Supplemental Digital Content 1: Candidate variables for multiple imputation model

Several variables that met the criteria for inclusion in the multiple imputation model were removed from the final model. The data set included over 2.5 million cases. Four variables were considered essential: disease, age, sex, and geographic area. These variables are frequently included in analyses, with race and Hispanic ethnicity, as risk factors for STIs. There were 57 geographic area categories, five disease categories and two sex categories. Because of the size of the data set and the number of categories among the essential variables, the MICE procedure took several hours to complete. Due to memory allocation problems, good candidates for inclusion in the imputation model were excluded from the final model. This document describes the criteria for inclusion in the multiple imputation model and the candidate variable that were excluded from the final model.

Predictors in analyses that use multiply imputed race and Hispanic ethnicity data

To select variables that should be included in the multiple imputation model, we must consider what analyses will include race and Hispanic ethnicity as a predictor. For example, if an analysis to compare the differences in rates of gonorrhea between married and unmarried adults was adjusted for race and Hispanic ethnicity, the multiple imputation model would need to include the marital status of the gonorrhea cases. Otherwise, the imputed values wouldn't preserve the relationship between marital status and race and Hispanic ethnicity.

Predictors of race and Hispanic ethnicity

Variables that help predict race and Hispanic ethnicity should be included in the imputation model. To evaluate potential predictors of race and Hispanic ethnicity for inclusion in the

multiple imputation model, we used chi-square tests. Categorical predictors were considered for inclusion in the multiple imputation model if the distribution of cases in race and ethnicity categories differed significantly (with type I error equal to 0.05) across categories of the predictor.

Predictors of missingness

We created a binary variable that indicated if race and Hispanic ethnicity was missing or recorded. To evaluate variables that were predictors of nonresponse for inclusion in the multiple imputation model, we used chi-square tests. Because of the large number of cases in the data set, even when differences between observed and expected cell frequencies were relatively small, chi-square tests indicated that the categorical variable was significantly associated with nonresponse. Therefore, variables with the lowest p-values were prioritized for inclusion in the multiple imputation model.

Description of variables

Rates of STIs vary by sex of sex partners. Research studies using data from the National Notifiable Diseases Surveillance System (NNDSS) often include sex of sex partner as a risk factor for STIs. NNDSS includes several variables to collect data on sexual behavior. Current sex, coded as male or female, is recorded for over 99% of cases. A variable for coding sexual partners has four possible responses: male, female, both, and refused. Two additional variables with yes/no responses ask if the patient had sex with men and if the patient had sex with women. For men, the responses to these variables are combined to create three categories: men who have sex with men only, men who have sex with women only, men who have sex with

men and women. Unfortunately, 87.7% of male cases in 2019 did not have complete information for the sexual behavior variables. Sexual behavior categories were removed from the multiple imputation model due to the amount of missing data and the time required to impute their missing values.

The National Center for Health Statistics (NCHS) classifies counties into six categories based on population density[1]. Some reports based on NNDSS data, like the gonorrhea objective for Health People 2030, compare cases per 100,000 population by metropolitan or non-metropolitan areas[2]. If the county of residence is not recorded, the NCHS urban-rural classification scheme can't be applied. County data was missing for less than one percent of NNDSS cases in 2019. The urban-rural categories were included in preliminary imputation models that did not provide results due to memory allocation problems.

HIV status is recorded for some STI case reports. The HIV status variable has five possible values: positive, negative, equivocal, refused to answer, and did not ask. Cases with positive or negative HIV tests had a lower percentage of nonresponse for the race and Hispanic ethnicity data but 88.5% of cases did not have a recorded HIV status. HIV status was also removed from the multiple imputation procedure because of the high percentage of missing data.

An NNDSS variable classifies the type of facility that performed the testing or diagnosis of the STI. Trends in the number of cases reported to different facility types are displayed in annual reports published by CDC. The variable has over 20 possible values and 15.2% of cases had unknown or invalid response types. Fourteen binary variables were created to check for associations between the most common diagnosis or testing facilities and race and Hispanic

ethnicity or nonresponse. These binary variables included HIV counseling and testing sites, STD clinics, family planning facilities, other health department clinics, private physicians or HMOs, hospitals or emergency rooms, correctional facilities, laboratories, prenatal clinics, school-based clinics, other hospitals, the Indian Health Service, military facilities, and other facilities. Five of the fourteen binary variables were included in preliminary imputation models. Cases that were diagnosed or tested at Indian Health Service facilities were likely to be associated with American Indian or Alaska Native race. Four diagnosing and testing facilities were associated with nonresponse: STD clinics, family planning facilities, other hospitals, and other facilities. Cases diagnosed or tested at other facilities were more likely to have missing race and Hispanic ethnicity data compared to cases that were not diagnosed or tested at other facilities. Cases diagnosed or tested at STD clinics, family planning facilities, or other hospitals were less likely to have missing race and Hispanic ethnicity data compared to cases not tested or diagnosed in those locations.

A case status variable classifies each record as confirmed, probable, or suspect. This variable was complete for over 99% of NNDSS STI case notifications in 2019. Based on Pearson chi-square tests, case status was a good predictor of nonresponse. Over 98% of cases had confirmed or probable case status. A Pearson chi-square test also indicated that the distribution of race and Hispanic ethnicity was significantly different between confirmed and probable cases. Although case status was a candidate variable for inclusion in the multiple imputation model, most cases (95.6%) were confirmed. The output from multiple imputation model would not change drastically if case status was excluded. We chose not to include case status to simplify the model and avoid memory allocation issues.

REFERENCES

1. Centers for Disease Control and Prevention. Data Access - Urban Rural Classification Scheme for Counties. Accessed November 14, 2022. https://www.cdc.gov/nchs/data_access/urban_rural.htm

2. Office of Disease Prevention and Health Promotion. Reduce Gonorrhea Rates in Male Adolescents and Young Men — STI-02. Accessed October 17, 2023. https://health.gov/healthypeople/objectives-and-data/browse-objectives/sexually-transmitted-infections/reduce-gonorrhea-rates-male-adolescents-and-young-men-sti-02

Supplemental Digital Content 2: Creation of subsets for model evaluation and methods for estimating variance and confidence intervals

Let N represent a population from a specific project area with age a and sex s. Each race/Hispanic ethnicity category in N has a population of size n_i where i indicates the race/Hispanic ethnicity category. There are Y cases of a specific nationally notifiable sexually transmitted infection that are reported from the project area for one year. Among the Y cases, X are reported with race and ethnicity data. The number of cases with race/Hispanic ethnicity category i is represented by x_i. We impute race/Hispanic ethnicity for the $Y - X$ cases with missing values. After imputation, we sample data for X individuals, with replacement, and estimate the incidence rates and their confidence intervals by race/Hispanic ethnicity category. The sample will include approximately $100 \times (Y - X)/Y$ percent of cases with imputed race/Hispanic ethnicity data. The incidence rates that are calculated from the multiply imputed data are represented by $\hat{r}_i$ with i indicating the race/Hispanic ethnicity category.

The number of cases from imputation l with race/Hispanic ethnicity category i is represented by $\hat{x}_{il}$. The natural log (ln) of the cases and rates are used to avoid negative values in the lower confidence intervals. With m representing the number of imputations, $ln(\hat{r}_i)$ are calculated as

$$\frac{1}{m}\sum_{l=1}^{m} ln\left(\frac{\hat{x}_{il}}{n_i}\right)$$

The variance of $ln\left(\frac{\hat{x}_{il}}{n_i}\right)$ is estimated in Appendix 1 below and is a component of the within imputation variance.

The confidence intervals for $ln(\hat{r}_i)$ are calculated using the following parameters: $\overline{U}_i$ = the within imputation variance, B_i = the between imputation variance, T_i = the total variance, and λ_i = the variation attributable to the missing data.

$$\overline{U}_i = \frac{1}{m}\sum_{l=1}^{m}\left(\frac{1}{\hat{x}_{il}}\right)$$

$$B_i = \frac{1}{m-1}\sum_{l-1}^{m}\left(ln\left(\frac{\hat{x}_{il}}{n_i}\right) - ln(\hat{r}_i)\right)^2$$

$$T_i = \overline{U}_i + B_i + \frac{B_i}{m}$$

$$\lambda_i = \frac{B_i + B_i/m}{T_i}$$

The 95% confidence intervals for $ln(\hat{r}_i)$ are calculated using a t distribution with v_i degrees of freedom.

$$ln(\hat{r}_i) \pm t_{v_i,0.975}\sqrt{T_i}$$

$$v_i = \frac{m-1}{\lambda_i^2}$$

We exponentiate the confidence intervals and compare them to the rates calculated from the unimputed case counts, x_i/n_i. We expect that x_i/n_i will be contained within the exponentiated 95% confidence intervals for $ln(\hat{r}_i)$.

To avoid dividing by zero or taking the natural log of zero, $\hat{x}_{il}$ is replaced by 1 in the formulas above if no cases from imputation l had race/Hispanic ethnicity category i. If there is no

variance between imputations, the normal distribution replaces the t distribution with v_i degrees of freedom.

Appendix 1: Delta Method

We will estimate the variance of variance of $ln\left(\frac{\hat{x}_{il}}{n_i}\right)$ using the delta method[1].

Let the rate $\hat{r}_{il} = \frac{\hat{x}_{il}}{n_i}$ have expected value $r_i = \frac{\mu_i}{n_i}$ and variance $\sigma_i^2 = \frac{u_i}{n_i^2}$.

Then, for large n_i, $\hat{r}_{il} - r_i \sim N(0, \sigma^2)$.

A function of $\hat{r}_{il}$ is denoted by $f(\hat{r}_{il})$.

Let $f(\hat{r}_{il}) = ln(\hat{r}_{il})$.

According to the delta method, the variance of $f(\hat{r}_{il})$ is approximately $\sigma_i^2[f'(r_i)]^2$

Using the delta method, the variance of $ln(\hat{r}_{il}) \cong \frac{u_i}{n_i^2}\frac{n_i^2}{u_i^2} = \frac{1}{u_i} \cong \frac{1}{\hat{x}_{il}}$.

REFERENCES

1. Hosmer DW, Lemeshow S, May S. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. 2nd ed. John Wiley & Sons; 2008:355-358:chap Appendix 1. *Wiley Series in Probability and Statistics*.

Supplemental Digital Content 3: Percentage of cases by race and Hispanic ethnicity category in the multiply imputed and original data sets

The following tables display the percentage of cases in each race and Hispanic ethnicity category for the cases with known race and Hispanic ethnicity compared to the sample of cases with multiply imputed race and Hispanic ethnicity. For each subset described in Table 2 of the main manuscript, race and Hispanic ethnicity data was multiply imputed. See Supplemental Digital Content 2 for a description of the subset selection. Let X represent the number of cases in the subset with reported race and Hispanic ethnicity data. The number of cases from imputation l with race/Hispanic ethnicity category i is represented by $\hat{x}_{il}$. The average number of cases in race/Hispanic ethnicity category i across m = 15 imputations is

$$\bar{x}_i = \frac{1}{m}\sum_{l=1}^{m} \hat{x}_{il}$$

The percentages of cases in each race/Hispanic ethnicity category, shown in the last column of the following tables are calculated as

$$100 * \frac{\bar{x}_i}{X}$$

Table 1. Percentage of cases in each race and ethnicity category for 2,127 reported cases of chlamydia in men aged 30-34 years in Georgia, 2019

Race/Hispanic Ethnicity	Reported Cases	Imputed Cases[a]
NH-American Indian/Alaska Native	0.1	0.1
NH-Asian	0.5	0.7
NH-Black/African American	74.2	73.4
NH-Native Hawaiian/Pacific Islander	0.1	0.0
NH-White	18.9	19.5
NH-Multiracial	0.2	0.3
Hispanic/Latino	6.0	6.0

[a]See Supplemental Digital Content 2 for a description of the subset selection. Among 2,127 sampled cases, 338 (15.9%) had missing race and Hispanic ethnicity.

Table 2. Percentage of cases in each race and ethnicity category for 2,995 reported cases of chlamydia in women aged 20-24 years in Kansas, 2019

Race/Hispanic Ethnicity	Reported Cases	Imputed Cases[a]
NH-American Indian/Alaska Native	1.5	1.1
NH-Asian	2.0	2.0
NH-Black/African American	19.6	19.9
NH-Native Hawaiian/Pacific Islander	0.3	0.2
NH-White	58.6	56.0
NH-Multiracial	0.0	1.0
Hispanic/Latino	18.0	19.8

[a]See Supplemental Digital Content 2 for a description of the subset selection. Among 2,995 sampled cases, 834 (27.8%) had missing race and Hispanic ethnicity.

Table 3. Percentage of cases in each race and ethnicity category for 665 reported cases of gonorrhea in women aged 15-24 years in Massachusetts, 2019

Race/Hispanic Ethnicity	Reported Cases	Imputed Cases[a]
NH-American Indian/Alaska Native	0.2	0.4
NH-Asian	2.3	2.3
NH-Black/African American	37.9	33.3
NH-Native Hawaiian/Pacific Islander	0.2	0.4
NH-White	32.2	39.9
NH-Multiracial	0.8	1.5
Hispanic/Latino	26.6	22.1

[a]See Supplemental Digital Content 2 for a description of the subset selection. Among 665 sampled cases, 248 (37.3%) had missing race and Hispanic ethnicity.

Table 4. Percentage of cases in each race and ethnicity category for 269 reported cases of primary and secondary syphilis in men aged 25-39 years in Nevada, 2019

Race/Hispanic Ethnicity	Reported Cases	Imputed Cases[a]
NH-American Indian/Alaska Native	0.0	1.0
NH-Asian	1.9	1.7
NH-Black/African American	22.7	24.1
NH-Native Hawaiian/Pacific Islander	0.0	0.4
NH-White	28.6	29.2
NH-Multiracial	0.7	0.6
Hispanic/Latino	46.1	43.0

[a]See Supplemental Digital Content 2 for a description of the subset selection. Among 269 sampled cases, 59 (21.9%) had missing race and Hispanic ethnicity.

Table 5. Percentage of cases in each race and ethnicity category for 2,689 reported cases of chlamydia in women aged 25-29 years in Washington, 2019

Race/Hispanic Ethnicity	Reported Cases	Imputed Cases[a]
NH-American Indian/Alaska Native	4.1	3.5
NH-Asian	3.9	3.3
NH-Black/African American	8.5	9.8
NH-Native Hawaiian/Pacific Islander	3.3	2.6
NH-White	50.7	51.2
NH-Multiracial	3.8	3.4
Hispanic/Latino	25.7	26.2

[a]See Supplemental Digital Content 2 for a description of the subset selection. Among 2,689 sampled cases, 941 (35.0%) had missing race and Hispanic ethnicity.