**Appendix: Multiple Imputation of Missing Data**

*Imputation of demographic variables*

We addressed missing values by multiple imputation. Overall, about 51% of all patients had missing data for at least one demographic variable. Patients with missing data differed from those with complete data by important characteristics including race/ethnicity, gender, marital status, anti-HCV testing, and anti-HCV positivity. The percentages of missing data by variable were as follows: gender (<1%), race/ethnicity (20%), marital status (29%), and income (23%). We originally planned to impute gender, race/ethnicity, and marital status for the current analysis. Census tract level income was primarily imputed to meet the needs of other researchers sharing the BEST-C retrospective dataset. We assumed that clinical variables were fully observed; i.e., if there was no documented history of hemophilia or HIV, we assumed the patient did not have the condition (or that no record of the condition was available to the physician to guide targeted testing). To maintain statistical power and to avoid potential bias in statistical estimates [[1-3](#_ENREF_1)], we performed multiple imputation to replace missing values using the method of sequential regression multivariate imputation implemented in SAS-callable IVEware version 0.1[[4](#_ENREF_4)]. We chose this method because it does not require monotone missing pattern and is also designed to accommodate imputation of variables with distributional assumptions other than normal. We assumed a missing-at-random mechanism and performed imputation separately by site [[2](#_ENREF_2)] due to variations in demographic composition, missing data patterns, and availability of variables used for imputation. To make the missing-at-random assumption plausible and to preserve correlations between variables, the imputation model for each site consisted of all analytic variables in addition to other variables from the dataset which were associated with missing information [[1](#_ENREF_1), [2](#_ENREF_2), [5-7](#_ENREF_5)] (see **Appendix Table 1**). Anti-HCV status was coded as “positive,” “negative,” or “not tested.” We specified appropriate distributions for all variables being imputed and requested 10 iterations per imputation to stabilize estimates. We determined that 10 imputed datasets were enough to yield an efficiency of at least 97% for each parameter estimate, relative to an infinite number of imputations. We checked the fit of the imputation model by re-fitting the model to each imputed dataset and plotting Pearson residuals against fitted values [[8](#_ENREF_8), [9](#_ENREF_9)]. Imputed datasets were individually analyzed with standard software and summarized per Rubin’s method with adjustment for degrees of freedom [[1](#_ENREF_1), [10](#_ENREF_10)]. We used Proc CROSSTAB (SUDAAN version 10.0.1) to analyze and summarize all proportions; Proc GLIMMIX (SAS version 9.3) was used to implement multilevel logistic regression models and estimates were combined using Proc MIANALYZE (SAS 9.3).

*Expected prevalence*

We also estimated expected anti-HCV prevalence by extending multiple imputation to anti-HCV status among patients who were not tested. Specifically, we assigned imputed values to each patient who was not tested for anti-HCV, fully conditional on all data observed for that patient. The overall percentage of patients not tested for anti-HCV was 91.6%; anti-HCV values were coded as missing for those not tested. We created 20 datasets to achieve a relative efficiency of approximately 96% [[1](#_ENREF_1)]. The same variables used to impute missing demographic variables were used to impute missing values for anti-HCV, with the exception of income which was excluded due to concerns about using census tract level income to impute anti-HCV status at the individual level. Sensitivity analysis suggested that excluding census tract level income did not have an appreciable effect on expected prevalence estimate.

|  |  |  |
| --- | --- | --- |
| **Appendix Table 1. Variables used for multiple imputation of missing data** | | |
| **Variable description** | **Distribution (specification in IVEware)** | **Applicable to which site?** |
| Gender | Categorical (dichotomous) | All |
| Marital status | Categorical (polytomous) | All |
| Race and ethnicity | Categorical (polytomous) | All |
| Census tract income | Categorical (polytomous) | All |
| Year of birth | Continuous | All |
| Approximate age in years | Continuous | All |
| HAV vaccination status | Categorical (dichotomous) | HFH, MSMC, UTH |
| HBV vaccination status | Categorical (dichotomous) | HFH, MSMC, UTH |
| HIV positive status | Categorical (dichotomous) | All |
| HIV positive status before testing | Categorical (dichotomous) | All |
| History of hemodialysis | Categorical (dichotomous) | HFH, MSMC, UAB |
| History of hemophilia | Categorical (dichotomous) | All |
| HCV test result | Categorical (polytomous) | All |
| Test for elevated ALT | Categorical (dichotomous) | All |
| Elevated ALT test result | Categorical (dichotomous) | All |
| Test for elevated AST | Categorical (dichotomous) | All |
| Elevated AST test result | Categorical (dichotomous) | All |
| History of IDU | Categorical (dichotomous) | All |
| Blood transfusion before 1992 | Categorical (dichotomous) | UTH, MSMC |
| Total number of visits | Count | All |
| Duration in months (first to last visit) | Count | All |

**References**

1. Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc **1996**; 91(434): 473-89.

2. Ma JH, Akhtar-Danesh N, Dolovich L, Thabane L, Investigators C. Imputation strategies for missing binary outcomes in cluster randomized trials. Bmc Med Res Methodol **2011**; 11.

3. Janssen KJM, Donders ART, Harrell FE, et al. Missing covariate data in medical research: To impute is better than to ignore. J Clin Epidemiol **2010**; 63(7): 721-7.

4. Raghunathan T. A Multivariate technique for multiply imputing missing values using a sequence of regression models. Surv Methodol **2001**; 27(1): 79-90.

5. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivar Behav Res **1998**; 33(4): 545-71.

6. Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Am Stat **2007**; 61(1): 79-90.

7. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Brit Med J **2009**; 339.

8. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med **2011**; 30(4): 377-99.

9. Abayomi K, Gelman A, Levy M. Diagnostics for multivariate imputations. J R Stat Soc C-Appl **2008**; 57: 273-91.

10. Barnard J, Rubin DB. Small-sample degrees of freedom with multiple imputation. Biometrika **1999**; 86(4): 948-55.