Assessment of the Performance of Imputation Techniques in Observational Studies with Two Measurements

Urko Aguirre[1,*], Inmaculada Arostegui[2,5], Cristóbal Esteban[3] and Jose María Quintana[4]

[1]*Research Unit, REDISSEC: Health Services Research on Chronic Patients Network, Hospital Galdakao-Usansolo, Galdakao, Bizkaia, Spain*

[2]*Department of Applied Mathematics, Statistics and Operational Research; Red de Investigación en Servicios Sanitarios y Enfermedades Crónicas. Faculty of Science and Technology, Leioa, Spain*

[3]*Pneumology Department, REDISSEC: Health Services Research on Chronic Patients Network, Hospital Galdakao-Usansolo, Galdakao, Bizkaia, Spain*

[4]*Research Unit, REDISSEC: Health Services, Research on Chronic Patients Network, Hospital Galdakao-Usansolo, Galdakao, Bizkaia, Spain*

[5]*BCAM – Basque Center for Applied Mathematics*

Abstract: In observational studies with two measurements when the measured outcome pertains to a health related quality of life (HRQoL) variable, one motivation of the research may be to determine the potential predictors of the mean change of the outcome of interest. It is very common in such studies for data to be missing, which can bias the results. Different imputation techniques have been proposed to cope with missing data in outcome variables. We compared five analysis approaches (Complete Case, Available Case, K- Nearest Neighbour, Propensity Score, and a Markov Chain Monte Carlo algorithm) to assess their performance when handling missing data at different missingness rates and mechanisms (MCAR, MAR and MNAR). These strategies were applied to a pre-post study of patients with Chronic Obstructive Pulmonary Disease. We analyzed the relationship of the changes in subjects HRQoL over one year with clinical and socio-demographic characteristics. A simulation study was also performed to illustrate the performance of the imputation methods. Relative and standardized bias was assessed on each scenario. For all missingness mechanisms, not imputing and using MCMC method, both combined with mixed-model analysis, showed lowest standardized bias. Conversely, Propensity Score showed worst bias values. When missingness pattern is MCAR or MAR and rate small, we recommend using mixed models. Nevertheless, when missingness percentage is high, in order to gain sample size and statistical power, MCMC is preferred, although there are no bias differences compared with the mixed models without imputation. For a MNAR scenario, a further sensitivity analysis should be made.

Keywords: HRQoL, Imputation, Missing data, Pre-post design.

INTRODUCTION

Missing data are a common problem among various types of medical design studies [1]. It is especially problematic in longitudinal studies that require repeated measures, in which the missingness rates can become high and relevant. This loss of information may be related to patient dropout, unavailability, death, or other reason, and can often reach missingness rates up to 30-40%. Missing data yield a form of selection bias that is related to the cohort design. Many methodologic difficulties can arise from missing data: (a) loss of efficiency and power due to reduction of the sample size; (b) complications in data handling and analysis; (c) presence of bias due to differences between cases with observed and lost data [2]. These problems are common in studies evaluating health related quality of life (HRQoL). Some observational HRQoL studies are designed to identify important associations between patients' socio-demographic or clinical characteristics and the response variable related to a specific disease. For these studies, at least two measurements are sufficient to detect these relationships: the first at baseline and the other at a predefined point during follow-up. In this situation, missing data of the outcome may arise due to patient death or worsening of disease. In addition, some patients with good baseline quality of life may abandon the study. Thus, the patients with complete information on quality of life are not usually representative of the population under study. Many researchers tend to ignore missing data and analyze the subsample with complete information, which may bias results. The appropriate statistical approach for analyzing the entire population, not just those with complete data, is a relevant problem that statisticians are now addressing. This area of research has become very active in last couple decades [3-6]. Little and Rubin [3] defined three possible missingness mechanisms. Briefly, when missing data are unrelated to the outcome, the missingness pattern is classified as

*Address correspondence to this author at the Research Unit, Hospital Galdakao-Usansolo, Barrio Labeaga s/n, 48960 Galdakao, Bizkaia, Spain;
Tel: +34 944007105; Fax: +34 944007132;
E-mail: urko.aguirrelarracoechea@osakidetza.eus

E-ISSN: 1929-6029/15

missing completely at random (MCAR). When the probability of missing data is related to patient characteristics and the observed outcome, the pattern is classified as missing at random (MAR). Finally, if the probability of missing data is also related to unobserved values of the outcome, then the pattern is missing not at random (MNAR). Several statistical methods have been proposed as appropriate techniques for handling missing data, providing alternatives to naive methods that deleted cases with missing values (known as complete case (CC) analysis) and to procedures that discard missing variable values (known as available case (AC) analysis). One of the suggested approaches involves imputing values. Imputation replaces missing data with numerical values obtained from the subset of observed values. The most commonly used imputation methods are based on single imputation. The mean imputation method and the K-nearest neighbor imputation (K-NNI) methods are good examples of single imputation methods used as alternatives to CC and AC analysis. Mean imputation replaces the missing value by the average value of the non-missing data, whereas K-NNI is a donor-based method in which the imputed value is either a value that was measured for another record in the dataset or the average value obtained from a subset of records. In general, single imputation methods are easy to implement. However, as the missingness rate increases, they introduce bias. Multiple imputation (MI) methods are based on more complex strategies, such as the propensity score (PS) or the Markov chain Monte Carlo (MCMC) method, both of which impute more than one set of values (generally 3 to 10) for the missing data using sampling variability and uncertainty [7]. The most recent literature in the field suggests that imputation is always better than ignoring missing data [8,9]. Knowing the missing data mechanisms is advisable to identify the most appropriate method of analysis. However, this knowledge is unavailable in many studies. Previous studies have evaluated imputation methods such as single methods (including the null value compared to either complete data or the mean value) and MI methods in different settings: longitudinal studies, and handling missing outcomes or covariates [10-12]. However, the published literature evaluating different methods or recommending any of them in pre-post observational HRQoL studies is still scarce. The special case of pre-post observational studies is a simple design where recommending that issue would be very helpfull for researchers.

The goal of our study was to compare the performance of various imputation techniques under different missingness mechanisms and rates when handling missing data of the outcomes in a pre-post study design. The selected imputation methods we employed were K-NNI, PS, and MCMC (widely used procedures for missing entries) under the three mentioned missingness patterns with 10% and 30% of data missing.

The rest of the manuscript is divided into four sections. In Section 2 we describe the three imputation methods. Section 3 focuses on the results after applying the imputation methods to the HRQoL-COPD study. Section 4 describes the simulation study and presents its results. Section 5 discusses the results and presents some general conclusions.

MATERIAL AND METHODS

K-Nearest Neighbor Imputation Method

K-NNI imputation is part of a machine learning and classification method [13,14]. It is based on the existence of similarities among the cases within a dataset: the more similarities between two cases, the closer the cases are.

Assume that a data set of size n can be expressed as the following pattern:

$$\{(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{r}_i)\} \quad \text{for} \quad i = 1,\ldots,n \tag{1}$$

where $\boldsymbol{x}_i = [\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots \boldsymbol{x}_{ip}]^T$ is the vector of p covariates, $\boldsymbol{y}_i$ is the outcome of interest, and $\boldsymbol{r}_i$ is the missingness indicator, for each subject $i = 1,\ldots,n$.

Given a subset where $\boldsymbol{r}_i = 0,$ for each variable in the subset and based on a metric distance, *K* closest cases to variables are selected. Once these K cases are selected, missing values are replaced upon the character of the variable to be imputed: mean average is used for the continuous data and mode for qualitative data.

Two parameters are needed in the *K*-NNI algorithm: the metric distance and the number of neighbors (*K*). The distance between two observations, say $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$, is defined as:

$$d(\boldsymbol{x}_a, \boldsymbol{x}_b) = \sqrt{\sum_{j=1}^{p} d_j(x_{aj}, x_{bj})^2} \tag{2}$$

where

$$d_j(x_{aj},x_{bj}) = \begin{cases} d_{cat}(x_{aj},x_{bj}) \text{ if } x_{aj} \text{ and } x_{bj} \text{ are categorical} \\ d_{cont}(x_{aj},x_{bj}) \text{ if } x_{aj} \text{ and } x_{bj} \text{ are continous} \end{cases} \quad (3)$$

The distance d_{cat} function for categorical variables assings a value of 0 if the discrete values are the same; otherwise, the value is 1 [13]. In addition, d_{cont} unction for continous instances is the Euclidean metric distance [14,15].

Multiple Imputation Methods

Using the MI methods, several imputation values (*M*) are calculated for every missing value [3]. With the *M* imputations, *M* complete data sets are generated, and for each data set created in this way, standard statistical procedures are performed. Finally, the results obtained from the *M* complete data sets (e.g., the regression coefficient) are pooled into a summary statistic. The point estimate of the summary statistic is calculated as the average of the *M* imputations whereas the variance of the summary statistic is calculated from two components: the within-imputation variance (*W,* the average of the variances of the summary statistics of the *M* imputations) and the between-imputation variance (*B,* the difference between the summary statistic of each imputation and the average of the summary statistics of the M imputations). Then, the total covariance matrix is given by

$$V = W + \left(\frac{M+1}{M}\right)B \quad (4)$$

There are many techniques for performing MI. These approaches assume that the missingness pattern is MAR. Two of the most-used MI algorithms are described in the following subsections.

Markov Chain Monte Carlo (MCMC) Method

The MCMC method is a popular approach within MI methodology. Given a general missingness pattern, this method is used to generate pseudorandom draws from multidimensional and otherwise intractable probability distributions by means of Markov chains [7]. Supposing that data come from a multivariate normal distribution, the MCMC method involves two basic steps: I-step (imputation) and P-step (posterior predictive distribution). By repeating these two steps, MCMC works as follows: For the I-step, a mean and covariance matrix is computed from the complete dataset. Values for missing data are then simulated randomly, selecting a value from the available data. The P-step comes next. In this phase, the posterior population mean vector and covariance matrix is simulated and updated, given the observed values (Y_{obs}) and the obtained simulated missing values (Y_{miss}) in the I-step.

Both the I-step and P-step provide in the first iteration a starting value, and posterior iterations create a stochastic Markov chain which converges in a function that approximates the real unknown $P(\theta, Y_{miss}|Y_{obs})$. To produce multiple imputations, one iterates over the I-step and P-steps and over Y_{miss} repeatedly until the distribution is stationary. This means that the mean vector and covariance matrix (θ) is unchanged. When this process converges, all the missing values in the dataset are substituted by imputed values.

Propensity Score Method

Rosenbaum and Rubin [16] introduced this statistical tool as a bias reduction technique to estimate the treatment effect in non-randomized studies of causal effects. By definition, PS method is defined as the conditional probability that the patient receives the treatment in the presence of its baseline characteristics. A formal mathematical expression of this method is denoted as:

$$e(\boldsymbol{X}) = Pr(Z = 1|\boldsymbol{X}) \quad (5)$$

$\boldsymbol{X}$ is the set of considered covariates and $Z = 1$ denotes if the subject receives the treatment and $Z = 0$ otherwise. PS is basically estimated by means of multiple logistic regression.

As the PS method uses many covariates, in observational studies a large proportion of subjects could have missing values in at least one covariate. Several approaches have been suggested for handling missing data in covariates using the PS method. It could also be applied for imputing missing data in outcomes when the missingness pattern is monotone ($Z = 1 \leftrightarrow R = 1$). In this missingness setting, the following steps are used to fill in unknown values: Firstly, a logistic regression model is fitted, indicating the probability of being missing:

$$logit(Pr(\boldsymbol{R}_i|\boldsymbol{X}_i,\beta)) = (1,\boldsymbol{X}_i)'\beta \quad (6)$$

where $\boldsymbol{R}_i = 1$ if Y_i is missing, $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots X_{ip})$ is the set of covariates for the *i*th observation, β, the coefficients for the $\boldsymbol{X}_i$ design matrix and $logit(p) = log\left(\frac{p}{1-p}\right)$.

After that, a propensity score for each observation to estimate the probability of being missing is created.

Based on propensity score values, the observations are divided into a fixed number of groups and an approximate Bayesian bootstrap is applied to each group [17]: Given a k group, suppose that Y_{obs} denotes the n_1 observations with nonmissing Y_j values and Y_{mis} denotes the n_0 observations with missing Y_j. The approximate Bayesian bootstrap imputation first draws n_1 observations randomly with replacement from Y_{obs} to create a new data set Y_{obs}^*. The process then draws the n_0 alues for Y_{mis} randomly with replacement from Y_{obs}^*.

Modeling and Imputation Procedures

To assess the influence that covariates may have on the change of the response variable, we framed this problem as a longitudinal data analysis or a repeated measures framework. Under this scenario, both the measurement at baseline and one-year follow-up are considered as outcome measures in the statistical approach. The most appropriate method to handle this issue is called the general linear mixed-effects (GLMM) model. The GLMM model is a multivariate regression model that generalizes the analysis of variance and general linear regression models. This is a statistical tool for modeling continuous outcome measures as a function of fixed (population factors) and random effects (individual parameters) [18]. Furthermore, we defined dummy variables for each categorical variable considered in the modelling approach. These statistical procedures were implemented using SAS System for Windows 9.3 release [19].

The three methods (K-NNI, PS, and MCMC) were developed using SAS System 9.3 release. In the case of the K-NNI methodology, two parameters were established: the distance and the number of neighbors (*K*) considered for the imputation. We used the previously described metric distance. As regards the *K* value, three different values -1, 5, and 7- were evaluated: This approach was developed using an ad hoc SAS Macro. With regard to the PS and MCMC methods, *M* = 20 imputations were considered. Two procedures available in the SAS system- PROC MI (for imputation) and PROC MINIANALYZE for summarizing the estimates - were applied.

APPLICATION: THE HRQoL-COPD STUDY

The HRQoL-COPD study was conducted between 2003 and 2004, patients diagnosed with chronic obstructive pulmonary disease (COPD) who visited an outpatient clinic were consecutively recruited [20]. This study was approved by the research committee of the Hospital Galdakao-Usansolo. Patients provided informed consent to take part in the study.

In addition to providing socio-demographic data, clinical variables were also collected and participants completed questionnaires related to their HRQoL at baseline and one year later. A subset of patients with no missing values in the clinical and socio-demographic variables was selected for this work. We applied three methods, *K*-NNI, PS, and MCMC to impute missing values to this subset in the outcome.

The outcome of interest is the change in the total score of the HRQoL status measured by the St. George's Respiratory Questionnaire [21]. This score measures the impact on overall health, daily life and perceived well-being in patients with COPD. The total score ranges from 0 (best status) to 100 (worst status) points. As covariates, apart from the above mentioned, we also selected the patients' baseline HRQoL status and their scores on the health status [20], which prognosticates the severity status of patients with stable COPD. The higher score in the health index indicates better overall clinical condition.

We aimed to assess the effects of the health status on the change of the aforementioned outcome in the observed data. We performed a multivariate model using GLMM models and the St George's Respiratory Questionnaire measured at both time points (baseline and at one year after) as outcome. Covariates related to the outcome or to the missingnes of the dependent variable (a logistic regression was performed to determine the variables) with a p-value < 0.20 [22], time, and the interaction of between each covariate and time effect - the change in HRQoL - were considered as fixed effects. Time was set as categorical. Random effects within the model was the intercept for individuals participants. In addition, the proposed imputation methods were applied in order to evaluate differences in the beta value of the interaction term between the health status variable and the time effect in the model when handling missing data using different approaches.

RESULTS

Relevant clinical and socio-demographic characteristics of the sample are shown in Supplementary Table **1**. More detailed information can be found in [20].

Table 1: Mixed Model Analysis Results for the HRQoL-COPD Study without Imputation (n = 543)

	AC (original analysis)		CC	
	$\hat{\beta}$ (s.e.)	p-value	$\hat{\beta}$ (s.e.)	p-value
Intercept	98.41 (6.00)	<0.001	97.27 (6.39)	<0.001
Age	-0.27 (0.08)	0.002	-0.24 (0.09)	0.005
health status	-6.85 (0.35)	<0.001	-6.85 (0.38)	<0.001
Adm1	-0.53 (1.88)	0.78	-1.34 (2.03)	0.51
Adm2	3.20 (2.34)	0.17	4.29 (2.60)	0.10
time effect	-7.09 (5.45)	0.19	-6.66 (5.52)	0.23
time x Age effect	-0.02 (0.07)	0.76	-0.03 (0.07)	0.68
time x health status effect	1.02 (0.32)	0.002	1.02 (0.33)	0.002
time x Adm1 effect	1.10 (1.73)	0.53	1.41 (1.76)	0.42
time x Adm2 effect	3.62 (2.20)	0.10	3.20 (2.25)	0.16

AC: Available Case. **CC**: Complete Case. $\hat{\beta}$ (s.e.): Beta regression coefficient (standard error). Adm1: presence only one previous hospital admission. Adm2: presence of two or more previous hospital admission. Time x (Age, health status, Adm1, Adm2) effect : Follow-up coefficients.

One year after the start of the study, of the 543 participants in the study, 63 (11.60%) did not complete the St. George's Respiratory Questionnaire, and consequently their COPD-related HRQoL could not be measured.

In the final model, baseline HRQoL status, health status, patient's age and previous hospital admissions were included. Shown in Tables **1** and **2** are the estimates of the intercept and slope from the different analyses for handling missingness and modeling their COPD-related HRQoL as a function of its baseline value and the health status, adjusted by patient's age and previous hospital admissions. Health status and the time×health status effects were found to be significant in all performed analyses. More precisely, the beta estimates obtained from the analysis of the original data set without imputing (AC, CC) and after the application of the proposed methods of imputation (1-NNI, 5-NNI, 7-NNI, PS and MCMC) showed different values. They ranged between 1.02 (CC) and 2.39 (7-NNI). Standard errors varied across the three imputation methods: 7-NNI and PS showed highest values.

SIMULATION STUDY

To illustrate and compare the performance of the three imputation techniques, we conducted a simulation study. It allowed us to evaluate the variability of the results obtained from the methods described earlier.

We used the subsample of the HRQoL-COPD study as the natural framework for designing the simulation study. We first assigned each subject $i, i = 1,\ldots,n$, the generated X_i variables with the same sizes as observed in the subsample (n = 400). For each measurement $t, t = 0,1$, we created longitudinal responses *via* the linear model $Y_{it}^{*} = \beta_0 + \beta_1 t + \sum_{i=2}^{5} \beta_i X_i + \sum_{i=6}^{9} \beta_i X_{i-4} t + \epsilon_{it}$ where ϵ_{it} is an error component. An unconstrained covariance matrix $\sum_y$ was used to generate the error components for the responses, such that the upper triangular part of the covariance matrix $\sum_y$ is defined by $vech(\sum_y) = (243.78, 149.57)'$. As described in previous sections, our main of interest was in making inference about the β estimate of the interaction between the health status and time effect.

HRQoL outcomes are, by definition, bounded variables: their score values range from 0 to 100. Because of that, Y_{it}^{*} simulated variable was truncated as follows:

$$Y_{it}^{*c} = \begin{cases} 0, & \text{if } Y_{it}^{*} < 0, \\ Y_{it}^{*c}, & \text{if } 0 \le Y_{it}^{*} \le 100, \\ 100, & \text{if } Y_{it}^{*} > 100. \end{cases}$$

with Y_{it}^{*c} as the truncated simulated variable.

After that, values of the covariates were associated to the simulated outcome, using the same structure and correlations of the HRQoL-COPD study.

The next step was to generate the different missingness mechanisms and rates to assess the variability of the results. Given the database, the missingness probability, p_i, was defined as a function

Table 2: Point Estimate Results for the HRQoL-COPD Study after Handling Imputation Techniques (n = 543)

K-Nearest Neighbour Imputation (K-NNI) methods	**1-NNI**		**5-NNI**		**7-NNI**	
	$\hat{\beta}$ **(s.e.)**	**p-value**	$\hat{\beta}$ **(s.e.)**	**p-value**	$\hat{\beta}$ **(s.e.)**	**p-value**
Intercept	98.41 (6.04)	<0.001	98.41 (5.95)	<0.001	98.41 (6.89)	<0.001
Age	-0.27 (0.08)	0.001	-0.27 (0.08)	0.001	-0.27 (0.09)	0.004
health status	-6.85 (0.35)	<0.001	-6.85 (0.35)	<0.001	-6.84 (0.40)	<0.001
Adm1	-0.53 (1.89)	0.78	-0.53 (1.86)	0.78	-0.53 (2.16)	0.81
Adm2	3.20 (2.35)	0.17	3.20 (2.31)	0.17	3.20 (2.68)	0.23
time effect	-7.68 (5.19)	0.14	-7.67 (5.01)	0.13	-0.16 (7.10)	0.98
time x Age effect	-0.01 (0.07)	0.90	-0.02 (0.07)	0.79	-0.30 (0.09)	0.001
time x health status effect	0.96 (0.30)	0.02	1.07 (0.29)	<0.001	2.39 (0.41)	<0.001
time x Adm1 effect	1.45 (1.62)	0.37	1.26 (1.57)	0.42	1.04 (2.22)	0.64
time x Adm2 effect	2.10 (2.02)	0.30	2.57 (1.95)	0.19	-0.79 (2.76)	0.78
Multiple Imputation (MI) methods			**PS**		**MCMC**	
			$\hat{\beta}$ **(s.e.)**	**p-value**	$\hat{\beta}$ **(s.e.)**	**p-value**
Intercept			98.41 (6.30)	<0.001	98.41 (6.03)	<0.001
Age			-0.27 (0.08)	0.002	-0.27 (0.08)	0.001
health status			-6.85 (0.37)	<0.001	-6.85 (0.35)	<0.001
Adm1			-0.53 (1.97)	0.79	-0.53 (1.89)	0.78
Adm2			3.20 (2.45)	0.19	3.20 (2.34)	0.17
time effect			-15.73 (6.73)	0.02	-6.79 (5.47)	0.21
time x Age effect			0.04 (0.09)	0.66	-0.03 (0.07)	0.70
time x health status effect			1.79 (0.41)	<0.001	1.04 (0.329	0.001
time x Adm1 effect			0.86 (2.12)	0.69	1.25 (1.78)	0.48
time x Adm2 effect			0.97 (2.64)	0.71	3.38 (2.27)	0.14

(1,5,7)-NNI: 1,5,7 –Nearest Neighbour Imputation. **PS:** Propensity Score. **MCMC**: Markov Chain Monte Carlo. $\hat{\beta}$ (s.e.): Beta regression coefficient (standard error). Adm1: presence only one previous hospital admission. Adm2: presence of two or more previous hospital admission. Time x (Age, health status, Adm1, Adm2) effect : Follow-up coefficients.

of a random process, of covariates and observed values of the outcome at the baseline measurement, or of covariates and the response missing variable. Specifically, $p_i = P(\eta_i)$, where η_i depends on the missingness patterns:

$$MCAR: \eta_i = \alpha$$
$$MAR: \eta_i = \boldsymbol{X}\alpha$$
$$MNAR: \eta_i = \boldsymbol{X}\alpha + \gamma Y_i^{*c}$$

where α, γ are vectors of coefficients of the $\boldsymbol{X}$ covariate design matrix and the Y_i^{*c} censored simulated variable.

We analyzed two specific missingness rates, 10% and 30%, within each pattern, varying the values of the α and γ parameters in the defined probability.

After that, the proposed methods were applied to obtain a complete dataset in order to estimate the beta coefficients and standard errors by means of generalized mixed-effects models. This simulation process was developed in two ways: (1) *N* = 500 and (2) *N* = 1000 runs with the aforementioned characteristics were used.

Given that our objective was to assess the performance of AC, K-NNI, PS, and MCMC imputation methods and their effects on the estimation and significance of beta values over time, we focused on the regression coefficient of the interaction between the time effect and the covariate (time×health status). We compared the obtained results from the different simulated scenarios ($\hat{\beta}$) with the true values (β, observed in the original sample without missing values), computing the following parameters [23]:

Table 3: Bias and Relative Width Assessment Parameters Obtained after *N* = 500 and *N* = 1000 Simulations

Missingness mechanism	Missingness Rate	Imputation Method	N = 500 simulations			N = 1000 simulations		
			RB	SB	Relative Width	RB	SB	Relative Width
MCAR	10%							
		CC	9.90	23.18	1.49	10.63	24.89	1.49
		AC	16.78	39.28	1.48	17.51	40.98	1.48
		1-NNI	13.10	30.67	1.41	13.84	32.41	1.41
		5-NNI	17.99	42.11	1.38	18.73	43.84	1.38
		7-NNI	21.01	49.19	1.38	21.74	50.91	1.38
		PS	67.51	162.72	1.70	70.19	164.32	1.70
		MCMC	17.84	41.77	1.47	18.58	43.49	1.46
	30%							
		CC	14.08	32.96	1.71	14.77	34.58	1.71
		AC	21.36	50.01	1.66	22.05	51.62	1.66
		1-NNI	22.48	52.63	1.42	23.21	54.33	1.42
		5-NNI	22.41	52.46	1.25	23.15	54.19	1.25
		7-NNI	25.80	60.41	1.24	26.52	62.09	1.24
		PS	172.14	402.99	1.99	172.70	404.31	1.99
		MCMC	19.67	46.04	2.13	19.67	46.04	2.14
MAR	10%							
		CC	2.90	6.99	1.51	3.75	8.78	1.51
		AC	-9.98	-23.37	1.49	-9.21	-21.57	1.49
		1-NNI	16.28	38.11	1.44	16.99	39.79	1.44
		5-NNI	10.94	25.62	1.39	11.67	27.33	1.36
		7-NNI	10.35	24.24	1.36	11.09	25.95	1.36
		PS	68.02	159.24	1.83	252.86	591.97	2.16
		MCMC	-4.61	-10.79	1.51	-3.88	-9.09	1.51
	30%							
		CC	18.09	42.34	1.75	18.88	44.20	1.74
		AC	20.87	48.85	1.67	21.66	50.70	1.67
		1-NNI	58.55	137.07	1.37	59.39	139.05	1.37
		5-NNI	62.74	146.87	1.23	63.54	148.75	1.23
		7-NNI	70.35	164.77	1.23	71.14	166.55	1.22
		PS	252.35	590.78	2.16	252.86	591.97	2.16
		MCMC	21.27	49.79	1.64	22.06	51.64	1.64
MNAR	10%							
		CC	6.06	14.18	1.49	6.97	16.32	1.49
		AC	3.91	9.14	1.47	4.76	11.14	1.47
		1-NNI	7.13	16.69	1.41	8.73	20.44	1.41
		5-NNI	13.52	31.65	1.37	14.56	34.09	1.37
		7-NNI	15.25	35.71	1.36	16.21	37.95	1.36
		PS	65.94	154.37	1.70	67.12	157.14	1.70
		MCMC	5.09	11.92	1.48	6.00	14.06	1.48

(Table 3). Continued.

Missingness mechanism	Missingness Rate	Imputation Method	N = 500 simulations			N = 1000 simulations		
			RB	SB	Relative Width	RB	SB	Relative Width
	30%							
		CC	26.68	62.46	1.77	27.29	63.89	1.77
		AC	25.32	59.28	1.70	25.84	60.49	1.70
		1-NNI	64.34	150.64	1.39	64.95	152.05	1.39
		5-NNI	83.11	194.57	1.24	83.27	194.95	1.24
		7-NNI	82.12	192.24	1.23	82.07	192.13	1.23
		PS	224.83	526.34	2.21	225.27	527.39	2.21
		MCMC	27.70	64.85	1.73	28.46	66.62	1.73

MCAR: Missing Completely at Random; **MAR**: Missing at Random; **MNAR**: Missing Not at Random. **RB**: Relative Bias; **SB**: standardized bias. **(1,5,7)-NNI**: 1,5,7 – Nearest Neighbour Imputation. **PS:** Propensity Score. **MCMC**: Markov Chain Monte Carlo.

Relative bias (RB): Relative bias was calculated by dividing the raw bias (difference between the mean value over simulation results and the true parameter) by the true value.

$$\frac{\bar{\hat{\beta}} - \beta}{\beta} \tag{7}$$

Standardized bias (SB): We compared the true value of the beta regression coefficient of the considered interaction factor in the model with the corresponding value obtained with each of the analyzed methods, relative to the standard error of the simulated value. Standardized bias was calculated as it follows:

$$\frac{\bar{\hat{\beta}} - \beta}{SE(\hat{\beta})} \tag{8}$$

where $\hat{\beta}$ is the value of the estimates obtained in the simulation study, $\bar{\hat{\beta}}$ the average estimate of interest over the performed simulations, β the true value and $SE(\hat{\beta})$ defined as $\sqrt{[1/(N-1)]\sum_{i=1}^{N}(\hat{\beta}_i - \hat{\beta})^2}$. As useful rule, a standardized bias exceeding 50% in a negative or positive direction, the bias is having a considerable impact on efficiency, coverage and error rates [24].

Coverage: The coverage of a confidence interval is the proportion of times that the obtained $\hat{\beta}_i \pm Z_{1-\alpha/2}SE(\hat{\beta}_i)$ confidence interval contains the true β specified parameter value. If the coverage value is below the 90%, the performance of the interval procedure will be troublesome.

Relative width: If one procedure has a similar or higher rate of coverage than another but yields intervals that are substantially shorter, then it should be preferred. Shorter intervals translate into greater accuracy and higher power.

Density plots of the different distributions generated by the proposed imputation methods were also depicted. To this end, for each imputation method within each simulation scenario, we summarized the values obtained in the replications using the median, since the distributions were skewed for many scenarios. All statistical analyses were conducted with SAS for Windows Version 9.3, and graphical displays were obtained with R 3.0 release [25].

Simulation Results

Relative and standardized bias of the change over time of the covariate of interest (time×health status) as well as coverage and relative width for *N* = 500 and *N* =1000 replicates are presented in Table **3**. At 10% of missingness, and for the MCAR setting, CC method showed lowest RB and SB values whereas PS method presented the highest values. Focusing on the MAR scenario, MCMC and CC methods showed SB values lower than 50% (6.99% and -10.79%, respectively). As for MNAR missingness mechanism, SB and RB values obtained from AC analysis were the lowest.

Increasing the missingness rate at 30%, 7-NNI and PS yielded higher SB than did MCMC, AC or even the CC method in MCAR and MAR scenarios. Moreover, CC and AC approaches showed a SB value lower than 50% (42.34% and 48.85%, respectively). As for the K-NNI method, they provided a bias exceeding 50% in a

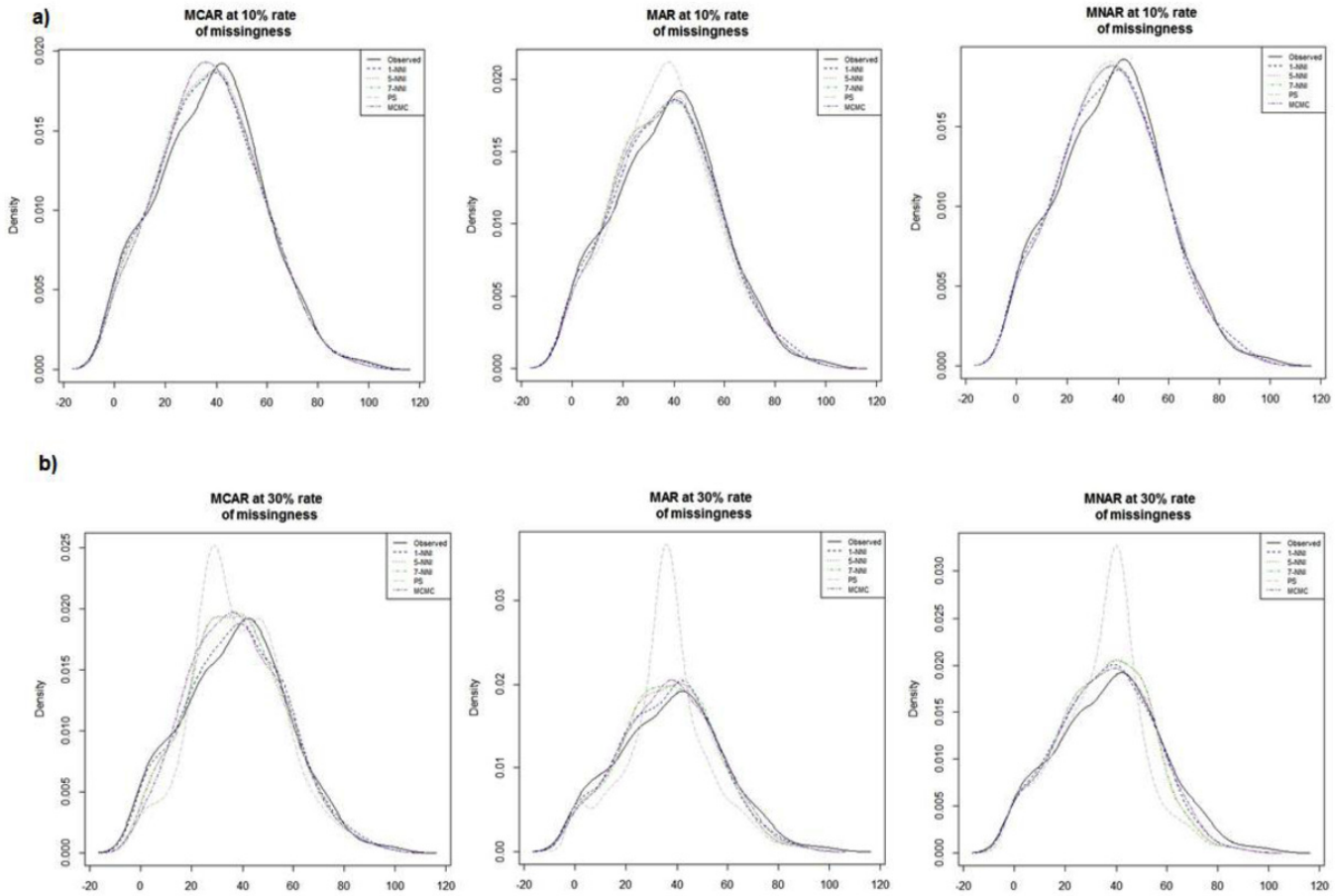

Figure 1: a) Density plots of the imputed distributions according to the missingness mechanisms (MCAR, MAR, MNAR) at 10% of loss of information. **b**) Density plots of the imputed distributions according to the missingness mechanisms (MCAR, MAR, MNAR) at 30% of loss of information. MCAR= missing completely at random, MAR = missing at random, MNAR = missing not at random.

positive or negative direction regardless of the K values. When missingness was considered MNAR at 10% of loss of information, all the studied methods except PS presented a nonsignificant SB value.

Regarding the coverage, in all studied methods at different scenarios (10% and 30% missingness rate, MCAR, MAR and MNAR mechanisms) coverage values were higher than 99. Figure **1** shows density functions of the imputed distributions, related to different missingness mechanisms and percentages of loss of information with *N* =1000 replicates. At 10% rate of missingness, and for the MCAR setting, all three imputation methods showed similar goodness of fit. As regards to the MAR pattern, MCMC and K-NNI had better results than the PS technique. In the MNAR setting, however, density plots corresponding to the three imputation methods are far from the observed distribution, indicating a poor goodness of fit. Similar behavior can be observed at 30% of loss of information.

DISCUSSION

Observational studies, clinical trials, and other biomedical study designs must assess changes in the outcome(s) of interest over time and evaluate its relationship with other variables. These are the main objectives in observational HRQoL studies. The assessment becomes complicated when data are missing.

Several statistical approaches have been developed for longitudinal data or repeated measures analysis to account for missing data. Nevertheless, there is small literature published in this area applied to pre-post observational HRQoL studies. In this study, we explored different methods of handling missing data in a repeated measures study with only two measurements. We compared the results of five analysis approaches: two of them based strictly on observed data - CC, AC- and the other three based on imputation of the unobserved data - K-NNI, PS, and MCMC to assess their performance in handling missing

data in HRQoL outcomes. The comparison of methods for the treatment of missing data in this context is very useful as a guide for the real practice. One major problem with missing data is that it is usually not known how non-response is generated. Thus, it is generally recommended to perform sensitivity analyses (i.e., a set of analyses showing the effect of the proposed methods according to different values of parameters on which the analyses are based) to determine the effectiveness of the statistical methods. In our simulation, we generated the missingness pattern assuming that the mechanism is known. This makes it easier to assess the performance of the evaluated imputation methods across the three missingness classifications. The percentage of lost information should also be taken into account when evaluating results obtained from different methods. In this way, one could apply to the population covering all real and possible loss information cases, especially when a small number of complete cases were used. We also assessed the bias yielded from different approaches for handling missing data and determined the technique that minimized bias.

To evaluate the performance of the various statistical methods and to compare their results over different missingness mechanisms, we computed the relative and standardized bias in a simulation study. Relative bias indicated the degree of deviation of the coefficient estimated by each method from the original beta estimate relative to the latter value. We also calculated the standardized bias to assess whether the relative bias was significant. We observed large differences in both of these parameters. In the MCAR scenario, CC and AC analyses showed low relative bias values regardless of the missingness rates (10% or 30%) and less biased coefficients than the K-NNI or PS methods. Deleting cases with missing data often leads to a decrease in the sample size and a loss of power and efficiency in the estimates. This loss of efficiency affects the significance of the estimates [3] and leads to larger standard errors in CC or AC methods, hence the benefit of obtaining smaller bias values. The MCMC multiple imputation method yielded similar relative and standardized biases compared to the CC and AC analyses. The main difference between the MCMC and CC/AC methods was due to the loss of participants, since the MCMC uses the whole sample without deleting cases. In the MAR setting, the performance of CC and AC methods is as good as in the MCAR setting, particularly with a 10% missingness rate. When the missingness rate increases to 30%, the loss of power is reflected in the results. Low and nonsignificant bias values at a 10% rate of missing data seems to result from applying the general linear mixed-models model approach. Moreover, when increasing the percentage of missing data, SB values of these three methods (CC, AC and MCMC) for handling missing data under the assumed mechanism (MAR) were found to be close to a significant bias. In previous studies for handling missing data in longitudinal studies with more than two measurements, MCMC was the preferred method [26-29]. However, this is not the case in our study. In our case, using MCMC there is no an important bias decrease compared to CC/AC methods, which is supported by other researchers' findings [30,31].

When there are factors related to both the outcome and missingness not included in the model, the missingness mechanism is considered to be MNAR. All methods showed important bias values suggesting performing more senstivity analysis.

The PS imputation method overestimated beta values, yielding biased results. This occurred in all missingness percentages and patterns. Although the missingness in the outcome variable is monotone, it is should be pointed out that this method, as it is focused on the propensity score values, uses a fixed number group division of the observations. In our case, as our scenario is based on a pre-post design, observations were not divided as the method usually prescribes. This agrees with known theoretical findings, that PS can give biased estimates of coefficients when data on predictor variables are missing [32,33]. Schafer [34] found that the PS technique is not appropriate for analyses involving relationships among variables. It would appear that this recommendation to not use PS is supported by the results presented here.

We also evaluated the K-NNI imputation method. Regardless the missingness mechanism and rate, the SB value increased as the K value increased.

Although there is currently no consensus about the appropriate acceptable loss rate in repeated measures or longitudinal studies [35], it is generally recommended that 80% should be used as the minimum acceptable follow-up rate (i.e., dropout of 20%) [36]. In many cases, though, the dropout rate fluctuates between 20% and 30%. In our simulated data, as the missingness rate increased from 10% to 30%, the relative and standardized bias also increased, especially when the type of non-response followed

MAR or MNAR pattern. These methods introduce more bias, despite of correcting it and consequently obtaining efficient estimates by means of correction methods.

We also considered the modeling approach in our analyses. In this study we applied general linear mixed-effects models to all analyses. With these models, tests tend to be more powerful than with standard methods such as general linear models. In fact, general linear mixed-effects models take into account how measurements of the same study unit are correlated. Moreover, compared to generalized linear models, general linear mixed-effects models provide more flexibility in handling missing data and tend to yield results that are more powerful when testing and making inferences compared with generalized linear models or even absolute change tests.

Our results could be affected by the simulation procedure. Despite the fact that correlations between variables and the simulated outcome were very close to those computed with the original response variables, a small bias could have been introduced into the final results. To the best of our knowledge, however, there is no consensus to carry out a simulation in scenarios like the ones we presented.

The generalizability of our findings is limited to settings similar to the scenarios considered in the simulation study. First, we assumed that the outcome of interest, the total St. George's Respiratory Questionnaire score, followed a normal distribution. However, HRQoL outcomes tend to be bounded variables - usually between 0 and 100 - and do not fit a normal distribution. Furthermore, for the performance of this study, we assumed that there were no missing values in covariates - which is in some situations may be unrealistic - and thus our results could only be extrapolated to studies with such features.

CONCLUSIONS

Our simulation study shows the comparison of imputation techniques for handling missing data in HRQoL continuous and bounded outcomes in a pre-post setting. It can be concluded that MCMC method did not show lower bias than AC but it provides more power when handling data. When missingness follows MCAR or MAR mechanism, we recommend using AC analysis combined with generalized linear mixed models when missingness rate is small. Nevertheless, when missingness percentage is high, MCMC imputation method is preferred: after imputing, increases sample size and therefore, statistical power gets higher. There is no method to determine if the dropout is MNAR. In that case, a further sensitivity analysis should be made.

ACKNOWLEDGEMENTS

This research was supported by Ministerio Español de Educación y Ciencia y fondos FEDER (MTM2010-14913), Fondo de Investigación Sanitaria (PI020510), Universidad del País Vasco UPV/EHU (GIU10/21, UFI11/52), Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco (IT620-13) and Departamento de Sanidad del Gobierno Vasco (2012111008).

SUPPLEMENTAL MATERIALS

The supplemental materials can be downloaded from the journal website along with the article.

REFERENCES

[1] Altman DG. Missing data. BMJ 2007; 334(7590): 424. http://dx.doi.org/10.1136/bmj.38977.682025.2C

[2] Barnard J. Applications of multiple imputation in medical studies: from AIDS to NHANES. Stats Methods Med Res 1999; 8(1): 17-36. http://dx.doi.org/10.1191/096228099666230705

[3] Little RJA. Statistical analysis with missing data. Wiley; 2002. http://dx.doi.org/10.1002/9781119013563

[4] Laird NM. Missing data in longitudinal studies. Stat Med 1988; 7(1-2): 305-15. http://dx.doi.org/10.1002/sim.4780070131

[5] Robins J, Rotnitzky A. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 1994; 89: 846-66. http://dx.doi.org/10.1080/01621459.1994.10476818

[6] Robins J, Rotnitzky A. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc 1995; 90: 106-21. http://dx.doi.org/10.1080/01621459.1995.10476493

[7] Molenberghs G. Missing data in clinical Studies. West Sussex, England: John Wiley & Sons; 2007. http://dx.doi.org/10.1002/9780470510445

[8] Janssen KJM, Donders AR, Harrell J, Vergouwe Y, Chen Q, Grobbee DE. Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol 2010; 63(7): 721-7. http://dx.doi.org/10.1016/j.jclinepi.2009.12.008

[9] Xie H. Analyzing longitudinal clinical trial data with nonignorable missingness and unknown missingness reasons. Comput Stat Data An 2012; 56(5): 1287-300. http://dx.doi.org/10.1016/j.csda.2010.11.021

[10] Ibrahim JG. Missing data methods in longitudinal studies: a review. Test 2009; 18(1): 1-43. http://dx.doi.org/10.1007/s11749-009-0138-x

[11] Marshall A, Altman DG. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Med Res Methodol 2010; 10: 112. http://dx.doi.org/10.1186/1471-2288-10-112

[12] White IR. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med 2010; 29(28): 2920-31. http://dx.doi.org/10.1002/sim.3944

[13] Garcia-Laencina PJ, Abreu PH, Abreu MH, Afonoso N. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. Comput Biol Med 2015; 59: 125-33. http://dx.doi.org/10.1016/j.compbiomed.2015.02.006

[14] Saini I, Singh D, Khosla A. QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. J Adv Res 2013; 4(4): 331-44. http://dx.doi.org/10.1016/j.jare.2012.05.007

[15] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, *et al*. Missing value estimation methods for DNA microarrays. Bioinformatics 2001; 17(6): 520-5. http://dx.doi.org/10.1093/bioinformatics/17.6.520

[16] Rosenbaum P. The central role of the Propensity Score in Observational Studies for Causal Effects. Biometrika 1983; 7: 41-55. http://dx.doi.org/10.1093/biomet/70.1.41

[17] Rubin DB.& Schenker. Multiple Imputation for interval estimation from simple random samples with ignorable nonresponse. J Am Stat Assoc 1986; 81: 366-74. http://dx.doi.org/10.1080/01621459.1986.10478280

[18] Molenberghs G. Linear mixed models forl Longitudinal data. Springer; 2000. http://dx.doi.org/10.1007/978-1-4419-0300-6

[19] SAS/STAT 9.3 User's Guide (2nd Edn). Version 9.3. Cary NC: SAS Institute; 2009.

[20] Esteban C, Quintana J, Moraza J, Aburto M, Aguirre U, Aguirregomoscorta J, *et al*. BODE-Index vs HADO-score in chronic obstructive pulmonary disease: Which one to use in general practice? BMC Med 2010; 8: 28. http://dx.doi.org/10.1186/1741-7015-8-28

[21] Jones PW, Quirk FH, Baveystock CM. A self-complete measure of health status for chronic airflow limitation. The St. George's Respiratory Questionnaire. Am Rev Respir Dis 1992; 145(6): 1321-7. http://dx.doi.org/10.1164/ajrccm/145.6.1321

[22] Vittinghoff E, Glidden DV, Shiboski SC. Regression Methods in Biostatistics. Linear, Logistic, Survival, and Repeated Measures Models. Second Edition. Springer; 2012.

[23] Burton A, Altman DG, Royston P&HR. The design of simulation studies in medical statistics. Stat Med 2006; 25(24): 4279-92. http://dx.doi.org/10.1002/sim.2673

[24] Collins LM, Schafer JL. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001; 6(4): 330-51. http://dx.doi.org/10.1037/1082-989X.6.4.330

[25] R: A language and environment for statistical computing, reference index version 3.0. Version 3.0. Viena, Austria: R Foundation for Statistical Computing; 2014.

[26] Spratt M, Carpenter J, Sterne JAC, Carlin JB, Heron J, Henderson J&TK. Strategies for multiple imputation in longitudinal studies. Am J Epidemiol 2010; 172(4): 478-87. http://dx.doi.org/10.1093/aje/kwq137

[27] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, *et al*. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009; 338: b2393.

[28] Mazumdar S, Tang G, Houck PR, Dew MA, Begley AE, Scott J, *et al*. Statistical analysis of longitudinal psychiatric data with dropouts. J Psychiatr Res 2007; 41(12): 1032-41. http://dx.doi.org/10.1016/j.jpsychires.2006.09.007

[29] Jorgensen AW, Lundstrom LH, Wetterslev J, Astrup A, Gotzsche PC. Comparison of results from different imputation techniques for missing data from an anti-obesity drug trial. PLoS One 2014; 9(11): e111964.

[30] Peters SA, Bots ML, den Ruijter HM, Palmer MK, Grobbee DE, Crouse JR, III, *et al*. Multiple imputation of missing repeated outcome measurements did not add to linear mixed-effects models. J Clin Epidemiol 2012; 65(6): 686-95. http://dx.doi.org/10.1016/j.jclinepi.2011.11.012

[31] Twisk J, de Boer M, de Vente W. Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. J Clin Epidemiol 2013; 66: 1022-8. http://dx.doi.org/10.1016/j.jclinepi.2013.03.017

[32] Allison PD. Multiple imputation for missing data: A cautionary tale. Socio Meth Res 2000; 28: 309-10. http://dx.doi.org/10.1177/0049124100028003003

[33] Satty A. Imputation methods for estimating regression parameters under a monotone missing covariate pattern: a comparative analysis. South African Statistical Journal 2012; 46: 327-56.

[34] Schafer JL. Multiple imputation: a primer. Stats Methods Med Res 1999; 8(1): 3-15. http://dx.doi.org/10.1191/096228099671525676

[35] Kristman V, Manno M. Loss to follow-up in cohort studies: how much is too much? Eur J Epidemiol 2004; 19(8): 751-60. http://dx.doi.org/10.1023/B:EJEP.0000036568.02655.f8

[36] Altman DG. Statistics in medical journals: some recent trends. Stat Med 2000; 19(23): 3275-89. http://dx.doi.org/10.1002/1097-0258(20001215)19:23<3275::AID-SIM626>3.0.CO;2-M

Received on 07-07-2015 Accepted on 07-08-2015 Published on 19-08-2015

http://dx.doi.org/10.6000/1929-6029.2015.04.03.1