Correction of Bias from Non-random Missing Longitudinal Data Using Auxiliary Information

Cuiling Wang[1], Charles B. Hall[1]
[1]Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY 10461

Abstract

Missing data is common in longitudinal studies due to drop-out, loss to follow-up, death, etc. The likelihood-based mixed effects model for longitudinal data gives valid estimates when the data are ignorably missing, that is, the parameters for the missing data process are separate from that of the main model for the outcome and the data are missing at random (MAR), i.e., the missing data process can depend on the observed data but not on the unobserved data, an assumption that is not testable without further information. There are occasions when additional information, an auxiliary variable, known to be correlated with the outcome of interest, is available when the outcome of interest is missing. Availability of such auxiliary information provides us an opportunity to test the MAR assumption. If the MAR assumption is violated, such information can be utilized to reduce or eliminate bias when the missing data process depends on the unobserved outcome through the auxiliary information and the observed outcome. We apply and compare two methods of utilizing the auxiliary information, joint modeling of the outcome of interest and the auxiliary variable, and multiple imputation. Even when the missing data process further depends on the unobserved outcome through other factors, the methods considered might reduce the bias comparing to the naive analysis. Cautions in applying these methods are also discussed.

Key Words: Missing data, linear mixed effects model, MNAR, joint modeling, multiple imputation (MI), auxiliary variable MAR (A-MAR)

1. Introduction

Longitudinal studies are widely used in epidemiological research to study the pattern of change of certain outcomes denoted by Y. The linear mixed effects model is one of the most popular statistical methods used for analyzing longitudinal data (Laird and Ware, 1982). For simplicity we only consider continuous outcome of interest in this paper. But the methods can be easily applied to discrete outcomes using the generalized linear mixed effects model. For continuous Y using i as the index for subject, the following linear mixed effects model (Laird and Ware, 1982) is often used:

$$Y_i = X_i\beta + W_i b_i + \varepsilon_i,$$

where Y_i is the vector of outcomes for subject *i,* X_i is the design matrix for the fixed effects, b_i is avector of random effects, W_i is the design matrix for the random effects, usually a subset of X_i, and ε_i is the random error. The parameter vector β for the fixed effects is often of primary interest.

In the presence of missing data, we denote the outcome of interest Y into two parts, $Y = (Y^o, Y^m)$, with Y^o and Y^m denoting the observed and missing part, respectively. Let R denotes the vector of the observation indicator of Y. According to Rubin (1976) and Little and Rubin (2002), three missing mechanisms are defined as follows:

- Missing completely at random (MCAR) if $R \perp (Y^o, Y^m)$
- Missing at random (MAR) if $R \perp Y^m \mid Y^o$
- Missing not at random (MNAR) if R depends on $Y^m \mid Y^o$.

The likelihood for the observed data (Y^o, R) is

$$f(y^o, r) = \int f(y^o, y^m) f(r \mid y^o, y^m) dy^m .$$

When (a) The missing mechanism is MAR, that is,

$$P(R = 1 \mid Y^o, Y^m) = P(R = 1 \mid Y^o), \quad (1)$$

the part for R can be factored out of the integral so that $f(y^o, r) = f(y^o) f(r \mid y^o)$. Furthermore, when (b) the parameters govern the missing data process, i.e., the distribution of R, and the parameters for the outcome *Y* are disjoint, the part of the likelihood for the missing data process $f(r \mid y^o)$ can be ignored. Thus when (a) and (b) hold, inference for the parameters in the model for Y can be based only on $f(y^o)$. The missing data is called ignorable when both condition (a) and (b) hold, otherwise the missing data process is non-ignorable or informative. Throughout we assume condition (b) always holds so that ignorable means MAR and informative means MNAR.

The likelihood-based mixed effects model is often used to analyze longitudinal data. As shown above, such analysis is built on the crucial assumption that the data is ignorably missing. Without additional information, the MAR assumption is unverifiable. When a violation of MAR is suspected, usually all we can do is either assuming a generally non-identifiable model for the informative missing process and model it together with the outcome, or performing a sensitivity analysis to evaluate to what extent the missing data process affect the results of interest.

Fortunately in some studies there is available additional auxiliary information which is correlated with the outcome of interest. This auxiliary information can be used to test the MAR assumption, and it can be utilized to eliminate or reduce bias if the missing data depend on the unobserved outcome through the auxiliary information.

Denote the auxiliary information as Z, where Z and *Y* are correlated. For simplicity, we assume Z is fully observed. The model is easily extended to the case that Z is also subject to missingness, as will be shown in section 2. Suppose that the missing data mechanism is MNAR, i.e., $P(R = 1 \mid Y^o, Y^m)$ depends on Y^m. Assume further that conditional on Y^o, R depends on Y^m only through Z. Then

$$P(R = 1 \mid Y^o, Y^m, Z) = P(R = 1 \mid Y^o, Z). \quad (2)$$

The missing data assumption (2) is called auxiliary variable MAR (A-MAR) by Daniels and Hogan (2007).

If (2) holds, then $P(R = 1 \mid Y^o, Y^m) = \int P(R = 1 \mid Y^o, Z) P(Z \mid Y^o, Y^m) dZ$. If, conditional on Y^o, Z is correlated with Y^m and R depends on Z, then $P(R = 1 \mid Y^o, Y^m)$ will depend on Y^m so that when *Y* is the only outcome data considered in the analysis, the missing process is not at random. By the definition of the auxiliary variable, the condition that *Y* and Z are correlated holds, thus $P(R = 1 \mid Y^o, Z) = P(R = 1 \mid Y^o)$ becomes a necessary condition for the MAR assumption for Y (and also a sufficient condition if R depends on Y only through Z in addition to Y^o). Under A-MAR condition (2), this is a testable assumption. If the data shows that R depends on Z conditional on Y^o, then MAR assumption is violated for Y. The information on Z can then be utilized to eliminate or reduce the bias in the estimation of the parameter vector of interest β. We consider two intuitive and easily applied methods of utilizing the auxiliary variable Z under A-MAR: joint modeling of the outcome of interest *Y* and the auxiliary variable Z, and multiple imputation of *Y* based on a model that includes *Z*. In section 2 the two methods are described. Results from simulation studies are presented in section 3. In section 4 a data example using a dementia screening study is applied. We conclude the paper with a discussions in section 5.

2. Methods

2.1 Joint modeling of the outcome of interest and the auxiliary variable

Denote $Y^* = (Y, Z)$, with $R^* = (R_Y, R_Z)$ the observation indicator for Y and Z. Denote the observed and missing part of Y^* as $Y^{*o} = (Y^o, Z^o)$ and $Y^{*m} = (Y^m, Z^m)$, respectively. Then the MAR assumption for Y^* is

$$P(R = 1 \mid Y^{*o}, Y^{*m}) = P(R = 1 \mid Y^{*o}). \quad (3)$$

That is, for Y^*, the data are missing at random. In this case, the missing process is ignorable and thus consistent and efficient estimates can be obtained based on the likelihood function of Y^{*o}. This is the A-MAR assumption which reduces to (2) when Z is fully observed.

The A-MAR assumption (3) for the longitudinal outcome $Y^* = (Y, Z)$ is more flexible than the MAR assumption (1) because it allows R_Y depends on Y^m through Z^o conditional on Y^o. Thus by jointly modeling the outcome of interest and the auxiliary variable, the missing data assumption is relaxed from MAR to A-MAR, i.e., it allowed the observation process of the outcome of interest being non-randomly missing in the sense of the model of interest through the auxiliary variable. Ibrahim et al (2001) proposed a similar joint modeling approach in a generalized linear model setting. However, the advantage of this approach is not achieved without a price. By introducing the auxiliary variable as a component of the outcomes studied, assumptions regarding the joint distribution of the outcome of interest Y and the auxiliary variable Z are added to the model assumptions for Y. If the joint distribution of Y and Z is correctly specified, the joint modeling approach should yield the most efficient estimate because it is likelihood based. However, the estimates can be biased, and possibly even inconsistent, if the joint distribution of Y and Z is mis-specified. In practice, the distribution of the outcome of interest and the auxiliary variable should be carefully examined and flexible specification of the joint distribution of Y and Z is recommended.

2.2 Multiple Imputation

From the auxiliary variable MAR assumption (3), Y^{*m} has the same distribution as Y^{*o}, thus it can be consistently imputed from Y^{*o}. Specifically, we replace Y^{*m} randomly with plausible values using Y^{*o} by regression or other techniques. The imputed complete data set will be then used to estimate the parameter of interest. This step will be repeated m times. The results from these m imputed data sets are combined into a single inferential statement using arithmetic rules to yield estimates, standard errors and p-values that formally incorporate missing-data uncertainty to the modeling process. This multiple imputation (MI) technique (Rubin 1987; Schafer 1997) has been widely applied in statistical analysis. The idea of adding auxiliary variables to the imputation procedure in multiple imputation to correct bias, even though the auxiliary variables are not included in the main model of interest, has been proposed before in the literature of multiple imputation (Meng, 1994; Rubin, 1996) and also recently by Collins et al (2001).

Similar to the joint modeling approach, the multiple imputation method assumes A-MAR rather than MAR. The price payed for relaxing the missing data assumption using MI is the introduction of the imputation model. Because the regression approach we used for the imputation model only makes assumption regarding the mean structure of the missing outcome conditional on the observed outcome and the auxiliary variable, it is a much weaker assumption than the one made on the joint distribution in the joint modeling approach. In our simulation studies, the MI approach showed to be less prone to mis-specification than the joint modeling approach.

2.3 Other Methods

Another alternative approach to utilizing the auxiliary variables when the missing data process is A-MAR rather than MAR is to include the auxiliary variable as an additional covariate in the main model for the outcome of interest. This approach was considered by Collins et al (2001). However, this will totally change the meaning of the model, in particular the interpretation of the parameters. The parameters in the new model including the auxiliary variable as covariate will all be interpreted as conditional on the auxiliary variable. If the association of a risk factor with the outcome is of interest and the risk factor examined is associated with the auxiliary variable, then adding the auxiliary variable as covariate would change the magnitude of this association. For this reason, we do not consider this approach in this paper.

Our focus on the main model for the outcome is a linear mixed effects model. Another popular method for longitudinal analysis is the generalized estimating equations (GEE) (Liang and Zeger , 1986) approach. This approach is valid when the data are missing completely at random (MCAR). Robins et al (1995) showed that the inverse probability weighting (IPW) method will give consistent estimates when data are missing at random (MAR). If the auxiliary variables are added to the model for the probability of observation in the IPW approach, then the assumption on missing data is reduced to A-MAR.

3. Simulation Studies

We conducted simulation studies with 1000 replications for longitudinal studies with two visits. A sample size of 500 was used. The outcome of interest $Y=(Y_1,Y_2)$ and the auxiliary variable $Z=(Z_1,Z_2)$ were generated from multivariate normal distributions, with $\mathrm{Var}(Y_j)=\sigma_Y^2$, $\mathrm{Var}(Z_j)=\sigma_z^2$, $\rho=\mathrm{corr}(Z_1,Z_2)=\mathrm{corr}(Y_1,Y_2)$, $r=\mathrm{corr}(Y_j,Z_j)$, $j=1,2.$ The model for the mean of Y is $E(Y_{ij})=\beta_0+\beta_1 t_{ij}$, where t_{ij} is 0 if j=1 and 1 if j=2. The parameter β_1 measures the expected decline on Y at wave 2 compared to wave 1 and is the parameter of interest here. The similar model for the mean of Z is $E(Z_{ij})=\gamma_0+\gamma_1 t_{ij}$. All variables are completely observed except that Y_2, the second measurement of Y, can be missing. The observation indicator R for Y_2 is generated from the following logistic model:

$$\mathrm{logit}\{P(R=1\mid Y,Z)\}=\alpha_0+\alpha_1 Z_1+\alpha_2 Z_2+\alpha_3 Y_1+\alpha_4 Y_2.$$

We set $(\beta_0,\beta_1)=(0,-3)$, $(\gamma_0,\gamma_1)=(0,-2)$, $\sigma_Y^2=9,\sigma_z^2=4$, $\rho=0.5$, $r=0.8$, $(\alpha_0,\alpha_1,\alpha_3)=(1,0,0.3)$. Different scenarios on (α_2,α_4) for the missing data mechanism were considered. For the MAR case, we set $(\alpha_2,\alpha_4)=(0,0)$, i.e., the missing data process does not depend on Y_2 nor Z; for the A-MAR case, we set $(\alpha_2,\alpha_4)=(0.5,0)$ or $(0.1,0)$ so the missing data process does not depend on the unobserved Y_2 but can depend on Z; finally, for the MNAR case, we set $(\alpha_2,\alpha_4)=(0.5,0.1)$ in which the observation probability depends on the unobserved Y_2. Simulation bias (Bias), standard error (STD) and percentage that the 95% confidence intervals cover the true parameter for β_1 (Coverage) are presented in table 1 for the approaches considered under different scenarios.

Table 1. Simulation results comparing joint modeling and multiple imputation approaches with model for Y only

Missing data cases	Method	Bias	STD	Coverage
MAR $(\alpha_2,\alpha_4)=(0,0)$	Model for Y only	0.0003	0.1573	0.942
	Joint modeling of Y and Z	0.0009	0.1307	0.945
	Multiple imputation	0.0002	0.1410	0.956
A-MAR $(\alpha_2,\alpha_4)=(0.1,0)$	Model for Y only	0.1154	0.1589	0.887
	Joint modeling of Y and Z	0.0008	0.1329	0.949
	Multiple imputation	0.0019	0.1422	0.956
A-MAR $(\alpha_2,\alpha_4)=(0.5,0)$	Model for Y only	0.7404	0.1916	0.027
	Joint modeling of Y and Z	0.0001	0.1388	0.948
	Multiple imputation	0.0019	0.1661	0.948
MNAR $(\alpha_2,\alpha_4)=(0.5,0.1)$	Model for Y only	1.0370	0.2057	0.000
	Joint modeling of Y and Z	0.0826	0.1446	0.907
	Multiple imputation	0.1237	0.1779	0.883

Results from table 1 show that both the multivariate longitudinal model and multiple imputation approaches that utilize auxiliary information correct the bias from non-random missing longitudinal data under auxiliary MAR, while the regular model for the outcome Y only yields biased estimates. With the other parameters fixed, the value of α_2 in the observation model measures the extent that the MAR assumption is violated. The more α_2 deviates from 0, the larger the violation. The bias of the estimate using regular mixed effects model for Y increases with the extent that MAR assumption is violated. Under MAR, estimates from the linear mixed effects model for Y gives consistent estimate as well. It is not necessary to utilize auxiliary variable in this circumstance. However, the two approaches utilizing auxiliary information showed some improvement on the efficiency of the parameter estimate. Under MNAR, all methods give biased estimates. However, utilizing auxiliary information reduced the bias.

The above results are based on a correct model specification on the joint distribution of Y and Z in the joint modelling approach and a correct imputation model for the multiple imputation approach. To examine the effect of mis-specification of the joint distribution of Y and Z on the estimate of the parameter of interest, we performed another set of simulations similar to the first one but with $\rho_1 = \text{corr}(Y_1, Y_2) = 0.6$, $\rho_2 = \text{corr}(Z_1, Z_2)$, $\text{corr}(Y_1, Z_2) = \text{corr}(Y_2, Z_1) = r\sqrt{\rho_1 \rho_2}$ where $r = \text{corr}(Y_j, Z_j) = 0.5$. Values of 0.5 and 0.2 were considered for ρ_2 to represent different degrees of difference from ρ_1 . In the mis-specified joint model of Y and Z , we assume a common correlation coefficient $\rho = \rho_1 = \rho_2$. For the missing data process, we focus on A-MAR case so α_4 is set as 0, and values of 0.1 and 0.6 were considered for α_2 as measures of two different levels of deviation from the MAR assumption. Simulation bias (Bias), standard error (STD) and percentage that the 95% confidence intervals cover the true parameter for β_1 (Coverage) are presented in Table 2 using linear mixed effects model for Y , correct and mis-specified joint modeling, and multiple imputation methods.

Table 2. Simulation results on effects of mis-specified joint modeling and mis-specified MI

Parameter settings	Method	Bias	STD	Coverage
$\rho_2 = 0.5, \alpha_2 = 0.1$	Model for Y only	0.0570	0.1434	0.932
	mis-specified joint model for Y and Z	-0.0447	0.1321	0.941
	Correct joint modeling of Y and Z	-0.0036	0.1404	0.950
	Multiple imputation	-0.0049	0.1438	0.946
$\rho_2 = 0.5, \alpha_2 = 0.6$	Model for Y only	0.3803	0.1596	0.325
	mis-specified joint model for Y and Z	-0.1282	0.1365	0.875
	Correct joint modeling of Y and Z	0.0004	0.1525	0.926
	mis-specified joint model for Y and Z	0.0002	0.1605	0.957
$\rho_2 = 0.2, \alpha_2 = 0.1$	Model for Y only	0.0677	0.1476	0.921
	mis-specified joint model for Y and Z	0.1127	0.1397	0.900
	Correct joint modeling of Y and Z	-0.0011	0.1416	0.935
	mis-specified joint model for Y and Z	-0.0022	0.1441	0.955
$\rho_2 = 0.2, \alpha_2 = 0.6$	Model for Y only	0.4514	0.1567	0.177
	mis-specified joint model for Y and Z	0.0383	0.1462	0.960
	Correct joint modeling of Y and Z	0.0019	0.1581	0.920
	mis-specified joint model for Y and Z	0.0003	0.1666	0.950

Table 2 shows that the bias from the mis-specified joint modeling of Y and Z approach increases as the extent of mis-specification of the joint distribution of Y and Z increases. This bias might be bigger or smaller than the bias from linear mixed effects model for the outcome of interest Y depending on how much the MAR assumption is violated.

4. Data Example

In a dementia screening study in a primary care geriatrics practice (Grober et al 2008), decline in memory as measured by the Free and Cued Selective Reminding Test (FCSRT) (Grober and Buschke, 1987) between the follow-up visit and baseline is of interest. We used a subset of the data in which the primary care physicians' assessment of memory in the clinical demential rating system (CDR) was available at both baseline and follow-up (n=238). Baseline FCSRT ranges from 0 to 44 (mean=27.6, std=8.47). The follow-up FCSRT is missing for 59 (25%) subjects. The original CDR rating on memory impairment is graded on a scale of 0-3, with 0=no impairment; 0.5=memory impairment; 1= mild dementia; 2= moderate dementia; and 3=severe dementia. Because of the low prevalence of CDR values of 2 or 3 in this population, we combined them with 1 as a category for overall dementia. Two indicators were defined for the three-category CDR rating: CDRhalf for CDR=0.5 and CDR1P for CDR >=1. The physicians' CDR memory impairment rating is highly associated with FCSRT performance. At baseline, mean (STD) of FCSRT among

the groups with CDR=0, CDR=0.5 and CDR >= 1 are 30.15 (6.54), 23.78 (8.99) and 16.80 (9.97), respectively, with significant difference ($p < 0.0001$) among the CDR categories. Hence CDR memory rating can be used as an auxiliary variable for FCSRT.

We first used CDR memory impairment rating as an auxiliary variable to examine the missing data mechanism. A logistic model for missing the follow-up FCSRT in relation with baseline FCSRT and CDR memory rating at baseline and follow-up was fit and the result was shown in Table 3. It shows that subjects with impaired baseline CDR memory rating are more likely to have missing follow-up FCSRT compared to those with no impairment CDR memory rating at baseline (p=0.016). The likelihood ratio test for testing whether the CDR memory rating can be omitted from the logistic model shows that CDR memory rating is significantly associated with the missing data process adjusting for baseline FCSRT (Chi-square=9.666, degree of freedom=4, p-value=0.046). This suggests that the missing data process might be A-MAR rather than MAR.

Next, we estimate the decline in FCSRT using a linear mixed effects model for FCSRT only and the two methods that utilizing the auxiliary information CDR. The first one is a linear mixed effects model for only FCSRT. The others are the joint modeling and multiple imputation approach utilizing the auxiliary variable CDR. In the joint modeling approach, the multinomially distributed CDR memory rating and the multivariate normally distributed FCSRT were jointly
modeled using correlated random effects as described below.

$$\mathrm{FCSRT}_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + \epsilon_{ij},$$

$$\log\left\{\frac{P(\mathrm{CDR}_{ij} = 0.5)}{P(\mathrm{CDR}_{ij} = 0)}\right\} = \alpha_{01} + \alpha_{11} t_{ij} + b_{1i},$$

$$\log\left\{\frac{P(\mathrm{CDR}_{ij} \geq 1)}{P(\mathrm{CDR}_{ij} = 0)}\right\} = \alpha_{02} + \alpha_{12} t_{ij} + b_{1i},$$

where $i = 1,\ldots,n, j = 1,2,$ are the subject and time index, respectively; t_{ij} is 0 if j =1 and 1 if j =2.

(b_{0i}, b_{1i}) are the subject specific random effects distributed as bivariate normal with mean (0,0), marginal variance (σ_0^2, σ_1^2) and correlation coefficient ρ; ϵ_{ij} is the normally distributed error term for FCSRT which is independent of the random effects. The parameter of interest β_1 represents the decline of FCSRT at follow-up compared to baseline. SAS 9.1 (SAS Institute Inc., Cary, N.C) procedure NLMIXED was used to fit this model.

In the multiple imputation approach, a linear regression model for the observed follow-up FCSRT was fit using baseline FCSRT, baseline and follow-up CDR memory rating. New parameters were randomly drawn from the posterior distribution of the parameters using non-informative prior. The missing follow-up FCSRT was imputed using this new parameters and the baseline FCSRT and CDR memory rating at baseline and follow-up. This process was repeated 5 times. Each of the 5 imputed data sets was then used as a complete data to calculate the FCSRT decline using regular linear mixed effects model. The 5 sets of this parameter estimates were averaged to yield the point MI estimate. The standard errors of each parameter estimate and the variation among the 5 estimates were combined to calculate the variance of the MI estimate (Rubin 1987). SAS 9.1 procedures MI and MIANALYZE were used to obtain the MI estimate.

The results are shown in Table 4. Because subjects with poorer CDR memory rating tend to miss their follow-up visits for FCSRT, the linear mixed effects model which did not take account of the CDR information under-estimated the magnitude of FCSRT decline compared to that from the joint modeling or multiple imputation approach. The model we adopted for the joint modeling of FCSRT and CDR, is a plausible model for joint modeling of a longitudinal continuous variable and a longitudinal categorical variable. More flexible alternative joint models might need to be considered. The multiple imputation makes assumption only on the mean structure of the FCSRT and thus we believe its estimate of FCSRT decline is closer to the true value.

Table 3. Estimates from the logistic model for missing follow-up FCSRT

Effects	Estimates	Standard Error	p-value
Baseline FCSRT	-0.005	0.023	0.811
CDRhalf (baseline)	0.936	0.389	0.016
CDR1p (baseline)	0.056	0.726	0.938
CDRhalf (follow-up)	-0.001	0.417	0.998
CDR1p (follow-up)	0.659	0.584	0.259

Table 4. Estimates of FCSRT decline using different methods

Method	Estimate	Standard Error	p-value
Regular linear mixed effects model for FCSRT	-2.283	0.432	<0.0001
Joint modeling of FCSRT and CDR	-2.384	0.429	<0.0001
Multiple Imputation	-2.600	0.612	0.0014

5. Discussion

The auxiliary information is valuable in testing the MAR assumption for the main model of interest and eliminating or reducing the bias when the missing process for the main model is not missing at random. Collecting auxiliary information that might be related to missing values has been advocated (e.g., Little 1995). As did other statisticians, we recommend collection of auxiliary variables when designing research studies, and taking the auxiliary variables into account when analyzing the data even they are not of primary interest. However, it has to be kept in mind that new model assumptions are introduced when utilizing the auxiliary information to relax the assumption on the missing data process, and thus such information needs to be utilized with caution. Further research is indicated.

Acknowledgements

The authors are grateful to Dr. Ellen Grober for providing the data. This research was supported by National Institute of Aging grants P01-AG03949 (PI: Richard Lipton). Dr. Charles B. Hall was also supported by R01-AG017854 (PI: Ellen Grober).

References

Collins, L.M., Schafer, J.L. and Kam, C. M. (2001) A comparison of inclusive and restrictive strategies in modern missing data procedure. *Psychological Methods* 6, 330-351.

Daniels, M. J. and Hogan, J. W. (2008) Missing data in longitudinal studies: strategies for Bayesian modeling and Sensitivity Analysis. New York: Chapman & Hall.

Grober, E and Buschke, H. (1987) Genuine memory deficits in dementia. *Developmental Neuropsychology* 3, 13–36.

Grober, E., Hall, C., Lipton, R. B., Teresi, J.A. (2008) Primary Care Screen for Early Dementia. *Journal of the American Geriatrics Society* 56, 206–213.

Ibrahim, J. G., Lipsitz, S. R. and Horton, N. (2001) Using auxiliary data for parameter estimation with non-ignorably missing outcomes. *Applied Statistics* 50, 361-373.

Laird, N. M. and Ware, J. H. (1982) Random effects models for longitudinal data. *Biometrics* 38, 963-974.

Liang, K. Y., Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13-22.

Little, R. J. A. (1995) Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 90, 1112-1121.

Little, R.J.A. and Rubin, D. B. (2002) Statistical Analysis with Missing Data, 2nd edition, New York: John Wiley.

Meng, X. L. (1994) Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 9, 538-573.

Robins, J. M., Rotnitzky, A., Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 90, 106-121.

Rubin, D. B. (1976) Inference and missing data. *Biometrika* 63, 581-592.

Rubin, D. B. (1987) Multiple imputation for nonresponse in surveys. New York: Wiley.

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

Schafer, J. L. (1997) Analysis of incomplete multivariate data. New York: Chapman & Hall