HHS Public Access

Author manuscript

Sex Transm Dis. Author manuscript; available in PMC 2024 November 14.

Published in final edited form as:

Sex Transm Dis. 2024 November 01; 51(11): 719–727. doi:10.1097/OLQ.0000000000002047.

Multiple imputation of race and Hispanic ethnicity in national surveillance data for chlamydia, gonorrhea, and syphilis

Tracy Pondo, MSPH,

Elizabeth Torrone, PhD,

Melissa Pagaoa, MPH

Division of STD Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA

Abstract

Background—Disease burden of sexually transmitted infections such as chlamydia, gonorrhea, and syphilis is often compared across age categories, sex categories, and race and ethnicity categories. Missing data may prevent researchers from accurately characterizing health disparities between populations. This article describes the methods used to impute race and Hispanic ethnicity in a large national surveillance data set.

Methods—All US cases of chlamydia, gonorrhea, and syphilis (excluding congenital syphilis) reported through the National Notifiable Diseases Surveillance System (NNDSS) from the year 2019 were included in the analyses. We used fully conditional specification to impute missing race and Hispanic ethnicity data. After imputation, reported case rates were calculated, by disease, for each race and Hispanic ethnicity category using Vintage 2019 Population and Housing Unit Estimates from the US Census. We then used case counts from subsets that contained only complete race and Hispanic ethnicity information to investigate if the confidence intervals from the multiply imputed data included the observed number of cases in each race and Hispanic ethnicity category.

Results—Among the 2,553,038 cases reported in 2019, race and Hispanic ethnicity were multiply imputed for 9% of syphilis cases, 22% of gonorrhea cases and 33% of chlamydia cases. In the subset analyses, every non-zero rate of reported cases was contained within the confidence intervals that were calculated from multiply imputed data.

Conclusions—Confidence intervals that account for the uncertainty of the predictions are an advantage of multiple imputation over complete-case analysis because a realistic variance estimate allows for valid hypothesis testing results.

Short summary:

Correspondence: Tracy Pondo, MSPH, Division of STD Prevention, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, MS H24-4, Atlanta, GA 30329, Fax: 404-315-2000, tpondo@cdc.gov.

Conflict of interest and sources of funding: None declared.

OMB/CDC disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Missing race and ethnicity data are multiply imputed in a large national surveillance data set. The methods are described to assist researchers in applying multiple imputation procedures to similar studies.

Keywords

Multiple imputation; surveillance data; variance estimation

The Centers for Disease Control and Prevention (CDC) collects and maintains case notification data for several nationally notifiable sexually transmitted infections (STIs) including chlamydia, gonorrhea, and syphilis through the National Notifiable Diseases Surveillance System (NNDSS)[1]. Annual reported case rates of these infections are used to monitor progress towards disease prevention goals including the reduction of racial and ethnic disparities in disease burden[2,3].

Evaluating disparities between population groups requires collecting and analyzing the demographic characteristics of infected individuals. Reported case rates of chlamydia, gonorrhea, and syphilis are often compared across age categories, sex categories, and race and Hispanic ethnicity categories. Reliable comparisons require accurate and complete data collection; however, because of the large burden of reported STIs in the United States (e.g., in 2019 over 1.8 million cases of chlamydia were reported), many cases are not able to be investigated fully by local public health staff[4]. Consequently, case reports may only have the information contained on the laboratory test result. Although age and sex are usually complete, race and Hispanic ethnicity data are often missing. A common approach to missing data is complete-case analysis; cases with missing data are excluded. If the data are missing completely at random, complete-case analysis is equivalent to choosing a random sample of the cases for analysis. With complete-case analysis, there is a loss of sample size but point estimates are still valid. If the data are not missing completely at random, in other words, if the probability of being missing is not the same for all cases, results from complete-case analyses could be misleading. We propose a multiple imputation strategy as an alternative to excluding cases with missing data when race and Hispanic ethnicity are predictors in analyses using STI case notification data collected through NNDSS.

METHODS

Race and ethnicity coding

In the national STI data collected through NNDSS, cases may be reported with information on both race and Hispanic ethnicity. Since 2007, both race and Hispanic ethnicity have been collected through the following seven binary variables: American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Other Pacific Islander, Other race, and White. This format allows for the selection of both Hispanic ethnicity and a race, as well selection of more than one race per person. For display in national surveillance reports, a combined race/Hispanic ethnicity variable is derived from the seven binary race and ethnicity variables. The derived race/Hispanic ethnicity variable is assigned one of the following seven values: Hispanic, non-Hispanic (NH) Multiracial, NH-

American Indian or Alaska Native, NH-Asian, NH-Black or African American, NH-Native Hawaiian or Other Pacific Islander, and NH-White. Cases reported as Hispanic are classified as Hispanic, regardless of their reported race, and include cases with unknown race. Cases reported as non-Hispanic or of unknown Hispanic ethnicity are considered non-Hispanic and categorized based on available race data. A small number of case notifications are provided to CDC through NNDSS using outdated race-coding categories that do not distinguish NH-Asian from NH-Native Hawaiian or Other Pacific Islander and do not allow for reporting of multiple races. The few cases reported in the legacy "Asian/Pacific Islander" category are re-coded to missing. Therefore, the race and Hispanic ethnicity data from jurisdictions using outdated race-coding categories are likely less reliable than data from jurisdictions using the current race-coding.

Variable selection

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. By including all the variables in models that will be applied to the data after imputation, we maintain the relationships between race and Hispanic ethnicity and the other analysis variables[5-7]. STI case notifications collected through NNDSS are frequently compared by age in years, current sex (coded as male or female), and geographic area of the residence of the person diagnosed with the STI. Therefore, these variables were considered essential to include in the imputation model.

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

Variables related to nonresponse were also evaluated for inclusion in the multiple imputation model. 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 Pearson 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. Details about these variables are described in Supplemental Digital Content 1.

A set of candidate variables for the multiple imputation model was reduced to a final model. Decisions to exclude variables from the final model were based on model complexity and computing time. The data set included over 2.5 million cases and the geographic areas, age, sex, and disease categories alone resulted in a multiple imputation model that required several hours of computation time. Some of the candidate variables had more missing responses than the race and ethnicity data and their inclusion would have further increased the time required to complete the imputations. Although additional variables were good

candidates for inclusion in the imputation model, due to memory allocation problems, they were not included in the final model. In the final imputation model, we limited the predictors to age, sex, disease category, and geographic area. See Supplemental Digital Content 1 for additional details on candidate variables that were removed from the final multiple imputation model.

Geographic area was categorized as the 57 jurisdictions funded for sexually transmitted disease (STD) prevention programs by CDC, including all 50 states, the District of Columbia, and six directly funded cities (Chicago, IL; San Francisco, CA; Los Angeles, CA; Baltimore, MD; Philadelphia, PA New York City, NY). If a state had one or more directly funded cities, the state data excluded the directly-funded city data (e.g., New York state data excludes New York City data).

Sex was reported as binary (male/female). Age was coded in years. Disease included five categories: chlamydia, gonorrhea, primary and secondary syphilis, early non-primary non-secondary syphilis, and unknown duration or late syphilis. The imputation model also included the disease-by-sex interaction term.

Imputation

The imputation procedure was performed using the MICE algorithm in R[5,6,8]. We ran 15 imputations with 10 iterations for convergence of the MICE algorithm.

We experimented with three possible models for predicting race and ethnicity. The default method for categorical variables in MICE is a polytomous logistic regression model. This method did not perform well in areas with sparse data. For example, rates of disease for NH-American Indian or Alaska Native or NH-Native Hawaiian and Other Pacific Islanders were overestimated. For this analysis, sparse data is roughly defined as strata that include less than two percent of cases with known race and Hispanic ethnicity. We next tried using linear discriminant analysis. This improved the prediction of rates in areas with sparse data but the estimates for those rates often had no between imputation variance. We decided on predictive mean matching as the best solution for predicting race/Hispanic ethnicity for categories with sparse data.

In MICE's predictive mean matching algorithm, a set of parameter estimates for predicting race and Hispanic ethnicity is drawn from the posterior distribution of the specified multiple imputation model. A donor is randomly selected from a set of three candidates with observed race and Hispanic ethnicity. The candidate set includes the race and Hispanic ethnicity from observations with predictions from the multiple imputation model that are closest to the prediction from the model generated from the sampled parameter estimates[5].

We also tried two methods for imputing missing sex and age. We originally imputed sex by logistic regression and age by predictive mean matching. We then chose to impute the missing values by randomly sampling from observed values regardless of the values of their other covariates because it greatly improved the speed of the imputations and because we were not concerned with the small number of cases with missing data in these variables (less than one percent).

We also experimented with different numbers of imputations. We originally imputed five data sets. The amount of time to run the imputations and the size of the data sets were the motivation for limiting the output to five imputations. Because the between variance estimates were not large enough with only five imputations, we tested a model with 30 imputations. Some researchers have recommended that the number of imputations be similar to the percentage of cases that are incomplete[7]. The number 30 was based on the percentage of missing race and ethnicity data for chlamydia cases (33%). For the final model, we imputed 15 data sets with 10 iterations for convergence of the MICE algorithm. We chose not to generate more than 15 imputations to limit the amount of storage space that would be required for the output, the amount of time needed to run the imputation model, and the time involved in running an analysis that combines data from multiple imputed data sets.

Analysis of imputed data

After imputation, rates of reported cases were calculated, by disease, for each race and Hispanic ethnicity category using Vintage 2019 Population and Housing Unit Estimates from the US Census[9]. We use the phrase “multiply imputed data” to refer to all cases, those with recorded and those with imputed race and ethnicity. The mean and variance of the rates were calculated using Rubin’s rules for pooling results from multiple data sets[10-12]. The natural log of the cases and rates were used for estimating the mean and variance to avoid negative values in the lower confidence intervals. Estimates were then exponentiated back to their original scale.

Model evaluation

Analyses performed on the data set with multiply imputed values allow for hypothesis testing with variance estimates that account for the uncertainty that results from the multiply imputed, synthetic, values. The multiple imputation process was designed so that relationships that exist in the cases with complete data were preserved in the cases with imputed data. We selected a set of cases with no unknown race and Hispanic ethnicity data and summarized the number of cases in each race and Hispanic ethnicity category. We then sampled, with replacement, a set of cases from the full surveillance data set with similar values of disease, age, geography, and sex. Among the sampled cases, a portion had missing race and Hispanic ethnicity. If the multiple imputation model replicates the relationships that exist in the cases with complete data and accounts for the uncertainty in the synthetic values, then the 95% confidence intervals associated with the sampled cases should contain the number of cases that were summarized from the complete-case data set, of the same size, with the same characteristics regarding disease, geography, age, and sex. Five subsets were selected with no unknown race and Hispanic ethnicity data. The subsets represent the data that would be included in a complete-case analysis. We compared case counts from the complete-case analysis to investigate if the confidence intervals from the multiply imputed data include the observed number of cases in each race and Hispanic ethnicity category. See Supplemental Digital Content 2 for a description of the sampling of cases for the subset analyses.

The following subset analyses were selected from age categories with the highest risk of disease to represent areas with relatively high proportions of missing race and Hispanic

ethnicity from various geographic regions: chlamydia cases among men aged 30-34 years in Georgia, chlamydia cases among women aged 20-24 years in Kansas, gonorrhea cases among women aged 15-24 years in Massachusetts, primary and secondary syphilis cases among men aged 25-34 years in Nevada, and chlamydia cases among women aged 25-29 years in Washington.

After imputation, records that had unknown race and Hispanic ethnicity in the original data set, have an imputed race and Hispanic ethnicity distribution that depends on the distribution of race and Hispanic ethnicity among cases with similar values of age, sex, and geographic area. If records with missing race and Hispanic ethnicity do not have the same distribution of age, sex, and geographic area as the records with complete race and Hispanic ethnicity data, then the imputation results are not directly comparable between the two groups. Our sampled cases are directly comparable to the complete-case subsets by matching the values of age, sex, and geographic area.

For each subset analysis, we estimated case rates using Poisson regression. Population estimates were included as the model offsets. We pooled the parameter estimates and variance-covariance matrices from each imputation and compared the results with the original data. We expected the 95% confidence intervals for the multiply imputed case rates to contain the case rates based on the original data. See Supplemental Digital Content 2 for additional details on the calculations involved in the estimation of case rates and confidence intervals for the subset analyses.

RESULTS

The full data set included 2,553,038 cases of reported chlamydia, gonorrhea, and syphilis (excluding congenital syphilis). Race and Hispanic ethnicity data were missing for 9% of syphilis case reports, 22% of gonorrhea case reports and 33% of chlamydia case reports (Table 1). However, less than one percent of case reports had missing age or sex for chlamydia, gonorrhea, and syphilis in 2019.

Figures 1, 2, and 3 display the rates of reported cases per 100,000 population for chlamydia, gonorrhea, and primary and secondary syphilis with 95% confidence intervals calculated using the 15 imputations. The race and Hispanic ethnicity categories with smaller populations have larger variances. For comparison, the rates of reported cases from the complete-case analysis are plotted in light grey next to the rates from the multiply imputed data in dark gray. Excluding cases with missing race and Hispanic ethnicity has more impact for chlamydia (Figure 1) and gonorrhea (Figure 2) compared with primary and secondary syphilis (Figure 3).

In Figures 4a and 4b, grouping the data by US state (excluding Washington, DC), we estimated the rate of reported chlamydia by race and Hispanic ethnicity with 95% confidence intervals. On the y-axis we plotted the proportion of cases with missing race and Hispanic ethnicity for the state. On the x-axis, we plotted the ratio of the width of the 95% confidence interval to the rate of reported cases. Figure 4a includes 3 panels for the 3 categories of race and Hispanic ethnicity with the largest number of cases. For the race and

Hispanic ethnicity categories in Figure 4a, the width of the confidence intervals is usually smaller than the rate of reported chlamydia cases. However, Figure 4b shows that the widths of most confidence intervals are greater than the rate of reported chlamydia cases for race and Hispanic ethnicity categories with sparse data. The width of the confidence intervals generally increases as the percent of data with missing race and Hispanic ethnicity increases. For cases with NH-White race and ethnicity, no states had confidence interval widths that exceeded the rate of reported cases.

Subset analyses for model evaluation

Five subset analyses were performed to assess the bias and coverage of the multiple imputation model. For each subset, a random sample of size X was drawn from the total cases, Y. The size of the random sample was equal to the number of complete cases, cases in the subset with known race and Hispanic ethnicity. The characteristics of each subset are described in Table 2.

A rate of reported cases (r_i) was estimated for each race and Hispanic ethnicity category from the cases with complete data where i indicates the race and Hispanic ethnicity category. The confidence intervals, calculated from a sample with some imputed data, should contain the rate r_i. Tables 3-7 show the results from the five subset analyses. Supplemental Digital Content 3 includes a table for each subset analysis comparing the proportion of cases by race and Hispanic ethnicity category in the multiply imputed and original data sets.

The percent of cases with missing data ranged from 16% in the GA subset (Table 3) to 37% in the MA subset (Table 5). Every non-zero rate of reported cases was contained within the confidence intervals that were calculated from the sampled data. For race and Hispanic ethnicity categories with sparse data, the 95% confidence intervals were excessively wide. For example, the 95% confidence interval for reported rates of chlamydia among NH-Multiracial females aged 20-24 years in the KS subset ranged from 1.3 to 16,986.5 cases per 100,000 population (Table 4). Although the KS subset included more cases than the other four subsets, it did not have any cases in the NH-multiracial category. The 15 imputed data sets for the KS subset analysis had an average of 29 NH-multiracial cases which represents about 1% of the total cases in the KS subset. No NH-multiracial cases were imputed in 8 of the 15 data sets for the KS subset analysis. The subset analyses demonstrate the challenges of imputing counts in categories with sparse data but show reasonable predictions and variance estimates for race and Hispanic ethnicity categories with higher case counts.

DISCUSSION

Multiple imputation of race and Hispanic ethnicity data allows us to include all reported STI cases in a statistical analysis without excluding cases that have missing race and Hispanic ethnicity. If a large proportion of cases are excluded and data are not missing at random, it is unlikely that the estimated model coefficients will generalize to all reported cases. Therefore, the estimates are especially useful for chlamydia and gonorrhea data with 33% and 22% missing race and Hispanic ethnicity information in the 2019 NNDSS data. Because primary, secondary and early non-primary non-secondary syphilis cases, representing recent infections, are usually investigated by public health staff through patient and provider

follow-up, case data for these stages are most likely to have complete information for race and Hispanic ethnicity.

Confidence intervals that account for the uncertainty of the predictions are an advantage of multiple imputation over complete-case analysis because a realistic variance estimate allows for valid hypothesis testing results. The variance estimates from the imputed data sets do not ignore the uncertainty associated with the missing race and ethnicity data and therefore allow for valid comparisons of rates by race and Hispanic ethnicity category between geographic areas, sexes, or age categories.

We compared observed rates to multiply imputed rates from subsets with similar distributions of age, sex, and geographic area. In each subset analysis with non-zero rates of reported cases, the confidence interval contained the rate that was calculated using the subset of data having no cases with missing race and Hispanic ethnicity. The width of the confidence interval indicates the amount of uncertainty associated with the predictions.

We noticed that confidence intervals for rates of reported cases in four race and ethnicity categories with sparse data are excessively wide. If there are truly no cases within a race and ethnicity category, or if the population is so small that no cases would be observed, our multiple imputation model is likely to overestimate the rate of reported cases. The wide confidence intervals for these estimates indicate when the data may be too sparse for reasonable comparisons between race and Hispanic ethnicity categories with or without multiple imputation.

Although the multiple imputation methods account for the uncertainties associated with the race and ethnicity estimates, additional uncertainties are not accounted for. NNDSS does not capture all persons infected with chlamydia, gonorrhea, and syphilis. Some infections are asymptomatic and are not diagnosed and reported. The confidence intervals presented in this analysis are for reported cases and not for true disease incidence. Therefore, we use the terminology “rates of reported cases” or “case rates” as opposed to “disease incidence” or “rates of chlamydia”. This analysis doesn’t account for the uncertainty due to underreporting.

We chose to include the derived race/Hispanic ethnicity category in the imputation model as opposed to the seven binary race and ethnicity variables. The strategy for creating the derived race/Hispanic ethnicity categories is included in the methods above. By creating the derived race/Hispanic ethnicity category before imputing the data, we avoid several complications related to the sum of positive responses to the seven binary variables. For example, if none of the imputed responses are positive, race and Hispanic ethnicity remain unknown.

There are alternative strategies for summarizing race and ethnicity data. For example, we have not separated race and ethnicity and therefore we do not report categories such as “Hispanic Whites”, “Hispanic Blacks” etc. Cases that could be grouped into the American Indian or Alaska Native, Asian, Black or African American, Multiracial, Native Hawaiian or Other Pacific Islander, or White categories are categorized as only Hispanic or Latino. We lost some information about race to maintain consistency with the race and ethnicity

categorizations in publicly available reports such as CDC's NCHHSTP AtlasPlus[13] and the annual STD Surveillance Report[14]. Future analyses may explore different strategies for categorizing the race and ethnicity data from NNDSS.

Multiple imputation of race and Hispanic ethnicity in STI case notification data from NNDSS allows us to use all cases in our rate comparisons. Complete case analysis of NNDSS case reports from 2019 would exclude over 30% of chlamydia cases and over 20% or gonorrhea cases. The confidence intervals generated from the multiply imputed data provide a valuable assessment of the uncertainty associated with the rates for specific race and Hispanic ethnicity groups. As we continue to assess disparities in disease burden across race and ethnicity categories, multiple imputation is a useful tool.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

REFERENCES

1. Centers for Disease Control and Prevention. National Notifiable Diseases Surveillance System (NNDSS). Accessed April 25, 2024. https://www.cdc.gov/nndss/index.html
2. Department of Health and Human Services. STI National Strategic Plan Overview. Accessed April 25, 2024. https://www.hhs.gov/programs/topic-sites/sexually-transmitted-infections/plan-overview/index.html
3. Office of Disease Prevention and Health Promotion. Healthy People 2030 Framework. Accessed April 25, 2024. https://health.gov/healthypeople/about/healthy-people-2030-framework
4. Centers for Disease Control and Prevention. Sexually Transmitted Disease Surveillance 2019. Accessed April 25, 2024. https://www.cdc.gov/std/statistics/2019/std-surveillance-2019.pdf
5. van Buuren S. Flexible Imputation of Missing Data. CRC Press; 2012.
6. van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate imputation by chained equations in R. Journal of statistical software. 2011;45:1–67.
7. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine. 2011;30(4):377–399. [PubMed: 21225900]
8. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2021. Accessed April 25, 2024. http://www.R-project.org/
9. United States Census Bureau. Population and Housing Unit Estimates Datasets. Accessed April 25, 2024. https://www.census.gov/programs-surveys/popest/data/data-sets.html
10. Little RJ, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. Wiley series in probability and statistics. John Wiley & Sons; 2002.
11. Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley Series in Probability and Statistics. John Wiley & Sons; 1987.
12. Rubin DB. Multiple imputation after 18+ years. Journal of the American statistical Association. 1996;91(434):473–489.
13. Centers for Disease Control and Prevention. NCHHSTP AtlasPlus. Accessed April 25, 2024. https://www.cdc.gov/nchhstp/atlas/index.htm
14. Centers for Disease Control and Prevention. STD Data & Statistics Archive. Accessed April 25, 2024. https://www.cdc.gov/std/statistics/archive.htm

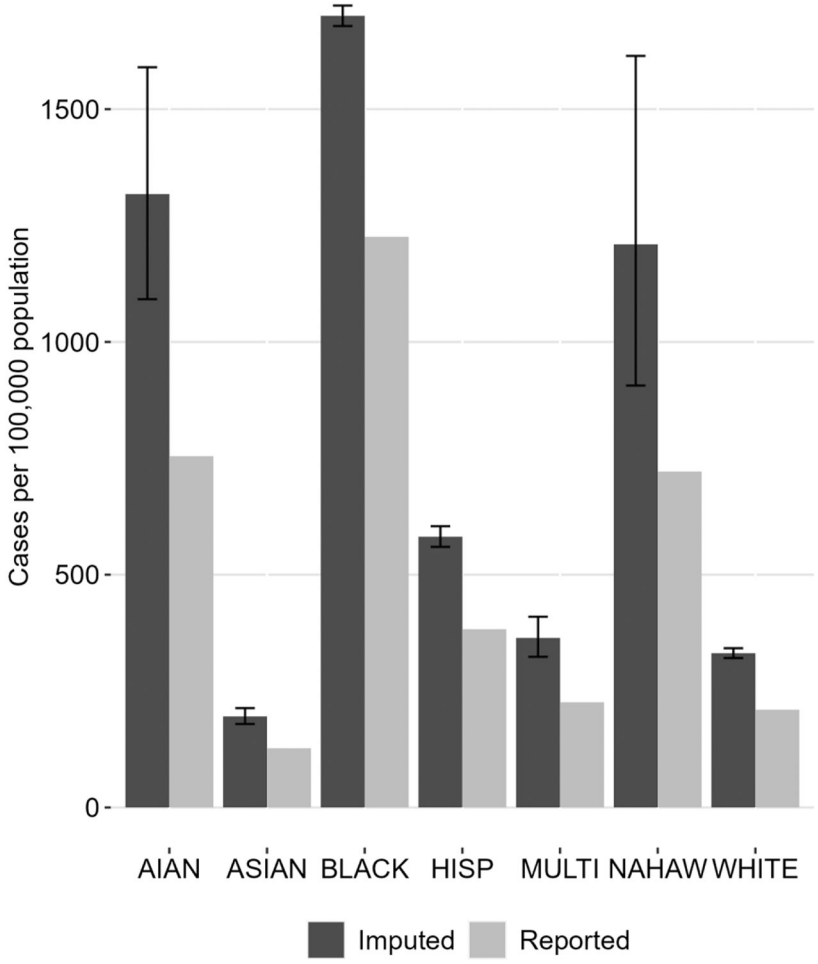

Figure 1:
US chlamydia cases reported to NNDSS in 2019 per 100,000 population from 15 imputed data sets with 95% confidence intervals vs. rates for cases with reported race and Hispanic ethnicity. AIAN, non-Hispanic (NH) American Indian or Alaska Native; ASIAN, NH-Asian; BLACK, NH-Black or African American; HISP, Hispanic/Latino; MULTI, NH-Multiracial; NAHAW, NH-Native Hawaiian or Other Pacific Islander; WHITE, NH-White

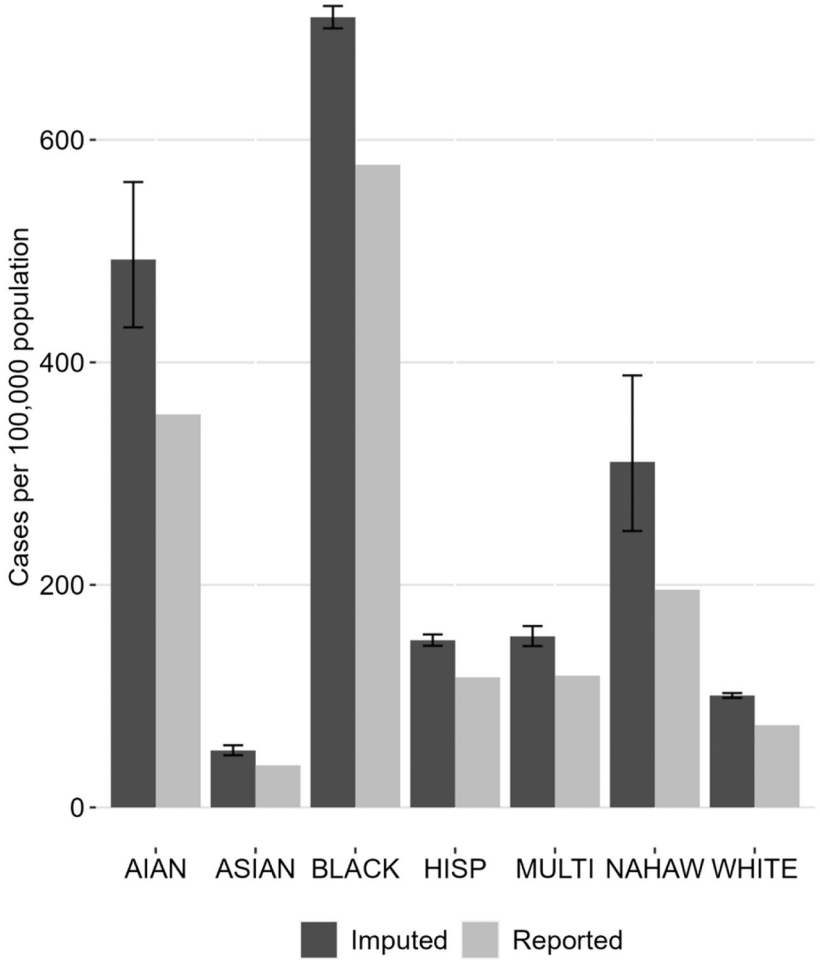

Figure 2:

US gonorrhea cases reported to NNDSS in 2019 per 100,000 population from 15 imputed data sets with 95% confidence intervals vs. rates for cases with reported race and Hispanic ethnicity. AIAN, non-Hispanic (NH) American Indian or Alaska Native; ASIAN, NH-Asian; BLACK, NH-Black or African American; HISP, Hispanic/Latino; MULTI, NH-Multiracial; NAHAW, NH-Native Hawaiian or Other Pacific Islander; WHITE, NH-White

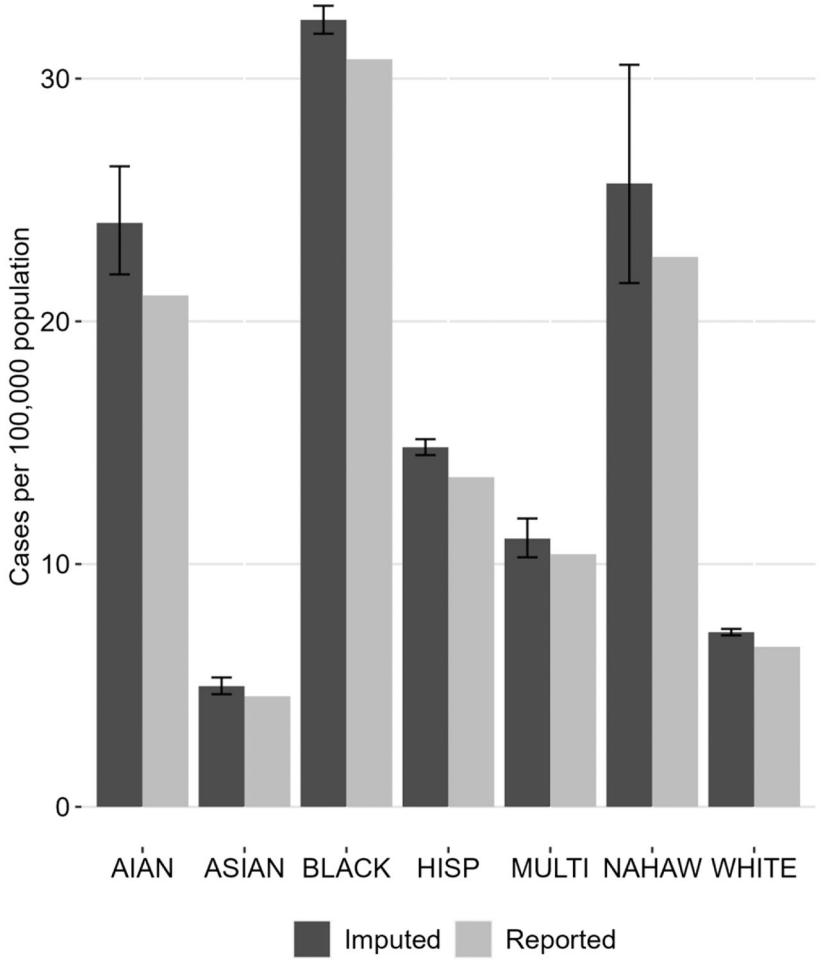

Figure 3:

US primary and secondary syphilis cases reported to NNDSS in 2019 per 100,000 population from 15 imputed data sets with 95% confidence intervals vs. rates for cases with reported race and Hispanic ethnicity. AIAN, non-Hispanic (NH) American Indian or Alaska Native; ASIAN, NH-Asian; BLACK, NH-Black or African American; HISP, Hispanic/Latino; MULTI, NH-Multiracial; NAHAW, NH-Native Hawaiian or Other Pacific Islander; WHITE, NH-White

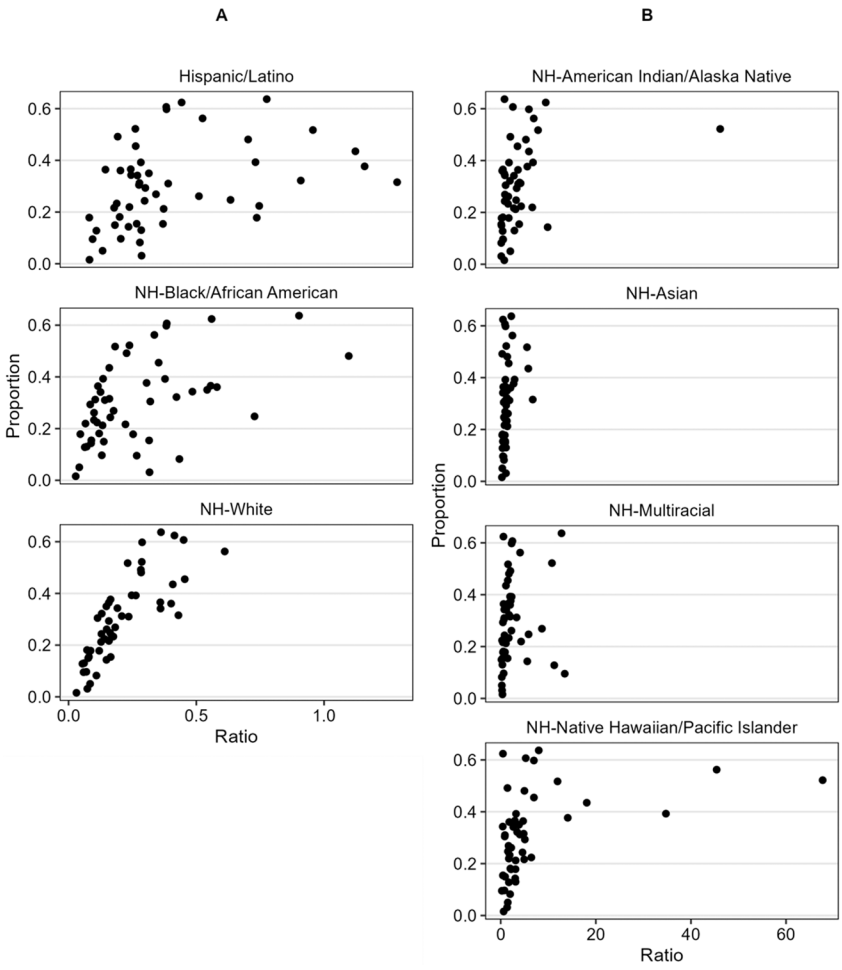

Figure 4:

Width of 95% confidence interval compared to proportion of cases in geographic area with missing race and Hispanic ethnicity data for chlamydia cases reported to NNDSS in 2019. Ratio on x-axis is width of 95% confidence interval divided by the reported chlamydia cases per 100,000 population. Confidence intervals for chlamydia cases per 100,000 population calculated from 15 imputed data sets for each of 50 states. A) For race and Hispanic ethnicity categories with the largest number of cases, most ratios do not exceed 1.0. B) For race and Hispanic ethnicity categories with smallest number of cases, most ratios exceed 1.0.

Table 1:

Number and percent of STIs reported to the National Notifiable Diseases Surveillance System in 2019 with missing race/Hispanic ethnicity

STI	Cases	Missing race	Percent missing
Chlamydia	1,808,703	594,933	32.9
Gonorrhea	616,392	136,429	22.1
Total Syphilis	127,943	11,021	8.6
Primary and Secondary	38,992	2,822	7.2
Early Non-Primary Non-Secondary	41,655	2,284	5.5
Unknown Duration or Late	47,296	5,915	12.5

Table 2:

Characteristics of cases in the five subset analyses

STI	Area	Age (years)	Sex	Cases	Subset[a]	Percent missing
Chlamydia	GA	30-34	Male	2,547	2,127	16.5
Chlamydia	KS	20-24	Female	4,083	2,995	26.6
Gonorrhea	MA	15-24	Female	1,061	665	37.3
Primary and Secondary Syphilis	NV	25-39	Male	363	269	25.9
Chlamydia	WA	25-29	Female	4,086	2,689	34.2

[a]See Supplemental Digital Content 2 for a description of the subset selection. Size of subset is equal to the total number of cases with reported race and Hispanic ethnicity data.

Table 3.

Rate of reported cases of chlamydia per 100,000 men aged 30-34 years in Georgia, 2019

Race/Hispanic Ethnicity	Rate - Reported[a]	Rate - Imputed[b]	95% CI[c]
NH-American Indian/Alaska Native	265.6	265.6	15.7 - 1,938.1
NH-Asian	57.8	82.7	28.0 - 204.7
NH-Black/African American	1,384.3	1,369.4	1,274.2 - 1,470.9
NH-Native Hawaiian/Pacific Islander	1,071.4	357.1	50.3 - 2,535.4
NH-White	232.3	239.1	187.9 - 301.6
NH-Multiracial	79.0	135.6	14.5 - 684.2
Hispanic/Latino	296.5	295.7	219.5 - 393.7

[a]Reported rates are based on 2,127 cases with reported race and Hispanic ethnicity

[b]See Supplemental Digital Content 2 for a description of the rates calculated from the imputed case counts. Among 2,127 sampled cases, 338 (15.9%) had missing race and Hispanic ethnicity.

[c]The 95% confidence intervals should contain the reported rates.

Table 4.

Rate of reported cases of chlamydia per 100,000 women aged 20-24 years in Kansas, 2019

Race/Hispanic Ethnicity	Rate - Reported[a]	Rate - Imputed[b]	95% CI[c]
NH-American Indian/Alaska Native	4,322.8	3,106.0	1,334.9 - 6,398.0
NH-Asian	1,229.7	1,253.3	571.9 - 2,492.7
NH-Black/African American	9,329.7	9,502.7	7,013.8 - 12,674.5
NH-Native Hawaiian/Pacific Islander	6,569.3	4,379.6	1,967.6 - 9,748.4
NH-White	2,516.1	2,403.3	1,949.0 - 2,940.8
NH-Multiracial	0.0	734.0	1.3 - 16,986.5
Hispanic/Latino	3,486.1	3,833.9	2,584.7 - 5,532.4

[a]Reported rates are based on 2,995 cases with reported race and Hispanic ethnicity

[b]See Supplemental Digital Content 2 for a description of the rates calculated from the imputed case counts. Among 2,995 sampled cases, 834 (27.8%) had missing race and Hispanic ethnicity.

[c]The 95% confidence intervals should contain the reported rates.

Table 5.

Rate of reported cases of gonorrhea per 100,000 women aged 15-24 years in Massachusetts, 2019

Race/Hispanic Ethnicity	Rate - Reported[a]	Rate - Imputed[b]	95% CI[c]
NH-American Indian/Alaska Native	115.9	339.9	13.9 - 2,879.1
NH-Asian	37.5	38.8	11.3 - 105.1
NH-Black/African American	653.4	575.1	457.7 - 717.6
NH-Native Hawaiian/Pacific Islander	409.8	1,092.9	45.9 - 8,330.0
NH-White	70.7	87.8	70.2 - 109.1
NH-Multiracial	35.5	68.7	18.3 - 206.8
Hispanic/Latino	246.0	204.6	147.0 - 280.3

[a] Reported rates are based on 665 cases with reported race and Hispanic ethnicity

[b] See Supplemental Digital Content 2 for a description of the rates calculated from the imputed case counts. Among 665 sampled cases, 248 (37.3%) had missing race and Hispanic ethnicity.

[c] The 95% confidence intervals should contain the reported rates.

Table 6.

Rate of reported cases of primary and secondary syphilis per 100,000 men aged 25-39 years in Nevada, 2019

Race/Hispanic Ethnicity	Rate - Reported[a]	Rate - Imputed[b]	95% CI[c]
NH-American Indian/Alaska Native	0.0	92.4	5.3 - 730.4
NH-Asian	18.0	16.8	5.8 - 45.8
NH-Black/African American	171.9	183.1	135.9 - 245.2
NH-Native Hawaiian/Pacific Islander	0.0	43.8	5.6 - 307.4
NH-White	51.7	53.0	37.8 - 73.3
NH-Multiracial	17.5	13.4	1.7 - 86.0
Hispanic/Latino	111.2	104.2	82.6 - 130.9

[a]Reported rates are based on 269 cases with reported race and Hispanic ethnicity

[b]See Supplemental Digital Content 2 for a description of the rates calculated from the imputed case counts. Among 269 sampled cases, 59 (21.9%) had missing race and Hispanic ethnicity.

[c]The 95% confidence intervals should contain the reported rates.

Table 7.

Rate of reported cases of chlamydia per 100,000 women aged 25-29 years in Washington, 2019

Race/Hispanic Ethnicity	Rate - Reported[a]	Rate - Imputed[b]	95% CI[c]
NH-American Indian/Alaska Native	2,822.7	2,419.0	1,099.7 - 4,760.2
NH-Asian	295.4	252.0	101.6 - 545.0
NH-Black/African American	1,841.1	2,122.5	996.6 - 4,146.9
NH-Native Hawaiian/Pacific Islander	3,643.1	2,810.8	1,961.3 - 3,963.9
NH-White	789.4	796.3	638.1 - 985.0
NH-Multiracial	727.1	658.0	269.0 - 1,391.5
Hispanic/Latino	1,634.3	1,667.8	1,278.8 - 2,149.9

[a]Reported rates are based on 2,689 cases with reported race and Hispanic ethnicity

[b]See Supplemental Digital Content 2 for a description of the rates calculated from the imputed case counts. Among 2,689 sampled cases, 941 (35.0%) had missing race and Hispanic ethnicity.

[c]The 95% confidence intervals should contain the reported rates.