Biol Res 38: 7-12, 2005

Multiple imputation procedures allow the rescue of missing data: An application to determine serum tumor necrosis factor (TNF) concentration values during the treatment of rheumatoid arthritis patients with anti-TNF therapy

IRENE SCHIATTINO[1], RODRIGO VILLEGAS[1], ANDREA CRUZAT[2], JIMENA CUENCA[2], LORENA SALAZAR[2], OCTAVIO ARAVENA[2], BÁRBARA PESCE[2], DIEGO CATALÁN[2], CAROLINA LLANOS[2,3], MIGUEL CUCHACOVICH[3] and JUAN C. AGUILLÓN[2]

[1] School of Public Health, Faculty of Medicine, University of Chile.
[2] Disciplinary Program of Immunology, ICBM, Faculty of Medicine, University of Chile.
[3] Rheumatology Section, Department of Medicine, University of Chile Clinical Hospital.

ABSTRACT

Longitudinal studies aimed at evaluating patients clinical response to specific therapeutic treatments are frequently summarized in incomplete datasets due to missing data. Multivariate statistical procedures use only complete cases, deleting any case with missing data. MI and MIANALYZE procedures of the SAS software perform multiple imputations based on the Markov Chain Monte Carlo method to replace each missing value with a plausible value and to evaluate the efficiency of such missing data treatment. The objective of this work was to compare the evaluation of differences in the increase of serum TNF concentrations depending on the –308 TNF promoter genotype of rheumatoid arthritis (RA) patients receiving anti-TNF therapy with and without multiple imputations of missing data based on mixed models for repeated measures. Our results indicate that the relative efficiency of our multiple imputation model is greater than 98% and that the related inference was significant (p-value < 0.001). We established that under both approaches serum TNF levels in RA patients bearing the G/A –308 TNF promoter genotype displayed a significantly (p-value < 0.0001) increased ability to produce TNF over time than the G/G patient group, as they received successively doses of anti-TNF therapy.

Key terms: multiple imputation, mixed model, TNF polymorphism.

INTRODUCTION

Longitudinal studies oriented toward evaluating patients' clinical responses to specific therapeutic treatments are frequently affected by a lack of significant elements of laboratory data. The primary cause that generates incomplete data in a clinical study is the scientists' inability to obtain patient blood samples at defined times of the protocol. Unlike the conventional statistic procedures that use only complete cases, the inferential analysis allows statistical evaluations in spite of partial loss of scientific information (missing data) (Lavory *et al*., 1995).

When a variable is studied over time, an imputation procedure allows us to predict plausible values for those unavailable figures. The new analysis will be based on those known values for the variable and the statistical evaluation will proceed as if the

Corresponding author: Dr. Juan C. Aguillón, Programa Disciplinario de Inmunología, ICBM, Facultad de Medicina, Universidad de Chile, Independencia 1027, Santiago. Casilla 13898, Correo 21. Tel: (56 2) 678-6724, Fax: (56 2) 735-3346. E-mail: jaguillo@med.uchile.cl

Recieved: March 2, 2004. In Revised Form: November 2, 2004. Accepted: April 22, 2005

information were complete. However, the application of the simple imputation analysis could generate slanted estimations, underestimated standard errors, and distortional hypothesis tests (Rubin, 1976).

Multiple imputation methods developed in recent years combine the simple imputation benefits with the incorporation of a multiple component to capture the uncertainty of absent data (Rubin, 1987). This component is obtained by the creation of *k* (typically 5 to 10) databases that contain plausible values for the incomplete cases. Each imputed file is analyzed separately by a standard statistical procedure, and the results obtained are combined in a single group with the estimated parameters that reflect the uncertainty provided for the missing data in the original database. The multiple imputation (MI) and multiple imputation analysis (MIANALYZE) procedures of the SAS software offer an interesting possibility to implement this strategy (Darmawan, 2002).

This study aims to apply this statistical methodology to a study with repeated measures and missing data. We evaluated differences in the increase of serum TNF concentrations depending on the –308 TNF promoter genotype of rheumatoid arthritis (RA) patients receiving anti-TNF therapy. The successful rescue of seven missing serum TNF levels allows us to establish that RA patients with G/A genotype displayed higher serum amounts of the cytokine than those of the G/G patient group during the treatment.

MATERIAL AND METHODS

Patients and Laboratory Determinations

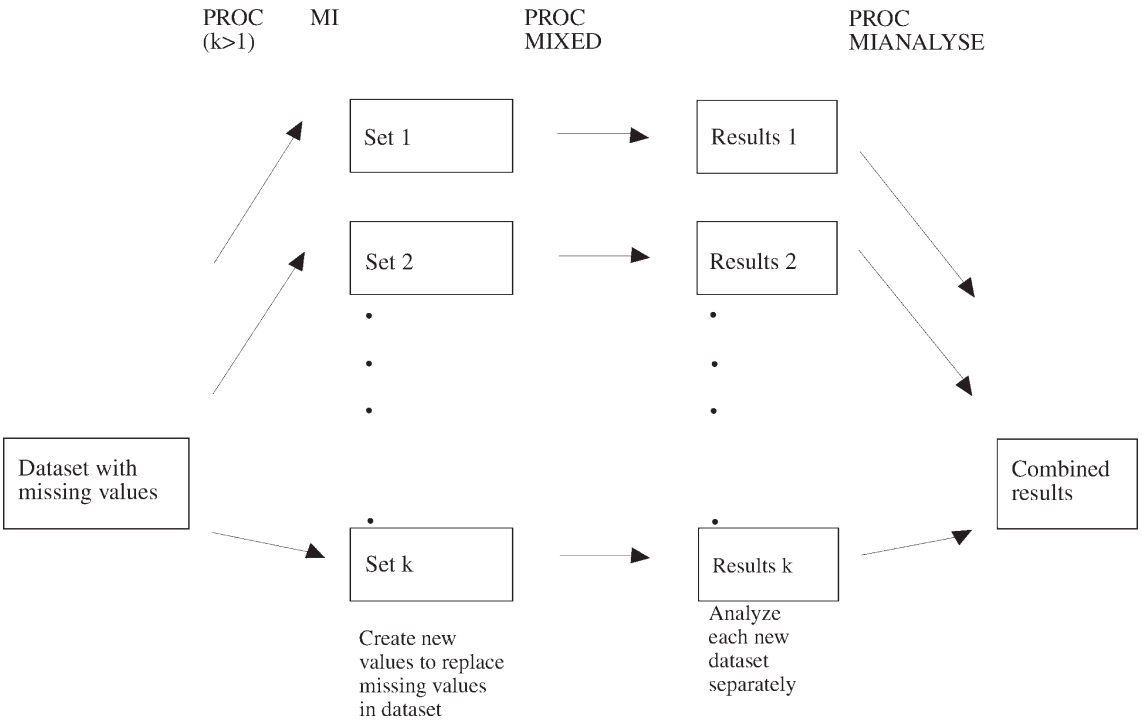
Twenty RA patients, as defined by the American College of Rheumatology criteria of diagnosis, were treated with anti-TNF therapy (Infliximab®) to determine the influence of the –308 (G/A) TNF promoter polymorphism on the responsiveness to the treatment. For this purpose 10 G/A and 10 G/G RA patients received Infliximab® (3 mg/kg of weight) at the beginning of the treatment (week 0) and at weeks 2, 6 and 14 thereafter (Cruzat et al., 2003; Cuchacovich et al., 2004). Before each Infliximab® infusion serum TNF concentrations were measured, with 4 dependent variables (y1, y2, y3 and y4) per individual. In this setup we had four individual dependent observations; from a statistical point of view we had a repeated measure problem. – 308 TNF genotypes were performed by polymerize chain reaction followed by a restriction fragment length polymorphism (Cuenca et al., 2001; 2003).

Statistical Methodology

The statistical treatment of repeated measures implies a multivariate dependence structure. We also had a treatment-fixed effect and a patient random factor, therefore suggesting that a mixed model analysis would be a valid alternative. The procedures, options, and sentences used in this work are documented at http:// support.sas.com. The Mixed Models method was considered to be the standard procedure, and it is assumed that missing data would not only form an arbitrary pattern, but would occur in any variable without any order restriction and depending only on the same variable (Darmawan, 2002; Der and Everitt, 2002). The MI procedure uses Markov Chain Monte Carlo Method (MCMC) to generate random variables in multiple chains, producing *k* sets of complete data. Parameter estimates are computed *k* times by the MIXED procedure, generating valid statistic inferences. The MIANALYZE procedure reads the estimated parameters and the associated covariance matrix.

Missing data are assumed to be partially or completely randomly distributed, and the procedures for this application are summarized in Figure 1 (Littell et al., 1996; Chantala and Suchindran, 2003).

For the multiple imputation procedure with the *k* imputed and analyzed sets, we define: $\hat{Q}_i$ = Estimation for the i^{th} analyzed group (i = 1, 2, 3,……, k). $\hat{U}_i$ Variance for the i^{th} analyzed group. $\overline{Q} = \frac{1}{k}\sum_{i=1}^{k} \hat{Q}_i$ corresponds



Complete Analysis of the datasets imputed datasets

Figure 1: Schematic representation for the application of MI, Mixed, and MIANALYSE procedures. The analysis is performed under the assumption that missing data are partially or completely distributed at random.

to the multiple imputation' s estimated point and represents the estimated average for each analyzed group.

The total estimated variance associated with $\bar{Q}$ is (Rubin, 1987):

$$T = \frac{1}{k}\sum_{i=1}^{k}\hat{U}_i + \frac{k+1}{k}\left[\frac{1}{k-1}\sum_{i=1}^{k}(Q_i - \bar{Q})^2\right]$$

The first term corresponding to the "within imputations variance" and the second term to the "between imputations variance."

RESULTS AND DISCUSSION

Means and standard deviations of age for G/A and G/G RA patients were 47.0 ± 10.6 and 53.0 ± 12.6 years, respectively. All G/A patients and 90% of G/G patients were females. No significant differences on age or sex were observed between G/A and G/G groups.

Table I shows descriptive statistics (minimum, maximum, mean, and standard deviation) for each of the four serum TNF concentrations in the G/A and the G/G patient groups. The third column displays the corresponding number of missing data. In this case the variables correspond to TNF measured immediately before patients began the treatment (y1) and before patients received the second (y2), the third (y3) and fourth (y4) Infliximab® doses. Differences between G/A and G/G groups for serum TNF concentrations at time 1 and 2 were not significant. Since most of missing data for TNF measurements were concentrated at time 3 and 4, it is difficult to evaluate differences between groups for those times.

Considering a total of seven missing values (35%) for the TNF evaluation on time, the MI procedure was applied. MI procedure uses Markov Chain Monte Carlo Method (MCMC) to generate random

variables in multiple chains for the imputation analysis performed (Li, 1988). In this application, the procedure completed 50 iterations before the imputed dataset was created. Since there is no a priori information about the media and covariance estimation, by default a non-informative a priori Jeffrey's distribution was used to estimate the media and covariances that will be used in the next step (Schafer, 1997).

In our model, complete data were available in 75% of the observations, and the last measurement was missing in four out of twenty patients. Table II shows the three missing data patterns observed in our original data with the corresponding absolute and relative frequencies. The fraction of missing information (λ) and the relative efficiencies $(1+\lambda/k)^{-1}$ for $k = 5$ imputations for each of the variables with missing data are shown in Table III. The multiple-imputation parameter estimates for each variable are displayed in Table IV. The inference for the "estimated parameters" are based on *t* distribution, and in this case, all three are significantly different than zero ($p<0.001$).

Next the MIXED procedure with the affirmation BY _IMPUTATION_ was applied for each of the five complete data sets created. The MIANALYZE procedure was applied using the previous results. The estimations and the multiple imputation inference of estimated parameters proved to be significant to the variable time. A similar result was obtained when a patient from the original database belonging to the G/G genotype group and whose serum TNF concentration was only determined at the beginning of treatment was excluded (Table V).

The missing information in our original database was greater than 5% of total data. In this situation, Roth (1994) and SAS Institute Inc. (1999) recommend the use of multiple imputation methods instead of simple imputation methods.

TABLE I

Descriptive statistics and number of missing data in both groups of patients for each variable

Patient Groups and Variable (TNF Levels at Time of Determination)		N	N° of Missing Data	Serum TNF Concentration (pg/ml)			
				Min	Max	Mean	Standard deviation
G/G:	**y1**	10	0	4.0	92.6	18.3	28.8
	y2	9	1	6.9	159.0	68.7	44.8
	y3	9	1	10.0	316.0	133.1	100.3
	y4	8	2	7.3	409.0	202.5	171.9
G/A:	**y1**	10	0	4.0	52.5	12.8	16.5
	y2	10	0	4.0	196.0	76.3	66.3
	y3	10	0	4.2	148.0	72.8	44.7
	y4	7	3	4.0	242.0	98.9	90.9

TABLE II

Observed missing data patterns (.), frequency and percentage of the groups

Pattern	y1	y2	y3	y4	Frequency	Percentage (%)
1	√	√	√	√	15	75.00
2	√	√	√	.	4	20.00
3	√		.	.	1	5.00

TABLE III

The fraction of missing information (λ) and the relative efficiencies for *k*=5 imputations for each one of the variables with missing data

Variable	Fraction of missing information (λ)	RelativeEfficiency = $(1+\lambda/k)^{-1}$
y_2	0.05	0.989
y_3	0.04	0.992
y_4	0.12	0.975

TABLE IV

Multiple-Imputation Parameter Estimates. Estimated mean and standard error for each variable. The inferences are based on the *t* distribution

Variable	Mean	Std Error Mean	95% Confidence Limits		DF	t for H0: Mean=Mu0
y2	73.195	12.674	46.327	100.063	16	5.775*
y3	100.529	18.070	62.222	138.838	16	5.563*
y4	156.798	33.726	84.463	229.132	14	4.649*

* Significant (p-value<0.05)

TABLE V

Final information after the procedure was applied

	The Mixed Procedure without Multiple-Imputation				
Variable	**Mean**	**Std ErrorMean**	**DF**	**t value**	**Pr > |t|**
Intercept	-30.09	11.52	18	-2.61	0.02
Group	-0.39	8.34	18	-0.05	0.96
Time	46.09	9.04	18	5.10	<0.001
	The MIANALYZE Procedure: Multiple-Imputation Parameter Estimates of our original database (20 patients).				
Intercept	-28.84	11.07	4985	-2.61	0.009
Group	1.46	8.06	14426	0.18	0.857
Time	45.11	8.64	10635	5.22	<.001
	The MIANALYZE Procedure: Multiple-Imputation Parameter Estimates of the database without the patient whose serum TNF concentration was only determined at the beginning of treatment (19 patients).				
Intercept	-28.99	11.74	5176	-2.47	0.01
Group	-0.86	8.85	1076	-0.09	0.92
Time	45.75	8.92	101448	5.12	<0.01

The assumption of normality for applying this methodology could be managed by the use of normalizing transformations available in the MI procedure or alternatively by using the multiple imputation for non-normal distributions model, which assumes that the slant produced is minimum (SAS Institute Inc., 1999).

Our results indicate that the relative efficiency of the multiple imputation model is greater than 99% for y2 and y3 variable values and 98% for the y4 value. Interestingly, the related inference was significant (p-value < 0.001). Based on Mixed Models, we demonstrated that serum TNF levels of RA patients bearing the G/A –308 TNF promoter genotype significantly (p-value < 0.001) increase over time as they received successive doses of anti-TNF therapy. This observation is qualitatively concordant with other results obtained without multiple imputation analysis.

Finally, we can conclude that the presence of missing data from a database does not create an unsolved problematic situation if they are treated with Mixed Models. The application of this methodology allows the construction of a complete database, avoiding the statistical analysis problem derived from repeating measures of missing data.

REFERENCES

CRUZAT A, CUCHACOVICH M, SALAZAR L, CATALÁN D, SCHIATTINO I, AGUILLÓN JC (2003) Treatment with anti-TNF monoclonal antibodies in rheumatoid arthritis and the –308 TNF promoter polymorphism influence. Biol Res 36(3-4): R62

CUCHACOVICH M, FERREIRA L, ALISTE M, SOTO L, CUENCA J, CRUZAT A, GATICA H, SCHIATTINO I, PÉREZ C, AGUIRRE A, SALAZAR-ONFRAY F, AGUILLÓN JC (2004) TNF-α levels and influence of –308 TNF-a promoter polymorphism on the responsiveness to infliximab in patients with rheumatoid arthritis. Scand J Rheumatol 33: 228-232

CUENCA J, PÉREZ C, AGUIRRE A, SCHIATTINO I, AGUILLÓN JC (2001) Genetic polymorphism at position –308 in the promoter region of the tumor necrosis factor (TNF): Implications of its allelic distribution on susceptibility or resistance to diseases in the Chilean population. Biol Res 34: 237-241

CUENCA J, CUCHACOVICH M, PÉREZ C, FERREIRA L, AGUIRRE A, SCHIATTINO I, SOTO L, CRUZAT A, SALAZAR-ONFRAY F, AGUILLÓN JC (2003) The –308 polymorphism in the tumor necrosis factor gene promoter region and ex - vivo lipopolysaccharide-induced TNF expression and cytotoxic activity in chilean patients with rheumatoid arthritis. Rheumatology (Oxford) 42: 308-313

CHANTALA K, SUCHINDRAN C (2003) Multiple Imputation for Missing Data. SAS OnlineDoc™: Version 8. www.cpc.unc.edu/services/computer/ presentations/ mi_presentation2.pdf

DARMAWAN IGN (2002) NORM software review: handling missing values with multiple imputation methods. Evaluation Journal of Australasia 2 (1): 20-24

DER G, EVERITT B (2002) A Handbook of Statistical Analyses using SAS. New York: Chapman & Hall

LAVORI PW, DAWSON R, SHERA D (1995) A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data. Statistics in Medicine, 14: 1913-1925

LI KH (1988) Imputation using Markov Chains. J Stat Comput Simul 30: 57-79

LITTELL RC, MILLIKEN GA, STROUP WW, WOLFINGER R (1996) SAS System for Mixed Models. SAS Institute Inc., Cary, NC, USA

ROTH P (1994) Missing data: A conceptual review for applied psychologist. Personnel Psychology, 47: 537-560

RUBIN DB (1976) Inference and missing data. Biometrika 63: 581-592

RUBIN DB (1987) Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, Inc.

SCHAFER JL (1997) Analysis of incomplete multivariate data. New York: Chapman & Hall

SAS Institute Inc. (1999) SAS Procedures Guide, Version 8, Cary, NC: SAS Institute Inc.