STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:3161–3175 (DOI: 10.1002/sim.1547)

A comparison of two methods for the estimation of precision with incomplete longitudinal data, jointly modelled with a time-to-event outcome

G. Touloumi[1,*,†], A. G. Babiker[2], M. G. Kenward[3], S. J. Pocock[3] and J. H. Darbyshire[2]

[1] *Department of Hygiene and Epidemiology, University of Athens Medical School, M. Asias 75, 115 27 Athens, Greece*
[2] *Medical Research Council Clinical Trials Unit, University College London Medical School, 222 Euston Road, London NW1 2DA, U.K.*
[3] *Medical Statistics Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, U.K.*

SUMMARY

Several methods for the estimation and comparison of rates of change in longitudinal studies with staggered entry and informative drop-outs have been recently proposed. For multivariate normal linear models, REML estimation is used. There are various approaches to maximizing the corresponding log-likelihood; in this paper we use a restricted iterative generalized least squares method (RIGLS) combined with a nested EM algorithm. An important statistical problem in such approaches is the estimation of the standard errors adjusted for the missing data (observed data information matrix). Louis has provided a general technique for computing the observed data information in terms of completed data quantities within the EM framework. The multiple imputation (MI) method for obtaining variances can be regarded as an alternative to this. The aim of this paper is to develop, apply and compare the Louis and a modified MI method in the setting of longitudinal studies where the source of missing data is either death or disease progression (informative) or end of the study (assumed non-informative). Longitudinal data are simultaneously modelled with the missingness process. The methods are illustrated by modelling CD4 count data from an HIV-1 clinical trial and evaluated through simulation studies. Both methods, Louis and MI, are used with Monte Carlo simulations of the missing data using the appropriate conditional distributions, the former with 100 simulations, the latter with 5 and 10. It is seen that naive SEs based on the completed data likelihood can be seriously biased. This bias was largely corrected by Louis and modified MI methods, which gave broadly similar estimates. Given the relative simplicity of the modified MI method, it may be preferable. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: observed information; missing data; longitudinal studies; standard errors

*Correspondence to: G. Touloumi, Department of Hygiene and Epidemiology, University of Athens Medical School, M. Asias 75, 115 27 Athens, Greece
†E-mail: gtouloum@med.uoa.gr

Received May 2002
Accepted February 2003

Copyright © 2003 John Wiley & Sons, Ltd.

1. INTRODUCTION

Many cohort studies and clinical trials use repeated measurements of laboratory markers to track disease progression and to evaluate new therapies. A problem faced by practitioners in the analysis of such longitudinal studies is the incomplete marker series due to staggered entry and/or patient's death or loss to follow-up [1–3]. Following the missing data classification introduced by Rubin [4] and Little and Rubin [1], missing data can be classified as: completely at random (MCAR), when the probability of missing data does not depend upon the outcome variable; missing at random (MAR), when the probability of missing data at any time during the follow-up depends on an individual's previously observed outcome variable, and missing non-ignorable (MNI), when the probability of non-response depends on the unobserved outcomes. We will use the terminology non-informative for MCAR or MAR data and informative as synonymous for MNI data. While missing marker values due to loss to follow-up or late entry to the study may be considered non-informative, incomplete marker data due to the patient's death or disease progression are likely to be informative. For example, patients who die early or progress rapidly to disease may tend to have a faster rate of change in the disease marker. With a non-ignorable missing data mechanism, valid inferences can only be made if the non-response process is explicitly modelled [5]. Several methods for the estimation and comparison of marker rates of change in longitudinal studies with staggered entry and non-ignorable missing data due to drop-outs, have been recently proposed [6–14].

Those model based methods that use maximum likelihood (ML) as a method for estimation often use forms of the EM algorithm. The main idea behind the EM (and associated algorithms) is at each iteration, to 'fill in' the missing data using their conditional density, given the observed data and the current estimates of the model parameters [15, 16]. This typically, but not necessarily (see for example reference [14] or Section 3.2), takes the form of the repeated replacement of sufficient statistics by their expectation given current parameter estimates. The great advantage of such algorithms is the ability to exploit full data procedures in incomplete data problems. However the algorithm itself does not provide a method for obtaining the information matrix associated with the estimates. The estimated asymptotic variance-covariance matrix of the ML estimates is not given simply by the inverse of the information matrix obtained at convergence from the completed data, after imputing missing values. Such a method would systematically underestimate the associated variances since it treats the missing values as if they were known.

In principle, (minus) the information matrix adjusted for missing data (observed data information matrix) could be estimated by twice differentiating the log-likelihood of the observed data. However this defeats the object of most EM-based analyses: the avoidance of incomplete data calculations. Alternative solutions for a few special cases have been published [17]. Efron and Hinkley [18] provide some interesting insights on the relative merits of different estimators of the observed data information matrix. Meilijson [19] proposed a method of numerically computing the covariance matrix of the MLE, based on individual observation by-observation score functions of the incomplete data. Louis [20] has provided a general technique for computing the observed data information in terms of complete data quantities, within the EM framework. This works through taking expectations of score and information type quantities. As an alternative, the multiple imputation (MI) approach uses approximate expectations indirectly in terms of the estimates and their covariance matrix [21–23].

Copyright © 2003 John Wiley & Sons, Ltd.

The purpose of this paper is to develop, apply and compare these two methods in longitudinal studies where the source of missing data is either death or disease progression (informative) or end of the study (assumed non-informative). The methods are illustrated by modelling CD4 count data from an HIV-1 clinical trial. The model combines a linear random effects model for the underlying pattern of the CD4 count data with a log-normal survival model for the informative drop-out process [14]. Model parameters are estimated through a particular numerical solution (RIGLS) of the REML score equations, combined with a nested EM algorithm. The latter is necessary to deal with censored survival data. In particular, a modification of the MI method is developed and evaluated for this specific model. In Section 2 the Louis and the MI method are outlined. In Section 3 the Alpha trial example, together with the REML/RIGLS method applied in modelling the CD4 cell count in the Alpha trial are introduced. In Section 4, the adoption of the Louis and the MI methods in estimating the SEs in longitudinal studies with informative drop-outs is discussed. Both methods are applied to the Alpha trial CD4 cell count data. In Section 6 a description is given of a simulation based evaluation of the two approaches, with concluding remarks in Section 7.

2. IMPUTATION BASED ESTIMATES OF THE VARIANCE-COVARIANCE MATRIX: THE LOUIS AND MULTIPLE IMPUTATION FORMULAE

Denote by Y the complete data, which can be decomposed into observed (Y_o) and missing (Y_m) data, respectively. Assume that Y follows the distribution $f(Y;\theta)$, for some parameter vector θ. Inference about θ will be based on the distribution of the observed data

$$f(Y_o;\theta) \tag{1}$$

Louis [20] shows that the information matrix for the maximum likelihood estimator of θ from (1) can be expressed in terms of the complete data score and information:

$$I(\theta \mid Y_o) = E_{Y_m|Y_o}\{I(\theta \mid Y_o, Y_m)\} - \mathrm{cov}_{Y_m|Y_o}\{U(\theta \mid Y_o, Y_m)\}$$

for the usual definitions of the score and information

$$U_i(\theta \mid Y_o, Y_m) = \frac{\partial \ln f(Y;\theta)}{\partial \theta_i}$$

and

$$I_{ij}(\theta \mid Y_o, Y_m) = -\frac{\partial^2 \ln f(Y;\theta)}{\partial \theta_i \partial \theta_j}$$

Note that the expectations are taken over the conditional distribution of the missing data given the observed. If the missing data mechanism is informative then implicitly the conditioning also involves the event of this particular pattern of missingness [1, 5, 10, 14].

In the absence of analytical solutions for these expectations, Monte Carlo approximations can be used instead [24]. Suppose that $Y_m^{(1)}, \ldots, Y_m^{(K_L)}$ represent K_L independent draws from the conditional distribution of Y_m given Y_o, using the maximum likelihood estimate $\hat{\theta}$ of θ.

Copyright © 2003 John Wiley & Sons, Ltd.

Then the empirical estimate of the information matrix for $\hat{\theta}$ is given by

$$I_{\mathrm{L}} = \frac{1}{K_{\mathrm{L}}} \sum_{k=1}^{K_{\mathrm{L}}} I(\hat{\theta} \mid Y_{\mathrm{o}}, Y_{\mathrm{m}}^{(k)}) - \frac{1}{K_{\mathrm{L}} - 1} \sum_{k=1}^{K_{\mathrm{L}}} \{U(\hat{\theta} \mid Y_{\mathrm{o}}, Y_{\mathrm{m}}^{(k)}) - \bar{U}(\hat{\theta} \mid Y_{\mathrm{o}}, Y_{\mathrm{m}}^{(\cdot)})\} \times \{U(\hat{\theta} \mid Y_{\mathrm{o}}, Y_{\mathrm{m}}^{(k)}) - \bar{U}(\hat{\theta} \mid Y_{\mathrm{o}}, Y_{\mathrm{m}}^{(\cdot)})\}^{\mathrm{T}} \tag{2}$$

for

$$\bar{U}(\hat{\theta} \mid Y_{\mathrm{o}}, Y_{\mathrm{m}}^{(\cdot)}) = \frac{1}{K_{\mathrm{L}}} \sum_{k=1}^{K_{\mathrm{L}}} U(\hat{\theta} \mid Y_{\mathrm{o}}, Y_{\mathrm{m}}^{(k)})$$

The variance-covariance matrix of $\hat{\theta}$ is then estimated from the inverse of this, $V_{\mathrm{L}} = I_{\mathrm{L}}^{-1}$.

The multiple imputation variance formula of Rubin and Schenker [21] (see also references [22] and [25]) is also based on Monte Carlo simulation from the conditional distribution of Y_{m} given Y_{o}, but operates rather on the set of estimates obtained from the solutions of the pseudo-complete scores: $U(\hat{\theta} \mid Y_{\mathrm{o}}, Y_{\mathrm{m}}^{(k)})$, $k = 1, \ldots, K_{\mathrm{M}}$. The Bayesian framework is natural when considering multiple imputation, but to maintain a comparative base for Louis' formula above we use large sample arguments and assume a frequentist justification. The main difference is that the same value of $\hat{\theta}$ is used to generate each of the imputed data sets, in contrast to the Bayesian setting in which a new value is drawn for each imputation from the posterior distribution of θ. Let $\theta^{(k)}$ be the maximum likelihood estimate calculated from the kth pseudo-complete data set $\{Y_{\mathrm{o}}, Y_{\mathrm{m}}^{(k)}\}$, $k = 1, \ldots, K_{\mathrm{M}}$. The multiple imputation (MI) estimator of the covariance matrix of $\hat{\theta}$ is

$$V_{\mathrm{MI}} = \frac{1}{K_{\mathrm{M}}} \sum_{k=1}^{K_{\mathrm{M}}} I(\hat{\theta} \mid Y_{\mathrm{o}}, Y_{\mathrm{m}}^{(k)})^{-1} + \frac{K_{\mathrm{M}} + 1}{K_{\mathrm{M}}} \sum_{k=1}^{K_{\mathrm{M}}} (\hat{\theta}^{(k)} - \bar{\theta})(\hat{\theta}^{(k)} - \bar{\theta})^{\mathrm{T}} \tag{3}$$

for

$$\bar{\theta} = \frac{1}{K_{\mathrm{M}}} \sum_{k=1}^{K_{\mathrm{M}}} \hat{\theta}^{(k)}$$

Note that this estimate of the covariance matrix of $\hat{\theta}$ does not apply in as much generality as that provided by Louis' formula (2), but typically, when valid, it requires substantially smaller Monto Carlo sample sizes. For a thorough discussion of the theory underpinning the validity of MI, and related stochastic imputation procedures, we refer the reader to references [26, 27]. The factor $(K_{\mathrm{M}} + 1)/K_{\mathrm{M}}$ is intended to improve the behaviour of the covariance estimator for small values of K_{M}. To add some insight into the relationship between the Louis and MI methods, a formal small-sample connection between them is derived in the Appendix for the special case of the multivariate normal linear model with known covariance matrix.

3. THE ALPHA TRIAL EXAMPLE: BRIEF DESCRIPTION OF THE JOINT MULTIVARIATE RANDOM EFFECTS MODEL

3.1. The Alpha trial CD4 cells count data

The Alpha trial is a multinational multi-centre randomized double-blind clinical trial carried out to compare the efficacy and toxicity of low (L) and high (H) dose of didanosine (ddl)

Copyright © 2003 John Wiley & Sons, Ltd.

in HIV-1 positive individuals intolerant to zidovudine (AZT). The results of the trial showed no significant difference in survival between the two treatment groups [28]. The number of CD4 cells, one of the most important immunological markers of HIV disease progression, was determined at randomization, weeks 4, 8 and every 8 weeks thereafter. For the purpose of this analysis, all subjects with a baseline and at least one CD4 measurement available during the follow-up were included in the analysis (762 in L and 803 in H). The majority of patients entered with a very low CD4 count (66 per cent with count 50 cells/μl) and had advanced disease (60 per cent with AIDS and/or HIV-encephalopathy). We estimate and compare the rate of change of CD4 cell count between the two treatment groups. It should be noted that by the end of study 70 per cent of the subjects died, a substantial number of deaths being early (205 by week 16).

3.2. The joint multivariate random effects (JMRE) model

To allow for possible non-random drop-out associated with death, the JMRE model [14] was used to analyse the data. Briefly, the method combines a linear random effects model for the underlying trajectory of the log transformed CD4 cell counts with a log-normal survival time model for death. The underlying concept is that a subject's probability of death or disease progression is related to that subject's underlying pattern of change in CD4 counts. Such a model is a reasonable one when the drop-out is mainly attributable to death or disease progression. Similar models using either log-normal or proportional hazards model for the survival data have been extensively used in AIDS epidemiology [29–32]. Denote by Y_i the set of n_i log CD4 counts from the ith subject and by W_i the corresponding log survival time if observed or log censoring time if not. Let Y_i^{J} be the combined vector of observations $\{Y_i^{\mathrm{T}}, W_i\}^{\mathrm{T}}$. The combined observations are modelled using a conventional linear mixed model [33]

$$Y_i^{\mathrm{J}} \mid b_i \sim \mathrm{N}(X_i\alpha + Z_i b_i; \sigma^2 I) \tag{4}$$

and

$$b_i \sim \mathrm{N}(0, G)$$

The b_i represent random regression coefficients associated with the ith subject, and X_i and Z_i are design matrices associated, respectively, with the fixed effects a and random effects b_i. In this setting

$$G = \left[\begin{pmatrix} \Sigma_c & \sigma_{ct} \\ \sigma_{ct} & \sigma_t^2 \end{pmatrix}\right]$$

where matrix Σ_c includes variances and covariances of the random effects for CD4 count data, vector σ_{ct} covariances of CD4 random effects and survival residuals and σ_t^2 is the variance of the survival residuals. The explanatory variables associated with the disease marker take the value 0 for the corresponding survival part of the response variable while the reverse is true for the explanatory variables associated with survival, that is, separate fixed effects models apply to the disease marker and survival outcomes. The construction of these variables is described in detail elsewhere [14]. To fit the model, allowing also for censored survival times, a form of EM algorithm [15] is adopted, considering censored survival times as missing data.

Copyright © 2003 John Wiley & Sons, Ltd.

At each iteration both the survival part of the response variable and the survival component of the residual cross-products are replaced for the censored observations by their conditional expectations given the observed data and the current parameter estimates (E-step), and then new REML parameter estimates are obtained (M-step). As it has been described in detail in reference [14], in the JMRE model, instead of the full maximization of the joint likelihood of the response and the drop-out (survival) process, estimators are obtained via the RIGLS method modified appropriately to incorporate censored survival data. Therefore, the method is less computationally intensive and with no serious convergence problems, while it can be fitted using existing software for RE models, such as MLn, after the appropriate modifications.

The results obtained from fitting the JMRE model to the data from the Alpha trial have already been reported [14]. In summary, a linear random effects model was used to model log CD4 cells count in the low (L) group and a piecewise linear (with change in slope at 8 weeks) in the high (H) group, to reflect an initial rise observed in this treatment group. The baseline CD4 cell count varied significantly with age, sex and HIV stage (defined as AIDS or no AIDS at entry) but rate of CD4 changes were only related to HIV stage for the H group. On average, baseline CD4 count was higher in females, in older patients and in individuals without AIDS at entry. High dose ddI (group H) was associated with a short term rise in CD4 cell count followed by a steady decline. The ultimate rate of CD4 decline (after the 8th week) in group H was significantly larger than that in group L for both patients without and with AIDS at entry, indicating that the initial superiority of H is not maintained over time. Log survival time was modelled as a function of baseline HIV stage and age at randomization, but did not differ significantly by treatment group or sex. The estimated median survival time for subjects of mean age (38 years) was 21.5 and 10.5 months for patients without and with AIDS at entry, respectively. Survival time was inversely associated with age at randomization. Patient's survival times were significantly and positively correlated with patient's baseline CD4 cell count, subject-specific rate of CD4 drop in group L and subject-specific initial rise and subsequent rate of CD4 decline after the 8th week in group H (estimated correlation coefficients 0.39, 0.42, 0.37 and 0.46, respectively) indicating that longer survival is associated with higher baseline CD4 cell count, greater initial rise (H group) and less steep decline in CD4 cell count.

4. ESTIMATION OF THE COVARIANCE MATRIX FOR THE PARAMETER ESTIMATES OF THE JMRE MODEL IN THE ALPHA TRIAL

The JMRE model is an example of a normal linear mixed model. Let

$$Y_i^{\mathrm{J}} \sim \mathrm{N}(X_i\alpha; V_i), \qquad i = 1, \ldots, m$$

independently, where

$$V_i = Z_i G Z_i^{\mathrm{T}} + \sigma^2 I$$

Then, for given V_i and dropping the superscript J from Y_i^{J} for simplicity, the score for the kth pseudo-complete data set is

$$U_k(\alpha) = \sum_{i=1}^{m} X_i^{\mathrm{T}} V_i^{-1} (Y_i^{(k)} - X_i\alpha)$$

Copyright © 2003 John Wiley & Sons, Ltd.

The empirical mean and covariance matrix of these are

$$\bar{U}(\alpha) = \frac{1}{K}\sum_{k=1}^{K} U_k(\alpha) = \sum_{i=1}^{m} X_i^{\mathrm{T}} V_i^{-1}(\bar{Y}_i - X_i\alpha)$$

and

$$C = \sum_{i=1}^{m}\sum_{j=1}^{m} X_i^{\mathrm{T}} V_i^{-1} A_{ij} V_j^{-1} X_j$$

respectively, for

$$A_{ij} = \frac{1}{K-1}\sum_{k=1}^{K} (Y_i^{(k)} - \bar{Y}_i)(Y_{jk} - \bar{Y}_j)^{\mathrm{T}}$$

the sample covariance matrix of the $Y_i^{(k)}$s. This implies that Louis formula reduces in this case to

$$I_{\mathrm{L}} = \sum_{i=1}^{m} X_i^{\mathrm{T}} V_i^{-1} X_i - \sum_{i=1}^{m}\sum_{j=1}^{m} X_i^{\mathrm{T}} V_i^{-1} A_{ij} V_j^{-1} X_j$$

where the first term in the above formula gives the estimated covariance matrix obtained from the complete data information matrix at convergence of the RIGLS algorithm. In this setting the missing data are missing log survival times, and the random draws of these will be made with model parameters fixed at their maximum likelihood estimates. For each missing log survival time, W_i, we need to simulate a value conditional on this subject's observed disease markers Y_i, and conditional on this value exceeding this subject's censoring time C_i. The first conditional distribution (ignoring the second condition) is obviously itself normal with mean and standard deviation (say η_i and ξ_i, respectively) given by standard regression formulae, and the latter then requires Tobit type calculations: for R_i a standard uniform random variable, the random variable $\tilde{W}_i$, has the required conditional distribution where

$$\tilde{W}_i = \eta_i + \xi_i \Phi^{-1}\left[\Phi\left(\frac{\ln C_i - \eta_i}{\xi_i}\right) + R_i\left\{1 - \Phi\left(\frac{\ln C_i - \eta_i}{\xi_i}\right)\right\}\right]$$

and $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Table I shows the estimates of the SEs of the fixed-effects parameter estimates using the Louis method with $K_{\mathrm{L}} = 100$. The table also shows the percentage increase in these variances compared with the naıve variances obtained from the completed data information matrix at the end of the JMRE estimation procedure. The increase in the variances of the estimates of the disease marker model (that is, CD4 cell count model) is minimal, whereas the increase in the variances of the estimates of the log-survival model is substantial. This is expected because the missing data are confined to the survival times.

For the MI method, because the covariance matrix of the data needs estimation, we need to modify the basic formula given in (3). The average of the parameter estimates ($\bar{\theta}$) from the imputed data sets are consistent only for the fixed effects (α), not for the covariance parameters (G, σ^2), because these are not linear functions of the data in the likelihood [10, 14]. To accommodate this, two changes have been made to the MI formula (3). First, the within

Copyright © 2003 John Wiley & Sons, Ltd.

Table I. Estimates of the SEs of the fixed-effects parameters by the joint multivariate random effects (JMRE) and by the Louis method.

Parameter	SE			
	Estimate	JMRE	Louis	Per cent difference*
CD4 cell count model				
Intercept terms				
Constant (C)	**3.7970**	0.0995	0.1019	4.75
Sex (M/F)	**0.2757**	0.0789	0.0813	5.99
Age-38 (years)	**0.0112**	0.0031	0.0031	0.15
HIV stage (AIDS/non-AIDS)	**−1.1030**	0.0581	0.0581	0.03
Slope terms				
Low dose group	**−0.2388**	0.0103	0.0105	3.40
High dose group: non-AIDS at entry				
Initial	**0.1791**	0.0648	0.0653	1.76
After week 8	**−0.2757**	0.0146	0.0150	5.21
High dose group: AIDS at entry				
Initial	**−0.1222**	0.0564	0.0565	0.31
After week 8	**−0.2782**	0.0153	0.0155	2.26
Survival model				
Intercept	**6.4810**	0.0358	0.0417	35.69
HIV stage (AIDS/non-AIDS)	**−0.7196**	0.0455	0.0508	24.36
Age-38 (years)	**−0.0093**	0.0023	0.0025	19.39

*Per cent difference: 100[var(Louis method)-var(JMRE method)]/var(JMRE method)

imputation term

$$W_{\mathrm{MI}} = \frac{1}{K}\sum_{k=1}^{K} I(\hat{\theta} \mid Y_{\mathrm{o}}, Y_{\mathrm{m}}^{(k)})^{-1}$$

(that is, the first term in the right-hand side of (3)) has been replaced by the covariance matrix obtained from the complete data information matrix at convergence of the RIGLS algorithm, that is by

$$\sum_{i=1}^{m} X_i^{\mathrm{T}} V_i^{-1} X_i$$

Second, the fixed effects part of the between-imputation covariance term

$$B_{\mathrm{MI}} = \frac{K+1}{K}\sum_{k=1}^{K} (\hat{\theta}^{(k)} - \bar{\theta})(\hat{\theta}^{(k)} - \bar{\theta})^{\mathrm{T}}$$

(that is, the second term in the right-hand side of (3)) has been calculated about the maximum likelihood estimator of the fixed effects, that is, $\hat{\alpha}$

$$B_{\mathrm{MI}}^{\mathrm{mod}} = \frac{K+1}{K}\sum_{k=1}^{K} (\hat{\alpha}^{(k)} - \hat{\alpha})(\hat{\alpha}^{(k)} - \hat{\alpha})^{\mathrm{T}}$$

Following the suggestion in Rubin [22] the number of imputations was set at $K_{\mathrm{M}} = 5$. The results corresponding to those for the Louis approach are given in Table II. These are broadly

Copyright © 2003 John Wiley & Sons, Ltd.

Table II. Estimates of the SEs of the fixed-effects parameters using modified multiple imputation method with five imputations.

Parameter	Pooled estimate	SE	Per cent difference from JMRE*	Per cent difference from Louis†
CD4 cell count model				
Intercept terms				
Constant (C)	**3.7980**	0.1009	2.74	1.95
Sex (M/F)	**0.2748**	0.0804	3.65	2.20
Age-38 (years)	**0.0112**	0.0031	0.11	0.00
HIV stage (AIDS/non-AIDS)	**−1.1028**	0.0581	0.03	0.00
Slope terms				
Low dose group	**−0.2411**	0.0110	14.29	−9.75
High dose group: non-AIDS at entry				
Initial	**0.1815**	0.0657	3.10	−1.23
After week 8	**−0.2785**	0.0155	12.27	−6.78
High dose group: AIDS at entry				
Initial	**−0.1187**	0.0567	1.17	−0.71
After week 8	**−0.2826**	0.0184	44.07	−40.92
Survival model				
Intercept	**6.4940**	0.0401	25.54	7.53
HIV stage (AIDS/non-AIDS)	**−0.7256**	0.0481	11.48	10.35
Age-38 (years)	**−0.0091**	0.0024	9.18	7.84

*Per cent difference: 100[var(MI method)-var(JMRE method)]/var(JMRE method).
†Per cent difference: 100[var(Louis method)-var(MI method)]/var(Louis method).

consistent with those from the Louis method, but the modified MI approach does show a tendency to overestimate some of the SEs associated with the fixed effects estimators of the CD4 model.

Although Rubin [22] suggests that five imputations should be sufficient in most settings to examine whether the MI results are sensitive to the number of imputations, we repeated the procedure with $K_M = 10$. The results are set out in Table III and percentage differences are given between these variance estimates and those from Louis and for those from the MI approach with $K_M = 5$. The only non-negligible difference between the two imputation sizes occurs for the +8 week slope term in the high dose group, the variance estimate that differs notably also between the Louis and MI procedures. Differences from the Louis variance estimates are also noticeably larger (in the region of 10 per cent) for the three parameters from the survival model where the missing values occur.

5. A SIMULATION STUDY

We have seen that, with some minor exceptions, the Louis and modified MI method produce similar estimates of precision for the JMRE model used in the analysis of the Alpha clinical trial data. To complement these comparative results, we use a simulation study to examine the performance of the methods in longitudinal studies where the main focus is to estimate the rate of change of a disease marker and in which some subjects drop out prematurely (informatively) due to attrition, while others experience a non-informative drop-out process

Copyright © 2003 John Wiley & Sons, Ltd.

Table III. Estimates of the SEs of the fixed-effects parameters using modified multiple imputation method with ten imputations.

Parameter	Pooled estimate	SE	Per cent difference from MI(5)*	Per cent difference from Louis†
CD4 cell count model				
Intercept terms				
Constant (C)	**3.8012**	0.1013	0.79	1.09
Sex (M/F)	**0.2722**	0.0806	0.50	1.63
Age-38 (years)	**0.0112**	0.0031	0.00	0.03
HIV stage (AIDS/non-AIDS)	**−1.1031**	0.0581	0.00	0.13
Slope terms				
Low dose group	**−0.2415**	0.0110	0.00	−9.90
High dose-group: non-AIDS at entry				
Initial	**0.1793**	0.0654	−0.91	−0.27
After week 8	**−0.2784**	0.0155	0.00	−7.78
High dose group: AIDS at entry				
Initial	**−0.1200**	0.0566	−0.35	−0.58
After week 8	**−0.2814**	0.0177	−7.46	−30.33
Survival model				
Intercept	**6.4944**	0.0396	−2.48	9.50
HIV stage (AIDS/non-AIDS)	**−0.7269**	0.0478	−1.24	11.35
Age-38 (years)	**−0.0091**	0.0024	0.00	12.45

*Per cent difference: 100[var(MI(10) method)-var(MI(5) method)]/var(MI(10) method).
†Per cent difference: 100[var(Louis method)-var(MI method)]/var(Louis method).

(end of the study, withdrawal). A total of 1000 data sets were generated, each containing 200 subjects (100 in placebo and 100 in treatment group), the disease marker (say CD4 cell count) being measured every month, with a study duration of 2 years. The population average log CD4 cell count was assumed to decline linearly over time. Drop-out was assumed to be due to death and loss to follow-up. Subject-specific baseline values and slopes were assumed to be correlated with survival time, and these three quantities (intercept, slope and log survival time) were generated from the trivariate normal distribution. The underlying slope and intercept were assumed to be uncorrelated while the correlation coefficient of survival time with subject's baseline value and slope were assumed to be 0.36 and 0.40, respectively. The baseline median CD4 cell count was set to 200 cells/μl (5.2983 on log scale) for both treatment groups. A treatment difference in the marker rate of change between placebo and treatment group was assumed: the rate of marker change on a log scale was −0.0346 and −0.0229 per month in placebo and treatment group, respectively, corresponding to a 34 per cent and 24 per cent drop per year in CD4 cell count. In addition, a difference in the survival time, with mean log-survival being 2.89 (GM = 18 months) and 3.18 (GM = 24 months), respectively, was assumed. Censored survival times due to patients' withdrawal were assumed to follow the exponential distribution with a mean rate of censoring of 5 per cent per year. Thus, censoring times were generated using the relationship

$$C_i \simeq -\frac{\ln(U)}{0.0042}$$

Copyright © 2003 John Wiley & Sons, Ltd.

Table IV. Summary of results from the simulation study (second row: percentage differences between the variance estimates and the naıve estimate*).

Parameter	Average estimates		Empirical variance	Average estimated variance ($\times 10^4$)		
	REML	MI(10)	($\times 10^4$)	Naıve	Louis	MI(10)
CD4 model						
Baseline	5.301	5.301	58.45	59.08	59.09	59.24
			−1.08		0.01	0.27
Drop/month	−0.023	−0.023	0.454	0.452	0.454	0.466
			0.85		0.40	2.95
Treatment difference in drop	−0.011	−0.011	0.927	0.935	0.939	0.948
			−0.87		0.38	1.33
Survival model						
Constant	3.162	3.176	68.18	51.96	67.67	68.11
			31.22		30.22	31.08
Treatment difference	−0.280	−0.284	107.84	96.23	102.01	116.78
			12.06		24.81	21.36

*Per cent difference: 100[var-var(naıve)]/var.

for $U \sim U(0.1)$. All subjects still alive were censored at the end of the study. Censored survival times due to withdrawal and study termination were assumed to be non-informative. The JMRE model was fitted to each of the 1000 simulated data sets. For each simulation the Louis method was used to estimate the covariance matrix of the parameter estimates, as described earlier. The modified MI method was also used for each simulation, with K_M set equal to 10. Table IV summarizes the results, displaying the mean of the 1000 parameter estimators, the mean of the 1000 MI pooled parameter estimators, the mean of the estimators' variances as estimated naively from the completed data information matrix, as well as by the Louis and modified MI methods. The table also contains the empirical estimates of the sampling variances obtained over the 1000 simulations. In Table IV we also present the percentage of the increase in the variances obtained by the Louis and modified MI methods compared to the naıve variances obtained by the JMREM method. The empirical variances obtained over the 1000 replications are also shown. The behaviour of the Louis and MI methods can be seen to be similar, and reasonable approximations to the empirical variances. In contrast, the naıve variance estimates seriously underestimate the empirical values for the survival component of the model. This effect was more marked for the estimator of the average log-survival time in the treatment group than for the corresponding between-group difference. This was a consequence of the shorter survival times in the placebo group leading to less censoring; the percentages of subjects with censored survival time was 53 and 38 in the treatment and in the placebo group, respectively. Note that, for the disease marker parameter estimators, the estimated variances produced by both methods were similar to the empirical variances.

6. DISCUSSION

It is well known that, in the presence of missing data, the asymptotic variance-covariance matrix of the MLE estimators is not given by the usual inverse of the completed data information

Copyright © 2003 John Wiley & Sons, Ltd.

matrix, even when the non-response is ignorable. Estimates of precision should be based on the information for the observed incomplete data. Often, however, this can be a awkward to calculate directly and methods that allow its computation, or approximation, through complete data quantities provide useful alternatives. Louis [20] has provided a general technique for computing the observed data information in terms of completed data quantities, while the multiple imputation formula for combining within- and between-imputation variances to obtain variances directly gives an alternative, though related, route [21–23]. In this paper, we have applied the Louis method and a modified version of the MI approach to longitudinal data of a disease marker, where the source of missing data was either death or disease progression (informative) or end of the study or loss to follow-up (assumed non-informative). This model combines a linear random effects model for the underlying pattern of change in the disease marker with a log-survival model for the informative drop-out process caused by a patient's death. A nested EM algorithm was applied to deal with censored survival data. Compared with the Louis method, the SEs of the parameters of the log-survival model estimated by the completed data information matrix were seriously underestimated while those of the disease marker model were only slightly affected. The latter were almost identical to the Louis corrected SEs. This was expected, because the missing data were censored survival data. It should also be noted that fixed effects estimates of the disease marker model, when RE models are applied, are weighted estimates, with weights inversely proportional to each subject's covariance matrix, thus, a function of the individual follow-up time, and therefore adjusted for the different number of observations per subject. In the analysis of such data, when the main interest is focused on the disease marker model, the extra effort in adjusting the SEs may not be worthwhile. If, on the other hand, the interest is on both disease marker and survival components of the model, adjustment of SEs for missing survival data is necessary.

The standard MI method of imputing missing survival data, in which each REML estimate from a completed data set are averaged and the covariance matrix for the resulting estimate obtained through a combination of between- and within-imputation covariance matrices, is not appropriate when there are parameters estimated whose sufficient statistics are non-linear functions of the data [10, 14]. In this setting this would result in biased estimation (over the multiple imputations) of the variance-covariance matrix, that is, the random effect variance and covariance parameters. Consequently a modified version of the MI approach has been used in which the between-imputation variance-covariance is estimated as the mean square error matrix of the pseudo-complete data estimators about the REML estimates, and the within-imputation covariance matrix is replaced by the matrix obtained by the JMREM at convergence. The Louis and the modified MI methods gave broadly similar results. Given the relative simplicity of the modified MI method, it may be preferable. Further simulated studies to determine the sensitivity of the multiple imputation method to the number of imputed data sets under different scenario of missing data structure as well as of the degree of missing data are needed, although our results indicated that five to ten imputed data sets should be enough in the setting of the paper.

In conclusion, in the analysis of longitudinal studies with missing data, adjustment of the parameters SEs is needed if the method of estimation is based on some form of imputation or completion of the missing data. However, when the model for the disease marker allows for different observations per subject, it seems that the SEs that are mostly underestimated are the SEs of the non-response parameters. Therefore, when the main interest is focused on the disease marker model, the extra effort to adjust SEs may not be worthwhile. While

Copyright © 2003 John Wiley & Sons, Ltd.

the Louis method provides a general method to adjust SEs for missing data, in longitudinal studies where the source of missing data is either death or disease progression and end of study, given its relative simplicity, the modified MI method considered in this paper may be a reasonable alternative.

APPENDIX: THE LINK BETWEEN THE LOUIS AND MULTIPLE IMPUTATION FORMULAE FOR THE MULTIVARIATE NORMAL SETTING

The similarity of (2) and (3) is suggestive. Denoting the identity matrix by I, the essential link between these approaches is to be found in the relationship between $I-M$ and its inverse for any symmetric square matrix M with all eigenvalues lying between -1 and 1

$$\begin{aligned}(I+M)^{-1} &= I + \sum_{r=1}^{\infty} (-M)^r \\ &\simeq I - M \end{aligned} \tag{A1}$$

To illustrate this we take the MI expression (3) and invert it to provide a comparison with I_{L}. Provided the left hand term

$$W_{\mathrm{MI}} = \frac{1}{K} \sum_{k=1}^{K} I(\hat{\theta} \mid Y_{\mathrm{o}}, Y_{\mathrm{m}}^{(k)})^{-1} \tag{A2}$$

'dominates' the right hand term

$$B_{\mathrm{MI}} = \frac{K+1}{K} \sum_{k=1}^{K} (\hat{\theta}^{(k)} - \bar{\theta})(\hat{\theta}^{(k)} - \bar{\theta})^{\mathrm{T}} \tag{A3}$$

which will typically be the case because the information associated with the missing data will be a comparatively small compared with the information in the observed data, then using (A1)

$$V_{\mathrm{MI}}^{-1} = (W_{\mathrm{MI}} + B_{\mathrm{MI}})^{-1} = \{W_{\mathrm{MI}}(I + W_{\mathrm{MI}}^{-1} B_{\mathrm{MI}})\}^{-1} \simeq W_{\mathrm{MI}}^{-1} - W_{\mathrm{MI}}^{-1} B_{\mathrm{MI}} W_{\mathrm{MI}}^{-1}$$

The first term on the right of this expression approaches the complete data information

$$\sum_{i=1}^{m} X_i^{\mathrm{T}} V_i^{-1} X_i$$

the first term in the Louis formula. For a given pseudo-complete data set, say the kth

$$\hat{\theta}_k = \left(\sum_{i=1}^{m} X_i^{\mathrm{T}} V_i^{-1} X_i \right)^{-1} \sum_{i=1}^{m} X_i^{\mathrm{T}} V_i^{-1} Y_{k,i}$$

Approximately, for a value of $\hat{\theta}_0$ (estimated from some complete data set Y_0) close to $\hat{\theta}_k$

$$\hat{\theta}_k \simeq \theta_0 + (Y_k - Y_0)^{\mathrm{T}} \left. \frac{\partial \theta}{\partial Y} \right|_{Y_0}$$

Copyright © 2003 John Wiley & Sons, Ltd.

or

$$(\hat{\theta}_k - \hat{\theta}_0)(\hat{\theta}_k - \hat{\theta}_0)^{\mathrm{T}} \simeq \frac{\partial \theta^{\mathrm{T}}}{\partial Y}(Y_k - Y_0)(Y_k - Y_0)^{\mathrm{T}} \left.\frac{\partial \theta}{\partial Y}\right|_{Y_0}$$

However

$$\frac{\partial \theta}{\partial Y} = (X^{\mathrm{T}} V^{-1} X)^{-1} X^{\mathrm{T}} V^{-1}$$

or

$$\frac{\partial \theta^{\mathrm{T}}}{\partial Y}(Y_k - Y_0) = \left(\sum_{i=1}^{m} X_i^{\mathrm{T}} V_i^{-1} X_i\right)^{-1} \sum_{i=1}^{m} X_i^{\mathrm{T}} V_i^{-1}(Y_{k,i} - Y_{k,0})$$

Finally, because $\hat{\theta}$ is linear in Y, for given V

$$\bar{\theta} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k = \theta(\bar{Y})$$

for

$$\bar{Y} = \frac{1}{K} \sum_{k=1}^{K} Y_k$$

Using this, and equating Y_0 and $\bar{Y}$, we have

$$\begin{aligned}
\bar{W}_{\mathrm{MI}}^{-1} B_{\mathrm{MI}} \bar{W}_{\mathrm{MI}}^{-1} &\simeq \frac{1}{K-1}(X^{\mathrm{T}} V^{-1} X)(\theta_k - \bar{\theta})(\theta_k - \bar{\theta})^{\mathrm{T}}(X^{\mathrm{T}} V^{-1} X) \\
&\simeq \frac{1}{K-1}(X^{\mathrm{T}} V^{-1} X) \sum_{k=1}^{K} \left[\left(\sum_{i=1}^{m} X_i^{\mathrm{T}} V_i^{-1} X_i \right)^{-1} \right. \\
&\quad \times \left\{ \sum_{i=1}^{m} \sum_{j=1}^{m} X_i^{\mathrm{T}} V_i^{-1}(Y_{k,i} - Y_i)(Y_{k,j} - Y_j)^{\mathrm{T}} V_j^{-1} X_j^{\mathrm{T}} \right\} \\
&\quad \left. \times \left(\sum_{i=1}^{m} X_i^{\mathrm{T}} V_i^{-1} X_i \right)^{-1} \right] (X^{\mathrm{T}} V^{-1} X) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} X_i^{\mathrm{T}} V_i^{-1} \left\{ \frac{1}{K-1} \sum_{l=1}^{K} (Y_{k,i} - Y_i)(Y_{k,j} - Y_j)^{\mathrm{T}} \right\} V_j^{-1} X_j^{\mathrm{T}}
\end{aligned}$$

and this is the second term on the right hand side expression for I_{L} in (2). Note that this argument assumes that the covariance terms (V_i) are known.

ACKNOWLEDGEMENTS

The authors are most grateful to the Alpha International co-ordinating Committee for providing the data of the Alpha clinical trial. This research was partly supported through a scholarship award by the Greek State Scholarship Foundation to Giota Touloumi.

Copyright © 2003 John Wiley & Sons, Ltd.

REFERENCES

1. Little RJA, Rubin DB. *Statistical Analysis with Missing Data.* Wiley: Chichester, 1986.
2. Hogan JW, Laird NM. Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine* 1997.
3. Kenward MG, Molenberghs G. Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research* 1999; **8**:51–83.
4. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**:581–592.
5. Laird NM. Missing data in longitudinal studies. *Statistics in Medicine* 1988; **7**:305–315.
6. Wu MC, Bailey KR. Analysing changes in the presence of informative right censoring by modelling the censoring process. *Statistics in Medicine* 1988; **7**:175–188.
7. Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics* 1988; **44**:175–188.
8. Schluchter MD, Jackson KL. Log-linear analysis of censored survival data with partially observed covariates. *Journal of the American Statistical Association* 1989; **84**:42–52.
9. Mori M, Woolson RF, Woodsworth GG. Application of empirical Bayes inference to estimation of rate of change in the presence of informative right censoring. *Statistics in Medicine* 1992; **11**:621–631.
10. Schluchter MD. Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine* 1992; **11**:1861–1870.
11. Diggle PJ, Kenward MG. Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics* 1994; **43**:49–94.
12. De Gruttola V, Tu XM. Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics* 1994; **50**:1003–1014.
13. Little RJA. Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995; **90**:1112–1121.
14. Touloumi G, Pocock SJ, Babiker AG, Darbyshire JH. Estimation and comparison of rates of change in longitudinal studies with informative dropout. *Statistics in Medicine* 1999; **18**:1215–1233.
15. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1977; **39**:1–38.
16. McLachlan DB, Krishnan T. *The EM Algorithm and Extensions.* Wiley: New York, 1997.
17. Hartley HO, Hocking R. The analysis of incomplete data. *Biometrics* 1971; **27**:7783–7808.
18. Efron B, Hinkley DV. The observed versus expected information. *Biometrika* 1978; **65**:457–487.
19. Meilijson I. A fast improvement of the EM algorithm on its own terms. *Journal of the Royal Statistical Society*, *Series B* 1989; **51**:127–138.
20. Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, *Series B* 1982; **44**:226–233.
21. Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable non response. *Journal of the American Statistical Association* 1986; **81**:366–374.
22. Rubin DB. *Multiple Imputation for Nonresponse in Surveys.* Wiley: Chichester, 1987.
23. Glynn RJ, Laird NM, Rubin DB. Selection modelling versus mixture modelling with nonignorable nonresponse. In *Drawing Inferences from Self-selected Samples*, Wainer H (ed.). Springer Verlag: New York, 1986.
24. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice.* Chapman and Hall: London, 1996; 264–268.
25. Schafer JL. *Analysis of Incomplete Multivariate Data.* Chapman & Hall/CRC: London, 1997.
26. Wang N, Robins JM. Large-sample theory for parametric multiple imputation procedures. *Biometrika* 1998; **85**:935–948.
27. Robins JM, Wang N. Inference for imputation estimators. *Biometrika* 2000; **85**:113–124.
28. Alpha international co-ordinating committee. MRC/ANRS European/Australian randomized double-blind trial of two doses of didanoside in zidovudine-intolerant patients with symptomatic HIV disease. *AIDS* 1996; **10**: 867–880.
29. Pawitan Y, Self S. Modeling disease marker processes in AIDS. *Journal of the American Statistical Association* 1993; **88**:719–726.
30. Tsiatis AA, DeGruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. Application to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* 1995; **90**:27–37.
31. Faucett CL, Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine* 1996; **15**:1663–1685.
32. Bycott P, Taylor J. A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. *Statistics in Medicine* 1998; **17**:2061–2077.
33. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data.* Springer Verlag: New York, 2000.

Copyright © 2003 John Wiley & Sons, Ltd.