Plausible Multiple Imputation and Consolidate Matched Pairs Models:

Quality of Life Square Tables with Missing in Two Rounds

By

Dr. Shakir. Hussain

Division of Public Health,
School of Medicine, University of Birmingham, UK

Abstract.

The aim of this paper is to introduce a procedure that integrates the performances of the imputer and the analyst. Estimates produced under the existence of missing data of certain mechanism mostly generate biases and loss of precision. This procedure is illustrated using a real life example in which we compare the effectiveness of patients' self-management (PSM) of anticoagulation versus the standard care.

A general framework dealing with the Quality of Life (QoL) data of ordinal response type is presented using items collected at the base and end points of a patient's questionnaire. Multiple Imputations (MI) is conducted as a part of the editing process to deal with missing data in two rounds, i.e., before and after, that is common in such studies. We apply a hierarchical log-linear model to describe the non-responses mechanism for the item under consideration and multiply impute 10 sets of plausible data sets. The generated data sets can then be used by the data analyst to consolidate a matched pairs model for potential investigation of symmetry, conditional symmetry and quasi symmetry in a square table or any other statistical analyses.

This paper illustrates, stage-by-stage how the procedure helps with selecting the most proper model for the non-response mechanism, and solves issues regarding precision and biases in the estimates.

Key Words: Multiple imputations, hierarchical loglinear models, square table, matched pairs model

1. Introduction

Editing and imputation are usually integrated to improve the quality of the data. Missing values may occur in survey items, census, or in records of files in the public or private sector in different activities. Since the intended data modelling is designed to work with complete data Multiple Imputation (MI) is one way to fulfil the requirements. Imputing for missing values is undertaken to generate complete sets, which will satisfy different interest groups

such as data editors, analyst and others. Complete data sets usually generated by imputation should show important properties of imputed values close to the true, marginal distributions should be the same as the true corresponding distribution, and the MI procedure should lead to imputed values that are plausible. The most important condition is that estimation results with unbiased model parameters and inference should be efficient. This condition may be achieved if the same person impute and analyse the data under the same models.

Several inconsistencies may appear between the imputer and the analyst. For example, the first type of inconsistency between the imputer and the analyst occurs when the analyst assumes (or imposes restrictions) more than the imputer. Schafer (1997) shows that there is nothing wrong with the imputed dataset and inference using MI is still valid but interval estimates will be wider. This happens under the validity of the imposed assumption, otherwise imposing an invalid assumption can lead to biased results and any inferences and predictions based on them. In this case the analyst's model can be considered as a special case of the imputer's model. Note that this problem may be caused by the analyst (who omits relevant information from the analysis which can lead to some kind of model misspecification) and not the data imputer.

The second type of inconsistency arises when the imputer assumes (or put restrictions on the data) more than the analyst. This happens when the analyst's model is more general than the imputer's model. In this case, the imputer applies some assumptions to the complete data as opposed to the analyst. Even in the case, the practical consequences of the inconsistence will depend on whether the imputer's extra assumptions are true or valid. Several researchers investigated the case when these extra assumptions were true, and as stated by Meng (1995) and Rubin (1996), the point estimates were unbiased, more efficient and have shorter intervals than observed data estimates derived completely from the analyst's model. The authors explain that this happens because it incorporates the imputer's superior knowledge about the state of nature. That is, the additional information supported by the imputer cannot invalidate the model; on contrary they can only help. However, in the case that the imputer's extra assumptions are not true, the model will be erroneous and any MI data created under a misspecified model can lead to misleading results and conclusions. Hence, it can be concluded that the most serious case is when the imputer imposes more assumptions that are not valid and that will later be the subject of the analyst's inquiry.

In this paper we introduced a procedure to minimise possible inconsistency, between the imputer and the analyst or the data editor, that might happens when the data includes missing values. We then illustrate this procedure with a real life example in which we compare the effectiveness of patients' self-management PSM of anticoagulation versus standard care.

It is a common practice to conduct QoL studies to examine the treatment effects. A large number of QoL studies have also been conducted with regard to measuring the quality of life for individuals especially in the area of health. Almost all QoL studies are intending to measure the health of populations, assessing the benefit of alternative use of resources, comparing two or more interventions in a clinical trial and making a decision regarding treatment for an individual patient. Many other studies deal with this kind of analysis using traditional t- tests or non-parametric methods such as the Wilcoxon sign rank test. Moreover, they often sum the option choices for patients into one score, a matter that has serious drawbacks. In QoL data the responses are generally of ordinal nature, the questionnaire are often scored [yes=1, no=0] or [agree = 5, . . ., disagree=1] and the total score is then obtained from the responses. To compare the change in score, practitioners frequently subtract the baseline value from the endpoint value resulting in a "change score" for each individual under study. Processing in this manner, information about individual items will be lost as a result of this aggregation. Moreover, parameter estimates based on subtract analyses will be dependent on the nature of particular items scored and may obscure the true factor structure of the items and leading to biased estimation.

In the case of ordinal scales, this may replace a continuous underlying structure and the intervals between values on the scale are not necessarily equally spaced. On this basis, a subject whose score increases from 1 to 2 may have shown an absolute increase in the underlying QoL variable of interest that is greater than the absolute increase from 2 to 4 made by other subjects in the study. However, if one calculates a change score for each subject then the opposite trend will be observed: the first subject change score will be lower than that of the second subject. To compare these changes in scores between treatments are commonly done by means of Wilcoxon sign rank test. This nonparametric test subtracts the first measurement from the second measurement and then ranks the results to obtain a test statistic. This process is vulnerable due to the potential difference between the change in an observed score and the absolute change in the underlying variable.

The other important issue is regarding tackling the problem of missing data, i.e. partial missing. In QoL data analysis, the use of list-wise deletions or "complete case analysis" is quite common. This practice should be totally avoided since it results in many drawbacks. For example, it leads to low power due to the reduction in the number of observations and the resulted estimates of these analyses can be biased when missingness occur at random (MAR) or not at random (MNAR). One simple approach is to use of single imputation. However, almost all types of single-wise values imputations suffer from the problem of underestimation of the variance of any estimate, which in turn can affect any test based on the estimates. MI, on the other hand, retains many of the advantage of single-wise imputation and rectifies the underestimation problem of the variances.

The rest of the paper is organized as follows: Section 2 gives a brief presentation of the example data, while the methodology is discussed in Section 3. In Section 4 we present the estimated results. Finally, we give a short summary and conclusions in Section 5.

2. Data description

The data is collected in the Midlands region of UK for two groups of patients, one receiving self management of anticoagulation (PSM) and the other receiving standardized care (Control). After a period of 12 months the patients have been requested to answer a Spielberger Questionnaire (see the appendix). For more information bout the data we refer to Fitzmaurice et al (2005).

Speilberger Questionnaire has 4 options and we may consider the following questions:

1- Is there any patient who is more anxious at baseline than the on treatment patients?
2- Is there any change in anxiety at the end of the study (after 12 months)?

Note that we, here, only consider the first item from this questionnaire in our analysis to illustrate our procedure, the other items can be analyzed in a similar manner. Since there are many empty cells in last option of the first analyzed item, we collapse the total number of options to three (instead of four) by joining options three and four together. The counts for this questionnaire are presented in the following square Table 1. Any practitioner wishing to use the data from this table will face the fact that more than 25% of the data is missing, 18 observations in the first round and 46 observations in the second round. However, 9

observations are missing in both rounds. Now, if we mistakenly apply a list-wise deletion approach to tackle this missing, we, end up with a square table consisting of 158 observations only (see Table 2 below), instead of 213 observations as in Table 1.

Table 1. Participants counts in two rounds opinion study

Item A	1_b	2_b	3_b	NA_b	Total
1_a	95	9	10	22	136
2_a	13	13	2	12	40
3_a	6	7	3	3	19
NA_a	4	4	1	9	18
Total	118	33	16	46	213

1_a, 2_a, 3_a, NA_a is Round 1 levels and 1_b, 2_b, 3_b, NA_b is Round 2 levels

Table 2. List-wise deletion of Table 1

Item A	1_b	2_b	3_b	Total
1_a	95	9	10	114
2_a	13	13	2	28
3_a	6	7	3	16
Total	114	29	15	158

1_a, 2_a, 3_a is Round 1 levels and 1_b, 2_b, 3_b is Round 2 levels

3. Methodology and Results for the Imputer

In this section we present the methodology that we use in our procedure.

3.1. Multiple Imputations

Missing data is common in empirical research and often is tackled by list-wise deletion, e.g. delete cases with any missing data. This may lead to a considerable reduction in the data set. Imputing a single value for each missing datum and then analysing the completed data using standard techniques for complete data will end up with inappropriate but may be smaller standard error estimate, shorter confidence intervals and misleading p-value (i.e., too significant). Single imputation such as the mean or the regression imputations result in ignoring the uncertainty due to the missing data. Multiple imputations involving more than one set of imputation allows valid assessment of uncertainty that come from missing data, see Rubin (1978). This may results in wider but more accurate confidence intervals and less significant p-values.

According to Fay (1992), miss-matching between the imputer's model and the analyst's model can be harmful. On the other hand, Schafer (2003) stressed the importance for investigating of potential MNAR mechanism in the data at the first hand, which implies that

the analyst will not necessarily use the same model as the imputer. Note that, recent publications with regard to non-ignorable MI have used different types of models for MNAR mechanism, see Little and Rubin (2002). The nature of our data in this study is of discrete categorical type with a hierarchical log-linear model for the joint distribution of the categorical and indicator variables. This will help in revealing any possible association between missing indicators and responses, and the existence of such association implies that we have a MNAR mechanism. This model will in its tern be used to generate 10 MI sets of data under that specific MNAR mechanism. These data sets will then be ready for the analyst for any further modelling.

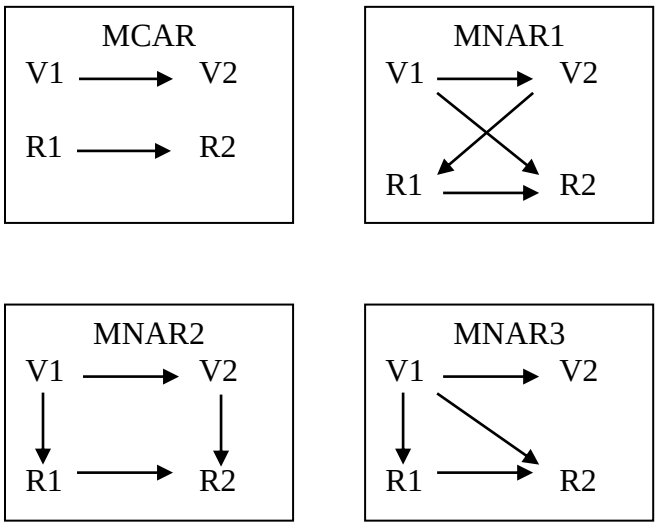
It is known that efficiency is a function of number of imputations (m) and that the efficiency gain rapidly diminishes when this number increases. Following Rubin (1987) it is sufficient to generate 3-5 MI sets to achieve enough efficiency. In our study, to avoid the above-mentioned problems and to achieve proper efficiency, we conduct 10 MI sets. We, however, also tried higher number of MI sets but the results did not indicate any differences when comparing to the results from the 10 sets.

3.2. Generate MI Square Tables

Since the missing values in QoL data occur both at the beginning and at the end of the study, we faced a problem of possible MNAR mechanism. On the other hand, when missing values depend on the variable itself, the missing is neither MCAR nor MAR, which means that the mechanism is nonignorable or MNAR. Hierarchical log linear model is commonly used for the detection of a possible MNAR mechanism. Little (1985) mentioned that when a categorical variable is missing on both rounds of the survey, a large number of models are available for describing the nonresponse mechanism. Models that describe the nonresponse mechanism face the problem of how to distribute patients into the joint distribution in the QoL data. In MI, a small number of plausible values are generated for each missing cell, Little and Rubin (2002) and Schafer (1997). These values are drawn from a plausible conditional (posterior) distribution of the missing value given the adopted model for non-response. The parameters of the plausible distribution are drawn from the estimated sampling distribution of their estimators.

When the variables are polytomous and have the same categories, we can arrange the data in a two-dimensional square table. Examples are found in health studies, such as the pre-test/post-test data, or the QoL items of Speilberger Questionnaire in this paper, see the appendix. Let V_1 and V_2 represent patient opinion at the beginning and at the end of the study respectively, R_1 indicates response to V_1 and R_2 indicates response to V_2. Since missing can happen either in the beginning or at the end of the study (V_1 or V_2) and it may happen in both (V_1 and V_2), all models that associate (R_1) "missing pattern in V_1" and (R_2) "missing pattern in V_2" with V_1 and/or V_2 are nonignorable, i.e., MNAR. When missing does not depend on either V_1 or V_2, then we have ignorable missing, i.e., MCAR. A saturated model (four way contingency table) that includes the joint distribution of (V_1, V_2, R_1 and R_2) is not estimable from the data. Following Little (1985), we conduct all possible MCAR and MNAR processes and end up with three possible models for the missing mechanisms. Figure 1 gives a graphical summary of our MCAR and three MNAR models that we will consider. The path diagrams for four models show different possible associations.

Figure 1: Showing the different mechanisms for missing data. Upper left: MCAR. Upper right: MNAR1 – Reverse time point effect. Lower left: MNAR2 – Current time point effect. Lower right: MNAR3 – Prior time point effect.



In the first row of Table 3 a test for MCAR (V_1V_2, R_1R_2) against MNAR1 (V_1V_2, R_1R_2, V_1R_2, V_2R_1) is conducted. The p-value is (0.035), which indicates that the reverse time point is significant, the missing in V_1 is associated with R_2, and also that the missing in V_2 is associated with R_1. The test for MCAR (V_1V_2, R_1R_2) against MNAR2 (V_1V_2, $R1R_2$, V_1R_1, V_2R_2), i.e., current time point effect on missingness indicates that the missing in V_1 is associated with R_1 and the missing in V_2 is associated with R_2, has a p-value (0.060). Finally,

the test for MCAR (V_1V_2, R_1R_2) against the prior time point effects model MANAR3 (V_1V_2, R1R2, V_1R_1, V_1R_2) is significant with p-value of (0.036).

Now, applying the above methodology of the imputer to our data and using the χ^2 test for goodness of fit we can verify the most proper model in Table 3 below. From this table we can verify that the second model MNAR2 is realistic and significant. However, MNAR1 and MNAR3 show significant difference from MCAR. One might question unrealistic models, like e.g. V_1R_2, that relate response to the values of variables at future point in time.

Table 3. Test results for the different models under MCAR against MNAR Mechanisms.

MODEL	χ^2	DF	P-value
MNAR1 VISIT.1:VISIT.2 R1:R2 VISIT.1:R2 VISIT.2:R1 "Reverse time point effect, Significant & Not Realistic"	6.678	2	0.035
MNAR2 VISIT.1:VISIT.2 R1:R2 VISIT.1:R1 VISIT.2:R2 "Current time point effect, Significant & Realistic"	5.602	2	**0.060**
MNAR3 VISIT.1:VISIT.2 R1:R2 VISIT.1:R1 VISIT.1:R2 "Prior time point effect, Significant & Not Realistic	6.631	2	0.036

R1 indicator for missing in visit 1, and R2 indicator for missing at visit 2.

Since we have three options in the square table, we generate counts for each cell in the ten of 3x3 square tables from a multinomial sampling following the three MNAR mechanisms in Table 3 indicated by the log-linear model.

Following the results from Table 3, we end up with the following parameters estimates of cell probabilities (multiplied by 10 000), in complete case estimate and in MNAR1, MNAR2 and MNAR3, see Table 4 below. These three sets of random draws reflect the uncertainty in the occurrence of non-response given in a model and the uncertainty about the model parameters. Estimates from the three models are very similar suggesting robustness in the final probabilities to alternative nonresponse models. Now, these 10 multiply imputed sets are ready for the purpose of analysis, i.e. consolidate inference that we cover in the following sub-section.

Table 4. The 3x3 Square Table Parameters Estimates of complete case and 3 MNAR

Parameter	Complete Cases Sample Size = 158	MNAR1 Sample Size = 213	MNAR2 Sample Size = 213	MNAR3 Sample Size = 213
π_{11*}	6011	4882	5540	5445
π_{21}	823	892	1033	1033
π_{31}	380	423	423	423
π_{12}	570	798	516	610
π_{22}	823	986	892	1174
π_{32}	443	610	610	376
π_{13}	633	751	657	516
π_{23}	127	282	188	188
π_{33}	190	376	141	235
Total	10000	10000	10000	10000

$*\pi_{11}$ denote count estimates of the cell number 1,1 in the 3x3 square table.

4. Methodology and Results for the Analyst

4.1 Rubin Rules

MI methods are among the most important techniques for missing data in multivariate analysis. Imputation is a generic term for filling the missing data with plausible values (Schafer 1997). MI manages to cover all cases in the construction of the model, thus maintain power and avoid biases. Missing data mechanisms, e.g. MCAR, MAR and MNAR, relate to the reasons for the missing observation and the associations between the observed and missing observation, Little & Rubin (2002). To keep an association between outcome and predictors, MI starts with plausibly generating (m) fixed number of complete data sets and then conducting proper analysis on all m completed sets. Each missing value will have m independent plausible values, each set will be analysed individually and then combined in one overall estimate. The variation among the m sets will compensate for missing uncertainty, thus may have larger confidence intervals bounds. Rubin (1987) introduced the rule "Rubin's Rules" (RR) for combining the estimate. The essence of RR is based on the technique to combine the imputed sets of the data. The Asymptotic normal approximation is assumed to perform RR, the single population parameter estimate $\hat{Q}_i$ and its variance U_i has the average of m overall estimate

$$\overline{Q} = \frac{1}{m}\sum_{i}^{m}\hat{Q}_i \quad \overline{U} = \frac{1}{m}\sum_{i=1}^{m}U_i \qquad (1)$$

The between imputation variance is,

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \overline{Q})^2 \qquad (2)$$

The total variance e.g. a sum of within and between,

$$T = \overline{U} + (1+\frac{1}{m})B \qquad (3)$$

is inflated by a factor (1/m) to reflect the extra variability due to imputing the missing data using a finite instead of an infinite number of imputations. The approximation

$$\frac{(Q-\overline{Q})}{\sqrt{T}} \sim t_\nu \text{ , with} \qquad (4)$$

$$\nu = (m-1)\left[1+\frac{\overline{U}}{(1+m^{-1})B}\right]^2 \qquad (5)$$

degrees of freedom, and a p-value for testing H_0: $Q = Q_0$ can be obtained by computing the following test statistics

$$W = \frac{(Q_0-\overline{Q})^2}{T} \qquad (6)$$

against the F distribution with k (number of parameter), and

$$\gamma = (m-1)(1+r^{-1})^2, \quad r = \frac{(1+m^{-1})B}{\overline{U}} \qquad (7)$$

degrees of freedom. Here *r* represents the relative increase in variance due to the missing data.

Thus, the p-value for testing the null hypothesis $Q = Q_0$ is

$$p = P(F_{k,\gamma} \geq W)\,. \qquad (8)$$

In this paper we consider the estimate, its uncertainty and the p-value from the above equations in our data analysis.

4.2. Problems with Complete Case Analysis

1. Loss of precision: let $\hat{\theta}_{cc}$ represent the estimate from the complete case (r), the increase of the variance relative to the estimate of data of size (n) with no missing e.g. $\hat{\theta}_{NM}$, is

$$Var(\hat{\theta}_{CC}) = Var(\hat{\theta}_{NM})(1+\Delta_{CC}) \qquad (9)$$

where $\Delta_{CC} = \frac{n-r}{r}$ is the proportion increase in variance from the loss of information. If r = n/2 then the variance is doubled.

2. Bias: If M_{CC} and M_{IC} denote the means in the two strata (Complete Case and Incomplete Case), the overall mean is

$$\mu = \pi_{CC}\mu_{CC} + (1-\pi_{CC})\mu_{IC} \qquad (10)$$

The Bias of the mean of complete case strata is

$$\mu_{CC} - \mu = (1-\pi_{CC})(\mu_{CC} - \mu_{IC}) \qquad (11)$$

Under MCAR, i.e. when $\mu_{CC} = \mu_{IC}$, the bias is equal to zero.

4.3. Matched Pair Model

4.3.1. Methodology

Responses of patient's opinions at the baseline and after treatment may be modeled using matched pair model, Agresti 1995. Patients in the QoL study (under PSM and control) are classified into categories of options that represent their opinion in a given question. The Spielberger questionnaire includes 6 items with each item consists of 4 categories (reduced to 3) to investigate the treatment effects after the patients participation in this PSM study (see the appendix). At this stage, we are mainly interested in investigating the following:

1. Testing whether the distribution between the 3 categories is the same.
2. Whether the after treatment choice is independent of the before treatment choice for the patients who change their mind.

3. Whether there is any symmetric/asymmetric association pattern between the before and after answers.

Many of the social and physiological studies are based on responses taken on the ordinal scale and may involve subjective opinion. In particular, an intermediate category will often be subjective to more misclassification than an extreme category because there are two directions in which to err from the extremes. A square table can be used to display joint rating of two occasions (before and after).

Tables of equal rows and columns with dependence structure have two matters. The first important issue is the difference in the marginal distribution, and the question that we like to have an answer for is "whether classification for after treatment tends to be higher than before treatment". The second important issue is regarding the extent of main-diagonal occurrence within the joint distribution of the rating, and the question one may ask is "whether there is agreement between the two occasions, e.g. non-significant of diagonal cells". Perfect agreement occurs when the total probability of agreement = 1.

It is well known that one cannot represent the models above in a log-linear form, so Agresti (2002) uses Generalized Linear Model (GLM) to investigate matched pairs models. Pre and post results obtained from these patient's data were analysed using GLM with Poisson family for the counts. The data is in square table format, and matched pairs models are used to analyse agreement between before and after treatments. Moreover, the structure of agreement and fitting different models is also investigated.

Now, we consider the patient's opinion before and after treatment. For any patient, indexed by s, suppose that the log probabilities for the k response categories are

$$\lambda_{s1},\ldots,\lambda_{sk} \qquad \text{before treatmen, and}$$
$$\lambda_{s1}+\tau_1,\ldots,\lambda_{sk}+\tau_k \qquad \text{after Treatment}$$

so that $(\tau_1,\ldots,\tau_k)$ is the treatment effect, assumed to be the same for all individual patients. To simplify the idea of modelling square tables, we introduce the main type of models we use in the analysis. The first is the symmetry model where we like to test the null hypothesis

$$H_0 \ : \qquad \pi_{ij}=\pi_{ji}.$$

In other words, we like to test if the cell probability on the one side of the main diagonal is a mirror image of that on the other side.

For expected frequency the logarithm value is

$$\log \mu_{ij} = \lambda + \lambda_i + \lambda_j + \lambda_{ij} \qquad (12)$$

Conditional symmetry model may be used when categories are ordered. This kind of models estimates an extra parameter τ for the off diagonal on structure of agreements. The main objective of this extra parameter is to quantify the effect on the structure of agreement. The symmetry model is a special case of the Conditional symmetry models when τ = 0. τ is equal to log (π_{ij} / π_{ji}). The following model represents a generalization that includes the condition when symmetry does not hold with ordered category

$$\log \mu_{ij} = \lambda + \lambda_i + \lambda_j + \lambda_{ij} + \tau I(i < j), \quad I(\bullet) \text{ is the indicator function} \qquad (13)$$

The quasi symmetry model implies some association and allows the main effects terms to differ so that

$$(\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}) \neq (\log \mu_{ji} = \lambda + \lambda_j^X + \lambda_i^Y + \lambda_{ij}) \qquad (14)$$

The main effect in (14) is different for rows (X) and columns (Y) and the resulting estimates are the differences in $\{\lambda_j^Y - \lambda_j^X\}$ for j = 1,2,…, . Now, set $\{\lambda_j^Y - \lambda_j^X = \beta u_j\}$, thus the difference in the effect of a category from one occasion to the other follows a linear trend in the category score.

The Quasi-Symmetry model is

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_j + \lambda_{ij} \quad , \qquad (15)$$

where u_j denotes order scores for the row and the column category. The greater the $|\beta|$ is, the greater is the difference between the joint probabilities π_{ij} *and* π_{ji}, that is the difference between the marginal row and column distributions.

The marginal homogeneity model compares marginal distributions in the square table, and one may test the hypothesis of marginal homogeneity by comparing the fit of the specific square table. It is known that a table satisfies both quasi symmetry and marginal homogeneity also satisfies symmetry and one can test marginal homogeneity by comparing goodness-of-fit statistics for the symmetry and quasi-symmetry models, see Bishop et al, (1975). The calculations in this study are performed using the statistical program packages, SPLUS-version 8.01.

4.3.2. Results

In this sub-section we present the most important results of this study with regards to the analyst and when the imputer generates 10 plausible data sets. We analyse the differences in responses between the two occasions, before and after. The results for the treatment group from the Spielberger questionnaire are presented in Tables 5 and 6. Since the hierarchical log-linear model indicated three types of MNAR mechanisms, the MCAR mechanism is violated. Hence, any analysis based on the list-wise deletion is not valid, i.e., the mechanism is not MCAR any more, and results will not be considered for any inferences since they are biased and have poor precision.

Following Agresti (2002), we first start with the symmetry model and we move on to the conditional symmetry and finally we try quasi symmetry. However, when symmetry is not rejected it is possible that the conditional and quasi symmetry models show no fit improvement. The symmetry model implies, for example, that the numbers of patients who change their opinion from 1→2 are “approximately” equal to those who change from 2→1, etc. The model has no interest in the number of patients who did not change their opinion between the two occasions, i.e., 1→1, 2→2 and 3→3. In the conditional symmetry model the main objective of the extra parameter τ for the off diagonal is to quantify the effect on the structure of agreement. Also, the symmetry model is a special case of the conditional symmetry models when $\tau = 0$. The quasi symmetry model permits marginal distributions to differ and allows for marginal heterogeneity. Symmetry model is a special case of quasi symmetry model in which the marginal distributions for the row and column variables are the same or when the score coefficients are equal to zero.

We present results for the matched pair model of square table with order categories for the three MNAR mechanisms in Table 5. These results are products of 10 consolidated estimates of the multiply imputed sets based on the Rubin rules. The most important message is that there exists a significant difference between the patients who changed their opinion from 1→2 against 2→1, but we are unable to reject the hypothesis of symmetry with regard to 1→3 against 3→1 and 2→3 against 3→2. The p-values show good fitting (0.481, 0.694, 0.396) in the 3 MNAR models. Estimates from the 3 models are very similar, suggesting robustness in the final probabilities to alternative nonresponse models. In the square table cells format, the counts in cell (2,1) show significant difference compared with the counts in cell (1,2) leading to rejecting symmetry in the counts of the two cells, see the fifth row (Sym [21⇔12]) in Table 5.

Table 5. Testing for Symmetry using Rubin rules under GLM.

Coefficients	**MNAR1**	**MNAR2**	**MNAR3**
Intercept	**1.445 (0.577)**	**1.574 (0.542)**	**1.470 (0.570)**
Sym [32⇔23]	**0.486 (0.653)**	**0.616 (0.631)**	**0.698 (0.652)**
Sym [31⇔13]	**0.895 (0.622)**	**0.952 (0.562)**	**0.881 (0.663)**
Sym [22⇔33]	**1.604 (0.644)**	**1.548 (0.642)**	**1.570 (0.617)**
Sym [21⇔12]	**1.372 (0.644)**	**1.175 (0.571)**	**1.294 (0.60)**
Sym [11⇔33]	**3.331 (0.585)**	**3.123 (0.568)**	**3.289 (0.578)**
P-value	**0.481**	**0.694**	**0.397**

"32⇔23" stand for testing the significant of the number of patients how change opinion from 3 to 2 against 2 to 3.

In Table 5 some of the modelling coefficients from the Spielberger questionnaire have mainly shown to be non significant although the fittings are rather good. Now, under MNAR2 model, we investigate visit 1 associated with pattern of missing in visit 1, and visit 2 with the pattern of missing in visit 2 which is more realistic than in the case of MNAR1 and MNAR3 in which we investigate visit 1 with visit 2 pattern of missing where one might question models that relate responses to the values of variables at future points of time, e.g. V_1R_2 in Figure 1 upper right and lower right. Accordingly, we select the mechanism of MNAR2 over MNAR1 and MNAR3 as the most proper approach for running the match pair model.

In Table 6, the estimated parameter of conditional symmetry and its uncertainty, $\tau = 0.066$ (0.345), implies that in general the number of patients below diagonal who positively changed

their opinion is not significantly different from those above the diagonal who negatively changed their opinion. The second column in Table 6 shows the results from the quasi symmetry model, the score coefficients and their uncertainties are β_1=0.156 (0.399), β_2=0.359 (0.505), the p-values of both models are (0.085, 0.040). The conditional and quasi symmetry models are rejected.

Table6. Testing for Conditional and Quasi symmetry of MNAR2.

Coefficients	**Conditional Symmetry**	**Quasi Symmetry**
Intercept	**1.744 (0.591)**	**1.386 (0.673)**
Sym [32⇔23]	**0.349 (0.747)**	**0.466 (0.681)**
Sym [31⇔13]	**0.831 (0.605)**	**1.020 (0.612)**
Sym [22⇔33]	**1.431 (0.676)**	**1.633 (0.734)**
Sym [21⇔12]	**0.885 (0.720)**	**1.197 (0.650)**
Sym [11⇔33]	**2.944 (0.619)**	**3.302 (0.677)**
τ	**0.066 (0.345)**	
β_1		**0.156 (0.399)**
β_2		**0.359 (0.505)**
P-value	**0.085**	**0.040**

"32⇔23" stand for testing the significant of the number of patients how change opinion from 2 to 3 against 3 to 2.

The imputer now uses the variables (V_1, V_2, R_1, R_2) in a 4-way contingency table format to investigate the model and then generate 10 plausible multiply imputed data sets. The analyst uses the first 2 of the 4 (Sym_{ij} in 2 occasions, $I_{i<j}$ indicator for the above and below diagonal and U_j for the difference in the categories) variables to test for symmetry and restrict against conditional and quasi symmetry. The first 2 variables are the same for the imputer and the analyst, but the imputer uses 2 extra variables (indicators) to help in imputation. The extra information used by the imputer is valid under the MCAR vs. MNAR missing mechanism and is not inconsistent with the analyst symmetric model since they have been used in the generation of the 10 MNAR sets. In fact the imputer's 2 extra indicators R_1 and R_2 have no valid interpretation in the modelling used by the analyst since there are no missing values any more in the 10 complete plausible sets.

5. Summary and conclusions

In this paper we introduced a procedure to integrate the performances of the imputer and the data analyst, which happens when the data includes missing values in 2 rounds. Hierarchical log-linear models, used by the imputer to identify the non-response mechanism for categorical data (in square table form), measured in two occasions are presented and discussed. Then the imputer uses multiple imputations to generate 10 plausible data sets of 3 MNAR mechanisms that are ready for the analysis. The analyst then applies GLM on the 3 MNAR sets to check the robustness of the model on the alternative nonresponse mechanisms. The consolidate estimates of the 10 sets using the Rubin rules methodology provides final estimates together with their uncertainty. The p-values of the goodness of fits for the 3 MNAR models shown to be reasonable. The most important result is that the differences between the patients who changed their opinion in the component ([1⇔2]) have shown to be significant, indicating the rejection of the null hypothesis of symmetry, unlike the other two components ([2⇔3] and [1⇔3]). The matched pair models of conditional and quasi symmetry are then investigated on the realistic MNAR2 mechanism by GLM. The estimated parameters are not significant and the two models goodness of fits p-values are poor.

Since the main aim of this paper is to introduce a procedure that integrates the performances of the imputer and the analyst or the data editor, this study shows how the imputer and the analyst form a base for QoL opinion study with different nonresponse mechanism.

References

Agresti, A. (1995). "Logit models and related quasi-symmetric loglinear models for comparing responses to similar items in a survey", *Sociological Methods in Research*, **50**, pp. 68-95.

Agresti, A. (2002). "*Categorical data analysis*", John Wiley and Sons, New York.

Bishop, Y. M. M. S. E. Fienberg, and P. W. Holland (1975). "Discrete Multivariate Analysis: Theory and Practice". *MIT Press, Cambridge, MA.*

Fay, R. E. (1992). "When are inferences from multiple imputation valid?" *Proceedings of the Survey Research Methods Section of the American Statistical Association, 227-232.*

Fitzmaurice, D. E Murray, S Hussain, S McCahon, R Holder, J Raftery, H Sandhar, F Hobbs. SMART: Self management of oral anticoagulation, A randomised trial". British Medical Journal 15 Oct 2005.

Little, R. J. A. and D. B. Rubin (2002). "*Statistical Analysis with Missing Data*", second edition, Wiley series in probability and statistics.

Meng, X. L. (1995). **"**Multiple-imputation inferences with uncongenial sources of input (with discussion)". *Statistical Science*, **10**, pp. 538-573.

Rubin, D. B. (1987). "*Multiple Imputation for Nonresponse in Surveys*", J. Wiley and Sons, New York.

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

Schafer, J. L. (1997). "*Analysis of Incomplete Multivariate Data*", Chapman & Hall/CRC.

Schafer, J. L. (2003). "Multiple Imputation in Multivariate Problems When yhe Imputation and Analysis Models Differ". *Statistica Neerlandica Vol. 57, nr. 1, pp.19-35.*

Appendix

Speilberger Questionnaire

	Not at all	Somewhat	Moderately	Very Much
A. I am worried	1	2	3	4
B. I feel calm	1	2	3	4
C. I am tense	1	2	3	4
D. I feel upset	1	2	3	4
E. I feel relaxed	1	2	3	4
F. I feel content	1	2	3	4