VERSITA

Journal of Official Statistics, Vol. 30, No. 1, 2014, pp. 147–161, http://dx.doi.org/10.2478/jos-2014-0008

The Relative Impacts of Design Effects and Multiple Imputation on Variance Estimates: A Case Study with the 2008 National Ambulatory Medical Care Survey

Taylor Lewis[1], *Elizabeth Goldberg*[1], *Nathaniel Schenker*[1], *Vladislav Beresovsky*[1], *Susan Schappert*[1], *Sandra Decker*[1], *Nancy Sonnenfeld*[1], *and Iris Shimizu*[1]

The National Ambulatory Medical Care Survey collects data on office-based physician care from a nationally representative, multistage sampling scheme where the ultimate unit of analysis is a patient-doctor encounter. Patient race, a commonly analyzed demographic, has been subject to a steadily increasing item nonresponse rate. In 1999, race was missing for 17 percent of cases; by 2008, that figure had risen to 33 percent. Over this entire period, single imputation has been the compensation method employed. Recent research at the National Center for Health Statistics evaluated multiply imputing race to better represent the missing-data uncertainty. Given item nonresponse rates of 30 percent or greater, we were surprised to find many estimates' ratios of multiple-imputation to single-imputation estimated standard errors close to 1. A likely explanation is that the design effects attributable to the complex sample design largely outweigh any increase in variance attributable to missing-data uncertainty.

Key words: Health survey; missing data; item nonresponse; fraction of missing information.

1. Background

The National Ambulatory Medical Care Survey (NAMCS) has been administered by the National Center for Health Statistics (NCHS) since 1973. While aspects of the sample design and survey instrument have evolved over the past twenty-five years, its objective has always been to collect and disseminate nationally representative data on office-based physician care. The ultimate sample unit is a doctor-patient encounter, drawn systematically from the terminus of a multistage, clustered sample design. Like many other surveys, the NAMCS is not immune to the potentially detrimental effects of missing data. As Figure 1 demonstrates, the (unweighted) item nonresponse rate for patient race, one of the most analyzed demographics, increased appreciably between 1999 and 2008. Such nonresponse on race has been experienced in the context of other NCHS health care surveys as well. For example, Kozak (1995) found that hospitals participating in the National Hospital Discharge Survey underreported race to varying degrees.

[1] National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782, U.S.A
Email: tlewis@survey.umd.edu
Acknowledgments: The findings and conclusions in this article 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. The authors thank the Editor, Associate Editor, and referees for comments that helped to improve the article.

Groves et al. (2002, Sec. 1.2) cited three issues that can arise with missing data due to nonresponse: (1) biases in point estimators; (2) inflation of the variances of point estimators; and (3) biases in customary estimators of precision. In this article, we focus on the third issue, and in particular the extent to which multiple imputation (Rubin 1987) results in estimates of precision that differ from those under single imputation in the context of the NAMCS with missing data on race.

Variance estimates for situations such as ours have been explored by Li et al. (2004), who used a bootstrap re-imputation scheme adapted to complex surveys (Shao and Sitter 1996) to account for missing-race uncertainty in the 2000 NAMCS. Li and her colleagues observed a few instances where the bootstrap re-imputation suggested standard errors should be inflated by up to 30%, but concluded most estimates necessitated an inflation of 6% or less. Their findings quelled concerns for a while, but as one can infer from Figure 1, the item nonresponse rate for race in the 2000 NAMCS was roughly half where it stood in 2008.

This article reports on research conducted at NCHS, using data from the 2008 NAMCS, to assess whether multiple imputation would better reflect the missing-data uncertainty than single imputation, which is currently used in the NAMCS, in light of the recent nonresponse rates of about 30% on race. Using a model-based imputation method with predictors similar to those used in the 2008 NAMCS cell-based procedure, we compared results under multiple imputation to those under single imputation, and we found that the increase in the estimated standard errors with multiple imputation tended to be small. We concluded that the extremely large design effects (Kish 1965) for estimates involving race tended to transcend the additional missing-data uncertainty that would be reflected by multiple imputation. This is discussed with the help of some basic theory partitioning the overall estimated variance increase into a component attributable to the complex survey design and a component attributable to missing-data uncertainty.

Section 2 of the article provides an overview of the NAMCS sample design and describes the imputation method used in our study. In Section 3, we present the major results from the comparison of multiple imputation with single imputation. Section 4 concludes the article with a brief discussion pointing out limitations and suggestions for further research.

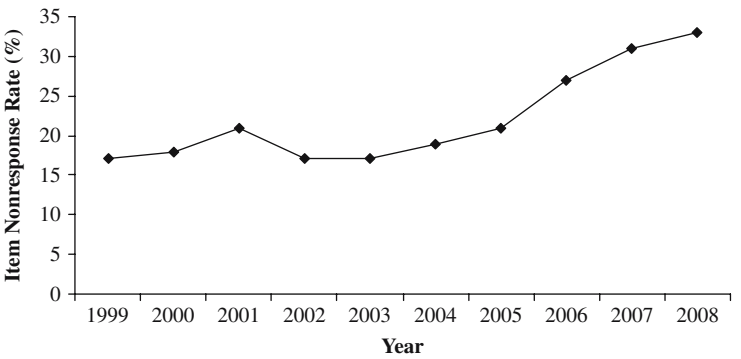

Fig. 1. Patient Race Item Nonresponse Rate Trend in the National Ambulatory Medical Care Survey, 1999 – 2008.

2. Data and Methods

2.1. NAMCS Sample Design

As previously noted, the NAMCS employs a multistage, clustered sample design. The primary sampling units (PSUs) consist of either single or grouped counties (or their equivalent), derived from a probability subsample of 112 PSUs from the 1985–1994 National Health Interview Survey (NHIS) design period. Within these PSUs, lists of non-federally employed physician practices obtained from the American Medical Association and American Osteopathic Association are stratified into fifteen specialty groups. A sample of physician practices is then selected from each stratum and randomly allocated into 52 subsamples, each corresponding to a week within the data collection period, the calendar year.

NCHS contracts with the U.S. Census Bureau to collect the patient visit information from sampled practices. Prior to their assigned one-week collection period, field representatives (FRs) meet with the physician or, more commonly, the physician's administrative staff, and analyze the expected count of pending patient visits. Based on this information, a systematic sampling interval is determined and utilized such that approximately thirty visits are selected over the course of the week. FRs try to recruit and train office staff to collect the sampled visits' data in real time, but more than half of the patient record forms (PRFs) are filled out by the FR using maintained patient files after the weeklong data collection period has concluded.

According to the 2008 NAMCS public-use data file documentation (NCHS 2009), a total of 3,319 physicians were selected, of whom 1,090 were ruled ineligible. Aside from having retired, common causes for ineligibility include a physician practicing in an institutional setting or as part of an emergency department outpatient facility. Of the 2,229 eligible physicians, 1,334 were contacted and agreed to participate, although 201 saw no patients during the data collection period randomly assigned. In the end, data were collected for 31,146 distinct visits. This number includes data from a supplemental sample of community health centers (CHCs) drawn with assistance from the Health Resources Services Administration and the Indian Health Service, of which a portion involved visits to non-physicians (e.g., nurse practitioners). Non-physician visits are excluded from the public-use file, which explains why the number of visits contained in the 2008 NAMCS public-use file (28,741) is fewer than analyzed in this article.

To compensate for the differential patient visit selection probabilities and physician-level nonresponse, a four-step weighting procedure yielded a final set of weights that can be used to better represent the target population. For more details on the weighting process, refer to Section I.K of NCHS (2009).

In addition to unit nonresponse caused by the fact that not all sampled physicians participate, the NAMCS is subject to item nonresponse in the returned PRFs. Some variables are more susceptible to missingness than others. Whereas most items' nonresponse rates are less than five percent, Section I.I.3 of NCHS (2009) lists specific rates for variables where the item nonresponse rate exceeds that threshold. Patient race has one of the highest rates: Of the 31,146 visits in the 2008 NAMCS, it is unknown for 10,149, or 32.6%.

The PRF extracts ethnicity and race from the physician records in accordance with the two-item format standardized by the U.S. Office of Management and Budget (1997).

The first item records whether the patient is Hispanic or Latino. Regardless of the response to the first, the second is a mark-all-that-apply with five races listed. A typical categorization for analysis breaks responses into six groups, cases where one and only one race was selected and a catch-all for individuals identifying with two or more races. Although we did investigate the imputation models' impact on the first question and the six racial categorizations, many are rare and yielded unstable estimates and standard errors. Because of this and for brevity purposes, we report a simplified, three-level race breakout: patients identified as white only, black only, or any other race (whether singly or in combination with white or black).

2.2. *Imputation Methods*

In this section we discuss the cell-based method used to impute missing race in the 2008 NAMCS, and contrast it with a model-based procedure that we felt was better suited to quantify the additional uncertainty reflected by multiple imputation. We also detail how we accounted for features of the complex sample design using this model-based approach.

The single-imputation method used in 2008 was based on a SAS® macro developed by Valverde and Marsteller (2007) that imputes missing race using a hybrid approach falling somewhere between a hot- and cold-deck (Andridge and Little 2010) and what Kalton and Kasprzyk (1986) term *hierarchical* imputation. When race is missing, the macro works dynamically to search for a donor on up to twenty-five matching criteria. For instance, the first criterion is to select a patient race randomly from a pool of donors within the same survey year, three-digit diagnosis code (see Section II.A.28 of NCHS 2009), and patient ZIP code. If no match can be found, the macro seeks a record of the same diagnosis code and patient ZIP code, but from the previous year's data.

Simply running the macro more than once to generate multiple imputations would not be prudent, since it ignores the imputation model's uncertainty. Rubin (1987) terms such a procedure *improper* (pp. 112–128). Rubin and Schenker (1986) offer the *approximate Bayesian Bootstrap* (ABB) as a way to perform proper multiple imputation in the cell-based setting. The ABB is akin to independently drawing a set of regression parameters from the posterior predictive distribution of an explicit imputation model prior to drawing each set of imputations. It was not immediately evident, however, what effect the hierarchical nature of the imputation macro would have on the theory underlying the ABB. We considered applying a bootstrap re-imputation scheme of the sort proposed by Efron (1994) and adapted to complex survey designs by McCarthy and Snowden (1985) and Shao and Sitter (1996), in the spirit of analyses undertaken by Li et al. (2004). In the end, we deemed a model-based multiple-imputation procedure most directly amenable to quantifying the increase in estimated variance in transitioning from single to multiple imputation.

The model-based procedure, sequential regression multivariate imputation (Raghunathan et al. 2001), was implemented using IVEware (http://www.isr.umich.edu/src/smp/ive/), free SAS-callable software developed by the Institute for Social Research at the University of Michigan, capable of imputing continuous, semicontinuous, categorical, and count variables. It uses an iterative algorithm which cycles through the variables with missing data, imputing the missing values of each variable conditional on the other

variables (Raghunathan et al. 2001). By imputing each variable in turn using those that came before or after, it builds interdependence among the data. Another useful feature is the ability to bound imputations within a specified range, something utilized in this and other NCHS imputation projects (e.g., Schenker et al. 2011).

In determining which covariates to include in the model-based procedure, we began by incorporating those utilized in the cell-based procedure and, based on input from subject matter experts, added variables we anticipated would help explain the missing data pattern and race itself, including patient age, sex, urban/rural indicator based on metropolitan statistical area (MSA), physician specialty group, reason for visit, natural logarithm of time spent with physician, and an indicator of who entered data into the PRF.

In addition to as many known covariates as possible, Rubin (1996) asserts imputations should be conditional on the sample design: "Minimally, major clustering and stratification indicators and sample design weights (or estimated propensity scores of being in the sample) should be included in imputation models" (p. 478). Indeed, a simulation by Reiter et al. (2006) exposes severe biases that can result from excluding such indicator variables when they explain the underlying mean function, even if the missingness mechanism is fully captured.

Nearly all the matching criteria in the cell-based method are at a finer level than PSU (i.e., ZIP codes generally lie within PSU boundaries). For the model-based method, we tried to include stratum and PSU indicators and sample weights as prescribed, but encountered convergence issues for the logistic regression parameters that did not cease until the PSU indicators were omitted. Reiter et al. (2006, p. 148) warn of such a problem:

> In some surveys the design may be so complicated that it is impractical to include dummy variables for every cluster. In these cases, imputers can simplify the model for the design variables, for example collapsing cluster categories, or including proxy variables (e.g., cluster size) that are related to the outcome of interest.

As a compromise, we incorporated local race distribution information from the U.S. Census Bureau's American FactFinder tool (http://factfinder2.census.gov/main.html). Specifically, we created two variables to house Census 2000 estimated proportions of non-Hispanic whites and non-Hispanic blacks at the ZIP code tabulation area level. For a portion of the cases (roughly 10%), patient ZIP code was unavailable. Where possible, we substituted physician practice ZIP code. For the remaining 3% of cases without a patient or physician ZIP, the race distribution variables were imputed, using IVEware's bounding feature to ensure proportions remained within [0, 1]. Kozak (1995) used a similar method at the county level, reporting: "Although not exact, the population distribution of a county appeared useful as a general indicator of the racial distribution of discharges from a hospital in the county" (p. 4).

2.3. Multiple-Imputation Inferences

In this section we introduce notation and formulas pertinent to inferences from multiply-imputed data as well as a few related metrics facilitating comparisons to singly-imputed data. Instead of a missing value being filled in once, multiple imputation calls for a missing value to be imputed M times ($M \geq 2$). In our study with the 2008 NAMCS, $M = 5$.

Each of the $m = 1, \ldots, M$ completed (observed plus imputed) datasets are analyzed individually and a particular quantity and its variance can be estimated through Rubin's (1987) straightforward combination rules given below.

If we let $\hat{Q}_m$denote the m^{th} completed-dataset estimate of a quantity Q, the quantity's overall multiple-imputation estimate is simply the average of the M estimates, $\bar{Q}_M = \frac{1}{M}\sum_M \hat{Q}_m$.

Let $\overline{U}_m$ denote the m^{th} completed-dataset estimated variance for $\hat{Q}_m$. The multiple-imputation estimated variance is the average of the M completed-dataset variances, $\overline{U}_M = \frac{1}{M}\sum_M \overline{U}_m$, plus a term reflecting the between-imputation variance of the estimate, $B_M = \sum_M \frac{\left(\hat{Q}_m - \bar{Q}_M\right)^2}{M-1}$.

After a finite imputation correction factor $\left(1 + \frac{1}{M}\right)$ is applied to the between-imputation variance component, the overall multiple-imputation variance formula is given by

$$T_M = \overline{U}_M + \left(1 + \frac{1}{M}\right) B_M. \tag{1}$$

A useful metric with a simple interpretation is the ratio of a quantity's multiple-imputation estimated standard error to its average single-imputation counterpart,

$$R = \sqrt{T_M / \overline{U}_M}. \tag{2}$$

The degree to which R exceeds 1 represents the percent increase in the estimated standard error attributable to *multiple* imputation.

Another related quantity is the *fraction of missing information* (FMI) (Rubin 1987, sec. 3.3; Wagner 2010), which can be approximated by the between-imputation variance component over the total variance,

$$FMI_{approx} = \left(1 + \frac{1}{M}\right) B_M / T_m. \tag{3}$$

Although the FMI typically depends to some extent on the percent of observations missing, it also depends on the analysis of interest and the extent to which the imputation model is predictive of the missing values. For example, if the imputation model is highly predictive, the FMI will tend to be substantially smaller than the item nonresponse rate.

3. Results

In an attempt to gauge the magnitude of missing-data uncertainty unaccounted for by single imputation, we calculated the ratio of multiple-imputation to average single-imputation estimated standard errors – Equation (2) in Subsection 2.3 – across a multitude of domains. For brevity, we present results from only a subset of those domains: the overall race distribution and the distribution by United States region, age group, and whether the patient has been diagnosed as diabetic. The estimated standard error ratios and other statistics related to these estimates are tabulated in Appendix.

The ratios for all domain estimates are plotted against their respective percent of observations missing in Figure 2. Most ratios exceed 1.0 only slightly, and just two surpass 1.1. These figures are in line with what was reported by Li et al. (2004), despite the patient race item nonresponse rate nearly doubling since the 2000 NAMCS data analyzed therein.

Intuition might lead one to expect the standard error ratios to increase with a higher item nonresponse rate. However, the plot exhibits no such trend. At least for the data at hand, the percent of missing observations alone does not predict the increase in estimated standard errors after multiply imputing. Estimates subject to 30% or more missingness are apparently no more severely underestimating the missing-data uncertainty by singly imputing than estimates with less than 30% missingness.

We followed numerous leads to explain the phenomenon, but most proved futile. For instance, we hypothesized the lopsided distribution of race might have triggered a software glitch. However, other than convergence issues discussed in Section 2, we concluded that IVEware performed soundly. As we will now discuss, the most reliable determinant of a small standard error ratio was found to be a large design effect in the underlying estimates.

Kish (1965, p. 193) defines a design effect as the ratio of the estimate's variance incorporating the complex design to the variance under a simple random sample of the same size

$$deff = \frac{\text{var}_{complex}(\hat{Q})}{\text{var}_{SRS}(\hat{Q})}. \tag{4}$$

The quantity we report in this article could perhaps more aptly be termed the *misspecification effect*, as it is the estimated variance accounting for the complex design features (i.e., stratification, clustering, and weights) over the estimated variance ignoring those features. Nonetheless, because these two terms are often colloquially exchanged for one another, we retain the more frequently utilized phrase.

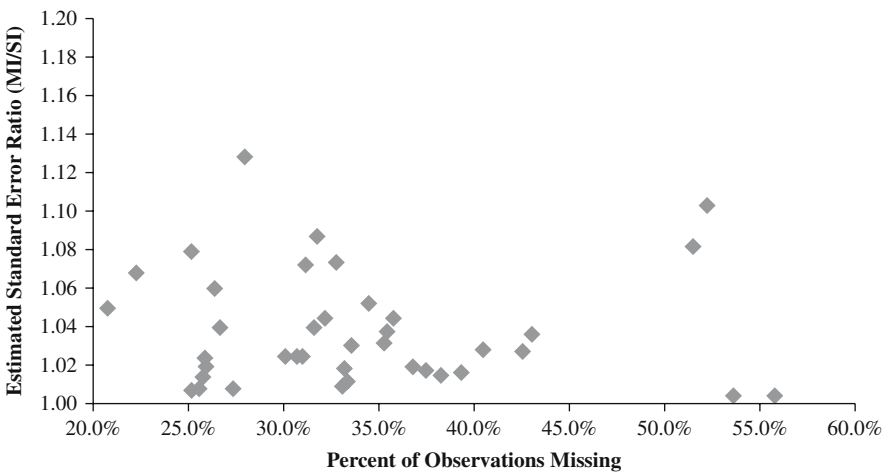

Fig. 2. The Relationship between the Percent of Observations Missing and the Ratio of Multiple-imputation (MI) to Average Single-imputation (SI) Estimated Standard Errors for Select Domain Estimates of Patient Race in the 2008 National Ambulatory Medical Care Survey.

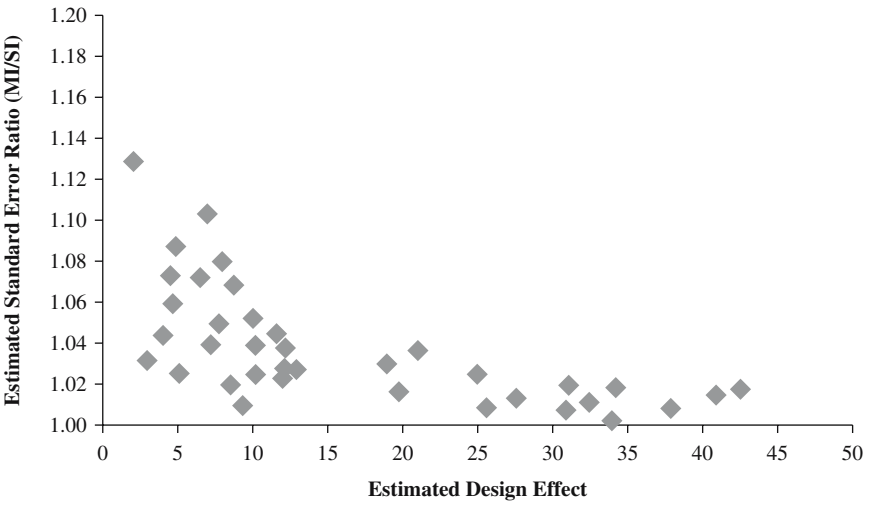

Fig. 3. The Relationship between the Average Completed-dataset Estimated Design Effect and the Ratio of Multiple-imputation (MI) to Single-imputation (SI) Estimated Standard Errors for Select Domain Estimates of Patient Race in the 2008 National Ambulatory Medical Care Survey.

Figure 3 illustrates the inverse relationship between the estimated standard error ratio and estimated design effect for the 2008 NAMCS data. Estimates with a larger design effect are clearly associated with a smaller increase in estimated standard error after multiple imputation. In one of Reiter et al.'s (2006) simulations a similar observation was made, where the multiple-imputation estimated standard error, even in the presence of a 30% item nonresponse rate, was only slightly larger than the complete data estimated standard error (i.e., the estimated standard error that would be obtained in the absence of nonresponse). The authors reason that the complex design "makes the within-imputation variance a dominant factor relative to the between-imputation variance. That is, the fraction of missing information due to missing data is relatively small when compared to the effect of clustering" (p. 146). Figure 3 demonstrates this concept over a range of design effects, using real data. Note that the x-axis scale was truncated at a design effect of 50 to allow for a clearer visualization of the patterns we wished to highlight. Although the truncation omits the two data points in the Appendix with the largest design effects – 70.38 and 97.34 – it does not substantively alter any of our observed patterns and conclusions. (A similar truncation is applied in Figure 4.)

Mentioned previously, an alternative gauge of missing-data uncertainty is the FMI (Wagner 2010). In fact, reproducing Figure 3 with FMI_{approx} of expression (3) on the vertical axis (not shown here) tells the same story. As the design effect increases, FMI_{approx} tapers. This occurs because the two metrics are monotonically related – our ratio of estimated standard errors is $\left(1 - FMI_{approx}\right)^{-\frac{1}{2}}$.

To further elucidate the relative impact of the design effect we can partition the increase in estimated variance into two components, that attributable to the complex sample design and that attributable to missing-data uncertainty as measured by using

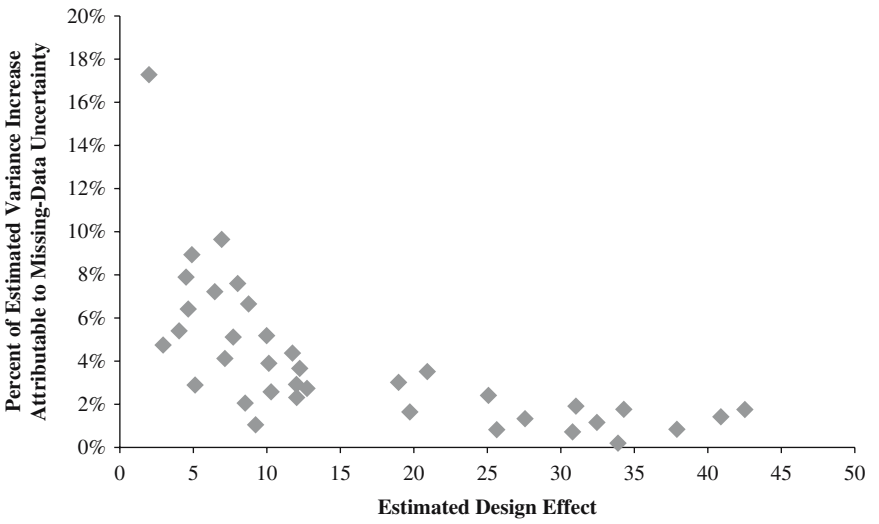

Fig. 4. The Percent of the Estimated Total Variance Increase Attributable to Missing-Data Uncertainty as a Function of the Average Completed-dataset Estimated Design Effect for Select Domain Estimates of Patient Race in the 2008 National Ambulatory Medical Care Survey.

multiple rather than single imputation. Specifically, we can conceptualize the term $\overline{U}_M$ in Equation (1) as being the product of $\overline{U}_{M(SRS)}$, the average completed-data-set variance assuming simple random sampling, and *deff*. Therefore, the approximate variance increase due to both the complex design and missing-data uncertainty can be written as

$$\Delta_M = \overline{U}_{M(SRS)} * (deff) + \left(1 + \frac{1}{M}\right) B_M - \overline{U}_{M(SRS)}. \tag{5}$$

The proportion of Δ_M attributable to missing-data uncertainty is simply the between-imputation term over the increase, or $\left(1 + \frac{1}{M}\right) B_M / \Delta_M$, whereas the proportion attributable to the complex design is the complement about 1, or $1 - \left(1 + \frac{1}{M}\right) B_M / \Delta_M$. We acknowledge, however, that this might not account perfectly for the two sources of increase, because the complex sample design could also affect the between-imputation term, B_M.

Figure 4 demonstrates the relationship between the design effect and the percent of the variance increase attributable to missing-data uncertainty as measured by multiple imputation. The pattern mirrors that appearing in Figure 3. In the presence of a larger design effect, the variance increase is dominated by the component attributable to the complex sample design. The figure suggests that, despite item nonresponse rates often exceeding 30%, a design effect of 10 or greater limits the impact of missing-data uncertainty to generally no more than 5% of the overall variance increase. Put another way, the portion of variance attributable to the complex design in these settings is at least 95% / 5% = 19 times greater than the portion attributable to missing-data uncertainty.

Another way to evaluate the relative increase in estimated variance due to the use of multiple rather than single imputation is to consider what the relative increase would be if the design effect were equal to 1. To approximate the answer, we fitted Lowess smoothers (Cleveland 1979), not shown here, to the data in Figure 3 with a variety of bandwidths that were large enough to avoid major jaggedness in the fitted curves. Extrapolating the curves to a design effect of one suggested a ratio of multiple-imputation to single-imputation estimated standard errors in the range of 1.08 to 1.1. Since, as mentioned earlier, the ratio equals $(1 - FMI_{approx})^{-\frac{1}{2}}$, it follows that the suggested range for FMI_{approx} is 14% to 17%. This range is consistent with a nonresponse rate of about 30% and an imputation model that is partially, not fully, predictive of the missing values.

4. Discussion

In this article, we presented results from a case study in which we evaluated the potential impact on estimated variances if a multiple-imputation strategy were adopted to handle instances of missing patient race in the 2008 NAMCS. The NAMCS sample design involves features such as clustering and highly variable analysis weights that result in extremely large design effects for estimates involving race. In these settings, we found multiple imputation increased estimated variances only modestly. Revisiting our key analytic quantity, the ratio of estimated standard errors in Equation (2), we can reason that as M goes to infinity, the ratio can be rewritten as

$$R \approx \sqrt{1 + \frac{B_M}{\overline{U}_M}}. \tag{6}$$

With a large design effect, the within-imputation component, $\overline{U}_M$, tends to be large relative to the between-imputation component, B_M, pulling the ratio towards 1.

At least among the domains investigated, the item nonresponse rate itself was not found to be predictive of the increase in estimated variance after multiply imputing the missing data. Even when the percent of imputed observations tops 30%, a large design effect can render multiple-imputation estimated standard errors only slightly greater than their single-imputation counterparts. For this reason, together with the increased complexity that multiple imputation poses to the typical NAMCS data user, it was decided to maintain a single-imputation approach for the NAMCS for the time being.

Despite the growing class of techniques available to compensate for missing data, the best way to handle nonresponse is to design data collection protocols preventing it from occurring in the first place (Lohr 1999). In mid-2009, NCHS raised FR awareness of the increased patient race item nonresponse rate, stressing the demographic's importance for analyses. The intervention appears to have been effective, as the item nonresponse rate for race dropped to 24% in the 2009 NAMCS and to 23% in the 2010 NAMCS. Albeit still high by many standards, at least the trend in Figure 1 has begun to reverse course.

Our study is not without limitations. For one, the domains analyzed herein are coarse in nature. It seems plausible that design effects may be attenuated for racial distributions

estimated for finer domains, which could produce scenarios where the proportionate increase in estimated variance due to using multiple imputation is larger than is reflected in this study.

Another limitation is that focus was restricted to only one variable, despite the fact that the NAMCS collects data on hundreds of other variables pertaining to the visit. In addition to feedback from NAMCS data users that patient race is a frequently utilized demographic, as previously mentioned, it is also subject to one of the highest item nonresponse rates. Although not presented here, we investigated another variable of key analytic interest, time spent with the physician, which was also susceptible to a high level of item nonresponse (26%) in the 2008 NAMCS. Similar findings were observed. Due to large design effects in the domains analyzed, multiple imputation increased estimated standard errors only slightly. As noted on page 18 of NCHS (2009), the item nonresponse rate for most other variables is 5% or less, so these are naturally of less concern.

A final limitation, noted in Subsection 2.2, is that we used a "compromise" method to reflect the features of the complex sample design in our imputation model. Had we accounted for those features perfectly, our results might have changed somewhat. However, we believe that our case study demonstrates an actual phenomenon for multiple reasons. First, variables related to the design features were included in the model. Second, as mentioned in Section 3, our case study yields results consistent with simulations reported in Reiter et al. (2006). Finally, if the survey clustering were more fully reflected in the imputation model, a likely result would be imputed values that are more differentiated, that is, less homogeneous, across the clusters. This might very well increase the design effects for each dataset completed by imputation, which, all else being equal, would accentuate the phenomenon displayed by our case study. Development of methods for reflecting design features parsimoniously in imputation models, such as by using random effects, is an important area for future methodological research.

Recent changes to the NAMCS sample design may prompt a re-evaluation at some point in the future. Beginning with the 2012 NAMCS, PSUs are no longer comprised of geographically clustered units. Instead, the universal list of physician offices is stratified by state and a sample selected within each, so the physician office now serves as the PSU. To the extent this new sample design alters the variability of weights or the heterogeneity of PSUs with respect to patient race, the magnitude of the design effects could change.

Aside from more empirical analyses such as the one discussed in this article, a simulation study and further theoretical research could foster a better understanding of the relationship between the design effect and the between-imputation component of variability reflected by multiply imputing missing data. Of particular interest would be to determine if and how the relationships we observed are moderated by how predictive the imputation model is and/or by alternative patterns of nonresponse.

Appendix

Appendix. Select Point Estimates, Estimated Standard Errors, Estimated Design Effects, and Indicators of Missing-Data Uncertainty from 2008 NAMCS Data Singly Imputed (SI) and Multiply Imputed (MI) by the Model-Based Method

Domain	Race	MI Estimate (%)	Average SI Estimated Standard Error (%)	MI Estimated Standard Error (%)	Estimated Standard Error Ratio (MI/SI)	Estimated Standard Error Assuming SRS (%)	Estimated Design Effect[1]	Percent Obs. Imputed (%)
Overall								
	White	84.6	1.243	1.265	1.018	0.212	34.283	33.2
	Black	10.5	0.949	0.961	1.013	0.181	27.618	25.8
	Other	4.9	0.837	0.851	1.017	0.128	42.538	37.5
Region								
Northeast	White	90.0	2.030	2.079	1.024	0.406	25.053	30.1
	Black	5.9	1.433	1.456	1.016	0.322	19.788	39.4
	Other	4.0	0.776	0.829	1.068	0.263	8.720	22.3
Midwest	White	88.9	1.316	1.349	1.025	0.412	10.195	30.7
	Black	9.0	1.285	1.314	1.023	0.371	12.014	25.9
	Other	2.1	0.341	0.352	1.031	0.198	2.978	35.3
South	White	82.5	2.203	2.217	1.007	0.396	30.903	25.2
	Black	15.5	2.192	2.197	1.002	0.376	33.965	18.8
	Other	2.0	0.211	0.238	1.128	0.148	2.037	28.0
West	White	81.3	3.943	3.956	1.003	0.470	70.380	55.8
	Black	5.3	0.710	0.783	1.103	0.269	6.937	52.3
	Other	13.4	4.007	4.024	1.004	0.406	97.340	53.7
Age								
Less than 15	White	79.5	3.200	3.261	1.019	0.574	31.087	36.8
	Black	13.5	2.965	2.989	1.008	0.481	37.939	27.4
	Other	7.0	1.318	1.354	1.027	0.369	12.771	42.6
15-24	White	81.5	2.400	2.420	1.009	0.785	9.351	33.1
	Black	13.7	1.946	2.042	1.049	0.699	7.743	20.8
	Other	4.8	1.512	1.555	1.028	0.434	12.126	40.5

Appendix. Continued

Domain	Race	MI Estimate (%)	Average SI Estimated Standard Error (%)	MI Estimated Standard Error (%)	Estimated Standard Error Ratio (MI/SI)	Estimated Standard Error Assuming SRS (%)	Estimated Design Effect[1]	Percent Obs. Imputed (%)
25-44	White	82.4	1.538	1.598	1.039	0.482	10.172	31.6
	Black	11.6	1.190	1.213	1.019	0.408	8.517	26.0
	Other	6.0	1.059	1.099	1.037	0.303	12.223	35.5
45-64	White	86.1	1.307	1.365	1.044	0.382	11.715	32.2
	Black	9.7	0.927	1.000	1.079	0.328	7.997	25.2
	Other	4.2	0.979	1.008	1.030	0.225	18.961	33.6
65-74	White	87.5	1.189	1.276	1.073	0.558	4.532	32.8
	Black	8.8	1.029	1.090	1.059	0.476	4.674	26.4
	Other	3.7	0.650	0.678	1.044	0.325	4.004	35.8
75 +	White	89.6	1.493	1.570	1.052	0.471	10.053	34.5
	Black	6.6	0.875	0.897	1.025	0.386	5.143	31.0
	Other	3.8	1.316	1.363	1.036	0.287	20.971	43.1
Diabetes								
No	White	85.0	1.262	1.276	1.011	0.221	32.507	33.4
	Black	10.0	0.950	0.957	1.008	0.188	25.613	25.6
	Other	5.0	0.858	0.870	1.014	0.134	41.006	38.3
Yes	White	81.4	1.842	1.975	1.072	0.723	6.501	31.2
	Black	13.9	1.685	1.751	1.039	0.628	7.188	26.7
	Other	4.7	0.957	1.041	1.087	0.433	4.895	31.8

Note:
[1]Average estimated design effect of the $M = 5$ completed datasets.

5. References

Andridge, R. and Little, R. (2010). A Review of Hot Deck Imputation for Survey Non-Response. International Statistical Review, 78, 40–64, DOI: http://www.dx.doi.org/ 10.1111/j.1751-5823.2010.00103.x.

Cleveland, W. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association, 74, 829–836, DOI: http://www.dx.doi.org/10.1080/01621459.1979.10481038.

Efron, B. (1994). Missing Data, Imputation, and the Bootstrap. Journal of the American Statistical Association, 89, 463–475, DOI: http://www.dx.doi.org/10.1080/01621459.1994.10476768.

Groves, R., Dillman, D., Eltinge, J., and Little, R., (Eds.) (2002). Survey Nonresponse. New York, NY: Wiley.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. Survey Methodology, 12, 1–16.

Kish, L. (1965). Survey Sampling. New York, NY: Wiley.

Kozak, J. (1995). Underreporting of Race in the National Hospital Discharge Survey. Advance Data from Vital and Health Statistics, No. 265. Hyattsville, MD: National Center for Health Statistics.

Li, Y., Lynch, C., Shimizu, I., and Kaufman, S. (2004). Imputation Variance Estimation by Bootstrap Method for the National Ambulatory Medical Care Survey, Proceedings of the Survey Research Methods Section of the American Statistical Association.

Lohr, S. (1999). Sampling: Design and Analysis. Pacific Grove, CA: Brooks/Cole.

McCarthy, P. and Snowden, C. (1985). The Bootstrap and Finite Population Sampling. Vital Health Statistics, 2(95). Hyattsville, MD: National Center for Health Statistics.

National Center for Health Statistics (2009). 2008 NAMCS Micro-Data File Documentation. Division of Health Care Surveys, National Center for Health Statistics, Centers for Disease Control and Prevention, U.S. Department of Health and Human Services, Hyattsville, MD, Available online at: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NAMCS/doc08.pdf (accessed January 2014).

Office of Management and Budget (1997). Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity, Federal Register 62FR58781-58790. Available at: http://www.gpo.gov/fdsys/pkg/FR-1997-10-30/pdf/97-28653.pdf (accessed January 2014).

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, 85–95.

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

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

Rubin, D. (1996). Multiple Imputation After 18 + Years (with discussion). Journal of the American Statistical Association, 91, 473–489.

Rubin, D. and Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. Journal of the American Statistical Association, 81, 366–374, DOI: http://www.dx.doi.org/10.1080/01621459.1986.10478280.

Schenker, N., Borrud, L., Burt, V., Curtin, L., Flegal, K., Hughes, J., Johnson, C., Looker, A., and Mirel, L. (2011). Multiple Imputation of Missing Dual-Energy X-Ray Absorptiometry Data in the National Health and Nutrition Examination Survey. Statistics in Medicine, 30, 260–276, DOI: http://www.dx.doi.org/10.1002/sim.4080.

Shao, J. and Sitter, R. (1996). Bootstrap for Imputed Survey Data. Journal of the American Statistical Association, 91, 1278–1288, DOI: http://www.dx.doi.org/10.1080/01621459.1996.10476997.

Valverde, R. and Marsteller, J. (2007). A Revised Matching Routine for Imputing Missing Race and Ethnicity in the National Ambulatory Medical Care Survey, Unpublished internal manuscript of the National Center for Health Statistics.

Wagner, J. (2010). The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data. Public Opinion Quarterly, 74, 233–243, DOI: http://www.dx.doi.org/10.1093/poq/nfq007.

Received November 2011
Revised January 2014
Accepted January 2014