Evaluating Imputation Methods and Software for Missing Vaccination Data in the National Immunization Survey

Taylor Lewis and Meena Khare[1]
National Center for Health Statistics, 3311 Toledo Road, Room 3115, Hyattsville, MD 20782

Abstract

The National Immunization Survey (NIS) is designed to collect vaccination information for U.S. children between the ages of 19 and 35 months, but due to nonresponse roughly 30% of the vaccination data is unascertained. The current method to compensate for missing provider data involves weighting adjustments. Using data from the 2008 NIS public-use data file, this paper evaluates four alternative imputation techniques and comments on some advantages and disadvantages of specific software used. All four imputation methods yielded slightly lower vaccination rates, although differences are small in magnitude. Single imputation is shown to underestimate standard errors as compared to the current weighting method and the other three multiple imputation (M = 5) methods. Standard errors between the three multiple imputation methods were all similar and mirror those from the weighting method.

Keywords: NIS, multiple imputation, missing data, weighting, nonresponse

1. Background

Since 1994, the Centers for Disease Control and Prevention (CDC) has sponsored the National Immunization Survey (NIS) to provide annual vaccination coverage estimates for children 19 to 35 months old residing in U.S. households. In the first stage of the survey, a list-assisted, random-digit dial (RDD) telephone sample is fielded for 67 geographical estimation areas (public-use variable ESTIAP08), strata, to screen households with one or more age-eligible children. The initial telephone interview captures numerous socio-demographic characteristics about the child and the mother, and concludes with a request for provider contact information where, in the second stage of data collection, an immunization history questionnaire is mailed to the identified provider(s) to collect the child's vaccination data. The providers are deemed the most reliable source of this information after a low correlation was discovered between household reports of vaccination and provider records (Ezzati-Rice et al., 1996; Khare et al., 2001a; Khare et al., 2001b).

Unfortunately, vaccination data are unascertained for roughly 30% of the children for whom data were collected at the RDD stage, because parents refuse to disclose or cannot furnish provider contact information or the provider does not respond even after contact information is obtained. Brick and Kalton (1996) refer to this form of missing data as "partial nonresponse," a murky middle ground between unit and item nonresponse where both weight adjustments and imputation seem viable compensation techniques. The current method compensating for provider nonresponse involves a detailed, multi-step weight adjustment method (see section 6.5 of Data User's Guide, U.S. Department of Health and Human Services, 2009), but we were curious to determine whether alternative imputation methods would produce substantively different results. A literature search on considerations when facing the partial nonresponse dilemma proved somewhat unsatisfactory, although Kalton (1986) suggests that, in the context of panel survey

[1] The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

nonresponse, imputation might be preferred over weight adjustment methods if known covariates are highly predictive of missing values. Previous in-house analyses suggested few noteworthy relationships between covariates used in the weight adjustments and vaccination status variables, and since the current imputation analysis draws from the same covariate pool, we did not expect wildly different point estimates. On the other hand, we were curious to determine what effect on measures of uncertainty the imputation methods might produce.

Section 2 describes the analysis data set in more detail and also outlines the specific imputation methods and software employed. Section 3 compares vaccination estimates under imputation and standard errors alongside nonresponse-adjusted weighted estimates. Section 4 concludes with reflections about the specific missing data software used in our research, and overall limitations.

2. Methods

2.1 Missing Vaccination Data

The key outcome variable susceptible to provider nonresponse that we consider in this paper is actually a composite of five distinct vaccine dose indicator variables. We focused on the up-to-date (UTD) status of a child in terms of the 4:3:1:3:3 series—meaning the child has received 4+ doses of diphtheria, tetanus acellular pertussis (DTaP), 3+ doses of poliovirus vaccine (polio), 1+ doses of a measles-containing vaccine (MCV), 3+ doses of *Haemophilus Influenza* type b (Hib), and 3+ doses of hepatitis B (HepB). It should be noted that while the key variable of interest is a combination of five other variables, the pattern of missingness for all six variables simultaneously is dichotomous. That is, all five vaccine doses (and, hence, the composite variable) were either all known or all missing. Because of this pattern, we chose to focus on imputing the composite UTD status for the 4:3:1:3:3 series in lieu of the underlying sequence of vaccine doses.

All data used in this paper come from the 2008 NIS public-use file[2]. The data set contains 25,948 observations and two sets of weights. The first analytic weight, RDDWT, weights the household-level responses obtained for 25,948 children after the first stage of the survey to the target population. The second weight, PROVWT, adjusts RDDWT for the subset of 18,430 children yielding complete vaccination data as obtained from providers[3]. There were also 85 respondents with partially complete provider data treated as missing for the provider weight adjustments, yet containing non-missing values for the key 4:3:1:3:3 UTD status variable. For our imputation analysis, these individuals were treated as respondents. Thus the rate of missingness is (25,948 – 18,430 – 85)/25,948 = 28.6%. Of this proportion missing, it is estimated that two-thirds of the cases are missing due to parent/guardian refusal or inability to supply provider contact information, and the other third because of provider non-contact or nonresponse (p. 43 of Data User's Guide, U.S. Department of Health and Human Services, 2009). It is certainly possible that different covariates are associated with the different reasons for missing data, but in this analysis we adjusted for all provider nonresponse in one step.

The available covariates for the 7,433 cases with missing 4:3:1:3:3 status that we felt best described the missingness mechanism included age of child, indicator of being first-born, race/ethnicity of child, sex of child, the mother's education level, the mother's marital status, the total number of children in the household, whether or not the child had a shot card (written record of child's vaccine doses given by provider to parent/guardian), and household income and poverty status indicator. There was a small degree of missingness present in these nine covariates

[2] Available at http://www.cdc.gov/nis/datasets.htm
[3] This includes 151 children who are voluntarily unvaccinated.

after data collection; a hot-deck imputation procedure was implemented to populate these data in the development of the public-use file. Because of the minor rates of missingness[4], these covariates were treated as known.

The other complicating factor in devising the imputation schemes was how to handle the complex sample structure of the survey data. Reiter et al. (2006) demonstrated the importance of modeling strata and clusters during imputation under complex surveys. NIS does contain clustering, but the rate is negligible: only 860 of the 25,076 (3.4%) distinct households had more than one age-eligible child. The bigger concern was stratification. The sample design involves 67 distinct estimation areas (strata, ESTIAP08 variable in the public-use file), each with a number of children ranging from 155 to 515. To allow for sufficient sample size needed to model the missingness mechanism, strata were collapsed with other geographically neighboring strata to form the ten Department of Health and Human Services regions. The ultimate respondent counts within region became between 1,295 and 4,231 children.

Under the stochastic view of nonresponse, where all potential respondents have a known, non-zero probability (propensity) of responding, ρ_i, the nonresponse bias that can impact unadjusted means is proportional to the covariance between the propensities and the outcome variable, y_i: $\frac{1}{N\bar{\rho}}\sum_{i=1}^{N}(\rho_i - \bar{\rho})(y_i - \bar{y})$ (see expression 1.2.3 in Bethlehem, 2002). A successful imputation scheme to mitigate this bias thus partitions respondents and nonrespondents into cells where the ρ_i*'s* and/or y_i's are very similar. Within such cells, the argument can be made that the data are then missing completely at random (Rubin, 1987), and imputing the missing nonrespondent's values with randomly chosen respondents is a legitimate compensation technique. Of course, the challenging, unverifiable task is forming cells allowing this assumption to hold. In the following subsection, we outline four distinct methods we evaluated.

2.2 Imputation Methods

The first imputation method was a standard hot-deck procedure carried out in the SOLAS software (Statistical Solutions, 2001). Donor cells were created by the cross-classification of aforementioned covariates. To the extent this cross-classification differentiates response propensities *and* the 4:3:1:3:3 UTD status distribution, nonresponse bias can be reduced. But the fact that a single, hot-deck imputation underestimates variances is well established. To account for this, we also compared a series of multiple imputation methods (Rubin, 1987), where missing 4:3:1:3:3 UTD status was filled in not once, but $M = 5$ times.

The second imputation method utilized IVEware, a SAS-callable set of macros developed by the Institute for Social Research at the University of Michigan (Raghunathan et al., 2001, 2002). The strategy here is to perform logistic regression on the 4:3:1:3:3 UTD status based on observed data and multiply impute for the unobserved data. After fitting the model, random parameter values are drawn from their posterior predictive distributions for each imputation independently and an estimated probability of being UTD is computed. This predicted probability is then evaluated against a random number chosen from a uniform [0, 1] distribution. If the random number is less than the predicted probability, the case is imputed as UTD; otherwise, the case is imputed as not UTD.

[4] Race/ethnicity of child was missing for 7.4% of cases, but all others covariates were missing less than 3%.

This method primarily seeks to minimize the $(y_i - \bar{y})$ portion of the bias formula above. If a predictive model can be constructed, this will at least reduce the variance of estimates; if the covariates in the model are also related to the propensities, bias reduction can be achieved as well (see Table 1 of Little and Vartivarian, 2005). We also comment here that, when assessing the model fit, it was determined there were significant interactions between region and the set of covariates. To account for this, ten separate models were fit, one for each region, prior to multiply imputing (*M* = 5) missing 4:3:1:3:3 UTD status.

The third imputation method employed was SOLAS' propensity score method, which first groups respondents into five cells based on ranked, estimated response propensities $\hat{\rho}_i$ from a logistic regression of a response indicator using the covariates as explanatory variables. Hence, this method primarily seeks to minimize the $(\rho_i - \bar{\rho})$ portion of the bias formula above, which gets more mileage in that it is applicable to any *y*, although our focus here remains strictly with 4:3:1:3:3 UTD status. Maintaining region has an independent, categorical predictor suggested a better model fit than including region interaction terms. Once grouped into quintiles, SOLAS applies the Approximate Bayesian Bootstrap (ABB) to properly account for the variance of the imputation model and multiply impute 4:3:1:3:3 UTD status *M* = 5 times.

Outlined by Rubin and Schenker (1986), the ABB is a nonparametric analog to drawing model parameters from their posterior predictive distribution as one would if using an explicit model. It is prudent to incorporate extra variability into a simple hot-deck since the hot-deck does not account for uncertainty in modeling the missing data mechanism—in Rubin's (1987) terminology, doing so allows for "proper" imputation.

The two-stage ABB proceeds as follows. If, within a cell, we define n_1 respondents and n_0 nonrespondents, each comprised of data vectors Y_{obs} and Y_{mis}, respectively, the first stage selects a vector of size n_1 with replacement from Y_{obs}. Using this new set, the second stage involves selecting n_0 values with replacement to impute the vector of missing outcome variables, Y_{mis}. This process is then repeated independently *M* times.

Our fourth imputation method attempted to incorporate the weights into the ABB imputation scheme. Like Method 2, we first performed a logistic regression on an indicator variable of response based on the covariates and grouped the propensities into quintiles. We experimented with a modified ABB, however, to impute the missing values. The first stage of the ABB was maintained, where a sample of n_1 units are selected from Y_{obs} with replacement. But instead of selecting an equal probability sample with replacement, Y_{mis} was filled in by selecting n_0 elements with probability proportion to size (PPS) with replacement, where the measure of size was RDDWT. This routine is essentially the "weighted random hot-deck" evaluated by Andridge and Little (2009) in an application involving data from the third National Health and Nutritional Examination Survey (NHANES III), originally motivated by Rao and Shao (1992). We wrote a SAS program to independently implement this procedure to produce *M* = 5 imputations.

Once the fully imputed data sets were created, all were merged into one SAS data set and PROC SURVEYMEANS was used to compute weighted estimates and standard errors, the latter approximated from the procedure's default Taylor series method (SAS Institute, 2008). For NIS weighted estimates, the PROVWT variable was used; for all imputation-derived estimates, RDDWT was used. For all analyses, the strata indicator variable was ESTIAP08 and the cluster variable SEQNUMHH, the household identifier. Rubin's rule (1987) for combining between- and within-imputation variance was applied via PROC MIANALYZE, also in SAS.

3. Results

Table 1 demonstrates a broad finding that imputed values derived from all four methods tended more toward children being not up-to-date with respect to the 4:3:1:3:3 series. That is, the covariate patterns of children whose vaccination data was missing matched somewhat closer the covariates of children who vaccination statuses were ascertained but who were not up-to-date. The differences, however, were small on an absolute scale—from the unweighted comparison below, one can note none are greater than two percentage points.

Table 1. Unweighted Percentages of Observed and Imputed 4:3:1:3:3 Up-To-Date Status

Imputation Method	Observed	Imputed
M1. Hot-Deck	78.5	77.1
M2: IVEware	78.5	76.3
M3: PUTD43133_SOLAS_PROP_5	78.5	76.6
M4: Propensity with PPS ABB	78.5	76.9

Figure 1 is another broad comparison like Table 1, though we now account for weights and show bounds of uncertainty via 95% confidence intervals. Though global in nature, this aptly summarizes our findings overall and within various one- and two-way breakouts examined. We tended to observe all four imputations methods led to slightly lower coverage estimates in comparison with the nonresponse-adjusted NIS estimates. Again, however, any differences were small on an absolute scale.

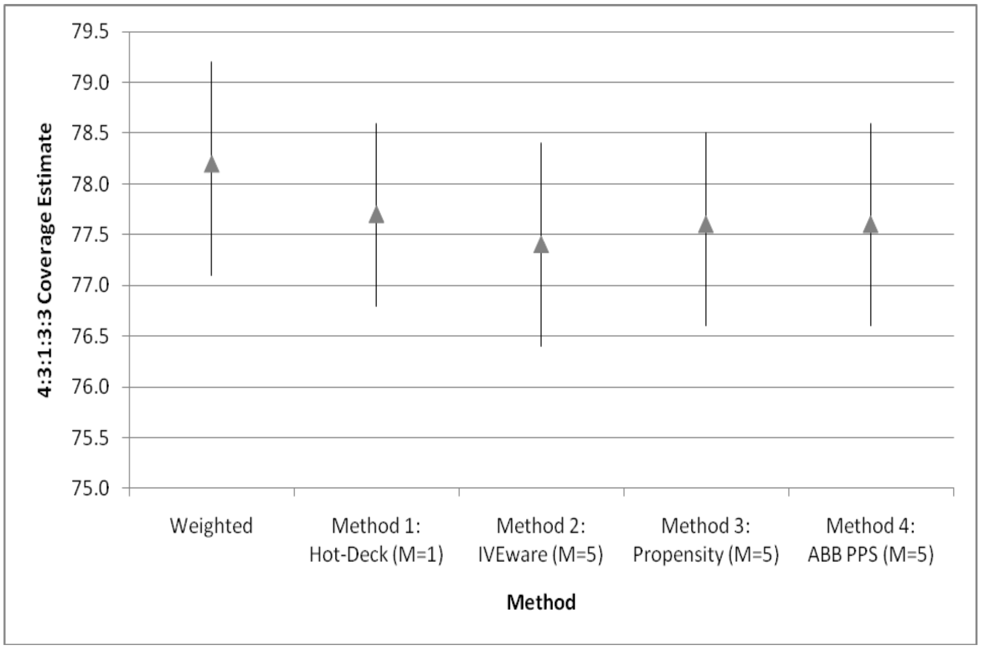

Figure 1. Nationwide 4:3:1:3:3 Up-to-Date Status Point Estimates and 95% Confidence Intervals by Missing Data Compensation Method

Figure 2a plots stratum-level standard error differences in 4:3:1:3:3 coverage estimates for each income and poverty status domain between the weighted and hot-deck methods. The fact that many of the points are above the horizontal zero difference line demonstrates the hot-deck's smaller standard errors in comparison with the weighted. The majority of points fall within a downward bias of 0.02.

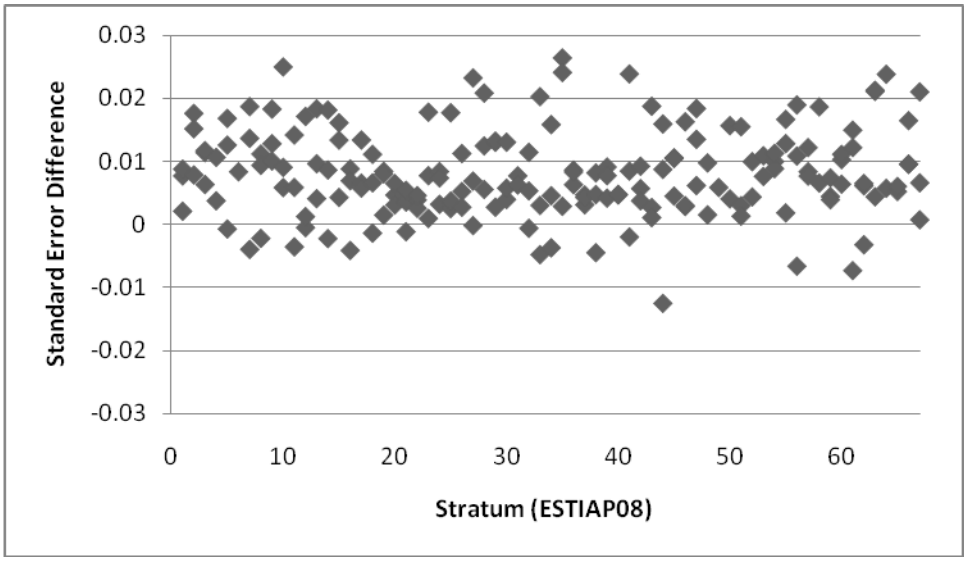

Figure 2a. Stratum-Level Standard Error Differences between NIS Weighted and Method 1 for 4:3:1:3:3 Status within Income/Poverty Status Domains.

In contrast, Figure 2b shows the same standard error differences between Methods 2 and 3. No pattern is evident, which reflects a general finding that standard errors between all multiple imputation methods were very similar. Though not shown, the standard errors between all multiple imputation methods were also very similar to those from the weighted estimates. Together, these findings show how the between-imputation variance term of Rubin's formula incorporates the needed additional uncertainty not achieved from single imputation and that we would not gain or lose precision moving from nonresponse-adjusted weights to multiple imputation.

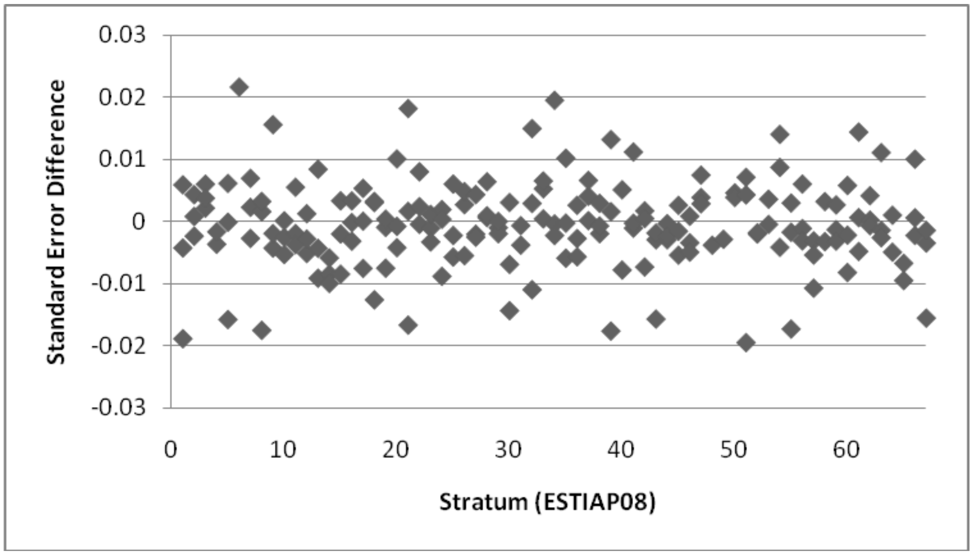

Figure 2b. Stratum-Level Standard Error Differences between Method 2 and Method 3 for 4:3:1:3:3 Status within Income/Poverty Status Domains.

We found the two propensity model methods, whether or not we modified the donor pool selection process to PPS, produced remarkably similar results compared to deviations between more disparate multiple imputation methods. Figure 3a plots the differences of stratum-level 4:3:1:3:3 child race/ethnicity domain estimates between Method 2 and Method 3. Figure 3b plots the same estimate differences on a common scale between Method 3 and Method 4. The differences in Figure 3b appear narrower, laying closer to the line of no difference than their counterparts in Figure 3a.

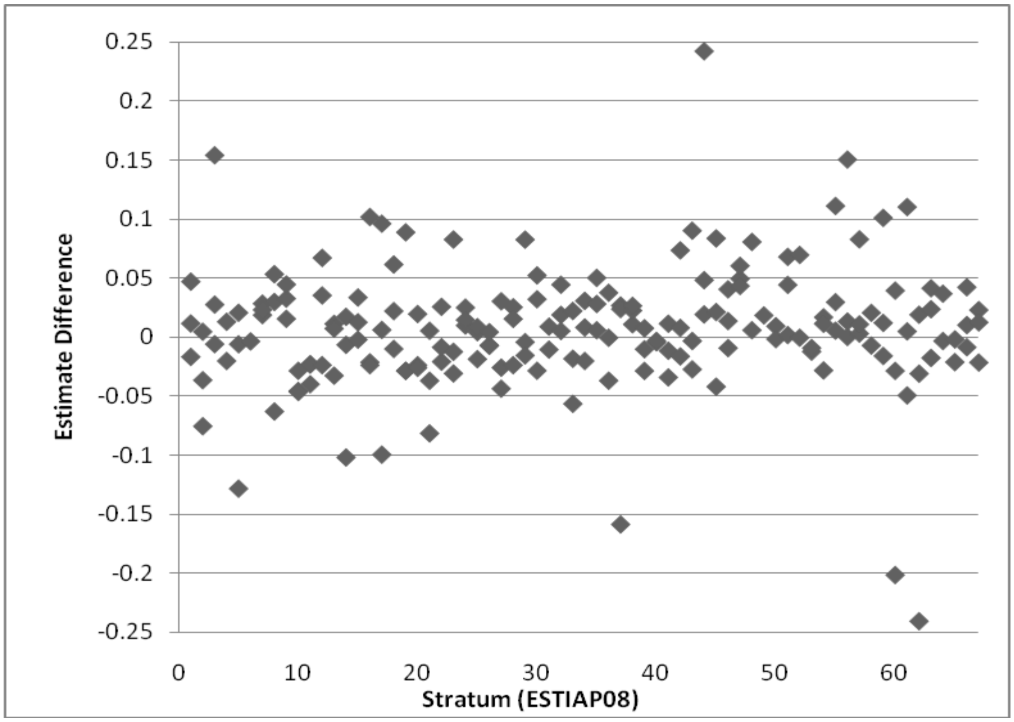

Figure 3a. Stratum-Level 4:3:1:3:3 Status Estimate Differences between Method 2 and Method 3 for Child Race/Ethnicity Domains.

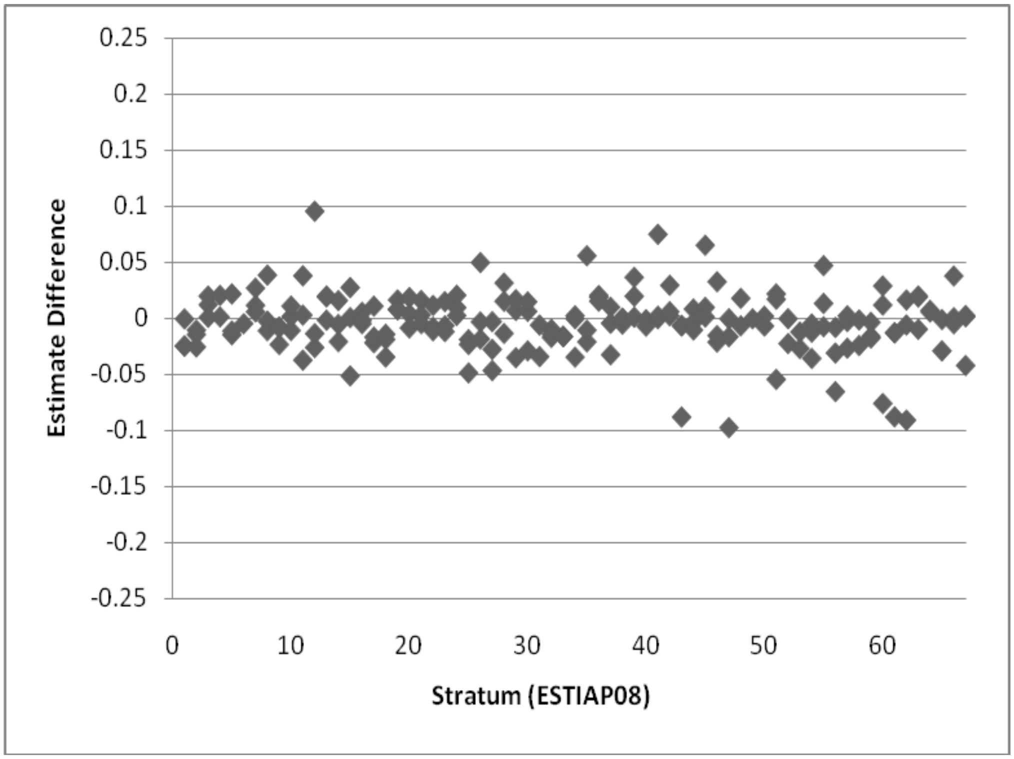

Figure 3b. Stratum-Level 4:3:1:3:3 Status Estimate Differences between Method 3 and Method 4 for Child Race/Ethnicity Domains.

4. Conclusion

4.1 Comments on Software

We acknowledge our focus in this research project was limited to methods geared toward monotone missingness for a single binary outcome variable and the use of SOLAS and IVEware, but certain deficiencies pervaded, even as we considered other routines (e.g., SAS's PROC MI, SUDAAN's PROC HOTDECK).

We were initially pleased to take note of the myriad resources available for carrying out imputation. Nearly every major statistical software has some built-in or user-created routine, so no matter what platform an analyst tends to utilize, there is likely a way to combat the missing data problem locally. What was astonishingly evident, however, is that there is great diversity in the types of methods offered; in fact, there is ambiguity in directly comparing two softwares' results because even seemingly uniform routines often differ by one or more subtleties.

The first broad comment is that there is a dearth of diagnostics output when the imputation routine concludes. There is a fine line between too much automation and not enough, and we certainly would not condone imputing via a "black box," but specious results can occur—especially when sample sizes for cells get small. An example diagnostic one auther prefers to inspect is the distribution of estimated response propensities, which may suggest additional refinements are necessary or that quasi-or complete-separation in the propensity model may be wreaking havoc. The ability to, say, output the propensities and/or propensity class indicators for a subsequent analysis would be helpful, along with the ability to modify propensity classification rules. We should point out that SOLAS does offer some leeway in propensity class adjustments and further allows for a refinement variable to match donors and recipients in a subset of the intial propensity class donor pool, potentially useful in the sample weight quandary. But, again, there is no easy way to output the propensities or verify donor pool characteristics.

An issue related to diagnostics is that most softwares do not insert imputation flag variables with the completed data set. These flags would be a welcomed inclusion, avoiding an additional programming step comparing the completed with original data sets.

Lastly, we were rather surprised to find no way to incorporate a straightforward (proper) hot-deck employing the ABB. Admittedly, this was not too arduous a task to manually program, but we found that the ABB was generally only callable as an embedded component within another routine. For instance, SOLAS utilizes the ABB but only as part of the propensity score method. To its credit, SOLAS does offer a straightforward single imputation hot-deck routine with a convenient cell-collapsing algorithm. Yet even if one repeated this independently *M* times, the variance will still be underestimated since the method does not explicitly employ the ABB.

4.2 Limitations and Further Research

The purpose of this research was to test alternative methods compensating for the fact that in the NIS, nearly 30% of all age-eligible children have unknown 4:3:1:3:3 UTD status. We evaluated four imputation methods against the current technique of adjusting the weights for provider nonresponse. We were interested in determining whether the point estimates and standard errors differed in any notable way when we attempted to fill in the missing data.

A major limitation with the present analysis is that, while the set of covariates between all four imputation methods is common, certain covariates were used in weighting but not imputation, and vice versa. For example, mother's race/ethnicity and metropolitan statistical area status were employed during the weighting process but are not available on the public-use data file. Thus, we can never know for certain which portion of any difference observed is attributable to the method and which to the varying set of covariates. Secondly, we effectively treat the RDDWT and PROVWT values as known and fixed when using Taylor series linearization to approximate variances, but different results may have been concluded during analyses of standard errors if we instead created replicate weights and applied the same weighting process to all replicates (Valliant, 2004).

Acknowledging the points above, we report a barely detectable tendency of point estimates from imputation to drift downward as compared to the current weighted estimates, which agrees with earlier findings by Khare and Yucel (2003) who, coincidentally, also used a slightly different set of covariates. It is somewhat unsurprising that point estimates varied little, given there are few variables in the public-use file that have extremely strong relationships with 4:3:1:3:3 UTD status. Unfortunately, no gold standard is available to conclude which missing data method achieves results nearest the truth.

Although this paper includes no formal tests of significance, it was tempting to conclude insignificance from Table 1 based on the fact that all four methods' confidence intervals overlap with that of the weighted estimate. However, as Schenker and Gentleman (2001) point out, such a judgement is akin to testing whether the difference in means is greater than the sum of the standard errors, when in fact the difference should be compared against the square root of the sum of squared standard errors. Moreover, since each method's point estimate is highly correlated, obvious when considering nearly 70% observations in the analysis dataset are identical, there is an ignored (positive) covariance term that would reduce a difference's variance even further.

Hence, even though we commented that differences in Table 1 are minuscule, we do not necessarily imply they lack statistical significance.

We confirmed the downward bias of standard errors from single imputation: nearly all stratum-level estimates had standard errors less than weighted, but typically differed no more than 0.02. Interestingly, in terms of the three multiple imputation (M = 5) methods compared, any efficiencies from a completely filled-in data set were counterbalanced by the between imputation variance component of Rubin's formula. In summary, it appears the multiple imputation methods do not provide an advantage over current weight adjustments.

Lastly, we cannot decipher much of a relationship between RDDWT and 4:3:1:3:3 UTD status. We speculate the efficacy of an amended ABB procedure such as Method 4 likely lies on a strong correlation between the survey weight and outcome variable, such as might be the case in PPS sampling. Without that relationship, any methods to incorporate the weights may lead to inefficiencies with no corresponding bias reduction. Toward the conclusion of the present research we discovered the recent work of Andridge and Little (2009, 2010) who suggest a better way to incorporate sample weights, at least in hot-deck imputation, is to use the sample weight as a stratifyer alongside auxiliary variables in the formation of cells, but after doing so, proceed as if there were no weights. This modification could serve as an avenue of future research.

References

Andridge, R., and Little, R. (2009). "The Use of Sample Weights in Hot Deck Imputation." *Journal of Official Statistics*, 25, pp. 21 – 36.

Andridge, R., and Little, R. (2010). "A Review of Hot Deck Imputation for Survey Non-Response." *International Statistical Review*, 78, pp. 40 – 64.

Bethlehem, J. G. (2003). "Weighting Nonresponse Adjustments Based on Auxiliary Information," Chapter 18 in Groves, R., Dillman, D., Eltinge, J., and Little, R. (eds.) *Survey Nonresponse*. New York: Wiley, 2002.

Brick, J.M. and Kalton, G. (1996). "Handling Missing Data in Survey Research." *Statistical Methods Medical Research*, 5, pp. 215 – 238.

U.S. Department of Health and Human Services (DHHS). National Center for Health Statistics. The 2008 National Immunization Survey. Hyattsville, MD: Centers for Disease Control and Prevention, 2009. Data User's Guide available online at ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NIS/NISPUF08_DUG.pdf .

Ezzati-Rice, T., Zell, E., Massey, J., and Nixon, M. (1996). "Improving the Assessment of Vaccination Coverage Rates with the Use of Both Household and Medical Provider Data." *Proceedings of the Survey Research Methods Section of the American Statistical Association*. Alexandria, VA.

Kalton, G. (1986). "Handling Wave Nonresponse in Panel Surveys." *Journal of Official Statistics*, 2, pp. 303-314.

Khare, M., Battaglia, M., Stockley, S., Wright, Robert A., and V. Huggins (2001a). "Quality of Immunization Histories Reported in the National Immunization Survey." *Proceedings of the International Conference on Quality in Official Statistics.* Stockholm, Sweden, May 14 – 15.

Khare, M., Ezzati-Rice, T., Battaglia, M., and Zell, E. (2001b). "An Assessment of Misclassification Error in Provider-Reported Vaccination Histories." *Proceedings of the Survey Research Methods Section of the American Statistical Association*. Alexandria, VA.

Khare, M., and Yucel, R. (2003). "A Comparison of Conventional Estimates of Vaccination Coverage with Estimates Obtained from Multiply Imputed Datasets Using Software Available for Multiple Imputation." *Survey Research Methods Section of the American Statistical Association.* Alexandria, VA.

Little, R., and Vartivarian, S. (2005). "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology,* 3, pp. 161–168.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J, and Solenberger, P. (2001). "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology,* 27, pp. 85 – 95.

Raghunathan, T., Solenberger, P., and Van Hoewyk, J. (2002). *IVEware: Imputation and Variance Estimation Software User Guide*. Ann Arbor, MI.

Rao, J., and Shao, J. (1992). "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation." *Biometrika,* 79, pp. 811 – 822.

Reiter, J., Raghunathan, T., and Kinney, S. (2006). "The Importance of the Sampling Design in Multiple Imputation for Missing Data." *Survey Methodology*. 32, pp. 143 - 150.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York, NY: Wiley.

Rubin, D.B., and Schenker, N. (1986). "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association,* 81, pp. 366 – 374.

SAS Institute. (2009). *SAS/STAT User's Guide, Version 9.2*. Cary, NC.

Schenker, N., and Gentleman, J.F. (2001), "On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals," *The American Statistician*, 55, pp. 182-186.

Statistical Solutions. (2001). *SOLAS 3.0 Imputation User Reference Manual*. Saugus, MA.

Valliant, R. (2004). "The Effect of Multiple Weight Adjustments on Variance Estimation." *Journal of Official Statistics*, 20, pp. 1-18.