Assessing the Impact of Missing Data on YRBS-Based Adolescent Suicide Research

Catalina Cañizares, PhD[1] Mark J. Macgowan, PhD[1] Raymond Balise, PhD[1]

[1] *Robert Stempel College of Public Health and Social Work, Florida International University*

[2] *Public Health Sciences, University of Miami*

Introduction

Adolescent suicide attempts (SA) are a critical public health issue in the United States, with rising rates over the past 15 years. According to CDC data from 2021, approximately 10% of adolescents in grades 9 to 12 have attempted suicide in the past year. This alarming trend underscores the urgent need to understand this phenomenon and develop effective prevention strategies to mitigate the tragic loss of young lives.

The Youth Risk Behavior Survey (YRBS) serves as a crucial data source for understanding adolescent health risk behaviors, particularly suicidal behaviors. However, the YRBS faces significant challenges due to missing data, especially regarding SA. Studies examining this issue have revealed that adolescents who self-identified as non-Hispanic Black had more than threefold higher odds of having a missing response to SA compared to their non-Hispanic White counterparts (Baiden, 2023). Additionally, adolescent males and those questioning their sexual identity also had higher odds of missing responses for SA (Baiden,2023).

Given these patterns of missingness, it is imperative that researchers address these missing responses statistically. However, few studies have investigated whether the missing data on SA threatens the validity of published research. Moreover, no systematic reviews have evaluated the methods used to address missing data when utilizing the YRBS.

Objective

This study aims to review the methods used in the published literature to address the missing data in the SA variable within the YRBS. Subsequently, the study will apply these methods to assess their impact on the results of a logistic regression analysis. By comparing the outcomes derived from different imputation techniques, this research seeks to determine the extent to which the choice of method influences the findings, thereby providing insights into best practices for handling missing data in studies of adolescent suicidal behaviors.

Methods

Literature Review

A search of major online databases, including PsycINFO, Google Scholar, and PubMed, was conducted using the terms “suicide attempts” and “YRBS”. Inclusion criteria required that studies utilized the YRBS as a data source and included one of the suicide related items as an outcome measure. No exclusion criteria were applied based on the year of the study or year of the YRBS cycle used.

Data Analysis

The analysis used the 2017 YRBS dataset. A logistic regression was conducted with 21 predictors encompassing demographic variables, traumatic events (bullying, cyberbullying, forced sexual relationships), substance use (alcohol, vaping, tobacco, hard drugs), risky behaviors (gun carrying, physical fights), and protective factors (sleep and sports). The outcome variable was SA. Both predictors and the outcome variable were based on binary transformations of each variable as provided by the CDC, except for the age variable.

Four imputation methods were employed in the analysis. The listwise deletion method included only cases with complete responses. Mode and mean imputation filled in missing categorical variables with the mode and continuous variables with the mean. Multiple Imputation using Chained Equations (MICE) was conducted with 20 iterations. Additionally, random forest and bagged imputations were performed. All imputations were carried out using RStudio, utilizing the MICE package and the Tidymodels framework.

Results

For the literature review, a total of 17 articles were examined. Eight articles used list-wise deletion to handle missing data, six likely eliminated instances with missing values based on the reported sample sizes, and three employed Multiple Imputation by Chained Equations (MICE).

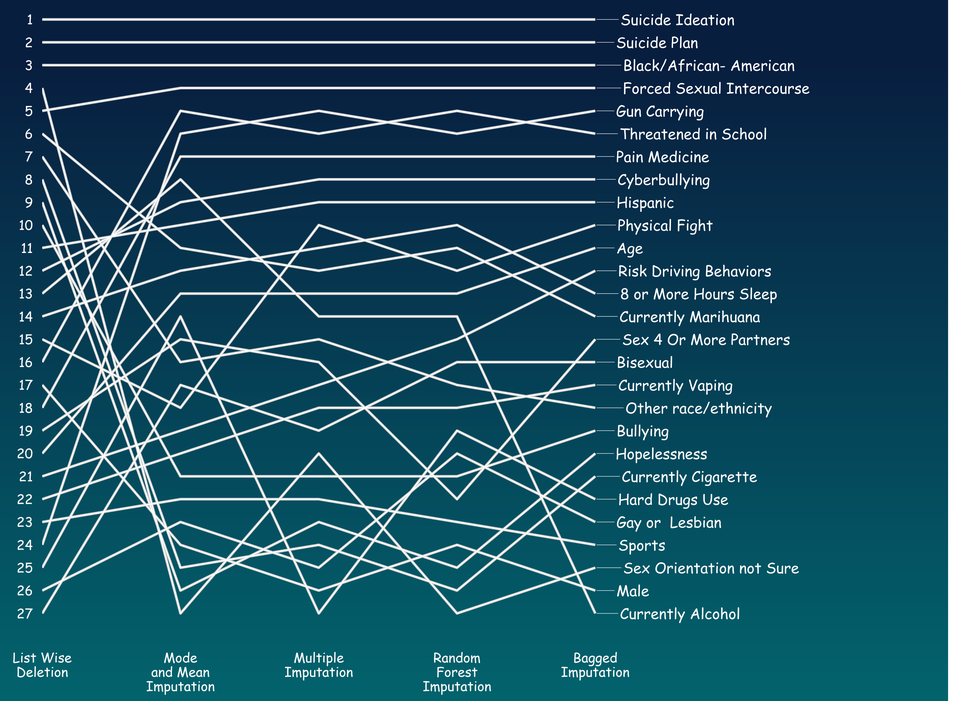
The **method used to address missingness** in the YRBS data **influences** the **effect sizes** of the predictors in the model, highlighting the **sensitivity of the results** to the **choice of imputation** technique.

The choice of **imputation method** has a **minimal impact** on the behavior of the odds ratios for predictors with **large effect sizes,...**

However, as the **effect size decreases,** the variation in effect sizes across different imputation methods becomes **more pronounced.**

The **variability** in effect sizes **across** different **imputation methods** suggests that researchers should conduct **additional analyses to detect patterns** of missingness.

Figure 1, Ranking of Effect Sizes by Imputation Method



FIU

For the logistic regression analysis, the sample included 12,551 adolescents, with 7.4% (n = 925) reporting suicide attempts (SA). However, 28% of the sample had missing data for this item. 9.3% were female and 5.1% were male. The highest proportion of SA was among Black/African American adolescents (9.8%), followed by Hispanic (8.2%), and non-Hispanic White (6.1%). Bisexual students had the highest rate of attempts (24%), followed by homosexual (19%) and heterosexual students (5.4%).

Figure 1 ranks effect sizes from highest to lowest. The results indicate that certain predictors—suicide ideation, planning, race, and being forced into sexual relationships—consistently increase the likelihood of suicide attempts, regardless of the imputation method used. These predictors do not vary in their rank of effect size across methods. However, predictors ranked 5th and below exhibit significant variability depending on the imputation method, particularly when comparing listwise deletion with the other methods.

Figure 2 displays Odds Ratios for each predictor, varying by imputation method, with multiple imputation yielding the highest effect sizes. Random forest, bagged imputation, and mode and mean imputation produced similar effect sizes.

Figure 2, Effect Sizes for Logistic Regression of Suicide Attempts by Imputation Method

	List Wise Deletion	Mode and Mean Imputation	Multiple Imputation	Random Forest Imputation	Bagged Imputation
Threatened in School	1.19	1.42	1.73	1.35	1.35
Suicide Plan	2.64	2.58	5.61	2.56	2.53
Suicide Ideation	4.4	3.64	14.98	3.77	3.7
Sports	1.09	1.06	1.12	1.06	1.04
Sex Orientation not Sure	1.55	1.02	1.14	1.01	1.04
Sex 4 Or More Partners	1.16	1.11	1.25	1.06	1.11
Risk Driving Behaviors	1.13	1.08	1.19	1.12	1.14
Physical Fight	1.22	1.1	1.46	1.15	1.17
Pain Medicine	1.17	1.31	1.52	1.25	1.22
Other race/ethnicity	1.35	1.11	1.25	1.1	1.08
Male	1.18	1.04	1.1	1.04	1.04
Hopelessness	1.31	1.04	1.12	1.04	1.06
Hispanic	1.28	1.22	1.48	1.2	1.2
Hard Drugs Use	1.05	1.13	1.02	1.09	1.05
Gun Carrying	1.08	1.32	1.84	1.33	1.45
Gay or Lesbian	1.03	1.05	1.1	1.07	1.05
Forced Sexual Intercourse	1.4	1.61	2.16	1.57	1.59
Cyberbullying	1.27	1.24	1.5	1.23	1.22
Currently Vaping	1.12	1.08	1.18	1.09	1.08
Currently Marihuana	1.4	1.19	1.29	1.17	1.13
Currently Cigarette	1.34	1.04	1.11	1.04	1.06
Currently Alcohol	1.23	1.26	1.27	1.13	1.04
Bullying	1.29	1.06	1.13	1.06	1.07
Black/African- American	1.56	1.69	2.93	1.73	1.7
Bisexual	1.02	1.11	1.15	1.1	1.09
Age	1.14	1.15	1.29	1.15	1.15
8 or More Hours Sleep	1.22	1.18	1.35	1.19	1.14

Discussion

This study aimed to evaluate methods for addressing missing data in studies using the YRBS and SA. The literature review revealed a prevalent use of listwise deletion or a lack of explicit mention or handling of missing data. The widespread use of listwise deletion without considering the randomness of missingness is concerning, as patterns in missing data may correlate with race and sexual orientation (Baiden,2023). This alignment with findings from other studies discourages the analysis of complete cases, as we cannot assume the data is missing completely at random (Rubin, 1976).

The logistic regression analysis supports the notion that ideation, planning, and racial factors are pivotal in SA, confirming results from past research (Franklin et al., 2017). However, the variability in effect sizes across imputation methods underscores the importance of consistently addressing missing data in suicide research using the YRBS. This consistency is crucial to improving external validity and reducing replicability issues.

This study’s limitations include its focus solely on the 2017 YRBS data and a restricted literature review. Future research should expand the scope of literature reviews and examine the randomness of missing data across different YRBS years to strengthen the evidence base for prevention programs.

FIU