Research on Youth Suicide and Sexual Orientation is Impacted by High Rates of Missingness in National Surveillance Systems

Peter Phalen[a,1] **and Aaron Kivisto**[b]

[a]Division of Psychiatric Services Research, University of Maryland School of Medicine; [b]College of Applied Behavioral Sciences, University of Indianapolis

This manuscript was compiled on June 16, 2023

Objective. The sexual orientation of young people who die by suicide in the United States is usually unknown. This study assessed how observed patterns of unknown sexual orientation are likely to affect research findings.

Methods. We analyzed the National Violent Death Reporting System (NVDRS) Restricted Access Dataset to assess whether sexual orientation among youth suicide decedents is disproportionately known for different demographics. We then assessed the degree to which estimated sexual minority rates would be affected if researchers were to assume either (a) that sexual orientation data is missing completely at random, or (b) that orientation information is missing at random after accounting for observed demographic patterns.

Results. Less than 10% of the sample had known sexual orientation. Sexual orientation was more frequently known for females, white people, and older people, and missingness varied by geography. The choice between modeling the data as missing completely at random versus at random conditional upon demographics had a more than 2-fold impact on estimated sexual minority rates among youth suicide decedents.

Conclusion. Research on sexual orientation and youth suicide is strongly impacted by how researchers account (or do not account) for missingness.

Suicidality | Sexuality | Adolescents | Young adults

Research on disparities in suicide related to sexuality is hampered by limited data on sexual orientation at time of death (1, 2). Posthumous identification of sexual orientation typically requires a psychological autopsy that relies mostly on third parties such as family members or close friends who may be unaware of the sexuality of the deceased person (3). This problem is compounded for younger age groups, as most people do not disclose to friends or close family members until approximately 16 years old on average, with many individuals waiting much longer (4).

The methodological issue of unknown sexual orientation of suicide decedents is so substantial that scientists have suggested that research in this area is inherently limited and "cannot reach top quality status" (3). This limitation increases data analysts' degrees of freedom in that researchers must make various assumptions and decisions about how best to handle missing data. It has been demonstrated that scientific results can be strongly influenced by researcher's subjective decisions regarding analytic strategies (5). For example, Silberzahn et al. (6) assigned 61 analysts from 29 teams to answer the same question using the same dataset and found high levels of variability in the resultant effect sizes (0.89 to 2.93) and whether significant effects were detected. The impact of these subjective analytic decisions is likely to be greater in certain situations such as when working with high rates of missing data (7).

Researchers have continued to use mortality data to estimate risk of suicide by sexual orientation, often without appropriate caveats about high rates of missing data or explicit discussion of associated analytic assumptions and their potential consequence. Estimates of the proportion of individuals who die by suicide who are sexual minorities have varied considerably, even when derived from the same data sources. For example, researchers have used the (relatively high-quality) National Violent Death Reporting System (NVDRS) dataset to publish estimates that lesbian, gay, and bisexual (LGB) individuals comprise as few as 2.5% of youth decedents (8) or as much as 8.4% (9). During the time period covered by these studies, approximately 6% of young people in the U.S. identified as LGBT (10), so these NVDRS-derived estimates suggest that sexual minorities were either substantially underrepresented or substantially over-represented among suicide decedents.

What analytic decisions were required to generate such estimates? Results were driven by multiple factors, but researchers' decisions about how to handle large amounts of missing sexuality information were notable. In Ream (9), the author simply restricted analyses to the 21% of cases with known sexuality—thus making the implicit assumption that sexual orientation was missing completely at random—and calculated an 8.4% rate of LGB status among NVDRS suicide decedents between the ages of 12 and 29. In Patten et al. (8) the authors attempted to more proactively account for missingness. First, the authors excluded all data from state/year combinations that had lower than 20% rates of known sexuality. (For example, if rates of missing sexuality reported by California in 2018 were 85%, then that year of data for that state would have been removed.) This step thereby removed 67.6% of all suicide decedents and left the researchers

Significance Statement

Over 90% of young people who die by suicide have unknown or unreported sexual orientation. The authors show that these patterns of missing data are non-random and have the potential to strongly bias research findings. Scientists and policy-makers should be cautious about using mortality data to study disparities in suicidality for people with different sexual orientations.

[1]To whom correspondence should be addressed. E-mail: pphalen@som.umaryland.edu

with data on 14 states and Puerto Rico (out of a possible 36 states that participated in NVDRS during the study period). Then, the authors removed all remaining cases with missing sexuality, thereby removing an additional 49.9% of the remaining sample. Thus, their final estimate that 2.5% of suicide decedents identified as LGB was derived from a subset of only 16% of all suicide decedents. In another study published using the same NVDRS dataset, Lyons et al. (11) used text-based analysis to look for affirmative evidence of sexual minority status and ultimately reported that only 0.5% of youth suicide decedents were LGB on the basis of such affirmative evidence being present. The authors of all these studies necessarily made very impactful analytic decisions about how to handle a lack of information about sexual orientation.

The accuracy of estimates of the sexual orientation of people who die by suicide has significant implications. The question of whether suicide rates are higher or lower for sexual minorities has been the subject of debate for decades (12) because the conclusion may cause or alleviate stigma, impact peoples' thoughts and feelings about themselves or their futures, and influence how limited public resources are directed (3). However, we are unaware of any attempts to quantify the degree to which analytic decisions about handling missing data may impact conclusions about sexual orientation and youth suicide. The present study evaluates how the decision between approaching sexual orientation information as (a) missing completely at random, or (b) missing at random (conditional upon demographic patterns) would impact estimated rates of sexual orientation among youth suicide decedents. We also discuss a third—in our view most likely but least quantitatively tractable—possibility that sexual orientation data is not missing at random.

Materials and Methods

We analyzed NVDRS Restricted Access Data—which includes data from 43 states, Puerto Rico, and the District of Columbia—for all suicides (viz., ICD-10 codes of X60-X84, Y87, U03) among youth aged 11-21 from 2015 to 2019. Sexual orientation was coded by NVDRS as heterosexual, gay, lesbian, bisexual, unspecified sexual minority, or missing/unknown, on the basis of law enforcement, coroner, and/or medical examiner reports (13).

We first estimated the rate of sexual minority status that would be calculated under the assumption that sexual orientation information is missing completely at random. If sexuality information is missing completely at random, then the best estimate of the true rate of sexual minority status among youth who die by suicide is simply the rate of sexual minority status after restricting the dataset to people with non-missing sexual orientation codes.

We then fit a multilevel regression model to assess whether sexuality was differentially known as a function of age, race, sex, and location (viz., state) of death. Finally, we assessed how estimated rates of sexual minority status would be affected if researchers were to assume that the data is missing at random conditional on observed demographics. If the data are missing at random conditional upon demographics, then estimating the rate of sexual minority status among youth who die by suicide requires statistical adjustment, which we carry out by imputing missing data using those observed variables (14). This study was deemed exempt by the University of Indianapolis IRB.

Table 1. Demographic characteristics and rates of non-missing sexual orientation codes among 12,117 youth suicide decedents

	Overall Sample		Coded for Orientation	
	N	%	N	%
Sex				
Male	9371	77.3	870	9.3
Female	2746	22.7	330	12
Age				
11 to 13	760	6.3	52	6.8
14 to 17	4010	33.1	396	9.9
18 to 21	7347	60.6	752	10.2
Race/Ethnicity				
White	8019	66.2	808	10.1
Black	1278	10.6	78	6.1
Hispanic	1603	13.2	189	11.8
Asian/Pacific Islander	516	4.3	42	8.1
American Indian	369	3.1	48	13
Multiracial	282	2.3	33	11.7
Unknown	50	0.4	*	*
Sexual Orientation				
Straight	972	8	-	-
Sexual minority orientation	228	1.9	-	-
Unknown	10917	90.1	-	-

*Indicates data are suppressed when counts < 10.

Please see the online supplement for a complete R markdown providing complete annotated data preparation steps and analyses.

Results

There were 12,117 youth suicide decedents in the study sample with ages ranging from 11 to 21 years old. As shown in Table 1, the sample was 77.3% male. The racial/ethnic composition was 66.2% White/non-Hispanic, 13.2% Hispanic, and 10.6% Black/non-Hispanic, with the remaining racial/ethnic groups individually comprising less than 5% of the sample.

9.9% (N=1,200) of the sample was successfully coded for sexual orientation. 19.0% of these 1,200 were sexual minorities.

The first approach that we considered was to model sexual orientation data as missing completely at random. Under this assumption, a researcher can simply restrict analyses to cases with non-missing sexuality information, which in this case corresponds to an estimated 19% rate of sexual minority status among youth suicide decedents.

However, multilevel modeling suggested that sexual orientation information was not missing completely at random. After controlling for other demographics, females were 38% (95%CI:18-61%) more likely to be coded for sexuality than males. Each additional year of age was associated with a 6% (95%CI:3-9%) increased likelihood of known sexual orientation. Black people (95%CI:12-48%) and Asian/Pacific Islanders (95%CI:1-54%) were both 32% less likely to be coded for sexuality than White people (Table 2). There was also substantial variation in rates of coding for sexuality by state (Figure 1), with raw rates ranging from less than 1% in several states (including California and New York) to approximately 50% in Wisconsin.

Given such strong demographic patterns, researchers might consider a second possibility, which is that the data are missing at random conditional upon observed variables. In such

Table 2. Adjusted Odds Ratios (AOR) of coding for sexuality, controlling for state and other demographics.

Variable	AOR	95%CI
Age, years	1.06	1.03-1.09*
Female (ref. Male)	1.38	1.18-1.61*
Race (ref. White)		
Hispanic	1.11	0.91-1.36
Black	0.68	0.52-0.88*
Asian/Pacific Islander	0.68	0.46-0.99*
American Indian	0.96	0.66-1.40
Multiracial	0.87	0.56-1.35
Unknown	0.42	0.10-1.81

*p<0.05. AORs are odds ratios for each variable after adjusting for all other variables reported in this table, as well as location of death.

a scenario, researchers would typically use imputation to estimate the "true" rate of sexual minority status. After imputation based on observed demographics, we find that the adjusted estimated rate of sexual minority status would be 44% (95%PPI:39-49%), as opposed to the 19% that would be estimated by simply restricting calculations to people with known sexual orientation. The adjusted rate would thus be more than 2 times larger (95%CI:2.04-2.5) than the rate that would be estimated if the data were assumed to be missing completely at random.

A third possibility is that the data are not missing at random. If this is the case, then both the naïve estimate and the demographically adjusted estimate are inappropriate. While we do not propose a specific statistical adjustments for such a scenario because nonignorable nonresponse models are often unidentifiable and at best require strong modeling assumptions and/or the inclusion of additional prior information (15), we find this third possibility to be the most plausible and describe it in detail later in the Discussion.

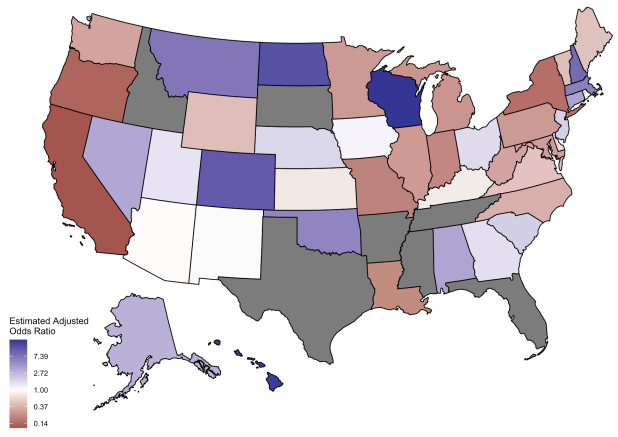

Fig. 1. Adjusted odds of coding for sexual orientation, by state. A value of one (1) corresponds to the average of the dataset after covarying for age, sex, and race. States that do not participate in NVDRS are colored grey.

Discussion

The current study finds that approximately 90% of youths who die by suicide have unknown sexuality at time of death, and that likelihood of known sexuality differs strongly by demographics.

Researchers attempting to derive conclusions about sexuality and youth suicide must choose between non-trivial assumptions in order to determine next steps. One potential assumption would be that sexual orientation is missing completely at random, i.e., that suicide decedents with known versus unknown sexuality are statistically indistinguishable. This appears to be the most standard approach taken in the literature, and, as applied to our 2015-2019 NVRDS dataset, would yield an estimated 19% rate of sexual minority status among youth suicide decedents, higher than the approximate 8% of young people in the United States who identified as LGBT over the same time period (10).

An alternative assumption that researchers might make would be to approach sexual orientation as missing at random after conditioning upon observed variables. Under such an assumption, researchers would be able to make valid inferences about suicide and sexuality after statistical adjustment (14), with our own modeling suggesting that adjustment for demographics would yield a more than two-fold increase in estimated rates of sexual minority status as compared to the estimate derived from simply restricting analyses to people with available data. Such large differences could be attributable to a range of demographic-related factors (e.g., variation in the acceptability of disclosure by group, greater self-knowledge of sexuality as people get older, etc.).

However, we do not find either the assumption that data are missing at random or that they are missing completely at random to be plausible. Instead, it seems overwhelmingly likely that rates of sexual minority status differ systematically between youth with missing versus non-missing orientation information even after accounting for all other observed variables. For example, given two children with the same age, gender, race, and location, a child who identifies as straight may feel more comfortable divulging their sexuality than a child who identifies as gay. This effect would cause relative underestimates in the rate of sexual minority status among people who die by suicide, because cases with missing sexuality information would be more likely to be LGBT+. Of course, other simultaneous effects are also plausible. For example, given two children with otherwise identical demographics, a child identifying as LGBT+ may be more likely to have their sexuality recorded by other people (such as police or coroners) for being noteworthy, and it seems reasonable to suppose that this effect could be particularly strong for younger children who are less likely to have known sexuality in general. Such an effect would lead to relative overestimates in the rate of sexual minority status among people who die by suicide because those with missing sexuality data would be more likely to be straight, and these overestimates would be stronger for younger people, thus illustrating how even research on variables obliquely related to sexuality may be impacted by nonignorable missing data (e.g., producing phantom findings about relationships between age and sexuality among youth suicide decedents). It is difficult to weigh these scenarios or make quantitative guesses about the plausible sizes of such potential effects. In general, if it is true that missing sexuality information is di-

rectly correlated with sexuality in these (or other) ways, then the data are not missing at random and statistical modeling cannot straightforwardly be used to correct any resulting bias (14, 15).

Our results are technically limited to the NVDRS, however, the NVDRS has a relatively comprehensive system for assessing sexuality, and we are concerned that bias in estimates of sexual orientation would be more pronounced in other public health datasets. Bias may be further compounded when attempting intersectional research on multiply marginalized sexual minority groups (16).

Importantly, we do not intend to suggest that all research on suicide risk and sexuality is futile. Our results have no direct bearing on studies with very different methodologies, such as research that follows people prospectively. For example, Feigelman et al. (17) carried out a prospective large-scale survey of living adolescents and adults and examined rates of completed suicide among participants. They measured an increased risk of suicide for Gay/Bisexual individuals of both sexes, although there was substantial uncertainty in their estimates and only the odds ratio for females reached statistical significance. Of course, studies of this kind may suffer their own methodological issues as noted by the authors.

Additionally, there may be considerable room for improvement in the ascertainment of sexual orientation information at time of death. The NVDRS is consistent in coding sexuality based on the law enforcement, coroner, and/or medical examiner reports provided by participating states (13), but these states have quite variable methodologies for producing those reports, with (e.g.) not all coroners or medical examiners systematically performing any psychological autopsy at all (18). Our present study accordingly found substantial variability in missingness of sexual orientation by state, and thus room for improvement at the state level (although, as can be seen in page 18 of the Online Supplement, there are substantial issues with missingness even after accounting for differences attributable to state-level variation).

Overall, our findings should serve to reinforce researchers and policy-makers in being cautious when considering posthumously derived estimates of suicide risk by sexuality, particularly for youth who are much less likely than adults to have disclosed their sexuality to friends or family (4). If it is assumed that sexual orientation information is missing at random, then estimates would require potentially large (>2-fold) adjustments to account for demographic patterns in missingness. Such adjustments have not even been attempted by any existing publication that we are aware of. Alternatively, if sexuality is assumed not to be missing at random (we find this possibility to be overwhelmingly likely) then such adjustments would not be warranted but nor would naive unadjusted estimates, thus yielding even greater uncertainty.

1. Haas AP, Lane AD, Blosnich JR, Butcher BA, Mortali MG (2019) Collecting sexual orientation and gender identity information at death. *American Journal of Public Health* 109(2):255–259.
2. Haas AP, Lane A (2015) Collecting sexual orientation and gender identity data in suicide and other violent deaths: A step towards identifying and addressing lgbt mortality disparities. *LGBT Health* 2(1):84–87. PMID: 26790023 PMCID: PMC4713015.
3. Plöderl M, et al. (2013) Suicide risk and sexual orientation: A critical review. *Archives of Sexual Behavior* 42(5):715–727.
4. Bishop MD, Fish JN, Hammack PL, Russell ST (2020) Sexual identity development milestones in three generations of sexual minority people: A national probability sample. *Developmental psychology* 56(11):2177–2193. PMID: 32833471 PMCID: PMC8216084.
5. Gelman A, Loken E (2014) The statistical crisis in science: data-dependent analysis–a "garden of forking paths"–explains why many statistically significant comparisons don't hold up. *American Scientist* 102(6):460–466.
6. Silberzahn R, et al. (2018) Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science* 1(3):337–356.
7. Groenwold RHH, Dekkers OM (2020) Missing data: the impact of what is not there. *European Journal of Endocrinology* 183(4):E7–E9. PMID: 32688333.
8. Patten M, Carmichael H, Moore A, Velopulos C (2022) Circumstances of suicide among lesbian, gay, bisexual and transgender individuals. *Journal of Surgical Research* 270:522–529.
9. Ream GL (2019) What's unique about lesbian, gay, bisexual, and transgender (lgbt) youth and young adult suicides? findings from the national violent death reporting system. *Journal of Adolescent Health* 64(5):602–607.
10. Newport (2018) In U.S., estimate of lgbt population rises to 4.5%. *Gallup.com*. section: Politics.
11. Lyons BH, et al. (2019) Suicides among lesbian and gay male individuals: Findings from the national violent death reporting system. *American Journal of Preventive Medicine* 56(4):512–521.
12. Remafedi G (1999) Suicide and sexual orientation: Nearing the end of controversy? *Archives of General Psychiatry* 56(10):885–886.
13. (2021) National violent death reporting system web coding manual, Technical report.
14. Bhaskaran K, Smeeth L (2014) What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology* 43(4):1336–1339. PMID: 24706730 PMCID: PMC4121561.
15. Little R, Rubin D (2002) *Statistical Analysis with Missing Data*, Wiley series in probability and statistics. (Wiley-Interscience). [Online; accessed 2022-12-30].
16. Cyrus K (2017) Multiple minorities as multiply marginalized: Applying the minority stress theory to lgbtq people of color. *Journal of Gay Lesbian Mental Health* 21(3):194–202. publisher: Routledge.
17. Feigelman W, Plöderl M, Rosen Z, Cerel J (2019) Research note on whether sexual minority individuals are over-represented among suicide's casualties. *Crisis* 41:1–4.
18. Rockett IRH, et al. (2018) Method overtness, forensic autopsy, and the evidentiary suicide note: A multilevel national violent death reporting system analysis. *PLOS ONE* 13(5):e0197805. publisher: Public Library of Science.