety of diagnostics were performed (as detailed below).

The second approach is based on the most straightforward reading of the original study's multiple imputation approach (with the exception of using the mice package in R instead of Stata's ice command). Accordingly, 10 datasets were generated and only the variables reported in Table 2 of the original study, along with the interview year and time to death, were included.

The third approach consists of a series of explorations of possible alternative specifications of the original. This approach (attempts 3 through 10 in Table 3) entailed some creativity: What if the imputation approach involved only a subset of the sample? What if the binary threshold for the construction of the structural stigma variable was adjusted, or fixed at the value described in the original study? For all approaches, incomplete binary variables were estimated using logistic regression, and incomplete continuous variables were estimated using predictive mean matching. All coding in R is available as a supplement on the journal's website.

Table 3.
Attempts to replicate original study's structural stigma measure proportion and hazard ratio. ($N = 914$).

Description of model	Top quartile threshold for PSU-level stigma (range 0–4)	Weighted proportion of sexual minorities residing in high-stigma PSU	Hazard ratio estimate of structural stigma on mortality
Complete cases only: no imputation	1.84	0.19	0.91 (0.38, 2.18)
Original study	1.77	0.12	3.03 (1.50, 6.13)
Attempt 1 (best practice): Imputation (passive) using all cases, with additional cumulative baseline hazard and individual-level political conservatism	1.75	0.19	0.89 (0.52, 1.50)
Attempt 2 (most straightforward replication): Imputation (passive) using all cases, all variables	1.74	0.19	0.96 (0.58, 1.57)
Additional attempts at replication			
Attempt 3: Same as 1a but using WTSALL (not WTSS) as weights	1.74	0.20	0.90 (0.54, 1.50)
Attempt 4: Imputation using only cases involving sexual minorities	1.50	0.26	0.96 (0.67, 1.37)
Attempt 5: Imputation of individual-level stigma items only	1.78	0.18	0.88 (0.42, 1.87)
Attempt 6: Remove passive imputation on stigma, impute all variable	1.78	0.18	0.96 (0.57, 1.62)
Attempt 7: Remove passive imputation, impute stigma items only	1.78	0.18	0.83 (0.42, 1.66)
Attempt 8: Same as #1, but fix cut-off at 1.77	1.77	0.18	0.97 (0.58, 1.64)
Attempt 9: Same as #2, but use top-quintile (20%) as threshold	1.84	0.15	1.13 (0.63, 2.02)
Attempt 10: Same as #2, but fix threshold so high-stigma PSU prop = 0.12	1.90	0.12	1.40 (0.77, 2.57)

Notes. The estimates in the right-most column seek to replicate the hazard ratio of structural stigma found in the original study's Table 2, Model 5. Attempt 1 employs 40 imputed datasets; Attempts 2–10 use 10 imputed datasets.

Table options

3. Results

3.1. Imputation model diagnostics

To prepare the data for the first (and "best") effort at replication, several imputation diagnostics were used to assess the results of the multiple imputation efforts. First, diagnostic plots aid in the assessment of the multiple imputation results for convergence and reasonableness. In order to assess algorithm convergence, Fig. 2a displays trace line plots for the proportions and standard deviations of the imputed values for each of the structural stigma items for 10 imputations, while Fig. 2b displays identical plots for 40 imputations. While there was initial trend, both sets of plots appear to show intermingling and trendless imputation streams after the fifth iteration, indicating healthy convergence after a brief burn-in period. Strip plot diagnostics of imputed values (red) and observed values (blue) for 40 imputations for each structural stigma value are available for examination as supplementary material at the journal's website (and at the end of this document). The overlap of imputed and observed dots suggests the results of the imputation appear reasonable.

Fig. 2.
a: Trace Plots for Assessing Convergence of Imputation Model: 10 imputations. b: Trace Plots for Assessing Convergence of Imputation Model: 40 imputations.

Figure options

Second, chi-square tests of imputed and observed values for each variable (not shown) revealed that there were no statistical differences for imputed and observed libhomo, colhomo, and spkhomo but that there were statistically significant differences for imputed and observed values of the variable homosex. The fact that it also had the highest missingness rate of 41.5% as well as a slightly higher unplanned missingness rate (see the right panel of Fig. 1) further called into question the MCAR assumption. Since the MAR assumption can be made more plausible by including additional variables in the imputation model, the imputation models in the best imputation attempt (#1) were refined by including individual-level conservatism (which was used in the construction of PSU-level conservatism) so that the difference in the rates for homosex is less statistically significant and is not a dramatic one (e.g., homosex = 1 in 73% of available cases versus 77% in the imputed data). Moreover, the imputed and observed distributions may be different but still missing-at-random and explainable by other variables in the dataset (see Abayomi et al., 2008 for further details).

Third, posterior predictive checking (PPC) was conducted to assess our best imputation approach against our most straightforward imputation approach (see He and Zaslavsky, 2012 and Nguyen et al., 2015). A Bayesian model-checking technique, PPC involves examining whether the analysis from the observed data looks "typical" of results obtained from the replicates produced by the imputation model by applying the analyses of interest to both the observed and replicated data, and measuring the discrepancies between estimates of a target quantity or quantities. The average estimates from the completed data and their replicates, as well as the posterior predictive p-values (estimated using 100 replications) are presented in Table 4.

Table 4.
Posterior predictive checking results for approaches 1 and 2 (best vs. most straightforward replication attempts).

Test quantity	Approach 1				Approach 2		
	Completed data	Replicated data	Discrepancy	Posterior predictive p-value	Completed data	Replicated data	Discrepanc
Top quartile PSU-level structural stigma value	1.75	1.75	−0.01	0.36	1.75	1.76	−0.01
Proportion living in high-stigma PSUs	0.25	0.25	−0.01	0.38	0.25	0.25	−0.01

Notes. Test quantities are unweighted. The posterior predictive p-value is the proportion of draws for which the test quantity computed with completed data is greater than the test quantity computed with replicated data.

Table options

Finally, the results of computing the PSU-level structural stigma variable using complete cases only are provided alongside the results of the original study and the best effort at replication (in Table 2 and Table 3). Sterne et al. (2009) emphasized the importance of careful comparison of multiple imputation results with the results available from complete-case analysis.

Additionally, the imputation command was adjusted according to each variable's level of measurement (i.e., continuous, dichotomous, or ordinal) to ensure proper estimation of population parameters given significant missing information. Passive imputation for the individual-level structural stigma variable (but not the other measures) was employed on eight (of 10) replication attempts. Passive imputation is a method for handling transformed, combined, or recoded versions of data during an imputation by ensuring that the transformations are always consistent with the data. Passive imputation for the individual-level structural stigma index ensures that it is consistently a sum of the four stigma items throughout the imputation process by forcing it to always depend on the most recently generated imputations of the four stigma items. Since the use of passive imputation is not settled (see Von Hippel, 2009 and Seaman et al., 2012), attempts both with and without passive imputation are included.

A basic comparison of the best replication attempt's variables (using multiple imputation) and the original study's variables appears in Table 1. Efforts to replicate the structural stigma variable issued in a lower top-quartile threshold for the measure, and hence a higher estimate for the share of respondents who reside in a high-prejudice PSU: 19 percent vs. the original study's 12 percent. Moreover, the best approach, the most straightforward attempt at replication, and the complete-cases-only approach all generate the same estimate (19 percent, in Table 3).

3.2. Predicting mortality of sexual minorities as a function of structural stigma

Table 2 displays estimates from Cox proportional hazard models seeking to replicate the original study's key Table 2, drawing on missing data that is imputed using passive imputation on the individual-level structural stigma index (using all cases), and follows White and Royston (2009) suggestion to use the estimate of the cumulative baseline hazard to the survival time as predictors and Royston's in imputation models. This approach employed 40 imputed datasets.

Instead of displaying all five original models and all five best replication models, only the third and fifth models are displayed, together with identical model estimates from completed cases only (i.e., no imputations) for comparison. Most of the original study effects are largely paralleled in the best replication attempt. The structural stigma estimate, however, is not statistically significant in any of the proportional hazard models in the best replication attempts displayed in Table 2. Even when it is regressed alone on mortality (not shown), its estimate is not significantly different from zero (HR = 1.40, 95% CI = 0.95–2.05). Whereas in the original study the hazard ratio of structural stigma appears not to weaken with the addition of covariates—and is at its largest in the final model—the hazard ratio consistently diminishes in the replication and complete-cases approaches. Thus the key predictor of mortality in the original study, and source of claims about an average 12-year reduction in life expectancy, is not associated with mortality in this effort at replication. (Even the zero-order correlation between structural stigma and mortality is only 0.11.)

The results employing only complete cases appear similar to the results of the replication effort using the best imputation approach. When structural stigma is included alone using only completed cases, no effect is apparent (HR: 1.37, 95% CI = 0.92–2.03). According to Sterne et al. (2009), “Where complete cases and multiple imputation analyses give different results, the analyst should attempt to understand why, and this should be reported in publications.” While multiple imputation can reduce bias and inefficiency compared to complete-case analysis, it is nevertheless unclear why there are such large differences between the complete-case analysis and the multiple imputation analysis in the original study but such small differences between the former and the multiple imputation analysis (using the best approach) in the present replication efforts.

3.2.1. Alternative imputation strategies and model results

Imputation of missing data tends to yield slight fluctuations in its estimates of population parameters, suggesting the wisdom of evaluating alternative attempts. Nine additional imputation approaches were attempted, with the idea that perhaps some variation or misspecification in imputation strategy may account for the different findings.

The best imputation strategy (from Table 2 and Table 3 “Attempt 1”) was altered in nine different ways prior to the generation of new Cox regression models, and the results are detailed in Table 3, where three different estimates for each attempt are displayed: (1) the top quartile threshold for PSU-level stigma, (2) the resulting weighted proportion of sexual minorities who reside in a high-stigma PSU, and (3) the hazard ratio estimate of structural stigma on mortality (as generated from a model identical to that found in the original study's fifth model in Table 2). The nine only differ by varying the manner in which the imputation was conducted.

Given that 40 imputations did not significantly improve upon 10 imputations, each of the subsequent nine attempts relied on 10 rather than 40 imputed datasets. Attempt 2 employs MICE using passive imputation on the individual-level structural stigma index, using all cases and all variables, and time-to-death as a covariate. This attempt can be considered the most straightforward replication approach based on a reading of the original study, but not as prudent as the best approach. Attempt 2, however, looks little different from the best approach, yielding a similar threshold for the binary stigma variable and identical share of respondents who live in a high-stigma PSU, as well as a hazard ratio estimate that is only marginally higher than the first attempt. The PPC results, displayed in Table 4, assess model fit by measuring the discrepancy between the

first two imputation models and the completed data with respect to the test quantities—the proportion of individuals living in high-stigma PSUs and the top-quartile value of the PSU-level structural stigma variable. If there is a misfit between an imputation model and the data with respect to these test quantities, extreme posterior predictive p-values of close to 0 or 1 would "flag" such a model. The results reveal that none of the discrepancies have extreme posterior predictive p-values, which means that both imputation models' performance for these quantities of interest are reasonable and similar.

Besides the first and second attempts, eight additional efforts at generating hazard ratio estimates by altering the manner in which missing data was imputed were pursued. They involved: (3) experimenting with using an alternate weight variable, (4) imputing the data using only cases involving sexual minorities rather than employing the complete sample, (5) imputing only the four structural stigma items (rather than other missing data as well) but retaining the passive imputation approach, and (6) imputing missing data from all variables (not just the stigma items) while removing the passive imputation approach. Removing the passive imputation means that missing data for the individual-level stigma variable was imputed for the summed index, not for the four individual stigma items (which would then be summed and averaged across PSUs). A similar effort (7) imputed

Table 3 reveals that the original study's 3.03 hazard ratio continues to appear dramatically distinctive—far larger and statistically significant—than each of these alternative estimates. Moreover, the variation in hazard ratios among these different efforts at replication is miniscule. No replication effort naturally resulted in a top-quartile measure of PSU-level structural stigma in which only 12 percent of the sample of sexual minorities lived.

The final three efforts attempted to mimic the second and most straightforward approach to replication, but did so by altering (or fixing) three different thresholds in the construction of the PSU-level stigma measure. Since the first seven attempts yielded top-quartile stigma thresholds that were slightly different than the original study's, the eighth attempt set a threshold at 1.77, to match that reported in the original study. That measure yielded 18 percent of the population living in a high-prejudice PSU, but a hazard ratio on mortality (in the full model) of 0.97. The ninth attempt shifted to the top quintile (20%) as a cut-off, and yielded 15 percent of the sample living in a high-prejudice PSU, but a hazard ratio on mortality of 1.13. Neither estimate was statistically significant. Finally, the threshold was fixed (at 1.90) in order to assure that only 12 percent of sexual minorities in the sample resided in a high-stigma PSU, in order to match that in the original study. This final attempt yielded a larger hazard ratio on mortality (1.40, with 95% confidence intervals of 0.77–2.57) than all previous attempts, but it too remained statistically insignificant.

3.2.2. Factor scores as measures of structural stigma

The authors of the original study note (on page 37) that they explored "alternative measures of structural stigma, including predicted factor scores at the PSU level and the average summed prejudice scores at the PSU level." They found that each of these produced stronger results—that is, more powerful effects on mortality, than the dichotomous measure they elected to use (for ease of interpretation). Replicating these was briefly explored as well, despite concerns in the psychometrics literature about such an approach and the variety of possible extraction and rotation methods the original study could have employed. Hence this approach is more exploratory. Indeed, factor score indeterminacy could issue in lots of sets of factor scores consistent with an identical set of factor loadings (Grice, 2001).

For the factor scores, polychoric/tetrachoric correlations were used, since the four stigma items are all treated as dichotomous. The default minimum residual (OLS) was employed as the factoring method, using varimax rotation (as well as experiments with promax, simplimax, and no rotation—it did not matter). This resulted in three extracted factors,

with factor scores obtained based on the factor with the highest proportion of variance. For the second effort—the average summed prejudice scores at the PSU level—the imputed four structural stigma items were summed to obtain an individual-level stigma score that was averaged for each PSU to create a continuous PSU-level structural stigma score.

As Hatzenbuehler et al. (2014: 37) discovered using this approach, so too in this replication: "(o)ur results … were stronger than the dichotomized measure." The Cox proportional hazard ratio for the alternative structural stigma measures were 1.49 for the first effort and 1.37 for the second. However, each estimate's 95% confidence intervals (0.67–3.32 and 0.52 to 3.65, respectively) suggest that neither is statistically significant, and the magnitudes remain much smaller than what is described in the original study.

3.2.3. Religiosity and PSU-percent black

Finally, the effect of a pair of potential confounds or alternative pathways of influence—the ones that prompted initial interest in this study in the first place—were explored. PSU-percent Black and an individual-level measure of religious service attendance were added separately, then together, to the most straightforward replication of the original

both models), displaying 95% confidence intervals of 0.01 and 0.39 in the model with religious attendance. Their inclusion did not affect the overall hazard ratio of structural stigma on mortality.

4. Discussion

Efforts to replicate Hatzenbuehler et al. (2014) study of the effects of structural stigma, as well as to improve upon its missing data imputation, failed to generate the original study's report of strong and statistically significant effects of structural stigma on the premature mortality of sexual minorities. Efforts to replicate the structural stigma measure following what could be called a "best practice" approach, as well as one following the most straightforward reading of the original study's description, each issued in results that indicated greater numbers of people living in "high" stigma PSUs as well as no effect of that stigma on the mortality of sexual minorities. Eight additional approaches to the imputation of missing data were attempted, none of which generated anything like the results reported in the original study. The same is true for the alternative measures—factor scores and average summed prejudice. Replication estimates appear similar to those generated using complete cases only.

Minimally, the findings of Hatzenbuehler et al. (2014) study of the effects of structural stigma seem to be very sensitive to subjective decisions about the imputation of missing data, decisions to which readers are not privy. Moreover, the structural stigma variable itself seems questionable, involving quite different types of measures, the loss of information (in repeated dichotomizing) and an arbitrary cut-off at a top-quartile level. Hence the original study's claims that such stigma stably accounts for 12 years of diminished life span among sexual minorities seems unfounded, since it is entirely mitigated in multiple attempts to replicate the imputed stigma variable.

The unavailability of the original study's syntax and the insufficient description of multiple imputation procedures leave unclear the reasons for the failed replication. It does, however, suggest that the results are far more contingent and tenuous than the original authors conveyed. This should not be read as a commentary on missing data or on the broader field of the study of social stigma on physical and emotional health outcomes, but rather as a call to greater transparency in science (Ioannidis, 2005). While the original study is not unique in its lack of details about multiple imputation procedures, future efforts ought to include supplementary material (online) enabling scholars elsewhere to evaluate and replicate studies' central findings (Rezvan et al., 2015). This would enhance the educational content of studies as well as improve disciplinary rigor across research domains.

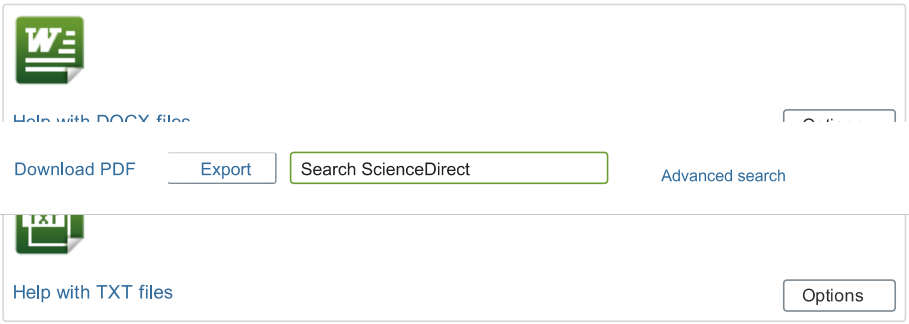
5. Conclusion

Repeated independent efforts were unable to replicate Hatzenbuehler et al. (2014)*Social Science & Medicine* article's key finding that structural stigma at the PSU level contributed to early mortality among a sample of sexual minority respondents. The obstruction to doing so rests in an insufficiently documented missing-data imputation process. However, numerous alternative missing data imputation approaches, performed in an effort to replicate the original study, each resulted in null effects of structural stigma on mortality.

Appendix A. Supplementary data

The following are the supplementary data related to this article:

Supplementary data related to this article, including original coding in R and additional imputation model convergence information, can be found on the journal's website.



References

Abayomi et al., 2008 K. Abayomi, G. Andrew, M. Levy
Diagnostics for multivariate imputations
Appl. Stat-J. Roy. St. C, 57 (2008), pp. 273–291 http://dx.doi.org/10.1111/j.1467-9876.2007.00613.x
Loading

Berglund, 2015 P. Berglund
Multiple Imputation Using Chained Equations: a Comparison of Stata, SAS, IVEware and R
Survey Methodology Program, Institute for Social Research (2015) http://www.misug.org/uploads/8/1/9/1/8191072/pberglund_mult_imputation_software_comparisons.pdf (Accessed 16 February 2003)
Loading

Bostwick, 2012 W. Bostwick
Assessing bisexual stigma and mental health status: a brief report
J. Bisex., 12 (2012), pp. 214–222 http://dx.doi.org/10.1080/15299716.2012.674860
Loading

Brand, 1999 J.P.L. Brand
Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets
Erasmus University, Rotterdam (1999)
Loading

Cumming, 2013 G. Cumming
The new statistics: why and how
Psychol. Sci., 25 (2013), pp. 7–29 http://dx.doi.org/10.1177/0956797613504966
Loading

Graham et al., 2007 J.W. Graham, A.E. Olchowski, T.D. Gilreath
How many imputations are really needed? Some practical clarifications of multiple imputation theory
Prev. Sci., 8 (2007), pp. 206–213 http://dx.doi.org/10.1007/s11121-007-0070-9
Loading

Grice, 2001 J.W. Grice
Computing and evaluating factor scores
Psychol. Meth., 6 (2001), pp. 430–450 http://dx.doi.org/10.1037/1082-989X.6.4.430
Loading

Hatzenbuehler, 2009 M.L. Hatzenbuehler
How does sexual minority stigma 'get under the skin'? A psychological mediation framework
Psychol. Bull., 135 (2009), pp. 707–730 http://dx.doi.org/10.1037/a0016441
Loading

Hatzenbuehler, 2011 M.L. Hatzenbuehler
The social environment and suicide attempts in a population-based sample of LGB youth
Pediatrics, 127 (2011) http://dx.doi.org/10.1542/peds.2010-3020 896e903
Loading

Hatzenbuehler et al., 2014 M.L. Hatzenbuehler, A. Bellatorre, Y. Lee, B.K. Finch, P. Muennig, K. Fiscella
Structural stigma and all-cause mortality in sexual minority populations
Soc. Sci. Med., 103 (2014), pp. 33–41 http://dx.doi.org/10.1016/j.socscimed.2013.06.005
Loading

Hatzenbuehler et al., 2009 M.L. Hatzenbuehler, K.M. Keyes, D.S. Hasin
State-level policies and psychiatric morbidity in lesbian, gay, and bisexual populations
Am. J. Pub Health, 99 (2009), pp. 2275–2281 http://dx.doi.org/10.2105/AJPH.2008.153510
Loading

He and Zaslavsky, 2012 Y. He, A.M. Zaslavsky

Loading

Herek and Garnets, 2007 G.M. Herek, L.D. Garnets
Sexual orientation and mental health
Annu. Rev. Clin. Psycho, 3 (2007), pp. 353–375
http://dx.doi.org/10.1146/annurev.clinpsy.3.022806.091510
Loading

Hummer et al., 1999 R.A. Hummer, R.G. Rogers, C.B. Nam, C.G. Ellison
Religious involvement and U.S. adult mortality
Dem, 36 (1999), pp. 273–285 http://dx.doi.org/10.2307/2648114
Loading

Ioannidis, 2005 J.P.A. Ioannidis
Why most published research findings are false
PLoS Med., 2 (8) (2005), p. e124 http://dx.doi.org/10.1371/journal.pmed.0020124
Loading

Ioannidis, 2008 J.P.A. Ioannidis
Why most discovered true associations are inflated
Epidemiology, 19 (2008), pp. 640–648 http://dx.doi.org/10.1097/EDE.0b013e31818131e7
Loading

Meyer, 2003 I.H. Meyer
Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence
Psych. Bul., 129 (2003), pp. 674–697 http://dx.doi.org/10.1037/0033-2909.129.5.674
Loading

Nguyen et al., 2015 C.D. Nguyen, K.J. Lee, J.B. Carlin
Posterior predictive checking of multiple imputation models
Biometrical J., 57 (2015), pp. 676–694
Loading

Rezvan et al., 2015 P.H. Rezvan, K.J. Lee, J.A. Simpson
The rise of multiple imputation: a review of the reporting and implementation of the method in medical research
Bmc Med. Res. Methodol., 15 (30) (2015) http://dx.doi.org/10.1186/s12874-015-0022-1
Loading

Rubin, 1987 D.B. Rubin
Multiple Imputation for Nonresponse in Surveys
Wiley & Sons, New York (1987)
Loading

Sakata et al., 2012 R. Sakata, P. McGale, E.J. Grant, K. Ozasa, R. Peto, S.C. Darby
Impact of smoking on mortality and life expectancy in Japanese smokers: a prospective cohort study
Brit. Med. J., 345 (2012), p. e7093 http://dx.doi.org/10.1136/bmje7093
Loading

Seaman et al., 2012 S. Seaman, J. Bartlett, I. White
Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods
BMC Med. Res. Meth., 12 (2012), p. 46
Loading

Shin and Raudenbush, 2010 Y. Shin, S.W. Raudenbush
A latent cluster-mean approach to the contextual effects model with missing data
J. Educ. Behav. Stat., 35 (1) (2010), pp. 26–53 http://dx.doi.org/10.3102/1076998609345252
Loading

Simmons et al., 2011 J.P. Simmons, L.D. Nelson, U. Simonsohn
False-positive psychology:Undisclosed flexibility in data collection and analysis allows presenting anything as significant
Psychol. Sci., 22 (2011), pp. 1359–1366 http://dx.doi.org/10.1177/0956797611417632

Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls
BMJ, 338 (2009), p. b2393 http://dx.doi.org/10.1136/bmj.b2393
Loading

van Buuren et al., 1999 S. van Buuren, H.C. Boshuizen, D.L. Knook
Multiple imputation of missing blood pressure covariates in survival analysis
Stat. Med., 18 (1999), pp. 681–694
Loading

Von Hippel, 2009 P. Von Hippel
How to impute interactions, squares, and other transformed variables
Soc. Meth., 39 (2009), pp. 265–291
Loading

White et al., 2011 I.R. White, P. Royston, A.M. Wood
Multiple imputation using chained equations: issues and guidance for practice
Stat. Med., 30 (2011), pp. 377–399 http://dx.doi.org/10.1002/sim.4067
Loading

White and Royston, 2009 I.R. White, P. Royston
Imputing missing covariate values for the Cox model
Stat. Med., 28 (2009), pp. 1982–1998 http://dx.doi.org/10.1002/sim.3618
Loading

Department of Sociology, University of Texas at Austin, 305 E 23rd St, A1700, Austin, TX 78712-1086, USA.

Note to users:

Corrected proofs are Articles in Press that contain the authors' corrections. Final citation details, e.g., volume and/or issue number, publication year and page numbers, still need to be added and the text might change before final publication.

Although corrected proofs do not have all bibliographic details available yet, they can already be cited using the year of online publication and the DOI , as follows: author(s), article title, Publication (year), DOI. Please consult the journal's reference style for the exact appearance of these elements, abbreviation of journal names and use of punctuation.

When the final article is assigned to volumes/issues of the Publication, the Article in Press version will be removed and the final version will appear in the associated

published volumes/issues of the Publication. The date the article was first made available online will be carried over.