ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

On Imputing Binary Data via Pairwise Associations and Corresponding Conditional Probabilities

İlişki Katsayıları ve Koşullu Olasılıkları Kullanarak İki Sonuçlu Veriler İçin Değer Atama

Irene B. HELENOWSKI, PhD,[a,b]
Hakan DEMİRTAŞ, Dr., Assoc.Prof.,[b]
Beyza DOĞANAY ERDOĞAN, PhD[c]

[a]Department of Preventive Medicine,
Feinberg School of Medicine,
Robert H. Lurie Comprehensive Cancer
Center, Northwestern University,
[b]Division of Epidemiology & Biostatistics,
School of Public Health,
University of Illinois-Chicago,
Chicago, IL, U.S.A.
[c]Department of Biostatistics,
Ankara University Faculty of Medicine,
Ankara

Geliş Tarihi/*Received:* 02.01.2012
Kabul Tarihi/*Accepted:* 01.02.2012

Yazışma Adresi/*Correspondence:*
Beyza DOĞANAY ERDOĞAN, PhD
Ankara University Faculty of Medicine,
Department of Biostatistics, Ankara,
TÜRKİYE/TURKEY
beyzadoganay@gmail.com

ABSTRACT Objective: In this work, we present a method for imputing binary data without making any multinomial or loglinear model assumptions. **Material and Methods:** Our approach employs principles of generating binary data from multivariate normally distributed values, as discussed in Emrich and Piedmonte. Specifically, a data set that follows a multivariate normal distribution is generated separately from observed data, using marginal binary proportions, and a matrix associated with pairwise tetrachoric correlations derived from phi coefficients. The same fraction of missing information is introduced in the generated data as in the original multivariate binary data, multiple imputation is then applied to the generated values via joint modeling under the normality assumption, and imputed values are dichotomized by quantiles corresponding to the binary proportions that are computed based on the original incomplete data. **Results:** Application of our imputation method to generated data in simulation studies and to real data examples led to promising results, as indicated by average estimates (AE) of pairwise correlation parameters comparable to the true correlation values associated with generated data and to the original correlation estimates involving real data, standardized bias (SB) values < 50%, small RMSE values associated with good accuracy and precision, coverage rates (CR) > 90%, and average widths (AW) of confidence intervals for correlation parameter estimates from the imputed data comparable to 95% confidence interval widths of true correlation values associated with generated data or original correlation estimates involving real data. **Conclusion:** Simulation studies and real data applications indicate that this new method is a promising approach in imputing binary data while relaxing multinomial and loglinear model assumptions.

Key Words: Missing data; multiple imputation; multivariate normal distribution; binary data

ÖZET Amaç: Bu çalışmada, iki sonuçlu değişkenler için multinomial ya da logaritmik doğrusal model varsayımları olmadan geliştirilen bir değer atama yöntemi tanıtılacaktır. Burada kullanılan yaklaşım, Emrich ve Piedmonte tarafından önerilen, çok değişkenli normal dağılım gösteren verileri kullanarak iki sonuçlu çok değişkenli veri türetme prensibine dayanmaktadır. **Gereç ve Yöntemler:** Gözlenen iki sonuçlu verilerden hesaplanan marjinal oranları ve iki sonuçlu değişkenler arası korelasyonları gösteren phi katsayılarından bulunan tetrakorik korelasyonları kullanılarak çok değişkenli normal dağılım gösteren bir veri seti türetilir. Türetilen veri setinden, gerçek veride gözlenen kayıp oranı kadar veri silinir ve çok değişkenli normal dağılıma dayalı bileşik modelleme yaklaşımı ile değer ataması yapılır. Sonra eksik veriler yerine atanan değerler, gerçek veriden hesaplanan oranlara karşılık gelen yüzdeliklere (kuantillere) göre iki sonuçlu hale dönüştürülür. **Bulgular:** Değer atama sonrasında bulunan ortalama korelasyon kestirimleri, benzetim çalışmasında belirlenen gerçek değerler ve gerçek veri setindeki kestirimler ile benzer bulunmuştur. Standardize yanlılık değerleri %50'nin altında, doğruluk ve kesinliğin yüksek olduğuna işaret eden hata kareleri ortalaması karekökü küçük, kapsama alanları da %90'ın üzerinde bulunmuştur. Değer atama sonucu bulunan güven sınırlarının ortalama genişlikleri, benzetim çalışmasından ya da gerçek veri uygulamasından elde edilen korelasyon parametre kestirimlerine ilişkin %95 güven sınırlarına benzer bulunmuştur. Yapılan benzetim çalışmaları ve gerçek veri seti uygulaması sonucunda elde edilen bulgular ışığında, önerilen değer atama yönteminin umut verici olduğu söylenebilir. **Sonuç:** Yapılan benzetim çalışmaları ve gerçek veri uygulaması, geliştirilen bu yeni yöntemin iki sonuçlu çok değişkenli veriler için multinomial ya da logaritmik doğrusal model varsayımları olmadan yapılan iyi bir değer atama yaklaşımı olduğunu göstermektedir.

Anahtar Kelimeler: Kayıp veri; çoklu değer atama; çok değişkenli normal dağılım; iki sonuçlu veri

Turkiye Klinikleri J Biostat 2012;4(1):1-9

This work focuses on approach for imputing binary data without making any assumptions associated with the multinomial or loglinear models conventionally used in imputing categorical data. Our method involves mapping binary data to normally distributed values, using principles described in Emrich and Piedmonte,[1] imputing the data via joint modeling under the normality assumption, and back-transforming the imputed values to binary data using quantiles based on marginal proportions. We recommend this approach for multiple imputation of binary data in scenarios where multinomial and loglinear assumptions are suspected to be violated and therefore should be relaxed.

Before discussing the imputation methods, we briefly introduce the mechanisms under which data can be missing as well as several different missing data patterns. The probability of missingness is independent of the observed and missing data under the Missing Completely at Random (MCAR) mechanism, only depends on the observed data under the Missing at Random (MAR) mechanism, and depends on the missing data and/or observed data under the Missing Not at Random (MNAR) mechanism.[2,3] Furthermore, missing data can involve patterns such as univariate, monotone, and arbitrary patterns. Assume that we have a data set with K variables. Then, in the univariate pattern, one variable has missing entries, while the other $K-1$ variables are completely observed. In the monotone pattern, the $(k+1)^{th}$ to the K^{th} variable have the same amount of missing information as the k^{th} variable plus an additional amount for $k = 1,\ldots,K$. Lastly, for the arbitrary pattern, missingness can occur anywhere in any of the K variables.[4]

MATERIAL AND METHODS

IMPUTING BINARY DATA VIA MULTINOMIAL AND LOGLINEAR MODELS

Schafer[5] discusses a joint modeling approach for imputing binary or categorical data which employs the EM (Expectation-Maximization), associated with taking the expectation of a complete and sufficient statistics, T, and maximizing this expectation, and DA (Data Augmentation) algorithms, in combination with the saturated multinomial model. The DA algorithm is comprised of the I-step, where values are drawn from a distribution conditional on the observed data and model parameters, and the P-step, where parameters are updated using a posterior distribution conditional on the observed and imputed data, as shown in (1) and (2), where Y_{mis*} and Y_{obs} correspond to missing and observed data, respectively, and $\theta = \{\theta_1, \theta_2, \ldots, \theta_D\}$ is the parameter vector involved with the saturated multinomial model.

$$Y_{mis*}^{(t+1)} \sim P(Y_{mis*} \mid Y_{obs}, \theta^{(t)}) \tag{1}$$

$$\theta^{(t+1)} \sim P(\theta \mid Y_{obs}, Y_{mis*}^{(t+1)}) \tag{2}$$

where $Y_{mis*}^{(t+1)}$ in (2) are the imputed data obtained via (1). The saturated multinomial model is based on a contingency table with cells denoted by subscript d=1,2,...D. The contingency table represents the distinct number of possible combinations of levels among the variables considered. For example, with three binary variables, the contingency table would have $D = 2^3 = 8$ cells. x_d $x = \{x_1,x_2,\ldots,x_D\}$ is then defined as the number of entries in cell d and θ_d, $\theta = \{\theta_1, \theta_2, \ldots, \theta_D\}$ is the probability that an entry is in cell d. Furthermore $\sum_{d=1}^{D} x_d = n$, where n if the total number of entries and $\sum_{d=1}^{D} \theta_d = 1$.[5] Therefore, the distribution function involving the multinomial model is:

$$P(x \mid \theta) = \frac{n!}{x_1! x_2! \ldots x_D!} \theta_1^{x_1} \theta_2^{x_2} \ldots \theta_D^{x_D} \tag{3}$$

When some terms, for example insignificant pairwise or higher-order interaction terms, can be eliminated, a loglinear model can be employed and parameter estimation can be achieved via the ECM (Expectation-Conditional Maximization) and DA-BIF (Data Augmentation-Bayesian Iterative Proportional Fitting) algorithms.

GENERATING BINARY DATA

To introduce our method for imputing binary data, we first explain the approach given in Emrich and Piedmonte[1] to generate binary data using normally distributed values. This approach involves the phi correlation (δ_{jk}), a special case of the Pearson correlation which measures the association between two binary variables. A cross-tabulation of two binary variables, Y_1 and Y_2, with cell counts that appear in equation (4) below, is given in Table 1. The phi coefficient, a derivative of the Pearson correlation,[6] can be calculated by:

$$\frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{10}+n_{11})(n_{00}+n_{01}) + (n_{00}+n_{10})(n_{01}+n_{11})}} \tag{4}$$

TABLE 1: Cross-tabulation of Y_1 vs. Y_2.

Y_1		Y_2
	0	1
0	n_{00}	n_{01}
1	n_{10}	n_{11}

where phi can range from:

$$\delta_{jk} \in \left\{ \begin{matrix} \max(-\sqrt{(p_j p_k / q_j q_k)}, -\sqrt{(q_j q_k / p_j p_k)}) \\ \min(+\sqrt{(p_j q_k / q_j p_k)}, +\sqrt{(q_j p_k / p_j q_k)}) \end{matrix} \right\} \tag{5}$$

$p_j = Pr(Y_j = 1),\ q_j = 1 - p_j$

This range ensures that all joint mass distribution values are nonnegative for all outcomes. We can then obtain the tetrachoric correlation ρ_{jk} using:

$$\Phi[z(p_j), z(p_k), \rho_{jk}] = \delta_{jk}(p_j q_j p_k q_k)^{1/2} + p_j p_k \tag{6}$$

[1,7] where Φ is the cumulative distribution function for a standard bivariate normal distribution. The tetrachoric correlation, a form of the Pearson correlation measuring the association between normally distributed values underling the binary variables, Y_1 and Y_2, is then used to generate a bivariate data set, Z, from bivariate normal distribution with mean vector (0,0), and 2x2 correlation matrix whose off-diagonal entries are ρ_{12}. Probabilities involving Y_1 and Y_2 pertaining to quantiles in the bivariate normal data in turn allow for the preservation of the same proportions observed in the original binary data.

This generation method can also be extended to multivariate data associated with a matrix comprised of pairwise tetrachoric correlations derived as in (6), given that the matrix is positive semi-definite. If the matrix is not positive semi-definite, a positive semi-definite closest to the derived matrix can be used.[8]

CONNECTION TO THE LURIE-GOLDBERG ALGORITHM

Lurie and Goldberg[9] developed an algorithm for generating multivariate continuous data associated with specifying marginal distributions by first generating normally distributed data and then applying the inverse marginal distribution function to the probability distribution function (PDF) values of the generated data. They first generate a matrix **X** ~ N(0, I) and multiply it to the transpose of the lower triangular matrix **L**, obtained via Cholesky decomposition from the correlation matrix, **R**, with the desired pairwise correlations between variables such that:

$$Y = XL^T \sim N(0, R) \tag{7}$$

They then obtain the standard normal PDF value for the entries of **Y**, i.e.:

$$U = \Phi(Y) \tag{8}$$

and apply the inverse function of the desired marginal distribution for each variable Y_j in the data set:

$$v_{ij} = F_j^{-1}(u_{ij}) \tag{9}$$

R', the correlation of the newly generated data set **V**, is then compared to the original correlation matrix **R**, via the RMSE:

$$RMSE = \sqrt{4SND/[k(k-1)]} \tag{10}$$

where *SND* is the squared norm of the absolute difference between **R** and **R'** and *k* is the the number of variables in the data set.

Convergence of the algorithm is determined when the RMSE falls below some constant *c* set for a desired accuracy. For example, with a desired accuracy of 0.01, the authors suggest setting $c = 0.005$.

Their algorithm therefore only requires information on the marginal distributions of the variables and pairwise correlations, but not of the joint distribution of the multivariate data which is often unknown. In the Lurie and Goldberg[9] approach, the lower triangular matrix derived via Cholesky decomposition from the matrix comprised of pairwise correlations is multiplied to the generated multivariate normally distributed data in order to preserve relationships among the variables.

Thus, the Lurie-Goldberg[9] algorithm is similar to the generation of binary data discussed in Emrich and Piedmonte[1] in that both algorithms involve generation of multivariate normal data and rely on information concerning pairwise correlations between variables. In the case of binary data, these pairwise correlations are the tetrachoric correlations derived from phi coefficients relating the binary variables, as given in equation (6). The Emrich-Piedmonte[1] algorithm differs from the Lurie-Goldberg[9] algorithm, in that mapping the normally distributed data to binary data is nonparametric, where quantiles corresponding to binary proportions are employed to dichotomize the normally distributed data. Contrarily, mapping of multivariate normally distributed data in the Lurie and Goldberg[9] algorithm involves inverse functions of marginal distributions.

IMPUTING BINARY DATA

Our algorithm for imputing binary data involves generating multivariate normally distributed data with a correlation matrix from pairwise phi correlation coefficients as described in Section 2.2. After generating these data, we then introduce the same fraction of missing entries in these data as found in the original data and calculate the proportions from the observed data.

Let R_1 and R_2 be missing indicator variables for Y_1 and Y_2. Suppose some proportion of Y_2 is missing with probability Pr($R_2 = 0$); we can therefore calculate:

$$\begin{aligned}&P(Y_1=1\,|\,Y_2=1)=P(Y_1=1,Y_2=1,R_2=1)\\&=P(Y_2=1\,|\,Y_1=1,R_2=1)P(Y_1=1,R_2=1)\\&=P(Y_2=1\,|\,Y_1=1,R_2=1)P(Y_1=1\,|\,R_2=1)P(R_2=1)\end{aligned}\qquad(11)$$

and obtain corresponding quantiles given by:

$$\begin{aligned}q_{11}&=Q_{p(Y_1=1|R_2=1)}(Z_1)\\q_{21}&=Q_{p(Y_2=1|Y_1=1,R_2=1)}(Z_2)\\q_{21}^*&=Q_{p(Y_2=1|Y_1=0,R_2=1)}(Z_2)\end{aligned}\qquad(12)$$

Assuming we have a bivariate data set where both Y_1 and Y_2 have missing entries, we define:

$$\begin{aligned}&P(Y_1=1,Y_2=1)=P(Y_1=1,Y_2=1,R_1,R_2)\\&=P(Y_2=1\,|\,Y_1=1,R_1,R_2)P(Y_1=1,R_1,R_2)\\&=P(Y_2=1\,|\,Y_1=1,R_1,R_2)P(Y_1=1\,|\,R_1,R_2)P(R_1,R_2)\\&=P(Y_1=1\,|\,Y_2=1,R_1,R_2)P(Y_2=1\,|\,R_1,R_2)P(R_1,R_2)\end{aligned}\qquad(13)$$

and obtain binary outcomes from imputed Z_1 and Z_2 values for variables Y_1 and Y_2 using quantiles based on:

$$\begin{aligned}q_{11}&=Q_{P(Y_1=1|Y_2=1,R_1=1,R_2=1)}(Z_1)\\q_{11}^*&=Q_{P(Y_1=1|Y_2=0,R_1=1,R_2=1)}(Z_1)\\q_{11}&=Q_{P(Y_2=1|Y_1=1,R_1=1,R_2=1)}(Z_2)\\q_{21}^*&=Q_{P(Y_2=1|Y_1=0,R_1=1,R_2=1)}(Z_2)\end{aligned}\qquad(14)$$

Quantiles are based on data corresponding to entries with both variables observed, i.e., R_1=1 and R_2=1. With these quantiles, we can compute proportions involving the generated bivariate normal data with variables Z_1 and Z_2 and check that the number of observed entries satisfies each condition are the same as the cell counts given in Table 1.

I.e.,

$$\begin{aligned}&\sum I(Y_1=1,Y_2=1\,|\,R_1,R_2)=\sum I(Z_1<q_{11},Z_2<q_{21}|\,R_1,R_2)\\&\sum I(Y_1=1,Y_2=0\,|\,R_1,R_2)=\sum I(Z_1<q_{11},Z_2>q_{21}|\,R_1,R_2)\\&\sum I(Y_1=0,Y_2=1\,|\,R_1,R_2)=\sum I(Z_1>q_{11},Z_2<q_{21}^*\,|\,R_1,R_2)\\&\sum I(Y_1=0,Y_2=0\,|\,R_1,R_2)=\sum I(Z_1>q_{11},Z_2>q_{21}^*\,|\,R_1,R_2)\end{aligned}\qquad(15)$$

Extending our method to the multivariate case involves basing our quantiles on probabilities:

$$\begin{aligned}&Pr(Y_k=y_k\,|\,Y_1=y_1,\dots,Y_{k-1}=y_{k-1},R_1,\dots,R_K)\\&y_k=0,1;k=1,\dots,K\end{aligned}\qquad(16)$$

where y_k = 0,1, k = 1, . . . ,$K \geq$ 3 and K is the number of variables in our data set.

Probabilities for all K variables can then be derived from the joint probability with corresponding quantiles obtained via

$$q_k=Q_{Pr(Y_k=1|Y_1,Y_2,\dots,Y_{k-1},R_1,R_2,\dots,R_k)}(Z_k)\qquad(17)$$

Here, $Z_1, \dots, Z_K$ comprise the normally distributed variables of a data set, Z, with mean **0**, and correlation matrix P, where the elements are pairwise tetrachoric correlations derived from the pairwise phi correlations via equation (6). We then impute data in the multivariate normal data set, Z, as described in Schafer[5], and dichotomize the newly imputed data for each variable k, k = 1,...,K via the quantiles obtained from equation (17).

Different measures can furthermore be utilized to assess the performance of multiple imputation such as the average estimate (AE), standardized bias (SB), root mean square error (RMSE), coverage rate (CR), and average width

(AW) of the confidence intervals for parameter estimates of θ_j's. The average estimate (AE) is defined as the average of all parameter estimates obtained from $m > 1$ imputed data sets and is computed as:

$$\overline{\theta} = \sum_{j=1}^{m} \hat{\theta}_j \tag{18}$$

This AE value can be compared to the true parameter via the T approximation:

$$T^{-1/2}(\overline{\theta} - \theta) \sim t_v \tag{19}$$

where T is the total variance comprised of the within-imputation variance, U, and between-imputation variance, B, given in (20) and (21), respectively.

$$\overline{U} = m^{-1} \sum U_j \tag{20}$$

$$B = (m-1)^{-1} \sum (\hat{\theta}_j - \overline{\theta})^2 \tag{21}$$

The T approximation follows a t distribution with degrees of freedom given in (22).

$$v = (m-1)\left[1 + \frac{\overline{U}}{(1+m^{-1})B}\right]^{-1} \tag{22}$$

Typically, $m = 10$ number of imputations is sufficient for most applications, but if $B >> U$, then more imputations may be needed.[3,5] SB, given in equation (23), measures the effect of bias on the parameter estimate obtained from the imputed data. Furthermore, SB values > 50% are considered to have adverse effects on parameter estimation.[2,10-12]

$$100 \times \left| \frac{E(\hat{\theta} - \theta)}{SE(\hat{\theta})} \right| \tag{23}$$

The RMSE given in equation (24) helps us to assess precision (efficiency) and accuracy.

$$\sqrt{E_\theta(\hat{\theta} - \theta)^2} \tag{24}$$

Coverage rate (CR) is defined as the percentage of times that the true parameter lies within the confidence interval of the parameter estimate; CR values < 90% are considered poor.[13] Lastly, average width (AW) is the average difference between lower and upper bounds of the parameter estimate obtained across $m > 1$ imputations. In this evaluation system, SB is an accuracy measure, AW is a precision measure, and CR and RMSE are integrated measures of accuracy and precision.

After computing pairwise correlations for the newly imputed binary data, we check if the updated phi matrix containing these pairwise elements is positive definite and if the matrix if fairly close to the original phi matrix, i.e., if for each element,

$$\left| \delta_{jk}^{imp} - \delta_{jk} \right| < c_{jk} \tag{25}$$

for some constant c_{jk}.

If the new phi matrix is not positive semi-definite, then we derive the 'nearest' positive semi-definite phi matrix and compare the elements of this matrix to those of the original phi matrix. We iterate the previously described procedures until the convergence criteria are met. This algorithm is summarized in (26), including the steps for multivariate normal data generation (MVN_generation), multiple imputation (MI), and dichotomization.

$$\begin{aligned} & Y_{(binary)} \\ & 1. \xrightarrow{MVN\_generation} Z_{(normal)} \\ & 2. \xrightarrow{MI} Z_{(normal)}^{imp} \\ & 3. \xrightarrow{Dichotomization} Y_{(binary)}^{imp} \end{aligned} \tag{26}$$

SIMULATION STUDY AND REAL DATA APPLICATIONS

In examining our method for the bivariate case with missing entries in the second variable, we first created data sets with two binary variables and 500 observations and either randomly deleted 125 entries in each variable to introduce 25% missingness in both variables under the Missing Completely at Random (MCAR) mechanism, or used a model where the probability of missingness in the second variable depended on the first variable to generate data missing under the Missing at Random (MAR) mechanism with 20% to 30% missingness. This model given in (27) was applied to the data first and was followed by introducing 25% missingness in the first variable by randomly deleting 125 out of the 500 entries.

$$\log(\frac{p_2}{1-p_2}) = -0.0125 + 0.125Y_1 \tag{27}$$

$$p_2 = Pr(R_{Y_2} = 0)$$

We applied our approach to create 10 imputed data sets at each of 1000 simulations and assessed the performance by looking at the average estimate (AE), standardized bias (SB), root mean-square error (RMSE), coverage rate (CR), and average width (AW).

Testing our method in the multivariate case with $K = 3$ variables, we generated binary data sets with 100 entries and induced either a 25% MCAR pattern in each variable or a 25% MCAR pattern in the first two variables, and a 20-30% MAR pattern in the third variable, where the missingness in this variable depended on the first variable, as shown in (28). We note that missingness in this third variable was introduced under the MAR mechanism before random deletion of the entries in the first variable was conducted.

$$\log(\frac{p_3}{1-p_3}) = -1.0 + 0.00125Y_1 \tag{28}$$

$$p_3 = Pr(R_{Y_3} = 0)$$

Again, we ran 1000 simulations, each involving $m = 10$ imputations. Convergence for each simulation was achieved when the absolute difference between each of the original pairwise correlations and the pairwise correlations obtained from the imputed data was less than some constant c_{jk}, with $j = 1,2$ and $k = 2,3$.

Our real data example comes from the NYC HANES (New York City Health and Nutrition Survey) database comprising of 831 men and 1168 women created to examine the association between disease prevalence and environmental factors in New York City. This example is motivated by the notion that most private health insurance policies in the US are offered through the workplace.[14-16] Employment-based health insurance affects aspects of employment such as job mobility[15] and the option of health packages that employers choose to offer.[14] Furthermore, it may be of importance to know if individuals with certain health conditions or infectious diseases have access to insurance.[17-19]

RESULTS

The measures given in Tables 2 and 3 demonstrate that the method works well for different correlated bivariate binary data sets in MCAR and MAR cases, respectively, as indicated by AE values comparable to true parameters, SB estimates < 50%, small RMSE values indicating good accuracy and precision, CR values > 90%, and AW estimates comparable to the confidence interval widths of the original parameters. Results in Table 4 also show the validity of our method for multivariate data via the assessment measures of AE, SB, RMSE, CR, and AW values.

TABLE 2: Results from applying the new imputation method to bivariate binary data missing under the MCAR mechanism, with 25% missingness in both variables.

	Data Set 1		Data Set 2	
Convergence Constant	0.0075		0.01275	
True δ	-0.7099		-0.4085	
Imputed δ	-0.7102		-0.4094	
SB	24.4278		40.1080	
RMSE	0.0012		0.0019	
CR	94.5486		94.4952	
AW	0.0876		0.1477	
	True p_1	**True p_2**	**True p_1**	**True p_2**
	0.5200	0.5040	0.4840	0.4942
	Imputed p_1	**Imputed p_2**	**Imputed p_1**	**Imputed p_2**
	0.5346	0.4947	0.4960	0.4993
	Data Set 3		Data Set 4	
Convergence Constant	0.0175		0.0075	
True δ	0.4083		0.7392	
Imputed δ	0.4078		0.7390	
SB	15.4593		14.3974	
RMSE	0.0024		0.0011	
CR	92.0796		94.3719	
AW	0.1492		0.0803	
	True p_1	**True p_2**	**True p_1**	**True p_2**
	0.5200	0.5165	0.5260	0.5156
	Imputed p_1	**Imputed p_2**	**Imputed p_1**	**Imputed p_2**
	0.5280	0.5277	0.5000	0.4906

AE: average estimate, SB: standardized bias, RMSE: root mean square error,
CR: coverage rate, AW: average width

TABLE 3: Results from applying the new imputation method to bivariate binary data missing under the MAR mechanism, with 25% missingness in the first variable and 20% - 30% missingness in the second variable.

	Data Set 1		Data Set 2	
Convergence Constant	0.00875		0.01275	
True δ	-0.7622		-0.4085	
Imputed δ	-0.7629		-0.4094	
SB	45.4977		40.1080	
RMSE	0.0013		0.0019	
CR	93.1269		94.4952	
AW	0.0740		0.1477	
	True p_1	**True p_2**	**True p_1**	**True p_2**
	0.5120	0.5040	0.4840	0.4960
	Imputed p_1	**Imputed p_2**	**Imputed p_1**	**Imputed p_2**
	0.5120	0.5076	0.4841	0.4993
	Data Set 3		Data Set 4	
Convergence Constant	0.01225		0.0075	
True δ	0.3492		0.7992	
Imputed δ	0.3488		0.7991	
SB	17.9689		7.1497	
RMSE	0.0018		0.0008	
CR	94.6627		95.3248	
AW	0.1559		0.0638	
	True p_1	**True p_2**	**True p_1**	**True p_2**
	0.4680	0.5600	0.5000	0.5200
	Imputed p_1	**Imputed p_2**	**Imputed p_1**	**Imputed p_2**
	0.4621	0.5599	0.4987	0.5116

AE: average estimate, SB: standardized bias, RMSE: root mean square error, CR: coverage rate, AW: average width

Table 5 gives the results from applying the new imputation method to a subset of 100 women from the database and indicates promise in our method as again shown by AE values comparable to the original estimates, SB estimates < 50%, small RMSE values implying good accuracy and precision, CR values > 90%, and AW estimates comparable to confidence intervals of original estimates.

Boxplots in Figures 1 and 2 show comparable ranges of mean and pairwise correlation estimates between 1000 generated data sets resembling the real data on average and 10 imputed data sets for each generated data set obtained from application of our method. Although variability was slightly higher for certain pairwise correlation estimates from imputed data, these values were still in an acceptable range. We furthermore note that the fraction of missing information in the generated data sets was equal to fraction of missing information in each corresponding variable of the original data and was introduced via the MCAR or MAR mechanism.

CONCLUSION

We introduce a novel procedure for imputing bivariate and multivariate binary data which allows us to

TABLE 4: Results from applying the new imputation method to multivariate binary data, with 25% missingness in all three variables under the MCAR mechanism and with 25% missingness in the first two variables and 20% - 30% missingness in the third variable under the MAR mechanism.

	MCAR Case		
Variable Pairs	(1,2)	(1,3)	(2,3)
Convergence Constant	0.025	0.025	0.05
True δ	0.0025	0.025	0.0325
Imputed δ	-0.7679	0.4419	-0.3793
SB	-0.7677	0.4414	-0.3786
RMSE	4.3744	11.6331	12.995
CR	0.0034	0.0035	0.0042
AW	91.4058	95.17	94.4604
	True p_1	**True p_2**	**True p_3**
	0.4600	0.4900	0.5200
	Imputed p_1	**Imputed p_2**	**Imputed p_3**
	0.4599	0.4766	0.5288
	MAR Case		
Variable Pairs	(1,2)	(1,3)	(2,3)
Convergence Constant	0.025	0.025	0.05
True δ	-0.7366	0.3917	-0.2889
Imputed δ	-0.7379	0.392	-0.2871
SB	28.3954	7.3437	39.6934
RMSE	0.0038	0.0038	0.0039
CR	91.5072	94.9118	95.207
AW	0.1852	0.3377	0.3645
	True p_1	**True p_2**	**True p_3**
	0.4600	0.4500	0.4300
	Imputed p_1	**Imputed p_2**	**Imputed p_3**
	0.4500	0.4603	0.4497

AE: average estimate, SB: standardized bias, RMSE: root mean square error, CR: coverage rate, AW: average width

TABLE 5: Description of variables and results from the new imputation method applied to a subset of 100 women from the NYC HANES database.

Variable	Label		Number (Percent) Missing
1	Herpes I (yes vs. no)		12 (12.0%)
2	Insurance Offered at Workplace (yes vs. no)		42 (42.0%)
3	Private Insurance (yes vs. no)		0 (0.0%)
Variable Pairs	(1,2)	(1,3)	(2,3)
Convergence			
Constant	0.0325	0.0325	0.0325
Original δ	-0.1422	-0.1376	0.5131
Imputed δ	-0.1435	-0.1388	0.5132
SB	22.5165	23.3595	3.8314
RMSE	0.0047	0.004	0.0043
CR	94.1559	95.0411	93.2211
AW	0.3907	0.3900	0.2964
	Original p_1	**Original p_2**	**Original p_3**
	0.77	0.55	0.68
	Imputed p_1	**Imputed p_2**	**Imputed p_3**
	0.76	0.54	0.68

AE: average estimate, SB: standardized bias, RMSE: root mean square error, CR: coverage rate, AW: average width

relax any assumptions associated with the saturated multinomial or loglinear model. This method involves imputing normally distributed values mapped from the original binary data and then dichotomizing these values using quantiles based on marginal proportions. This approach is semi-parametric in nature as imputing under the normality assumption of joint modeling involves the parametric portion and dichotomizing via quantiles constitutes the nonparametric portion. We have validated this method via simulation studies and real data applications under MCAR and certain MAR scenarios. Thus, we recommend this approach as a possible avenue for imputing binary data when multinomial or loglinear model assumptions may be violated.

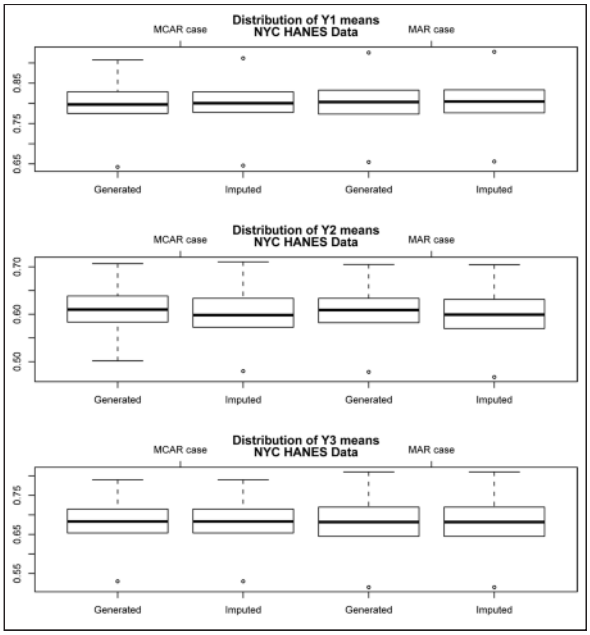

FIGURE 1: Boxplots of means for data sets resembling the NYC HANES data applied to the new imputation method for binary data generated under the MCAR or MAR mechanism.

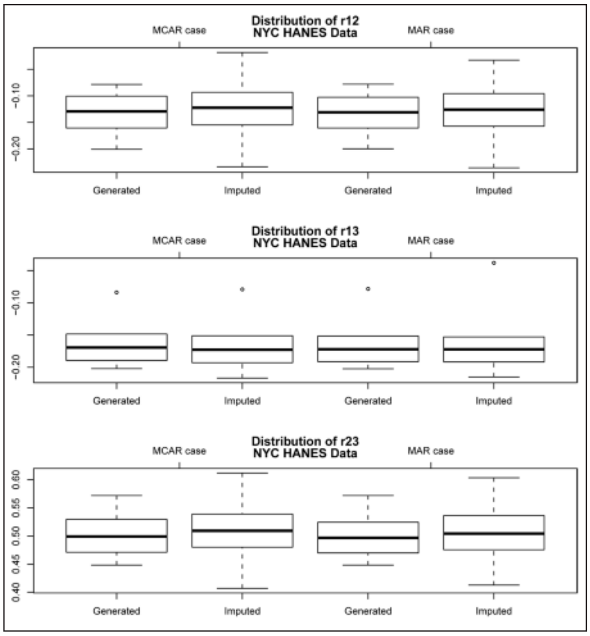

FIGURE 2: Boxplots of pairwise correlations for data sets resembling the NYC HANES data applied to the new imputation method for binary data generated under the MCAR or MAR mechanism.

REFERENCES

1. Emrich JL, Piedmonte MR. A method for generating high-dimensional multivariate binary variates. The American Statistician 1991;45 (4):302–4.
2. Demirtas, H. Simulation driven inferences for multiply imputed longitudinal data. Statistica Neerlandica 2004;58(4):466-82.
3. Schafer, JL, Olsen MK. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. Multivariate Behavioral Research 1998;33(4):545-71.
4. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7(2):147-77.
5. Schafer JL. Analysis of Incomplete Multivariate Data. 1st ed. London: Chapman and Hall; 1997. p.430.
6. Guilford JP. Psychometric Methods. 1st ed. New York: Mc Graw-Hill Book Company Inc; 1957. p.597.
7. Demirtas H, Doganay B. Simultaneous generation of binary and normal data with specified marginal and association structures. J Biopharm Stat 2012;22(2):223-36.
8. Higham N. Computing the nearest correlation matrix - a problem from finance. IMA J Numer Anal 2002;22(3):329-43.
9. Lurie PM, Goldberg MS. An approximate method for sampling correlated random variables from partially specified distributions. Management Science 1998; 44(2): 203-18.
10. Demirtas H, Hedeker D. Gaussianization-based quasi-imputation and expansion strategies for incomplete correlated binary responses. Stat Med 2007; 26(4): 782-99.
11. Demirtas H, Freels SA, Yucel, RM. Plausibility of multivariate normality assumption when imputing non-gaussian continuous outcomes: a simulation assessment. J Stat Comput Simul 2008;78(1):69-84.
12. Demirtas H, Hedeker D. Multiple imputation under power polynomials. Communications in Statistics-Simulation & Computation 2008; 37(8):1682-95.
13. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001;6(4):330-51.
14. Feldman R, Finch M, Dowd B, Cassou S. The demand for employment-based health insurance plans. J Hum Resour 1989;24(1):115-42.
15. Monheit AC, Cooper PF. Health insurance and job mobility: theory and evidence. Ind Labor Relat Rev 1994;48(1):68-85.
16. Gruber J, Madrian BC. Employment separation and health insurance coverage. J Public Econ 1997;66(3):349-82.
17. Bergner M. Measurement of health status. Medical Care 1985;23(5):696-704.
18. Ayanian JZ, Kohler BA, Abe T, Epstein AM. The relation between health insurance coverage and clinical outcomes among women with breast cancer. N Engl J Med 1993;329(5):326-31.
19. Trouiller P, Olliaro P, Torreele E, Orbinski J, Laing R, Ford N. Drug development for neglected diseases: a deficient market and a public-health policy failure. Lancet 2002;359 (9324): 2188-94.