Journal of Data Science **1**(2003), 293-312

Comparison of Two Multiple Imputation Procedures in a Cancer Screening Survey

Coen A. Bernaards[1], Melissa M. Farmer[1], Karen Qi[1], Gareth S. Dulai[1,2], Patricia A. Ganz[1], Katherine L. Kahn[1,3].
[1] *University of California, Los Angeles*
[2] *Greater Los Angeles Veterans Administration Healthcare System*
[3] *RAND*

Abstract: Commonly in survey research, multiple, different analyses are conducted by one or more than one researcher on the same data set. The conclusions from these analyses should be consistent despite the presence of missing data. Multiple imputation is frequently used to ensure consistency of analyses. Two methods for multiple imputation of missing data are a combination of hot deck and regression imputation, and multivariate normal multiple imputation. It is unknown whether these methods will give similar results in practical situations with large numbers of variables. We applied both multiple imputation methods to a cancer screening survey data with 2 continuous, 48 Likert scale items, and 74 binary response items. Correlations and variances of imputated data sets were compared in a first attempt to investigate similarity of the imputation methods. The results of both methods were found to be similar; either of the two methods are endorsed for surveys similar to the data set presented.

Key words: cancer screening, multiple imputation, nonresponse, survey.

1. Introduction

Colorectal cancer (CRC) is the second leading cause of cancer mortality in the United States of America with an estimated 56,700 deaths in 2002 (ACS, 2002; Jemal, Thomas, Murray, and Thun, 2002). In 2002, it is estimated that there will be 148,300 new cases of CRC, only one-third of which are diagnosed at a localized stage. The overall 5-year relative survival rate was 61% in 1992-1997, but varied significantly by stage at diagnosis: 90% for localized compared to 69% for regional and 8% for distant (ACS, 2002).

There is evidence that screening using the fecal occult blood test and/or sigmoidoscopy reduces the risk of CRC mortality, in part by detection of early stage disease (Winawer, Fletcher, Miller, Godlee, Stolar, Mulrow, Woolf, Glick, Ganiats, Bond, Rosen, Zapka, Olsen, Giardiello, Sisk, Van Antwerp, Brown-Davis, Marciniak and Mayer, 1997), yet screening rates for CRC remain low (Breen, Wagener, Brown, Davis and Ballard-Barbash, 2001). This research team is currently conducting a randomized effectiveness trial of a quality improvement intervention to increase CRC screening rates within provider organizations that contract with a large California Health Maintenance Organization (HMO). The two-year intervention is being delivered to the provider organizations and targets the primary care providers, nurses, and administrative staff to improve the rate at which enrolled patients utilize CRC screening tests.

As part of the baseline data collection for this randomized controlled trial, a survey was mailed to a stratified random sample of primary care providers from each of the provider organizations. Survey administration began in November 1999 and continued through 2001 and yielded a 67% response rate. The purpose of the survey was to gain a better understanding of provider beliefs and recommendation patterns for CRC screening as well as identify barriers and facilitators to screening. A follow-up survey will be conducted at the conclusion of the intervention trial in Fall 2002.

A common problem in statistical practice is nonresponse. A vast amount of literature is available on missing data methods, see e.g. Little and Rubin (1987), Schafer (1997). Some missing data methods are aimed at specific analyses. For example, the Expectation Maximization (EM) algorithm

(Dempster, Laird and Rubin, 1977), is a tool for maximizing a specified likelihood function. Typically, in a data set like the CRC survey, multiple, different analyses are conducted by one or more different researchers. Also, different analyses may study different scientific questions with, potentially, different outcome variables. Conclusions between analyses should be consistent, however. Multiple imputation is a missing data method that is not aimed at specific analyses. In multiple imputation, each of the missing values is replaced by a set of plausible values (e.g. five plausible values are generated for each missing value to give five completed data sets) that represent uncertainty about the missing data. Each of the completed data sets are then analyzed using readily available complete data methods. Finally, the estimates from each of the analyzed data sets are combined using standard formulae from Rubin (1987, 1996). Once multiply imputed data sets are obtained, all analyses can be carried out without rejecting respondents because of missingness. Conclusions of multiply imputed data sets of different manuscripts are then consistent.

Reports of multiple imputation in applications of statistical methods is increasing. Some recent applications include Heitjan and Little (1991), Heitjan and Landis (1994), Hediger, Overpeck, McGlynn, Kuczmarski, Maurer and Davis (1999), and Barnard and Meng (1999). Also, there has been increased interest in introducing multiple imputation to interdisciplinary audiences. For example, Little and Rubin (1989) direct their work toward social scientists, while Rubin and Schenker (1991), Schafer (1999a), Molenberghs, Burzykowski, Michiels and Kenward (1999), Bennett (2001), and Patrician (2002) have applications and introductions for medical research.

Research comparing multiple imputation methods for specific missing data problems is limited. Previous research include Reilly (1993) who focused on hot deck multiple imputation methods. Schenker and Taylor (1996) compare the regression method, predictive mean matching, and a regression method imputing an additional residual draw multiple imputation methods. Multivariate normal imputation is not included. Horton and Lipsitz (2001) compare several computer packages for multivariate normal and regression methods. However, their research is limited to a few continuous variables. They find that additional research is needed for cases where the normality assumption is violated such as with binary variables.

For purposes of initial research and directions for the follow-up survey, we originally planned to use a combination of hot deck and regression imputation for the CRC survey. Of note, these methods require programming effort, whereas free software is available for multivariate normal imputation. Multivariate normal imputation relies on the assumption of multivariate normality of the data, however.

In the present paper we compare two multiple imputation procedures by applying them to the survey from our CRC intervention study. For each multiple imputation strategy five completed data sets were created. For the first strategy, missing covariates were multiply imputed using hot deck. Next, for each of the multiply imputed data sets of covariates separately, missing response variables were imputed by regression on all covariates. This two-stage procedure resulted in five completed data sets. For the second strategy, multiple imputation by assuming a multivariate normal distribution for the data was applied. The present study is a first attempt at extending Horton and Lipsitz (2001) comparison of multiple imputation methods using real data.

Creating complete, multiply imputed data sets requires special software and methodology. The stand alone program SOLAS has a version of hot deck among other methods. We implemented hot deck and regression multiple imputation in S-plus. Software to produce multivariate normal multiply imputed data sets include S-plus 6, SAS 8.2, SPSS, and the standalone program NORM, Schafer (1999b). Here NORM was used. See Horton and Lipsitz (2001) for a recent overview of computer software for multiple imputation.

In section 1.1 the data set is introduced. Section 2 introduces the hot deck and regression imputation methods, including the S-plus functions. Section 3 introduces the multivariate normal multiple imputation. Section 4 has a brief overview on obtaining estimates from multiply imputed data sets. Imputations are compared in section 5. Discussion is in section 6.

1.2 Description of the dataset

An extensive description of the survey design, administration, and initial results can be found in Dulai, Farmer, Ganz, Bernaards, Qi, Dietrich, Bas-

tani, Belman, and Kahn (submitted). The provider survey questionnaire included 136 items, consisting of 12 provider characteristics (covariates), and 124 items addressing questions related to cancer screening practices.

The original questionnaire was mailed to 1340 potential respondents, of which 891 returned the questionnaire, a 67% response rate. Of the 891 completed surveys, 9 from respondents who indicated that they did not provide primary care were dropped from further analysis, reducing the sample size to 882. All returned questionnaires had at least 49 of 136 completed items. 95% of the respondents answered 109 items or more. Only 289 respondents completed all 136 items. Across all 882 respondents and all 136 items, there were 4888 missing item responses. Figure 1 presents an overview of all percentages of missing data for each variable separately of all 136 variables included in this study. Using listwise deletion for analysis of this data set would reduce the sample size by 67%.

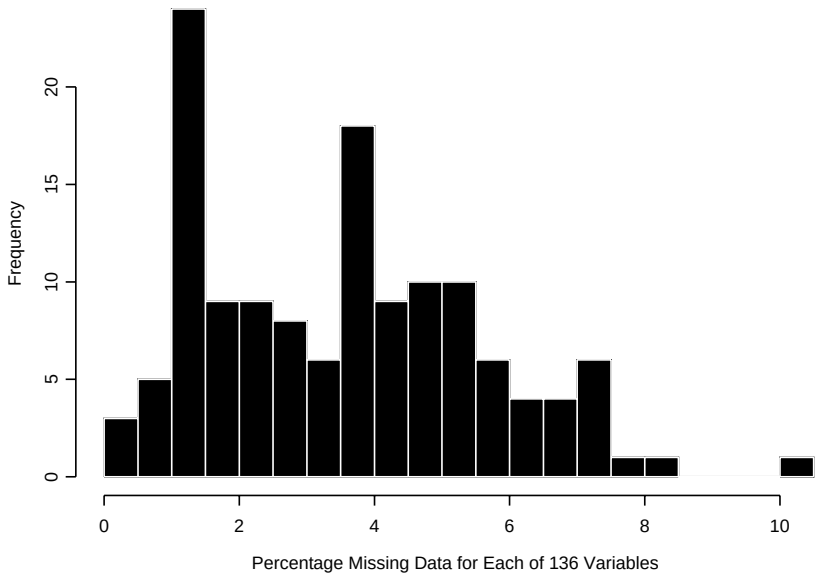

Figure 1. Histogram of the percentage of missing data for each of the 136 variables included in this analysis

Table 1: List of All 9 Binary, 1 Trivariate, and 2 Continuous Provider Characteristics (Covariates) Used in The Study.

Covariate	Type	Range
Gender	binary	0,1
White/Non White	binary	0,1
Ever Been Screened	binary	0,1
CRC Family History	binary	0,1
Specialty Training	binary	0,1
Gastroenterology Training	binary	0,1
Medical School Affiliation	binary	0,1
Other PCP in Office	binary	0,1
Independent Practice Association (IPA) or Medical Group	binary	0,1
Small/Medium/Large Group Size	integer	0,1,2
Number of Years in Practice	integer	$1,\ldots,52$
Age	integer	$18,\ldots,72$

The 12 provider characteristics included in the study (Table 1), consisted of nine binary, one trivariate, and two continuous variables. Missing data occured only on the binary provider characteristics.

The 124 items on cancer screening practices consisted of 74 binary response items, 48 Likert scale items and 2 continuous items. Binary response items were coded 0-1. All Likert scale items had five ordered response categories ranging from "Not Effective" to "Very Effective" or similar options, coded 1 through 5.

Because of the large number of variables there were 330 unique missing data patterns in the data. The missing data mechanisms were unknown; no assumptions were made about them.

2. Hot deck and regression multiple imputation

The first method used for multiple imputation consisted of hot deck and regression methods.

2.1 Hot deck imputation

We used hot deck multiple imputation to impute missing data on 12 provider characteristics (covariates). In this hot deck imputation, the donor group consist of $i = 1, \ldots, 831$ respondents who had completely observed covariates (i.e., no missing data). The remaining $j = 1, \ldots, 51$ respondents had at least one missing provider characteristic. For each complete respondent i and incomplete respondent j, a difference $DIFF_{ij}$ is computed across all provider characteristics that were observed for incomplete respondent j. That is,

$$DIFF_{ij} = \sum_k |z_{jk} - z_{ik}|$$

Provider characteristics that were unobserved for incomplete respondent j were left out of the equation. Thus, if, for example, an incomplete respondent j had only 10 provider characteristics observed then $DIFF_{ij}$ was computed across these 10 characteristics only. Note that $DIFF_{ij}$ takes on nonnegative values only since it is a sum of absolute differences. For each incomplete respondent j 831 values for $DIFF_{ij}$ were obtained. Of these 831, five donor cases that had the smallest values for $DIFF_{ij}$ were considered for imputation. For each completed data set we drew one donor case at random from the five donor cases that had the smallest value for $DIFF$. More than five donor cases were considered only if there were multiple equal smallest $DIFF$ values. Thus, for example, if the smallest three $DIFF$ values were unique, but there were six donor cases that all had the fourth smallest $DIFF$ value, then a total of nine donor cases were considered.

Nine of the covariates included were binary (0-1); the other three covariates included in the hot deck imputation were age, IPA or medical group, organization size (small/medium/large), and Number of Years in Practice, see Table 1. A ten year difference in age or number of years in practice would exceed a difference in all binary covariates combined. In order to have age and number of years in practice comparable to binary covariates, we first subtracted the minimum and, next, divided by the maximum. Thus, for age we used the transformed $(\text{Age}-28)/(72-28)$ rather than Age itself in the hot deck imputation, and for number of years in practice we used $(\text{Number of Years in Practice}-1)/(52-1)$.

2.2 Regression imputation

Three separate regressions methods were used for imputation: linear regression for Likert scale variables and continuous variables, logistic regression for binary outcomes, and generalized linear regression for multiple unordered categories.

We used one set of covariates for all regression models. This allowed us to conduct the regression imputation once, even if across analyses a variable changed from an outcome to a predictor variable. This practical efficiency is useful in a study like this where variables change roles and importance across a large and diverse set of analyses and research analyses. Since we had relatively few percent missing data for each variable that was imputed (less than 7.5 percent in 119 out of 124 variables, and only 1 variable exceeded 10 percent missing data), we thought it reasonable to use just one set of covariates for all regression models. The advantage is that imputations are easier to produce compared to building different regression models for different variables. We note, however, the cost of a small loss of efficiency (Rubin 1996, p. 478-479).

Multiple linear regression

Univariate multiple linear regression was used for all Likert items, and for continuous outcomes. Included provider characteristics were completed using the hot deck procedure. Missing data in each of the regressions therefore was present only on the response variable. The model was estimated using all observed responses, and, next, the missing responses were imputed by predicting it through the observed covariates. A random draw from a normal distribution with mean zero and residual standard error for the standard deviation was added to the predicted value. Finally, after addition, Likert scale data were rounded to the nearest integer from 1 through 5.

Logistic regression

Logistic regression was used for items with binary outcomes. The following procedure was similar to the multiple linear regression imputation.

Regression coefficients were estimated using all completed covariates, and, next, missing responses were predicted using covariates that corresponded to the missing item. For each missing response, the predicted value was a number between 0 and 1. The imputed value was a Bernoulli random variable that takes on the value 1 with probability equal to the predicted value.

Generalized linear regression for multiple outcomes

Finally, a regression method to impute multiple unordered outcomes was implemented. However, this method was not used when regression imputation is compared to imputation based on the multivariate normal distribution.

Generalized linear regression for multiple unordered categories was used for two items which had more than two nominal outcomes. The procedure was the same as for logistic regression imputation, except that predicted outcomes were actual categories. No additional drawing or rounding was needed.

3. Multivariate normal multiple imputation

Multiple imputation by assuming a multivariate normal distribution on all variables jointly is a common method for multiple imputation. Imputed data sets are obtained in two steps. First, the mean vector and the covariance matrix are obtained using the EM algorithm. Second, with the obtained estimates, data augmentation is carried out in order to obtain multiply imputed values. Details of the two steps are extensively documented in Schafer (1997).

We used the program NORM to obtain five multiply imputed data sets. By default, NORM imputes integers for variables that consists of integers only. Imputed integers ouside the range of the variables were set at the nearest integer allowed. Thus, Likert scale data ranged from 1 to 5; imputed values smaller than 1 were set at 1, and imputed values larger than 5 were set at 5.

Binary items also were assumed to be a subset of the multivariate normal

distribution. Binary imputed values were rounded to either a 0 or a 1 value. The multivariate normality of 74 binary variables is questionable, however, Schafer (1997) and Schafer and Olsen (1998) also proceed by rounding in similar fashion. Schafer and Olsen's (1998) application of NORM had 12 variables total, three of which were binary.

The two items with unordered multiple outcomes were not included in the multivariate normal imputation method because this led to a covariance matrix that was not invertible.

4. Computing point estimates from multiply imputed data sets

Once multiply imputed data sets are obtained, each is analyzed separately using standard complete data methods regardless of the multiple imputation procedure used. Here we briefly review how to combine the estimates from the separate analyses to obtain the final estimate. Details can be found in Rubin (1987, p.76) and Schafer (1997, p.112).

Suppose we want to obtain an estimate from five completed data sets for some quantity of interest Q in the population. For example, Q could be the mean of a variable, the correlation between two variables, or a regression coefficient. For the five completed data sets we first obtain the five point estimates for the quantity of interest, denoted $\hat{Q}_1, \ldots, \hat{Q}_5$, and their accompanying variances, denoted $\hat{U}_1, \ldots, \hat{U}_5$ (the squared standard errors). Thus, if $\hat{Q}_1$ is the sample mean of some variable in the first completed data set, then $\hat{U}_1$ is the sample variance of that same variable in the first completed data set divided by the sample size.

The final point estimate $\bar{Q}$, based on the five completed data sets equals

$$\bar{Q} = \frac{1}{5} \sum_{i=1}^{5} \hat{Q}_i.$$

The within variance $\bar{U}$, of the five completed data sets equals

$$\bar{U} = \frac{1}{5} \sum_{i=1}^{5} \hat{U}_i.$$

The between variance, B, is the variance between the point estimates,

$$B = \frac{1}{5-1}\sum_{i=1}^{5}(\hat{Q}_i - \bar{Q})^2 = \frac{1}{4}\sum_{i=1}^{5}(\hat{Q}_i - \bar{Q})^2.$$

The combined estimate for the variance of $\bar{Q}$ is based on both the within and the between variance. This combined variance estimate is called total variance, and denoted by T,

$$T = \bar{U} + (1 + (1/5))B = \bar{U} + 6/5 \cdot B.$$

The relative increase in variance r is defined as

$$r = (1 + (1/5))B/\bar{U} = 6/5 \cdot B/\bar{U}.$$

Interval estimates for $\bar{Q}$ are constructed using a t distribution with degrees of freedom equal to

$$\nu = (5-1)(1 + (1/r))^2.$$

Finally, the fraction of missing information about Q due to nonresponse is denoted γ,

$$\gamma = \frac{r + 2/(\nu + 3)}{r + 1}.$$

5. Evaluation of imputations

Without knowing what the missing values exactly would be if they had been observed, it is impossible to say if analyses from a multiply imputed data set equal the complete data set. Also, one-to-one comparisons of the two methods by simply comparing the actual multiply imputed data sets was not an option because of the random draws incorporated while imputing. Since we were interested in comparison of two multiple imputation methods we compared descriptive statistics in order to decide on equivalence of the two methods. First correlations are compared. Next, between and within imputation variances for sample means are compared.

We computed correlations between all 136 items involved here for both imputation strategies. In the terminology of section 4, Q is the correlation

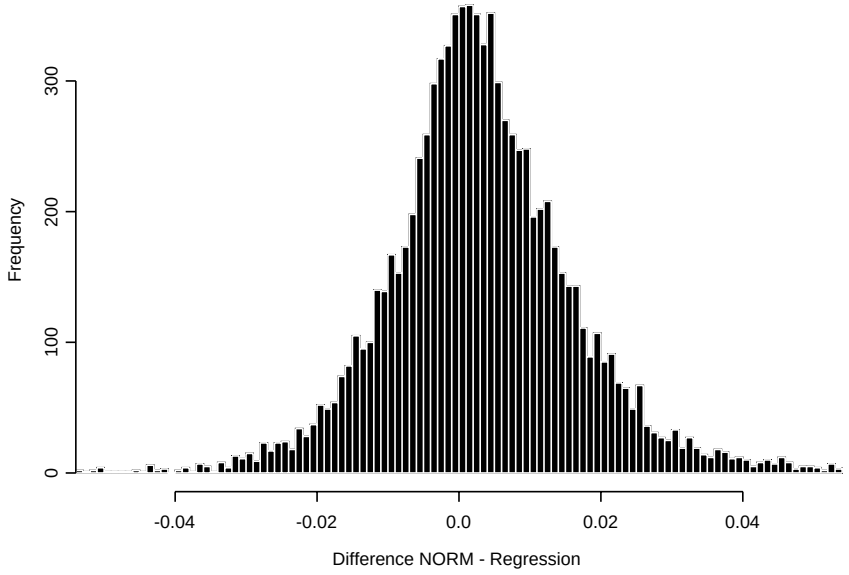

Figure 2. Differences in correlations between data imputed using NORM and using regression imputation. 29 differences below −0.05 and 88 differences exceeding 0.05 were left out.

coefficient between any two variables. Here we are interested only in the point estimates $\bar{Q}$. The differences between all $136 \times 135/2 = 9180$ correlations were computed by subtracting the regression imputed correlations from the correlations obtained through imputation by NORM. The mean difference was 0.0034, and standard deviation 0.0151. The minimum difference was −0.111, and the maximum was 0.171. Figure 2 is a histogram of all differences between correlations between −0.05 and 0.05. Additionally, of all 9180 differences involved, 28 were below −0.05, and 83 exceeded 0.05. When the histogram is approximated by a normal distribution, we find 163 correlations in the lower 2.5% and 257 correlations in the upper 2.5%. By chance alone 459 correlations may be expected to be in the extreme 5%. Similarly, in the lower 0.5% we find 61 correlations, and in the upper 0.5%

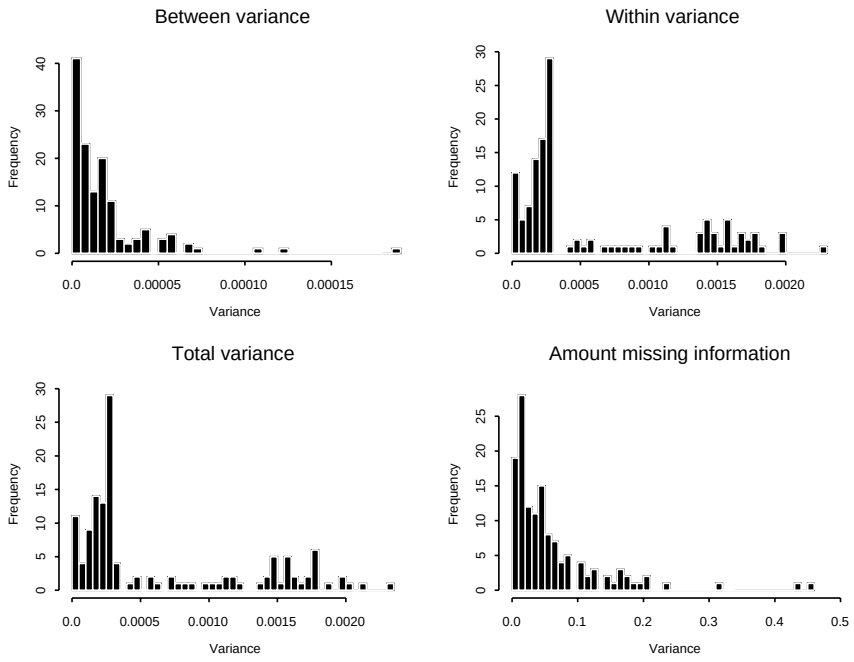

Figure 3. Between variance, within variance, total variance and amount of missing information of all variables using multivariate normal imputation.

we find 138 correlations. By chance alone 92 may be expected. In the upper and lower 0.1% we find 29 and 82 correlations respectively, whereas only 9 are to be expected by chance. This suggests that some correlations estimated from these two methods are really different. Practical implications will be limited because for the lower 0.1% the actual difference in correlations is 0.04, and for the upper 0.1% the actual difference is 0.05. Thus, although correlations for regression imputation tended to be lower than for imputation using NORM, for practical purposes there no differences in interpretation is to be expected.

We now turn to variance estimates for sample means. For all variables separately, point estimates for the means, between variance, within variance, total variance, and amounts of missing information under both imputation

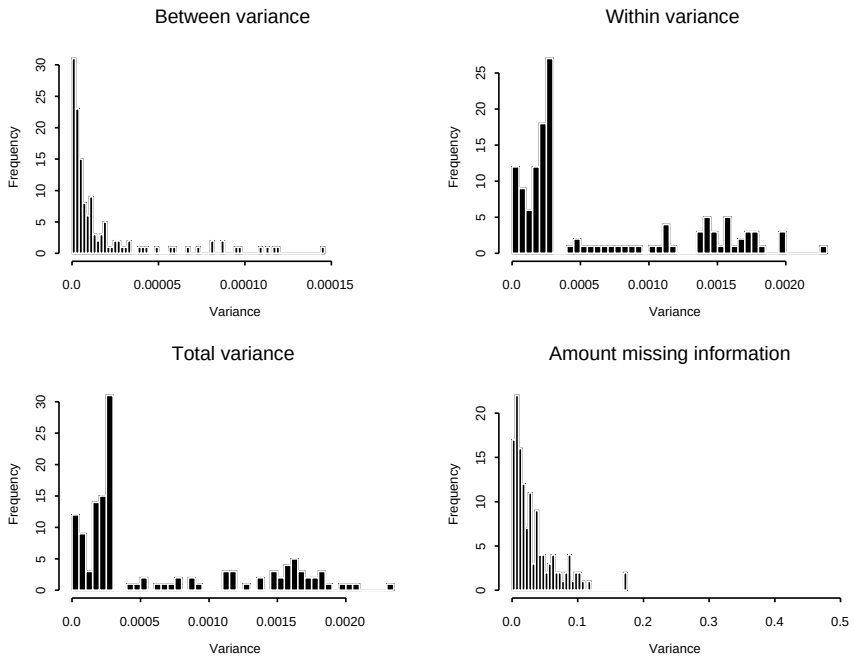

Figure 4. Between variance, within variance, total variance and amount of missing information of all variables using multivariate normal imputation.

methods were computed. In the terminology of section 4, Q is now the mean of a variable.

Point estimates for the means under both imputation methods were very close. For binary variables differences in variable means between the two methods ranged from -0.046 to 0.036. For Likert variables, the differences ranged from -0.006 to 0.016. For the two continuous variables the differences were 0.5 and 36.

Histograms for the variances and the amounts of missing information are in Figures 3 and 4. Although visually no striking differences are present, it can be seen that regression imputed variables (Figure 3) generally have less between imputation variance than multivariate normal imputation (Figure 4).

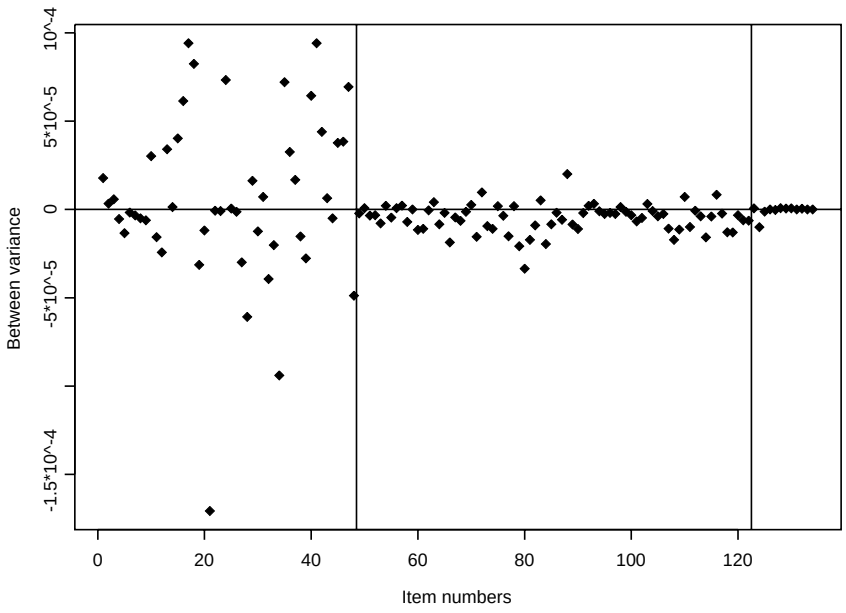

Figure 5. Differences of between variance of multivariate normal imputation and regression imputed data for Likert scale data (left panel), binary response data (middle panel) and covariates (right panel) separately. Values higher than zero had between variance higher for multivariate normal imputation. Values lower than zero had between variance higher for regression imputation.

Figure 5 shows the differences in between variance of multivariate normally imputed data and regression imputed data. Most Likert scale variables had approximately the same between variance for both imputation methods. One variable had higher regression imputation between variance than normal imputation between variance. This variable asked whether flexible sigmoidoscopy was not recommended to screen for colorectal cancer because other health concerns take precedence. We have no theoretical reason why this specific question resulted in a remarkable difference between the two imputation methods.

For binary variables, no remarkable differences in between variance occured. For covariates there was virtually no difference between normal imputation and hot deck imputation.

6. Discussion

We compared two imputation methods for imputation of a large number of variables. Typically, in survey research of this type, multiple analyses need to be done on the data. Consistency across analyses is desirable even though the dependent variable may vary across. Multiple imputation of missing data can facilitate this consistency. Two multiple imputation methods were applied and compared.

In the case considered here, small to moderate amounts of missing data were present. However, listwise deletion of the dataset would result in only 289 complete cases left to analyze from the original 882 returned surveys. We studied two multiple imputation methods in order to obtain five completed datasets. We found that regression imputation resulted in complete data statistics similar to the multivariate normally imputed data.

A limitation of the current research is that only two methods were compared. Methods were chosen for their ease of availability and implementation. S-plus programs for hot deck and regression that were used here can be downloaded free of charge from www.stat.ucla.edu/~coen/regress.php, and NORM can be obtained from the web free of charge as well. Other, more complicated multiple imputation schemes may require building a full Bayesian model. Regression and multivariate normal imputation are known to result in proper imputations (Rubin, 1987, p.118), if the model is correct for the missing data.

Another limitation is the lack of "correct" imputed values. This paper is based on real data with real nonresponse. It is not based on simulated data in which the missing values and the missing data mechanism are known.

For data similar to that presented in this paper, both imputation methods appear to give similar results, despite the large number of binary variables. Accordingly, we endorse the application of either of these two methods to missing data; we believe either would result in satisfactory multiply imputed data sets. Findings in analyses based on our multiply imputed

data sets do not appear to be an artifact of the chosen multiple imputation method. Since a large number of variables is common practice in survey research, with limited amounts of nonresponse both regression and multivariate normal appear viable options.

References

ACS (2002). Cancer Facts and Figures, 2002. Atlanta: American Cancer Society, publication no. 5008.02.

Barnard J., Meng X. L. (1999). Applications of multiple imputation in medical studies: from AIDS as NHANES *Statistical Methods in Medical Research*, **8**, 17-36.

Bennett D. A. (2001). How can I deal with missing data in my study? *Australian And New Zealand Journal of Public Health*, **25**, 464-469.

Breen N., Wagener D. K., Brown M. L., Davis W. W., Ballard-Barbash R. (2001). Progress in cancer screening over a decade: results of cancer screening from the 1987, 1992, and 1998 National Health Interview Surveys. *Journal of the National Cancer Institute*, **93**, 1704-1713.

Dempster, A. P. and Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, **39**, 1-38.

Jemal, A., Thomas, A., Murray, T. and Thun, M. (2002) Cancer Statistics, 2002. *Cancer Journal for Clinicians*, **52**, 23-47.

Dulai, G. S., Farmer, M. M., Ganz, P. A., Bernaards, C. A., Qi, K., Dietrich, A., Bastani, R., Belman, M., Kahn, K. L. (submitted). Barriers and facilitators to colorectal cancer screening test utilization by primary care providers in a managed care setting.

Hediger M. L., Overpeck M. D., McGlynn A., Kuczmarski R. J., Maurer K. R., Davis W. W. (1999). Growth and fatness at three to six years of age of children born small— or large—for—gestational age. *Pediatrics*, **104**, (3), G1-G6.

Heitjan D. F. and Landis J. R. (1994). Assessing secular trends in blood-pressure – a multiple-imputation approach. *Journal of the American Statistical Association*, **89**, 750-759.

Heitjan D. F. and Little R. J. A. (1991). Multiple imputation for the fatal accident reporting system. *Applied Statistics – Journal of the Royal Statistical Society Series C*, **40** 13-29.

Horton, N. J. and Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician*, **55**, 244-254.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Little, R. J. A. and Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods and Research*, **18**, 292-326.

Molenberghs G., Burzykowski T., Michiels B., and Kenward, M.G. (1999). Analysis of incomplete public health data. *Revue d Epidemiologie et de Sante Publique*, **6**, 499-514.

Patrician P. A. (2002). Multiple imputation for missing data *Research in Nursing & Health*, **25**, 76-84.

Reilly, M. (1993). Data analysis using hot deck multiple imputation. *The Statistician*, **42**, 307-313.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, **91**, 473-489.

Rubin, D. B. and Schenker, N. (1991). Multiple imputation in health-care databases: an overview and some applications. *Statistics in Medicine*, **10**, 585-598.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer J. L. (1999a). Multiple imputation: a primer *Statistical Methods in Medical Research*, **8**, 3-15.

Schafer, J. L. (1999b) NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2. Software for Windows 95/98/NT, available from the URL: http://www.stat.psu.edu/ jls/misoftwa.html.

Schafer J. L. and Olsen M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective *Multivariate Behavioral Research*, **33**, 545-571.

Schenker, N., and Taylor, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis*, **22**, 425-446.

Winawer S. J., Fletcher R. H., Miller L., Godlee F., Stolar M. H., Mulrow C. D., Woolf, S. H., Glick, S. N., Ganiats T. G., Bond, J. H., Rosen, L., Zapka, J. G., Olsen, S. J., Giardiello, F. M., Sisk, J. E., Van Antwerp, R., Brown-Davis, C., Marciniak, D. A., and Mayer, R. J. (1997). Colorectal cancer screening: clinical guidelines and rationale. *Gastroenterology*, **112**, 594-642.

Received July 31, 2002; accepted November 15, 2002

Coen A. Bernaards
UCLA, Division Cancer Prevention and Control Research
Jonsson Comprehensive Cancer Center
A2-125 CHS, 650 Charles Young Drive South
Box 956900
Los Angeles, CA 90095-6900
coen@stat.ucla.edu

Melissa M. Farmer
UCLA, Division Cancer Prevention and Control Research
Jonsson Comprehensive Cancer Center
A2-125 CHS, 650 Charles Young Drive South
Box 956900
Los Angeles, CA 90095-6900
farmermm@ucla.edu

Karen Qi
UCLA, Division Cancer Prevention and Control Research
Jonsson Comprehensive Cancer Center
A2-125 CHS, 650 Charles Young Drive South
Box 956900
Los Angeles, CA 90095-6900
xqi@ucla.edu

Patricia A. Ganz
UCLA, Division Cancer Prevention and Control Research
Jonsson Comprehensive Cancer Center
A2-125 CHS, 650 Charles Young Drive South
Box 956900
Los Angeles, CA 90095-6900
pganz@mednet.ucla.edu

Gareth S. Dulai
UCLA, Department of Medicine
Division of digestive diseases
44-138 CHS, 10833 Le Conte Ave
Los Angeles, CA 90095
gdulai@mednet.ucla.edu

Katherine L. Kahn
UCLA, Division of General Internal Medicine
911 Broxton Plaza
Los Angeles, CA 90095-1736
kkahn@ucla.edu