Journal of Biometrics & Biostatistics

ISSN: 2155-6180

Gebregziabher et al., J Biomet Biostat 2012, 3:3
DOI: 10.4172/2155-6180.1000138

Research Article

Open Access

Lessons Learned in Dealing with Missing Race Data: An Empirical Investigation

Mulugeta Gebregziabher[1,2,4]*, Yumin Zhao[1,4], Neal Axon[1,3,4], Gregory E. Gilbert[1], Carrae Echols[1] and Leonard E. Egede[1,3,4]

[1] *Center for Disease Prevention and Health Interventions for Diverse Populations, Ralph H. Johnson VA Medical Center, Charleston, South Carolina, USA*
[2] *Division of Biostatistics and Epidemiology, Medical University of South Carolina, Charleston, South Carolina, USA*
[3] *Division of General Internal Medicine and Geriatrics, Medical University of South Carolina, Charleston, South Carolina, USA*
[4] *Center for Health Disparities Research, Medical University of South Carolina, Charleston, South Carolina, USA*

Abstract

Background: Missing race data is a ubiquitous problem in studies using data from large administrative datasets such as the Veteran Health Administration and other sources. The most common approach to deal with this problem has been analyzing only those records with complete data, Complete Case Analysis (CCA) which requires the assumption of Missing Completely At Random (MCAR) but CCA could lead to biased estimates with inflated standard errors.

Objective: To examine the performance of a new imputation approach, Latent Class Multiple Imputation (LCMI), for imputing missing race data and make comparisons with CCA, Multiple Imputation (MI) and Log-Linear Multiple Imputation (LLMI).

Design/Participants: To empirically compare LCMI to CCA, MI and LLMI using simulated data and demonstrate their applications using data from a sample of 13,705 veterans with type 2 diabetes among whom 23% had unknown/missing race information.

Results: Our simulation study shows that under MAR, LCMI leads to lower bias and lower standard error estimates compared to CCA, MI and LLMI. Similarly, in our data example which does not conform to MCAR since subjects with missing race information had lower rates of medical comorbidities than those with race information, LCMI outperformed MI and LLMI providing lower standard errors especially when relatively larger number of latent classes is assumed for the latent class imputation model.

Conclusions: Our results show that LCMI is a valid statistical technique for imputing missing categorical covariate data and particularly missing race data that offers advantages with respect to precision of estimates.

Introduction

Much research in the past decade has focused on inequalities in the health and healthcare of different racial/ethnic groups [1]. For example, studies in the Veterans Affairs (VA) health system have revealed disparities in medication adherence, surgery and invasive procedures, and other care processes [2]. In these types of health disparities research, race/ethnicity is the key covariate of interest. Race/ethnicity is also a potential adjustment factor in most analyses of healthcare data. Thus, accurate race/ethnicity information is imperative for these types of studies to lead to high-quality health services research.

However, analysis of race/ethnicity data is hampered by the high proportion of missing race/ethnicity data [3,4]. For example, in data from 1997 to 2005 used by Sohn et al. [3], 45% of Veterans had missing or unknown race/ethnicity in their records. Compounding the problem, race/ethnicity data are not usually missing completely at random (MCAR) as patients with higher degrees of comorbidity and service connectedness are less likely to have missing race/ethnicity [5]. Such missingness in race/ethnicity data can bias results in health disparity studies unless properly accounted for. On the other hand, our experience analyzing both local and national VA data (with over 20% missing race information) indicates both complete case analysis (CCA) and multiple imputation (MI) did not result in different inferences [6,7] despite the fact that CCA is only valid under MCAR and MI is more appropriate under missing at random (MAR).

Several methods of handling missing covariate data are available in the literature. The default analysis in many software programs is Complete Case Analysis (CCA) which requires a strong assumption of MCAR and is known to lead to biased statistical inference if MCAR is violated. Another approach available in most commercial statistical software packages is multiple imputations which in standard implementations uses the distribution of the observed data to estimate a set of plausible values for missing data. It requires the assumption of MAR and is widely known to lead to unbiased estimates that are reflective of the uncertainty due to missingness [8,9]. The most commonly used imputations models for missing categorical data such as race are logistic regression and discriminant analysis models. Both are appropriate for imputing categorical variables with monotone missing data pattern. The latter is appropriate when the predictors are multivariate normal with equal within group covariance matrix. Related to MI are multiple imputations using chained equations (MICE) [10]. This is a fully conditional method where the imputation model for each variable with missing values is specified using a conditional distribution and draws from the conditional distribution are used to impute the missing values. A potential limitation is that a draw from each conditional distribution may not always lead to a draw from the joint distribution [11] and lack of theoretical basis [12].

***Corresponding author:** Mulugeta Gebregziabher, Division of Biostatistics and Epidemiology, Medical University of South Carolina, Charleston, USA, Tel: 843-876-1112; Fax: 843-876-1126; E-mail: gebregz@musc.edu

Received February 13, 2012; **Accepted** April 13, 2012; **Published** April 15, 2012

Citation: Gebregziabher M, Zhao Y, Axon N, Gilbert GE, Echols C, et al. (2012) Lessons Learned in Dealing with Missing Race Data: An Empirical Investigation. J Biomet Biostat 3:138. doi:10.4172/2155-6180.1000138

Finally, log-linear multiple imputation (LLMI) uses a fully saturated log-linear imputation model which is a logistic regression model with all two-way and higher order interaction terms included [13,14].

Latent class multiple imputation (LCMI) is an alternative multiple imputation approach [15,16]. It uses a latent class model to estimate the joint distribution of the observed data. It provides a latent classification of subjects (latent classes) which are explained by the relationship among the observed categorical variables. Data for subjects in each class are then used to impute the missing values of subjects within the same class. LCMI has been shown to produce estimates with minimal bias and smaller standard errors in the analysis of data with missing categorical covariates [15]. This report describes an empirical comparison of the performance of CCA, MI, LLMI and LCMI in dealing with missing race/ethnicity data in a dataset similar to those used by many health services researchers.

Motivating Data Example

Our data included a retrospective cohort of 13,705 veterans with type 2 diabetes recruited from a tertiary center and five community-based outpatient clinics in the southeastern United States. The diagnosis of diabetes was based on a previously validated algorithm for VA data [17]. Subjects were followed from September 1996 until death, loss to follow-up, or May 2006. Among these, 77% had observed race/ethnicity information while the remaining 23% did not have known race/ethnicity data. The details of the creation of the study data set are provided in an earlier paper [6,7]. The study was approved by our institutional review board (IRB) and local VA Research and Development committee.

Outcome measures: The outcome variable was annual mean HbA1c calculated from measurements taken in three-month intervals. It was categorized as good control (HbA1c ≤ 8%) or poor control (HbA1c>8%). When HbA1c values were not observed in a 3-month period, they were considered as missing values. For subjects with two or more HbA1c values in a given three-month time interval, the most recent HbA1c for that interval was used.

Predictor variables: Other risk factors (or covariates) included age, gender, race/ethnicity, marital status, employment status and co-morbidities. Based on age distribution in the VA, age was categorized into four groups (<50, 50-64, 65-74, 75 and above). Race/ethnicity was classified as non-Hispanic white and non-Hispanic black, Hispanic/Other and unknown race/ethnicity. Marital status was classified as never married, married, or separated/widowed/divorced. Employment was classified as employed, not employed, or retired. Comorbidity variables (Table 1) were defined based on enhanced ICD-9 codes using validated algorithms [18].

Simulation Study

Additionally, we used a limited simulation study to demonstrate and make comparisons among the different methods. The details of the simulation study design are as follows. We generated binary data to study how well LCMI performs compared to MI, LLMI and CCA. Especially, the simulation study addresses how the estimation of a

Variables	All (n=13705)	With Race (n=10551)	Without Race (n=3154)	P value
age (mean ± sd)	56.8 ± 11.6	56.6 ± 11.6	57.6 ± 11.5	<.0001
HbA1c (mean ± sd)	7.2 ± 1.5	7.3 ± 1.5	7.2 ± 1.4	0.0077
age group <50	3906 (28.5%)	3135 (29.7%)	771 (24.4%)	<.0001
50-64	5851 (42.7%)	4444 (42.1%)	1407 (44.6%)	
65-74	3112 (22.7%)	2328 (22.1%)	784 (24.9%)	
75+	836 (6.1%)	644 (6.1%)	192 (6.1%)	
Male	13293 (97%)	10233 (97%)	3060 (97%)	0.9228
Never Married	766 (5.6%)	627 (5.9%)	139 (4.4%)	<.0001
Married	8885 (64.9%)	6691 (63.4%)	2194 (69.8%)	
Divorced/separated/widowed	4037 (29.5%)	3228 (30.6%)	809 (25.7%)	
Employed/self employed	3236 (25.6%)	2433 (24.9%)	803 (28.3%)	<.0001
Not Employed	5744 (45.5%)	4635 (47.4%)	1109 (39.1%)	
Retired	3643 (28.9%)	2716 (27.8%)	927 (32.7%)	
HbA1c ≥ 8	3251 (23.7%)	2575 (24.4%)	676 (21.4%)	0.0006
anemia	694 (5.1%)	570 (5.4%)	124 (3.9%)	0.001
cancer	2048 (15%)	1682 (16%)	366 (11.6%)	<.0001
Cerebrovascular Disease	630 (4.6%)	527 (5%)	103 (3.3%)	<.0001
Congestive Heart Failure	1886 (13.8%)	1585 (15%)	301 (9.6%)	<.0001
Cardiovascular Disease	2392 (17.5%)	1944 (18.4%)	448 (14.2%)	<.0001
Hypertension	11368 (83.1%)	8914 (84.6%)	2454 (78%)	<.0001
Hypothyroidism	749 (5.5%)	573 (5.4%)	176 (5.6%)	0.7353
Liver Disease	730 (5.3%)	619 (5.9%)	111 (3.5%)	<.0001
Chronic Lung Diseases	2389 (17.5%)	1964 (18.6%)	425 (13.5%)	<.0001
Fluid/Electrolyte Disorders	940 (6.9%)	817 (7.8%)	123 (3.9%)	<.0001
Obesity	4165 (30.4%)	3291 (31.2%)	874 (27.8%)	0.0002
Other Diseases	2678 (19.6%)	2263 (21.5%)	415 (13.2%)	<.0001
Peripheral Vascular Disease	2037 (14.9%)	1709 (16.2%)	328 (10.4%)	<.0001
Depression	3513 (25.7%)	2898 (27.5%)	615 (19.5%)	<.0001
Psychoses	729 (5.3%)	648 (6.1%)	81 (2.6%)	<.0001
Substance Abuse	1309 (9.6%)	1121 (10.6%)	188 (6%)	<.0001

Table 1: Patient Characteristics by availability of race/ethnicity information among Veterans with type-2 diabetes.

few latent classes can improve upon standard multiple imputation techniques. We simulated data with a binary outcome variable (Y) where the Pr(Y=1) was determined using a logistic regression model given by, $\text{logit}(\Pr(Y = 1|X)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_5 X_5$. Where each of the X's are also binary variables with X_1 denoting race (0=white, 1=black), which can be missing and the other four covariates X_2 to X_5 that were generated jointly with X_1 are covariates were potential confounders of the relationship between X_1 and Y. For simulations, we set β_1=0.69 which is equivalent to an odds ratio of 2.0. After generating complete data according to the above model, data sets with missing race (X_1) were generated from the cohort with a 30% and 50% missing proportion. The probability of missingness was $\text{logit}\ (\Pr\ (M{=}1)){=}\gamma_0{+}\gamma_1 Y{+}\gamma_2 X_2$, where M is an indicator of missing X_1. We considered a wide range of missing scenarios broadly based on missing data mechanism classifications in Little and Rubin (2002) given as missing completely at random (MCAR) and missing at random (MAR). Further specification of the missingness within MAR was based on the dependence of the probabilities of missing X_1 on another covariate X_2 or on the outcome (Y). That is, missing X_1 may depend on X_2 (MAR(X_2)), on both X_2 and Y (MAR(X_2, Y)), or on Y only (MAR(Y)). For a 30% missingness to create MCAR, MAR(Y), MAR(X_2) and MAR(X_2, Y), we used gamma0=(-0.9,-1.9,-2.0,-2.2) with gamma1=(0, 0, 1.8,0, 0.8) and gamma2=(0,0, 0.8, 0.8).

Missing Data Analysis Using LCMI

LCMI is a multiple imputation approach where the imputation model is based on a latent class model. The latent class imputation model provides a latent classification of subjects (latent classes) which provide a sufficient representation of the joint distribution which explains the complex relationship among the observed categorical variables and is used as a tool for density estimation [16]. That is, the observed data of subjects within each of the estimated classes are used to impute the missing values of subjects within the same class. [15,16]. Advantages of LCMI include the ability to model complex associations between categorical variables– a distinct advantage over LLMI. The details of this approach are in Gebregziabher and Desantis (2010).

The LCMI is implemented in four steps as follows. Let $X_{i,obs}$ and K denote the observed data (all fully observed variables in table 1 including the outcome variable) and the estimated latent class respectively.

Step 1: Fit the latent class model to the observed data, $X_{i,obs}$.

Step 2: Sample from the posterior probability distribution of latent class given the observed data, $P(K_i{=}k|X_{i,obs}{=}x_{i,obs})$.

Step 3: Sample from the distribution of the missing data conditional on class, $P(X_{i,mis}|K_i{=}k)$.

Step 4: Use a within class posterior sampling to impute the missing category or value of X. In our case, the latent class model was fitted using proc LCA Version 1.1.5 [21,22]. PROC LCA is a SAS procedure for latent class analysis developed for SAS Version 9.2 for Windows and is used to estimate latent classes measured by categorical indicators. We used a full Bayesian MCMC approach to sample from the posterior distribution of the missing data model. Finally, after we imputed the missing categories we used likelihood and/or GEE methods to estimate the parameters (regression coefficients and their corresponding standard errors) of the HbA1c models as described below.

For each missing data analysis method, we performed two sets of analyses. First, for the crossectional data analysis we used logistic regression (PROC Logistic, SAS 9.1.3) to study the association between HbA1c control (1=HbA1c >8%, 0=HbA1c ≤ 8 %) and race/ethnicity with and without adjusting for demographic and clinical variables. For each subject, HbA1c control was defined as mean HbA1c being 8% or less over the entire study period.

Second, for the longitudinal data analysis we used a general estimating equations (GEE) approach [19,20] using PROC Genmod, SAS 9.1.3 to assess whether HbA1c control varied by race/ethnicity. Both unadjusted and covariate adjusted models were fitted with HbA1c control at each quarterly visit as response variable using time and race/ethnicity as primary variables of interest. The final model was adjusted for demographic variables and comorbidities.

Results

Table 1 shows the patient characteristics for the 13,705 Veterans with type 2 diabetes included in the study sample stratified by whether race/ethnicity information was available or not. The patients with missing race data had lower levels of comorbid conditions compared to those with race data. For instance, the prevalence of cancer was 11.6% in those without race compared to 16% in those with race. Similarly, the prevalence of CHF was 9.6% in those without race compared to 15% in those with race. Similar trends were observed for most of the medical and psychiatric comorbidities. Most Veterans (76.4%) had HbA1c values ≤8% and the proportions were 75.6% and 78.6% in those with race and without race respectively. The mean HbA1c was 7.3 (sd 1.5) in those with race and 7.2 (sd 1.4) in those without race information.

Table 2 shows parameter estimates from two different analysis scenarios. The first two columns indicate the method used and the categories of race/ethnicity. The first column lists the four missing data methods used for comparison with four row entries for LCMI that vary by the number of latent classes consider for the imputation model.The remaining four columns are odds ratio (OR) estimates and their corresponding standard error (SE) estimates. While the last two columns are from analysis of the longitudinal data with a binary

		Dichotomous HbA1c			
		crossectional		longitudinal	
Method	Race	OR	SE	OR	SE
	NHW(ref)	1.000	-	1.000	-
CCA	NHB	1.787	0.052	1.635	0.040
	Others	0.839	0.212	1.088	0.146
MI	NHB	1.815	0.051	1.624	0.039
	Others	0.874	0.200	1.070	0.137
LCMI-2	NHB	1.806	0.049	1.614	0.038
	Others	0.896	0.197	1.068	0.139
LCMI-3	NHB	1.827	0.047	1.629	0.037
	Others	0.921	0.184	1.087	0.134
LCMI-4	NHB	1.830	0.048	1.630	0.038
	Others	0.915	0.182	1.088	0.134
LCMI-5	NHB	1.827	0.046	1.628	0.036
	Others	0.917	0.181	1.083	0.134
LLMI	NHB	1.749	0.049	1.577	0.038
	Others	1.070	0.188	1.186	0.139

CCA=Complete case analysis, MI= Multiple imputation with logit imputation model, LCMI-k = Latent class imputation with k (k=2,..,5) classes, LLMI: Log-linear multiple imputation

Longitudinal=estimates from Proc Genmod with logit link, crossectional=logistic regression with dichotomized mean HbA1c of the repeated measurements over time

Table 2: Parameter estimates with corresponding standard error (SE) for comparing Non-Hispanic Black (NHB) and other race/ethnic groups with Non-Hispanic Whites (NHW) in the association study of Glycemic Control and race/ethnicity in Veterans with type-2 diabetes.

	CCA		MI		LLMI		LCMI	
Missing Mechanism	Bias	ASE	Bias	ASE	Bias	ASE	Bias	ASE
MCAR	0.09	0.51	0.05	0.42	0.03	0.42	0.03	0.40
MAR(X2)	0.10	0.53	0.04	0.44	0.03	0.45	0.02	0.41
MAR(Y)	0.09	0.55	0.05	0.43	0.03	0.43	0.03	0.42
MAR(Y,X2)	0.10	0.57	0.03	0.44	0.01	0.45	0.01	0.43

MCAR = Missing completely at random
MAR(X2) = Missing at random that depends on X2 (eg. age of the person)
MAR(Y) = Missing at random that depends on outcome Y (eg. HbA1c)
MAR(X2,Y) = Missing at random that depends on outcome Y (eg. HbA1c) and X2 (eg. Age)
CCA= Nomplete case analysis
MI = Multiple imputation
LLMI = Log-linear multiple imputation
LCMI = Latent class multiple imputation with three classes

Table 3: Bias in the mean log-odds ratio (Bias=estimated mean – 0.69) and asymptotic standard error (ASE) estimates of a logistic regression model from a simulation study with 30% missing race data (n=200, true log odds ratio=0.69).

	CCA		MI		LLMI		LCMI	
Missing Mechanism	Bias	ASE	Bias	ASE	Bias	ASE	Bias	ASE
MCAR	0.16	0.80	0.07	0.52	0.02	0.54	0.03	0.47
MAR(X2)	0.18	0.93	0.07	0.58	0.05	0.68	0.02	0.54
MAR(Y)	0.19	0.93	0.08	0.54	0.01	0.57	0.03	0.49
MAR(Y,X2)	0.17	0.92	0.04	0.54	0.03	0.60	0.01	0.50

MCAR = Missing completely at random
MAR(X2) = Missing at random that depends on x2 (eg. age of the person)
MAR(Y) = Missing at random that depends on outcome y (eg. HbA1c)
MAR(X2,Y) = Missing at random that depends on outcome Y (eg. HbA1c) and X2 (eg. Age)
CCA= Complete case analysis
MI = Multiple imputation
LLMI = Log-linear multiple imputation
LCMI = Latent class multiple imputation with three classes

Table 4: Bias in the mean log-odds ratio (Bias=estimated mean – 0.69) and asymptotic standard error (ASE) estimates of a logistic regression model from a simulation study with 50% missing race data (n=200, true log odds ratio=0.69).

outcome analyzed using Proc GENMOD, the middle two columns are OR and SE estimates from the analysis of the crossectional data with binary HbA1c analyzed using Proc LOGISTIC in SAS 9.2. The OR and SE columns show the odds ratio and their corresponding standard error estimates for NHB and Other races with NHW as the reference category. The odds ratio estimates are relatively similar across the different methods ranging between 1.75 using LLMI to 1.83 using LCMI-4 for NHB and between 0.84 using CCA and using LLMI for Others. The standard error estimates for CCA are slightly higher than the other imputation methods. For example, in the crossectional setting, SE for odds ratio comparing NHB and NHW using CCA was 0.052 and this was reduced to 0.046 using LCMI-5 representing a 12% improvement in precision. The same trend is observed in comparing Other to NHW in both the crossectional and longitudinal settings. In almost all cases, LCMI provided more precise estimates than the other methods. In all scenarios, LCMI-5 provided lower standard error estimates.

The simulation study results for a 30% and 50% missing data scenario are reported in Table 3 and Table 4 respectively. Both absolute bias (estimated value minus true value) and asymptotic standard error are reported. Under the 30% scenario, the biases for CCA were not very large and were similar across both MCAR and MAR mechanisms. However, the size of the bias was between two and three fold compared to MI, LLMI or LCMI. On the other hand, when the level of missingness was increased to 50% the bias in CCA increased substantially to up to 25% while the bias in LLMI and LCMI remained low. In summary, the simulation results indicate that LCMI leads to parameter estimates that are less biased and characterized by lower standard errors compared to CCA, MI and LLMI. A more detailed and rigorous simulation study about these comparisons is reported elsewhere [12].

Discussion

Health services researchers who examine large datasets require complete information on covariates in order to perform accurate analyses. For example, in health disparities research, race/ethnicity is the key covariate of interest. However, race data is substantially missing in some VA data sets as well as other data sources. Thus, robust statistical techniques are needed to deal with the problem of missing race/ethnicity data in studies of health disparities and in other applications. This report provides empirical evidence on the performance of multiple imputation techniques with varying imputation models in dealing with missing race data.

Imputation techniques are preferable to other approaches in many cases. It is often invalid to assume that race/ethnicity data is missing completely at random [5]. In the case of VA analyses, supplementing missing race/ethnicity information with data from other sources such as Medicare is not always possible. Moreover, it may result in higher rates of misclassification in non-Black minorities [23-25].

Among imputation methods, LCMI offers some advantages over MI and LLMI. Many datasets like ours have multiple variables with missing categorical data which make it difficult to satisfy the assumption of a multivariate normal distribution needed to perform MI. In LCMI, the imputation is based on a latent class model which does not require the same distribution assumptions. Thus, LCMI represents an alternative to MI that may perform better in certain datasets. The LLMI approach requires a saturated imputation model that includes all higher order associations among categorical variables. Because of this, LLMI may be computationally infeasible even for a moderately large number of variables. In contrast, LCMI represents a more computationally efficient approach. Importantly, LCMI performed comparably to these other techniques in a simulation study while providing more precise estimates with lower standard errors [15]. The main limitation of LCMI is that there are no proved approaches to determine the number of latent classes in a latent class imputation model that are sufficient to well approximate the joint distribution of the variables in the data. In the limited studies in the literature, it has been recommended to use as many latent classes as possible [15,16].

In summary, LCMI represents a valid statistical approach for the imputation of missing categorical data such as race/ethnicity data that may be of use to a variety of health services researchers working with large administrative datasets.

Disclosure

This study was supported by Grant # REA 08-261, Center for Disease Prevention and Health Interventions for Diverse Populations funded by Veterans Affairs Health Services Research and Development (PI – Leonard Egede).

Conflict of Interest

None of the authors have any financial disclosure or conflict of interest to report.

References

1. Smedley BD, Stith AY, Nelson AR. Institute of Medicine (2008) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. Unequal Treatment Confronting Racial and Ethnic Disparities in Healthcare. Washington, D.C.: National Academy Press; 2002. Centers for Disease Control and Prevention. National diabetes fact sheet: general information and national estimates on diabetes in the United States, 2007. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.
2. Saha S, Freeman M, Toure J, Tippens KM, Weeks C, et al. (2008) Racial and ethnic disparities in the VA health care system: a systematic review. J Gen Intern Med 23: 654-671.
3. Sohn MW, Zhang H, Arnold N, Stroupe K, Taylor BC, et al. (2006) Transition to the new race/ethnicity data collection standards in the Department of Veterans Affairs. Popul Health Metr 4: 7.
4. Long JA, Bamba MI, Ling B, Shea JA (2006) Missing race/ethnicity data in Veterans Health Administration based disparities research: a systematic review. J Health Care Poor Underserved 17: 128-140.
5. Meghani SH, Wiedemer NL, Harden P, Pulman-Mooar S, Garvin J, et al. (2008) Missing Race/Ethnicity Data in the VA Longitudinal Online Research Database: Implications for Health Disparities Research: Paper presented at: VA HSR&D National Meeting, Baltimore, USA.
6. Egede LE, Mueller M, Echols CL, Gebregziabher M (2010) Longitudinal Differences in Glycemic Control by Race and Ethnicity among Veterans with Type 2 Diabetes. Med Care 48: 527-533.
7. Gebregziabher M, Egede LE, Lynch CP, Echols C, Zhao Y (2010) Effect of trajectories of glycemic control on mortality in type 2 diabetes: a semiparametric joint modeling approach. Am J Epidemiol 171: 1090-1098.
8. Rubin DB (1987) Multiple Imputation for Nonresponse in Surveys. Wiley: New York.
9. Little RJA, Rubin DB (2002) Statistical analysis with missing data. (2nd edn.), Wiley: Hoboken NJ.
10. Van Buuren S, Oudshoorn K (2000) MICE: Multivariate imputation by chained equations. MICE V1.0 User's Manual, TNO Report PG/VGZ/00.038, TNO prevention and Health: Leiden.
11. Van Buuren S (2007) Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 16: 219-242.
12. White IR, Royston P, Wood AM (2011) Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 30: 377–399.
13. Schafer JL (1997) Analysis of Incomplete Multivariate Data. Chapman and Hall, London, UK.
14. Agresti A (2002) Asymptotic Theory for Parametric Models. Categorical data analysis (2nd edn.), Wiley: Hoboken, NJ.
15. Gebregziabher M, DeSantis S (2010) A latent class based multiple imputation approach for missing categorical data. J Stat Plan Inference 140: 3252-3262.
16. Vermunt JK, Van Ginkel JR, Van der Ark LA, Sijtsma K (2008) Multiple imputation of incomplete categorical data using latent class analysis. Sociol Methodol 38: 369-397.
17. Miller DR, Safford MM, Pogach LM (2004) Who has diabetes? Best estimates of diabetes prevalence in the Department of Veterans Affairs based on computerized patient data. Diabetes Care 27: B10–B21.
18. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, et al. (2005) Coding algorithms for defining comorbidities in ICD-9 and ICD-10 administrative data. Med Care 43: 1130-1139.
19. Verbeke G, Molenberghs G (2000) Linear Mixed Models for Longitudinal Data. New York: Springer-Verlag.
20. Fitzmaurice G, Laird N, Ware J (2004) Applied Longitudinal Analysis. New York: John Wiley and Sons.
21. Lanza ST, Collins LM, Lemmon DR, Schafer JL (2007) Proc LCA: a SAS procedure for latent class analysis. Struct Equ Modeling 14: 671–694.
22. Collins LM, Lanza ST (2010) Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences. New York: John Wiley and Sons: 331.
23. Haywood T, Tarlov E, Owens A, Weichle TW, Hynes DM (2009) Improving Patient Race and Ethnicity Information in VA Studies: Agreement between VA and Medicare Data. Paper presented at: VA HSR&D National Meeting, Baltimore MD In.
24. VIReC (2009) Race data quality update.
25. Lauderdale DS, Goldberg J (1996) The expanded racial and ethnic codes in the Medicare data files: their completeness of coverage and accuracy. Am J Public Health 86: 712-716.