



Published in final edited form as:

J Public Health Manag Pract. 2024 ; 30(5): 733–743. doi:10.1097/PHH.0000000000002014.

Estimating the Prevalence of Injection Drug Use Among Acute Hepatitis C Cases From a National Surveillance System: Application of Random Forest-Based Multiple Imputation

Shaoman Yin, MSPH PhD,

Kathleen N. Ly, MPH,

Laurie K. Barker, MSPH,

Danae Bixler, MD,

Nicola D. Thompson, PhD,

Neil Gupta, MD

Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, Georgia.

Abstract

Background: Injection drug use (IDU) is a major contributor to the syndemic of viral hepatitis, human immunodeficiency virus, and drug overdose. However, information on IDU is frequently missing in national viral hepatitis surveillance data, which limits our understanding of the full extent of IDU-associated infections. Multiple imputation by chained equations (MICE) has become a popular approach to address missing data, but its application for IDU imputation is less studied.

Methods: Using the 2019–2021 National Notifiable Diseases Surveillance System acute hepatitis C case data and publicly available county-level measures, we evaluated listwise deletion (LD) and 3 models imputing missing IDU data through MICE: parametric logistic regression, semi-parametric predictive mean matching (PMM), and nonparametric random forest (RF) (both standard RF [sRF] and fast implementation of RF [fRF]).

Results: The estimated IDU prevalence among acute hepatitis C cases increased from 63.5% by LD to 65.1% by logistic regression, 66.9% by PMM, 76.0% by sRF, and 85.1% by fRF. Evaluation studies showed that RF-based MICE imputation, especially fRF, has the highest accuracy (as measured by smallest raw bias, percent bias, and root mean square error) and highest efficiency (as measured by smallest 95% confidence interval width) compared to LD and other models. Sensitivity analyses indicated that fRF remained robust when data were missing not at random.

Correspondence: Shaoman Yin, MSPH, PhD, Division of Viral Hepatitis, Centers for Disease Control and Prevention, 1600 Clifton Road, H24-6, Atlanta, GA 30333 (wso3@cdc.gov).

The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. The authors have no conflicts of interest relevant to this article to disclose.

The authors declare no conflicts of interest.

Supplemental digital content is available for this article. Direct URL citation appears in the printed text and is provided in the HTML and PDF versions of this article on the journal's Web site (<http://www.JPHMP.com>).

Conclusion: Our analysis suggested that RF-based MICE imputation, especially fRF, could be a valuable approach for addressing missing IDU data in the context of population-based surveillance systems like National Notifiable Diseases Surveillance System. The inclusion of imputed IDU data may enhance the effectiveness of future surveillance and prevention efforts for the IDU-driven syndemic.

Keywords

injection drug use; missing data; multiple imputation by chained equations; random forest

Introduction

Injection drug use (IDU) is a primary contributor to the syndemic of overdose,¹⁻³ human immunodeficiency virus (HIV),⁴⁻⁶ hepatitis C,⁷⁻⁹ and other associated conditions.^{10,11} An estimated 3.7 million people (1.5% of the US adult population) injected drugs in 2018, about 5 times the 2011 estimate.^{12,13} The incidence rate of acute hepatitis C virus (HCV) infection in the United States has quadrupled from 2011 to 2021 (0.4-1.6 cases per 100 000), and IDU is the most commonly reported risk factor.¹⁴ There were 69 800 estimated acute HCV infections during 2021, which was more than 15 times the 2030 viral hepatitis national strategic goal of 4400.¹⁵

To identify potential outbreaks, understand transmission, and better address the syndemic, it is crucial to monitor and track IDU through national viral hepatitis surveillance data. Obtaining information on IDU is challenging due to limited resources to investigate acute cases and reluctance to disclose potentially stigmatizing behaviors. In the Centers for Disease Control and Prevention's (CDC's) National Notifiable Diseases Surveillance System (NNDSS), more than 60% of reported acute hepatitis C cases are missing IDU data.¹⁴ The large number of missing IDU data limits the usefulness of surveillance data to inform public health prevention efforts, such as linkage to treatment and harm reduction services.

Choosing the appropriate approach to handle missing data depends on the pattern of missingness: missing completely at random (MCAR, missingness is independent of any variable), missing at random (MAR, missingness is associated with other variables), and missing not at random (MNAR, missingness is associated with the missing variable itself).¹⁶ Listwise deletion (LD) relies on the MCAR assumption, but this assumption is often invalid leading to biased estimates. Imputation of missing data generally starts with the MAR assumption and then addresses the possibility of MNAR through sensitivity analysis.¹⁷

Multiple imputation by chained equations (MICE) is a principled and flexible approach to impute missing data.^{18,19} MICE generates multiple imputed datasets by iteratively predicting missing values for each variable through an imputation model based on other variables. To our knowledge, MICE has not been applied to impute missing IDU information for acute viral hepatitis cases from the national surveillance system though it has been applied to other data elements for other conditions.^{20,21} One study used a logistic regression (LR)-based imputation model to impute missing substance use information but was limited to a small cohort.²² As national surveillance data usually have a large number of variables

with complex relationship patterns such as nonlinearity, non-normality, and hierarchical interactions, specifying the appropriate parametric model for imputing missing IDU data could be challenging and labor-intensive.

Recent studies have shown that machine learning algorithms such as random forest (RF) could be powerful tools when they are incorporated with MICE.²³⁻²⁵ RF combines the outputs from multiple decision trees,²⁶ does not require determining an imputation model, and allows inclusion of many predictors for automatic screening while accounting for complex relationships among variables. RF-based imputation performs well under moderate to high missingness, even when data are MNAR in certain cases.²⁷

To better understand the contribution of IDU among acute hepatitis C cases in the United States, we aimed to establish a flexible and robust MICE approach to impute missing IDU information in the context of a large national surveillance dataset. Specifically, using NNDSS, we compared the performance of LD and 3 classes of models to impute missing IDU data through MICE: parametric LR, semi-parametric predictive mean matching (PMM),²⁸ and nonparametric RF model (standard RF [sRF] through R “mice” package,¹⁹ and fast implementation of RF [fRF] through R package “miceRanger”²⁹). Evaluation and sensitivity analyses were conducted to assess the performance of each imputation model.

Methods

Datasets and variables

CDC’s NNDSS is a nationwide collaboration among health departments to report and monitor nationally notifiable conditions in the United States.³⁰ During 2019-2021, a total of 13 852 confirmed (ie, HCV RNA-positive) acute hepatitis C cases were reported to NNDSS and included in this analysis. Each case was assumed to represent 1 unique person. For this analysis, the IDU information was obtained from 2 questions, one about whether the person injected drugs in the 2 weeks to 6 months prior to the onset of symptoms, and a second about ever having injected drugs. If the response was yes to either question, then IDU was coded yes. Both unknown and blank values were treated as missing. Other analytic variables from NNDSS included age (≤ 19 , 20-29, 30-39, 40-49, 50-59, and ≥ 60 years), sex (male or female), race/ethnicity (non-Hispanic American Indian/Alaska Native, non-Hispanic Asian/Pacific Islander, non-Hispanic Black, non-Hispanic White, Hispanic, or Other), and urbanicity (large central metro, large fringe metro, medium metro, small metro, micropolitan, and non-core). Because MICE can impute missing values for multiple variables sequentially, missing values for all the above analytic variables were also imputed. Considering other risk data available for acute hepatitis C cases within surveillance data (such as non-IDU, needlestick injury, incarceration status) are often correlated with IDU, they were included as auxiliary variables. The unknown values of these auxiliary risk factors were treated as nonmissing. Because the individual IDU information could be influenced by local social and environmental determinants, we sought county-level measures as additional auxiliary variables to further improve MICE imputation.¹⁶ We extracted all major county-level measures from 2 data sources: County Health Rankings & Roadmaps (2020 data, <https://www.countyhealthrankings.org>) and CDC Social Vulnerability Index (SVI, 2018 data, <https://www.atsdr.cdc.gov/placeandhealth/svi.html>), and merged these to the case-level

NNDSS data using county Federal Information Processing Standards codes. Overall, 134 predictor variables were included in this analysis (Supplemental Digital Content, available at <http://links.lww.com/JPHMP/B405>).

MICE imputation models

Four MICE imputation models including LR, semi-parametric PMM, and nonparametric sRF and fRF were conducted. From the 134 predictor variables, we first screened the variables for their association with IDU missingness through the “quickpred” function provided by R package “mice” using an absolute Pearson correlation threshold of ≥ 0.2 .¹⁹ Selected auxiliary variables and analytic variables (variables No. 1-29 in Supplemental Digital Content, Table, available at <http://links.lww.com/JPHMP/B405>) were included in the LR, PMM, and sRF models. For simplicity, this analysis did not account for the potential interactions in LR and PMM models. For fRF, we included all 134 variables for initial screening because it is designed to analyze high-dimensional data. For each imputation model, we generated 10 different imputed datasets, and the final estimate for the IDU prevalence was derived based on Rubin’s rule, which can account for the variance between and within imputed datasets.¹⁶ All analyses were conducted using RStudio (version 2023.03.0, RStudio Team), with “tidyverse,”³¹ “mice,”¹⁹ and “miceRanger” packages,²⁹ and their dependencies.

Evaluation analysis

To evaluate the performance of the MICE imputation models (ie, how well the imputed data estimated the “true” data) under variant scenarios, we used 3 datasets as the “true” datasets from NNDSS acute hepatitis C data. These include Dataset 1: NNDSS 2019-2021 data (excluded missing IDU data, $n = 4945$); Dataset 2: Partial NNDSS 2019-2021 data limited to 4 states where IDU missing proportion was $\leq 30\%$ (excluded missing IDU data, $n = 499$); Dataset 3: Partial NNDSS 2019-2021 data limited to 23 states where IDU missing proportion was $> 60\%$ (excluded missing IDU data, $n = 2268$). A total of 29 variables including IDU, age, sex, race/ethnicity, urbanicity, SVI quartile, and other county-level auxiliary variables were included in all the evaluation analyses.

The evaluation process is described in Figure 1. First, we established a logistic model for IDU missingness based on the original 2019-2021 national raw data, which included variables No. 2-29 in Supplemental Digital Content, Table 1, available at <http://links.lww.com/JPHMP/B405>, and accounted for the mechanism of MAR. Then this model was applied to the “true” dataset where some of the true IDU data were replaced with missing values based on model prediction. To create randomness, the dataset was further reshuffled into 50 testing datasets through bootstrap sampling with replacement. For each of these 50 testing datasets, we conducted the 4 MICE imputation models, separately. The prevalence of IDU and its 95% confidence interval (CI) from the imputed datasets were derived based on Rubin’s rule.¹⁶ As a comparison, the prevalence of IDU and its 95% CI (based on normal approximation) through LD were also derived. All estimates were averaged over the 50 testing datasets and compared to the “true” value of the initial dataset. Four evaluation measures were calculated including: (1) raw bias (the difference of estimated IDU prevalence minus the “true” IDU prevalence); (2) percent bias (the

percentage of the raw bias over the “true” IDU prevalence); (3) 95% CI width (the upper limit minus the lower limit of 95% CI); and (4) the root mean square error (RMSE), which measures the average magnitude of the error between the estimated and true values. The smaller the value of the raw bias, percent bias, or RMSE, the more accurate the estimates from the MICE imputation. The narrower the 95% CI width, the more statistically efficient the MICE imputation.

Sensitivity analysis

We conducted a sensitivity analysis to assess the stability of estimation from the selected MICE imputation when the missing data follows MNAR rather than MAR. The only difference from the MAR evaluation analysis is that we mimic the MNAR mechanism in generating the missing data (Supplemental Digital Content, Figure, available at <http://links.lww.com/JPHMP/B405>). We first derived a formula with coefficient values using outputs of the LR for IDU missingness in the evaluation study. Then, we added IDU to the formula as a predictor, varying its coefficients ($\beta = 0, .1, .2, .5, .8$, and 1.0). Finally, we applied this modified regression model formula to the “true” data to generate missing IDU data. Overall, this missing data generation process accounts for the association of IDU itself with the IDU missingness as defined by MNAR, eg, if case patients who inject drugs are less likely to provide IDU risk information than case patients who do not inject drugs. When $\beta = 0$, there is an equal chance of IDU data being present or missing. As the β coefficient increases from 0.1 to 1.0, the likelihood of IDU data being missing increases among persons with “true” IDU. Under these scenarios, we calculated the raw bias, percent bias, 95% CI widths, and RMSE to assess the robustness of MICE imputation against the MNAR assumption.

Results

Population characteristics

Table 1 shows the population characteristics overall and stratified by IDU missingness. Of 13 852 acute hepatitis C cases reported to NNDSS during 2019-2021, 8907 (64.3%) were missing IDU data. Most persons with acute hepatitis C were aged 20 to 39 years (57.3%), male (64.0%), and White (63.4%). Chi-square tests detected significant associations of IDU missingness with age, sex, race/ethnicity, urbanicity, and SVI quartile. For example, when compared to acute hepatitis C cases with complete IDU, cases with missing IDU were more frequent among persons who were aged ≥ 60 years (14.0% vs 8.1%), non-Hispanic African American (11.0% vs 5.8%), and Hispanic (10.5% vs 7.5%), residing in large central metro counties (34.9% vs 19.3%) and counties in the highest SVI quartile (27.4% vs 18.5%). In contrast, IDU was more frequently complete for persons who were aged 20 to 29 years (28.6% vs 23.7%), aged 30 to 39 years (34.8% vs 30.2%), female (40.0% vs 33.6%), non-Hispanic White (74.5% vs 57.3%), and residing in counties in the 2 lowest SVI quartiles (55.9% vs 43.3%). These findings indicate that IDU missingness was not MCAR among acute hepatitis C cases in NNDSS.

MICE model comparisons

Table 2 shows the prevalence of IDU among persons with reported acute hepatitis C infections overall and by characteristics from LD and 4 MICE imputations (LR, PMM, sRF, and fRF). Overall, 3140 persons with acute hepatitis C reported IDU in the raw data. After MICE imputations, this number increased more than 3-fold to 9015 by LR, 9263 by PMM, 10 524 by sRF, and 11 791 by fRF. The number of persons with imputed IDU also increased significantly among certain sociodemographic subgroups. For example, 42 persons aged ≥ 60 years reported IDU in the raw data; after fRF imputation, this age group had an estimated 1112 persons who injected drugs (PWIDs). Among non-Hispanic African American persons, 107 reported IDU in the raw data; after fRF imputation, non-Hispanic African American persons accounted for an estimated 1057 PWID.

The overall prevalence of IDU among persons with acute hepatitis C also increased from LD (63.5%; 95% CI, 62.1%-64.8%), to LR (65.1%; 95% CI, 62.6%-67.6%), PMM (66.9%; 95% CI, 64.2%-69.6%), sRF (76.0%; 95% CI, 74.0%-78.0%), and fRF (85.1%; 95% CI, 84.4%-85.8%). The prevalence of IDU among all demographic subgroups also increased from LD to fRF-MICE models. For instance, 75.0% (95% CI, 72.8%-77.0%) of persons aged 30 to 39 years from LD were PWID. This prevalence increased to 90.1% (95% CI, 89.2%-90.9%) using fRF. Large increases ($>25\%$) in the prevalence of IDU also were seen among persons who were aged 50 to 59 or ≥ 60 years, non-Hispanic American Indian/Alaskan Native, non-Hispanic Asian/Pacific Islander, non-Hispanic African American, Hispanic, or resided in large central metro counties or in counties in the 2 highest SVI quartiles.

MICE model evaluations

Table 3 shows the evaluation measures of raw bias, percent bias, 95% CI width, and RMSE across the 5 estimation approaches (LD, LR, PMM, sRF, and fRF) under 3 scenarios. For Datasets 1 and 3, fRF exhibited the best performance as indicated by the lowest percent bias (3.19% and 2.71%, respectively), compared to LD (15.30% and 23.90%, respectively), LR (6.64% and 9.21%, respectively), PMM (5.38% and 8.88%, respectively), and sRF (4.22% and 9.33%, respectively). These findings were also supported when we assessed the raw bias and RMSE. For Dataset 2, LR showed the worst performance with percent bias at 11.92%, followed by LD (5.36%), PMM (5.26%), fRF (4.81%), and sRF (3.42%). For all the Datasets 1 to 3, fRF showed the greatest efficiency of estimation as indicated by the smallest 95% CI widths compared to LD, LR, PMM, and sRF.

Sensitivity to Missing not at Random

Though we found that IDU missingness was associated with other observed variables, suggesting these data are MAR, MNAR was still possible. To assess the sensitivity of fRF to MNAR, we repeated the evaluation process but incorporated MNAR when missing IDU data were generated (Supplemental Digital Content, Figure, available at <http://links.lww.com/JPHMP/B405>). We tested scenarios with the beta coefficient for IDU missingness set to 0, .1, .2, .5, .8, and 1.0 (Table 4). Under the MAR assumption (beta coefficient at 0), the percent bias of fRF was 3.19%. Under the MNAR assumption (with beta coefficient varying from .1 to 1.0), the percent bias of fRF ranged from 1.57% to 2.93%. No specific pattern

was observed when beta coefficients increased. Consistently, other measures such as raw bias, 95% CI width, and RMSE remained stable across different scenarios.

Discussion

To address the high missing values of IDU in national surveillance data, this study assessed 3 imputation models through MICE. We found that the fRF-based MICE imputation produced the most accurate estimation of IDU prevalence while remaining robust to the MNAR assumption. The estimated IDU prevalence among acute hepatitis C cases in the United States during 2019-2021 was as high as 85.1% as compared to 63.5% from the LD approach.

Among the different approaches we evaluated for handling missing IDU data, LD had the most bias compared to other MICE imputations. This was expected as LD excluded a large amount of data which reduced statistical power, and the assumption of MCAR did not hold. In addition, we found that in certain circumstances (such as using Dataset 2 with smaller sampler size), LR showed worse performance than LD, which might be explained by the less optimized imputation model. In all the tested scenarios, RF-based MICE imputation, especially fRF, showed better performance and efficiency than LD and other models. These differences might reflect the different levels of robustness of imputation modeling when dealing with large surveillance data with complex patterns such as interactions and nonlinearities. A high-quality imputation model is necessary to allow MICE to have accurate and efficient estimation.³² Our analysis suggested that the data-driven machine learning such as RF was advantageous because it can automatically capture valuable underlying patterns from many data elements to increase estimation accuracy and efficiency. This approach is necessary and advantageous to improve the inference and interpretation of viral hepatitis surveillance data. Many jurisdictions do not currently have sufficient resources to obtain nor impute IDU information on all acute hepatitis C cases; some might lack the resources to conduct imputation themselves. A national-level imputation approach could generate information to inform jurisdiction hepatitis C surveillance and prevention program efforts, improving imputation by use of data from multiple jurisdictions.

While we have demonstrated significant associations for missing IDU information with the observed variables (ie, MAR), we cannot exclude the possibility that IDU missingness may be associated with other unobserved variables or IDU itself due to potential stigma or other concerns (ie, MNAR). To address this possibility, it is recommended to include as many related variables as possible in the imputation model.¹⁶ In our analysis, in addition to the case-level surveillance data, we added 108 county-level measures with diverse themes from public data sources. However, this also caused computational speed issues with sRF, but not for fRF, which is designed for high-dimensional data. Simulation studies showed that RF outperformed other machine learning mechanisms such as k-nearest neighbors and was robust when data were MNAR.²⁷ This finding is consistent with our sensitivity analysis.

The fRF-MICE may help us to better understand the prevalence and burdens of IDU-associated acute hepatitis infections in the United States. A previous study had used acute hepatitis C infection as a proxy measure for IDU, assuming IDU is the major risk for

the infection.³³ However, the true extent to which IDU contributed to acute hepatitis C has not been investigated from national surveillance system data. Another study limited to acute hepatitis C cases with available risk data estimated that the prevalence of IDU among these cases ranged from 58.5% in 2004 to 84.0% in 2014.⁹ Due to small samples used, these estimates were likely to be unstable and not nationally representative. Compared to the proxy-based approach, the multiple imputation-based approach from fRF-MICE, which considers the entire study population, might provide a more robust estimation on the IDU-associated infections.

Through the application of fRF-MICE, we conducted an analysis to compare the prevalence of IDU among different sociodemographic subgroups (Table 2). For example, the highest IDU prevalence of 93.0% was estimated for non-Hispanic Native American/Alaska Native persons compared to the lower value of 72.9% for non-Hispanic Asian/Pacific Islander persons, which might indicate potential racial disparities among IDU-associated infections. However, further studies are warranted due to the limited sample size at the subpopulation level by race/ethnicity. Another application of fRF-MICE is to study trends of IDU-associated infections over person, place, and time. Given the robustness of fRF-MICE, future research may explore the performance of this approach at a more refined geographic level, such as the county or city level, or within a specific subpopulation. This exploration would enhance the ability to detect clusters or outbreaks associated with IDU.

Application of fRF-MICE to address missing data and enhance understanding of surveillance data may also help public officials and health providers prioritize populations with a high prevalence of IDU, which can be used to support hepatitis prevention and elimination activities. For example, public officials and health providers may offer harm reduction services, including syringe services, education on safe injection and non-injection drug alternatives, medication for opioid use disorder, HIV and viral hepatitis testing and treatment, and hepatitis A and hepatitis B vaccination to sociodemographic populations and geographic areas with high prevalence of IDU where health care is not readily accessible. The fRF-MICE-based IDU estimation may help guide evidence-based strategies for targeted outreach and resource allocation, which can reduce morbidity, mortality, and further transmission of viral hepatitis.

This study has several limitations. First, RF does not have model specifications and may be more difficult to interpret than parametric models. However, RF can rank the importance of variables used for imputation, which may be helpful for understanding the model. Second, RF is usually computationally intensive when dealing with high-dimensional large datasets; this limitation can be alleviated by using more computationally efficient methods such as fRF-MICE. Third, like all other imputations, fRF-MICE assumes the missing data follow MAR. Sensitivity analyses should be conducted when MNAR is possible. Fourth, this analysis does not fully explore the selection bias introduced by variables or data sources used in the model. Further analysis is needed to explore potential bias and refine the model accordingly. Lastly, caution should be exercised when interpreting differences at the subpopulation level, especially when sample sizes are small. Potential bias might exist with the fRF-MICE due to limited information available to capture the characteristics and variability of small subpopulations. Further work is warranted to assess the validity of fRF-

MICE in addressing the issue of missing IDU data at the state or local, such as comparing the imputation analysis results with those obtained from enhanced surveillance studies that have more complete IDU data.

Conclusion

IDU is the most significant driver for acute hepatitis C in the United States; accurate and complete data are crucial for guiding hepatitis C elimination efforts. While efforts are underway to improve the completeness of surveillance data, imputation could serve as a supplemental approach to handle missing data appropriately when obtaining complete data is challenging. This analysis demonstrated that RF-based MICE imputation is an effective, flexible, and straightforward approach to address missing IDU data when analyzing high-dimensional national surveillance data with complex patterns like NNDSS. The utilization of advanced imputation methods, such as RF-MICE, has the potential to inform public health interventions aimed at addressing the syndemic of viral hepatitis, HIV, and drug overdose driven by IDU.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Hall EW, Rosenberg ES, Jones CM, et al. Estimated number of injection-involved drug overdose deaths, United States, 2000-2018. *Drug Alcohol Depend.* 2022;234:109428. [PubMed: 35364419]
2. Ciccarone D. The rise of illicit fentanyl, stimulants and the fourth wave of the opioid overdose crisis. *Curr Opin Psychiatry.* 2021;34(4):344-350. [PubMed: 33965972]
3. Ciccarone D. The triple wave epidemic: supply and demand drivers of the US opioid overdose crisis. *Int J Drug Policy.* 2019;71:183-188. [PubMed: 30718120]
4. Lyss SB, Buchacz K, McClung RP, et al. Responding to outbreaks of human immunodeficiency virus among persons who inject drugs—United States, 2016-2019: perspectives on recent experience and lessons learned. *J Infect Dis.* 2020;222(Suppl 5):S239-S249. [PubMed: 32877545]
5. Williams LD, Ibragimov U, Tempalski B, et al. Trends over time in HIV prevalence among people who inject drugs in 89 large US metropolitan statistical areas, 1992-2013. *Ann Epidemiol.* 2020;45:12-23. [PubMed: 32439148]
6. Alpren C, Dawson EL, John B, et al. Opioid use fueling HIV transmission in an urban setting: an outbreak of HIV infection among people who inject drugs—Massachusetts, 2015-2018. *Am J Public Health.* 2020;110(1):37-44. [PubMed: 31725317]
7. Ryerson AB, Schillie S, Barker LK, et al. Vital signs: newly reported acute and chronic hepatitis C cases—United States, 2009-2018. *MMWR Morb Mortal Wkly Rep.* 2020;69(14):399-404. [PubMed: 32271725]
8. Zibbell JE, Iqbal K, Patel RC, et al. Increases in hepatitis C virus infection related to injection drug use among persons aged ≤ 30 years—Kentucky, Tennessee, Virginia, and West Virginia 2006-2012. *MMWR Morb Mortal Wkly Rep.* 2015;64(17):453-458. [PubMed: 25950251]
9. Zibbell JE, Asher AK, Patel RC, et al. Increases in acute hepatitis C virus infection related to a growing opioid epidemic and associated injection drug use, United States, 2004 to 2014. *Am J Public Health.* 2018;108(2):175-181. [PubMed: 29267061]
10. Marks LR, Nolan NS, Liang SY, et al. Infectious complications of injection drug use. *Med Clin North Am.* 2022;106(1):187-200. [PubMed: 34823730]
11. Degenhardt L, Webb P, Colledge-Frisby S, et al. Epidemiology of injecting drug use, prevalence of injecting-related harm, and exposure to behavioural and environmental risks among people

- who inject drugs: a systematic review. *Lancet Glob Health*. 2023;11(5):e659–e672. [PubMed: 36996857]
12. Lansky A, Finlayson T, Johnson C, et al. Estimating the number of persons who inject drugs in the United States by meta-analysis to calculate national rates of HIV and hepatitis C virus infections. *PLoS One*. 2014;9(5):e97596. [PubMed: 24840662]
 13. Bradley H, Hall EW, Asher A, et al. Estimated number of people who inject drugs in the United States. *Clin Infect Dis*. 2023;76(1):96–102. [PubMed: 35791261]
 14. Centers for Disease Control and Prevention. Viral hepatitis surveillance report—United States. 2021. <https://www.cdc.gov/hepatitis/statistics/2021surveillance/index.htm>. Accessed December 15, 2023.
 15. U.S. Department of Health and Human Services. 2020. Viral hepatitis national strategic plan for the United States: a roadmap to elimination (2021–2025). Washington, DC. <https://www.hhs.gov/sites/default/files/Viral-Hepatitis-National-Strategic-Plan-2021-2025.pdf>. Accessed April 30, 2024.
 16. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley & Sons; 2019.
 17. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Med Res Methodol*. 2015;15(1):30. [PubMed: 25880850]
 18. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res*. 2011;20(1):40–49. [PubMed: 21499542]
 19. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45:1–67.
 20. Everett CJ, Gebregziabher M. Estimating missing concentrations of p,p'-DDT in the 1999-2004 National Health and Nutrition Examination Survey using multiple imputation. *Environ Sci Pollut Res Int*. 2019;26(6):5885–5891. [PubMed: 30613886]
 21. Li C, Miles TP, Shen Y, et al. Measuring bereavement prevalence in a complex sampling survey: the 2019 Georgia Behavioral Risk Factor Surveillance System (BRFSS). *BMC Med Res Methodol*. 2023;23(1):138. [PubMed: 37312061]
 22. Javanbakht M, Lin J, Ragsdale A, et al. Comparing single and multiple imputation strategies for harmonizing substance use data across HIV-related cohort studies. *BMC Med Res Methodol*. 2022;22(1):90. [PubMed: 35369872]
 23. Shah AD, Bartlett JW, Carpenter J, et al. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am J Epidemiol*. 2014;179(6):764–774. [PubMed: 24589914]
 24. Stekhoven DJ, Buhlmann P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112–118. [PubMed: 22039212]
 25. Eisemann N, Waldmann A, Katalinic A. Imputation of missing values of tumour stage in population-based cancer registration. *BMC Med Res Methodol*. 2011;11(1):129. [PubMed: 21929796]
 26. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
 27. Tang F, Ishwaran H. Random forest missing data algorithms. *Stat Anal Data Min*. 2017;10(6):363–377. [PubMed: 29403567]
 28. Little RJA. Missing-data adjustments in large surveys. *J Bus & Econ Stat*. 1988;6(3):287–296.
 29. Wilson SM: Multiple imputation by chained equations with random forests. <https://cran.r-project.org/web/packages/miceRanger/index.html>. Accessed May 23, 2023.
 30. Centers for Disease Control and Prevention. National Notifiable Diseases Surveillance System. <https://www.cdc.gov/nndss/index.html>. Accessed August 17, 2023.
 31. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *J Open Source Software*. 2019;4(43):1686.
 32. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2011;30(4):377–399. [PubMed: 21225900]
 33. Van Handel MM, Rose CE, Hallisey EJ, et al. County-level vulnerability assessment for rapid dissemination of HIV or HCV infections among persons who inject drugs, United States. *J Acquir Immune Defic Syndr*. 2016;73(3):323–331. [PubMed: 27763996]

Implications for Policy & Practice

- Information of injection drug use (IDU) is frequently missing in national viral hepatitis surveillance data. Properly handling missing IDU data is critical to identify potential outbreaks, understand transmission, and better address the IDU-associated syndemic.
- This study demonstrated that the random-forest based multiple imputation by chained equations (RF-MICE) has the highest accuracy and efficiency on missing IDU data imputation compared to other approaches using the 2019-2021 acute hepatitis C data from the National Notifiable Disease Surveillance System (NNDSS).
- The RF-MICE imputation could be a valuable approach in addressing missing IDU data in the context of population-based surveillance systems like NNDSS, where obtaining complete data is challenging.

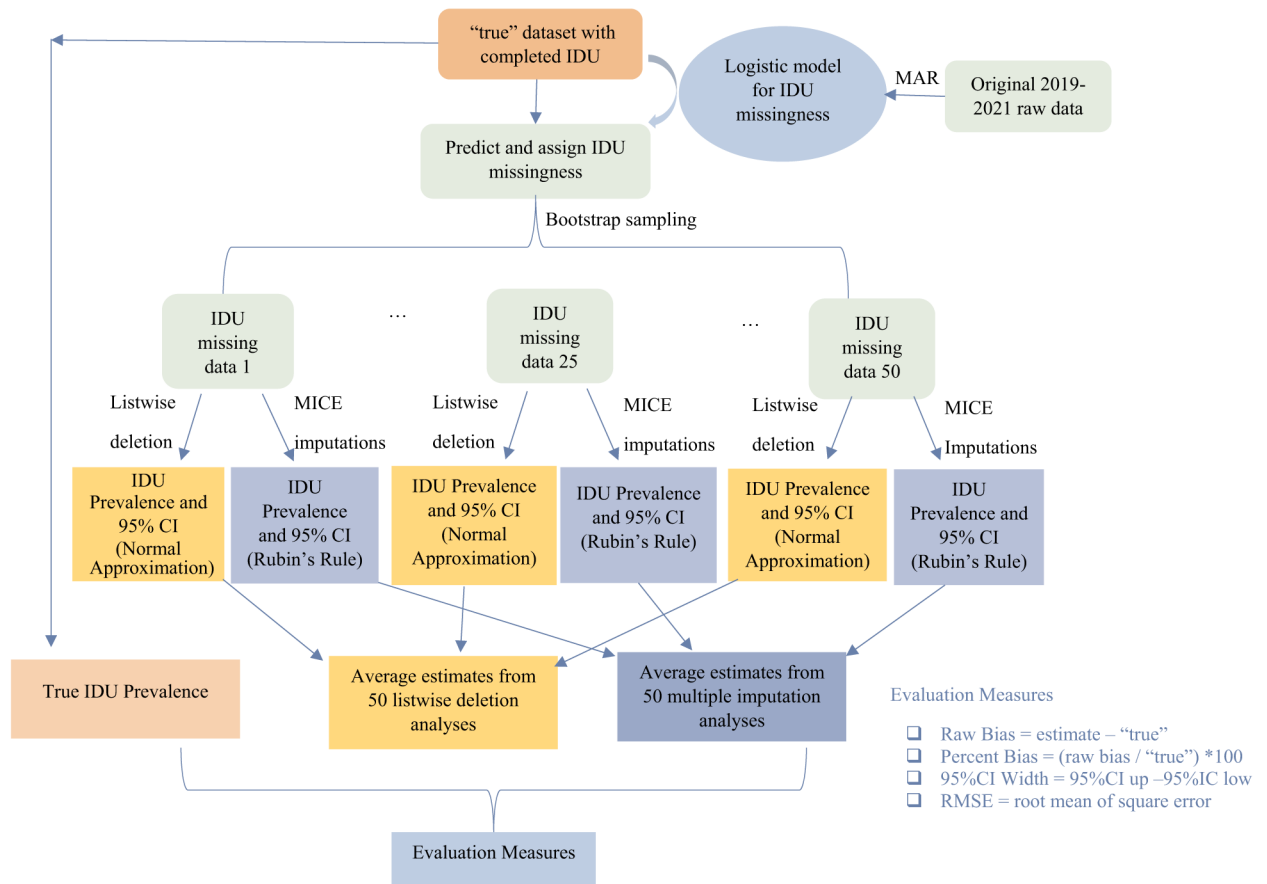


FIGURE 1. Evaluation of MICE for Estimating Prevalence of IDU Among Reported Acute Hepatitis C Cases, NNDSS. This figure is available in color online (www.JPHMP.com). Abbreviations: IDU, injection drug use; MICE, multiple imputation by chained equation; NNDSS, National Notifiable Diseases Surveillance System.

Characteristics of Reported Acute Hepatitis C Cases Overall and by IDU Missingness Status in the United States, NNDSS 2019-2021

TABLE 1

	IDU			P-value ^d
	Total	Complete (N = 4945)	Missing (N = 8907)	
IDU	(N = 13 852)			
Yes	3140 (22.7%)	3140 (63.5%)	0 (0%)	
No	1805 (13.0%)	1805 (36.5%)	0 (0%)	
Missing	8907 (64.3%)	0 (0%)	8907 (100%)	
Age groups (years)				<.001
≤19	187 (1.4%)	74 (1.5%)	113 (1.3%)	
20-29	3531 (25.5%)	1416 (28.6%)	2115 (23.7%)	
30-39	4410 (31.8%)	1721 (34.8%)	2689 (30.2%)	
40-49	2370 (17.1%)	830 (16.8%)	1540 (17.3%)	
50-59	1697 (12.3%)	498 (10.1%)	1199 (13.5%)	
≥60	1649 (11.9%)	403 (8.1%)	1246 (14.0%)	
Missing	8 (0.1%)	3 (0.1%)	5 (0.1%)	
Sex				<.001
Male	8859 (64.0%)	2966 (60.0%)	5893 (66.2%)	
Female	4969 (35.9%)	1976 (40.0%)	2993 (33.6%)	
Missing	24 (0.2%)	3 (0.1%)	21 (0.2%)	
Race/ethnicity				<.001
Non-Hispanic Native American/Alaskan Native	186 (1.3%)	68 (1.4%)	118 (1.3%)	
Non-Hispanic Asian/Pacific Islander	179 (1.3%)	48 (1.0%)	131 (1.5%)	
Non-Hispanic African American	1265 (9.1%)	286 (5.8%)	979 (11.0%)	
Non-Hispanic White	8789 (63.4%)	3686 (74.5%)	5103 (57.3%)	
Hispanic	1310 (9.5%)	373 (7.5%)	937 (10.5%)	
Other	507 (3.7%)	182 (3.7%)	325 (3.6%)	
Missing	1616 (11.7%)	302 (6.1%)	1314 (14.8%)	
Urbanicity ^b				<.001
Large central metro	4061 (29.3%)	952 (19.3%)	3109 (34.9%)	

	Total	IDU		P-value ^a
		Complete	Missing	
	(N = 13 852)	(N = 4945)	(N = 8907)	
Large fringe metro	2816 (20.3%)	1177 (23.8%)	1639 (18.4%)	
Medium metro	3013 (21.8%)	1169 (23.6%)	1844 (20.7%)	
Micropolitan	1237 (8.9%)	562 (11.4%)	675 (7.6%)	
Non-core	1012 (7.3%)	363 (7.3%)	649 (7.3%)	
Small metro	1466 (10.6%)	622 (12.6%)	844 (9.5%)	
Missing	247 (1.8%)	100 (2.0%)	147 (1.7%)	
SVI quartile ^c				<.001
Q1 (lowest vulnerability)	3325 (24.0%)	1398 (28.3%)	1927 (21.6%)	
Q2	3300 (23.8%)	1364 (27.6%)	1936 (21.7%)	
Q3	3620 (26.1%)	1167 (23.6%)	2453 (27.5%)	
Q4 (highest vulnerability)	3359 (24.2%)	916 (18.5%)	2443 (27.4%)	
Missing	248 (1.8%)	100 (2.0%)	148 (1.7%)	

Abbreviations: IDU, injection drug use; NNDSS, National Notifiable Diseases Surveillance System; Q, quartile; SVI, Social Vulnerability Index.

^aChi-squared test for the association between IDU missingness (complete or missing) and characteristics where a P-value < .05 was considered statistically significant.

^bUrbanicity was categorized according to the 2013 National Center for Health Statistics (NCHS) urban-rural classification scheme for counties and county-equivalent entities.

^cSVI reflected the Overall Social Vulnerability Index, which was the composite percentile index of 15 social and environmental domains in the 2018 CDC/ATSDR US Social Vulnerability Index data.

Prevalence of IDU Among Reported Acute Hepatitis C Cases Using LD and MICE Imputation, NNDSS 2019-2021

TABLE 2

Number and Percent (95% CI) of IDU = Yes ^a										
	LD	Parametric LR ^b		Semi-parametric PMM ^b		Nonparametric Random Forest (sRF) ^b		Nonparametric fRF ^c		
Total	3140	63.5 (62.1-64.8)	9015	65.1 (62.6-67.6)	9263	66.9 (64.2-69.6)	10524	76.0 (74.0-78.0)	11791	85.1 (84.4-85.8)
Age groups, y										
≤19	58	78.4 (67.3-87.1)	156	83.6 (75.6-91.7)	158	84.3 (77.2-91.5)	162	86.6 (80.1-93.2)	171	91.5 (87.4-95.5)
20-29	1123	79.3 (77.1-81.4)	2861	80.9 (77.7-84.2)	2879	81.5 (78.2-84.7)	3087	87.3 (85.8-88.9)	3238	91.6 (90.7-92.5)
30-39	1290	75.0 (72.8-77.0)	3315	75.1 (72.3-77.9)	3375	76.5 (73.8-79.2)	3760	85.2 (83.8-86.6)	3974	90.1 (89.2-90.9)
40-49	476	57.3 (53.9-60.7)	1464	61.7 (58.6-64.9)	1506	63.5 (59.6-67.4)	1833	77.3 (74.8-79.9)	2006	84.6 (83.1-86.1)
50-59	148	29.7 (25.7-33.9)	754	44.4 (39.4-49.4)	795	46.8 (42.4-51.3)	965	56.9 (49.8-64.0)	1289	75.9 (73.7-78.2)
≥60	42	10.4 (7.6-13.8)	464	28.2 (23.2-33.1)	550	33.3 (28.8-37.8)	716	43.4 (35.5-51.3)	1112	67.4 (64.1-70.8)
Sex										
Male	1256	63.6 (61.4-65.7)	3250	65.1 (61.8-68.4)	3338	66.9 (63.4-70.4)	3778	75.7 (73.0-78.4)	4193	84.0 (82.9-85.1)
Female	1883	63.5 (61.7-65.2)	5765	65.1 (62.4-67.8)	5924	66.9 (64.3-69.5)	6746	76.1 (74.2-78.1)	7598	85.7 (84.9-86.5)
Race/ethnicity										
Non-Hispanic Native American/Alaskan Native	55	80.9 (69.5-89.4)	176	82.7 (71.7-93.6)	166	80.9 (71.6-90.2)	169	86.2 (79.6-92.8)	174	93.0 (89.4-96.7)
Non-Hispanic Asian/Pacific Islander	11	22.9 (12.0-37.3)	73	32.2 (20.9-43.5)	84	41.1 (29.3-53.0)	147	66.2 (57.3-75.2)	131	72.9 (66.1-79.7)
Non-Hispanic African American	107	37.4 (31.8-43.3)	557	39.2 (32.4-46.0)	644	45.2 (39.8-50.6)	924	66.2 (62.2-70.1)	1057	81.9 (79.5-84.3)
Non-Hispanic White	2495	67.7 (66.2-69.2)	7004	71.0 (69.0-73.0)	7093	71.5 (69.4-73.7)	7750	78.0 (76.2-79.9)	8794	85.8 (85.1-86.5)
Hispanic	179	48.0 (42.8-53.2)	808	52.9 (46.4-59.4)	871	57.1 (50.5-63.7)	1070	70.6 (64.7-76.5)	1198	83.1 (80.4-85.9)
Other	125	68.7 (61.4-75.3)	397	66.7 (59.4-74.0)	405	69.9 (62.3-77.5)	463	78.2 (73.3-83.1)	436	86.0 (82.9-89.0)
Urbanicity ^d										
Large central metro	559	58.7 (55.5-61.9)	2287	55.6 (48.0-63.2)	2397	58.6 (51.1-66.1)	3199	75.9 (73.1-78.8)	3628	89.1 (87.8-90.5)
Large fringe metro	680	57.8 (54.9-60.6)	1842	63.6 (61.1-66.2)	1851	64.6 (61.9-67.3)	2025	71.0 (67.9-74.1)	2288	78.9 (77.3-80.6)
Medium metro	753	64.4 (61.6-67.2)	2077	68.2 (65.1-71.3)	2146	69.8 (66.3-73.2)	2344	76.6 (73.8-79.4)	2649	83.7 (82.3-85.2)
Micropolitan	444	71.4 (67.7-74.9)	1141	76.5 (73.4-79.6)	1149	76.3 (73.5-79.0)	1200	81.4 (78.6-84.2)	1286	87.6 (85.9-89.3)
Non-core	381	67.8 (63.8-71.6)	876	69.9 (65.9-73.9)	918	72.0 (68.2-75.8)	937	75.6 (72.3-78.9)	1041	84.1 (82.1-86.2)
Small metro	253	69.7 (64.7-74.4)	792	75.0 (71.3-78.7)	802	77.1 (72.9-81.2)	819	80.8 (77.4-84.2)	900	88.9 (86.9-90.8)
SVI quartile										

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Q1 (lowest vulnerability)	972	69.5 (67.0-71.9)	2500	73.2 (71.3-75.2)	2517	74.4 (72.0-76.7)	2678	79.7 (77.3-82.1)	2889	86.2 (85.1-87.4)
Q2	900	66.0 (63.4-68.5)	2273	68.2 (65.4-70.9)	2331	69.6 (67.0-72.1)	2593	77.3 (75.1-79.4)	2864	84.6 (83.4-85.8)
Q3	652	55.9 (53.0-58.7)	2202	60.5 (56.6-64.4)	2298	63.0 (58.6-67.4)	2691	72.5 (69.9-75.1)	3104	83.3 (82.0-84.6)
Q4 (highest vulnerability)	546	59.6 (56.3-62.8)	2040	58.9 (53.9-63.9)	2118	61.1 (55.8-66.4)	2561	74.8 (71.7-78.0)	2934	86.5 (84.8-88.2)

Abbreviations: CI, confidence interval; fRF, fast implementation of random forest; IDU, injection drug use; LD, listwise deletion; LR, logistic regression; MICE, multiple imputation by chained equations; NNDSS, National Notifiable Diseases Surveillance System; PMM, predictive mean matching; Q, Quartile; sRF, standard random forest; SVI, Social Vulnerability Index.

^aN denotes the number of injection drug use. Prevalence % and 95% CI were estimated based on Rubin's rule for MICE imputations.

^bMICE imputation models through LR, PMM, or sRF. A total of 29 variables including IDU, age, sex, race/ethnicity, urbanicity, SVI quartile, and other county-level auxiliary variables were included each of these imputation models.

^cMICE imputation models through fRF. A total of 134 variables including IDU, age, sex, race/ethnicity, urbanicity, SVI quartile, and other county-level auxiliary variables were initially included for the random forest selection.

^dUrbanicity was categorized according to the 2013 National Center for Health Statistics (NCHS) urban-rural classification scheme for counties and county-equivalent entities.

^eSVI reflected the Overall Social Vulnerability Index, which was the composite percentile index of 15 social and environmental domains in the 2018 CDC/ATSDR US Social Vulnerability Index data.

TABLE 3

Evaluation of the Performance of LD and MICE Imputations for Estimating Prevalence of IDU Among Reported Acute Hepatitis C Cases in the United States, NNDSS 2019-2021

	Raw Bias ^a	Percent Bias ^b	95% CI Width ^c	RMSE ^d
Dataset 1 (N = 4945, including 50 states and District of Columbia)				
LD	-9.71	15.30	3.47	9.75
LR	-4.22	6.64	4.24	4.38
PMM	-3.41	5.38	4.35	3.60
sRF	-2.68	4.22	3.90	2.89
fRF	2.03	3.19	2.75	2.22
Dataset 2 (N = 499, limited to data from 4 states where original IDU missing proportion was ≤30%)				
LD	-4.09	5.36	10.66	4.77
LR	-9.18	11.92	15.83	10.36
PMM	-3.53	5.26	13.28	4.87
sRF	1.31	3.42	8.79	3.17
fRF	2.78	4.81	7.22	4.50
Dataset 3 (N = 2268, limited to data from 23 states where original IDU missing proportion was >60%)				
LD	-13.32	23.90	5.15	13.39
LR	-5.13	9.21	6.76	5.41
PMM	-4.95	8.88	7.00	5.22
sRF	-5.20	9.33	5.86	5.43
fRF	0.02	2.71	4.24	1.89

Abbreviations: CI, confidence interval; fRF, fast implementation of random forest; IDU, injection drug use; LD, listwise deletion; LR, logistic regression; MICE, multiple imputation by chained equations; NNDSS, National Notifiable Diseases Surveillance System; PMM, predictive mean matching; RMSE, root mean square error; sRF, standard random forest.

To evaluate performance of MICE imputations under variant scenarios, 3 kinds of “true” datasets were used. Dataset 1: 2019-2021 NNDSS national acute hepatitis C dataset; Dataset 2: 2019-2021 NNDSS state-specific hepatitis C dataset, limited to 4 states with IDU missing proportion equal or less than 30%; Dataset 3: 2019-2021 NNDSS state-specific hepatitis C dataset, limited to 23 states with IDU missing proportion more than 60%. All missing IDU data were excluded when being used as the “true” datasets. A total of 29 variables including IDU, age, sex, race/ethnicity, urbanicity, SVI quartile, and other county-level auxiliary variables were included in all the evaluation analyses.

^aRaw bias is the difference of estimated IDU prevalence minus the “true” IDU prevalence.

^bPercent bias is the percentage of the raw bias over the “true” IDU Prevalence.

^c95% CI width is the difference between the upper limit and the lower limit of the 95% confidence interval around the estimated IDU prevalence.

p RMSE is the root mean square error, which measures the average magnitude of the error between estimates and true value.

TABLE 4

Sensitivity of the Performance of fRF-MICE to MNAR on Imputing Missing IDU Data, NNDSS

IDU Coefficient	Raw Bias ^a	Percent Bias ^b	95% CI Width ^c	RMSE ^d
0.0	2.03	3.19	2.75	2.22
0.1	1.86	2.93	2.68	2.03
0.2	1.42	2.24	2.69	1.58
0.5	1.19	1.89	2.70	1.39
0.8	0.89	1.57	2.72	1.19
1.0	-0.73	1.62	2.76	1.25

Abbreviations: CI, confidence interval; fRF-MICE, nonparametric implementation of random forest multiple imputation by chained equations; IDU, injection drug use; MNAR, missing not at random; NNDSS, National Notifiable Diseases Surveillance System.

The IDU coefficient (beta) measures the contribution of IDU in predicting IDU missingness. When beta = 0, there is an equal chance of IDU data being present or missing regardless of “true” IDU status. As the beta coefficient increases from .1 to 1.0, the likelihood of IDU data being missing increases among persons with “true” injection drug use. NNDSS 2019-2021 acute hepatitis C data (n = 4945, excluding missing IDU data) were used as the “true” dataset. A total of 29 variables including IDU, age, sex, race/ethnicity, urbanicity, SVI quartile, and other county-level auxiliary variables were initially included for fRF-MICE selection.

^aRaw bias is the difference of estimated IDU Prevalence minus the “true” IDU prevalence.

^bPercent bias is the percentage of the raw bias over the “true” IDU prevalence.

^c95% CI width is the difference between the upper limit and the lower limit of the 95% confidence interval around the estimated IDU prevalence.

^dRMSE is the root mean square error, which measures the average magnitude of the error between estimates and true value.