Current Journal of Applied Science and Technology

34(2): 1-10, 2019; Article no.CJAST.34913
ISSN: 2457-1024
(Past name: British Journal of Applied Science & Technology, Past ISSN: 2231-0843, NLM ID: 101664541)

Complete Case versus Inverse Probability Weighting Methods of Fitting Incomplete Longitudinal and Survival Data Joint Models

D. O. Nyaboga[1*], A. Mwangi[2] and D. Lusweti[3]

[1]Department of Statistics and Computer Science, Moi University, Eldoret, Kenya.
[2]Department of Behavioral Sciences, Moi University, Eldoret, Kenya.
[3]Department of Biological Sciences, Moi University, Eldoret, Kenya.

Authors' contributions

This work was carried out in collaboration among all authors. Author DON designed the study, performed the statistical analysis, wrote the protocol and wrote the first draft of the manuscript. Authors AM and DL managed the literature searches and the analyses of the study. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/CJAST/2019/v34i230121
Editor(s):
(1) Dr. Pengtao Sun, Associate Professor of Mathematics, University of Nevada Las Vegas, 4505 Maryland Parkway, Las Vegas, Nevada 89154, USA.
Reviewers:
(1) Prashant Kumar, Savitribai Phule Pune University, India.
(2) Rashmi Awad, School of Statistics, Devi Ahilya University, India.
Complete Peer review History: http://www.sdiarticle3.com/review-history/34913

Original Research Article

Received 17 April 2017
Accepted 08 July 2017
Published 30 March 2019

ABSTRACT

Missing data is a common problem in real word studies especially clinical studies. However, most people working with such data, often drop missing cases from individuals with incomplete observations that occur when patients do not complete the treatment or miss their scheduled visits. This may lead to misleading results and ultimately affect the decision of whether an intervention is good or bad for the patients under treatment. The comparison of Complete Case (CC) and Inverse Probability Weights (IPW) techniques of handling missing data in various models has been addressed, however little has been done to compare these methods when applied to joint models of longitudinal and time to event data. Therefore, this paper seeks to investigate the impact of assuming CC analysis on clinical data with missing cases, comparing it with IPW method when fitting joint models of longitudinal and survival data setting full data model as the baseline model. This paper made use of randomized aids clinical trial data. The model with Deviance Information

**Corresponding author: E-mail: nyabogadickson@yahoo.com;*

Criteria (DIC) close to that of full data joint model is considered the best. From the results, joint models from full data, CC and IPW had DIC of 10603.94, 8410.33 and 10600.95 respectively. The joint model obtained from IPW data had a DIC too close to that of full data joint model as compared to model from CC data.

Keywords: Joint model; missing data; longitudinal data; complete case; inverse probability weighting.

1. INTRODUCTION

The majority of clinical studies generate longitudinal and time to event data which sometimes depend on each other. However, in most cases, missing data occur in these data sets due to some patients not completing treatment or miss their scheduled visits. There are several ways of handling these missing cases. One is through Complete Case analysis (CC) where missing data is discarded and analysis only works with complete observations. This is only varied when data are missing completely at random [1]. This method is inefficient as it is biased.

The other methods are inverse probability weighting (IPW) [2] and multiple imputations (MI) [1]. These methods attempt to fill in the missing data randomly using the previously observed data.

These methods of handling missing data have been compared in several studies in terms of parameter estimation and standard errors. The comparison indicates that models fitted with incomplete data sets subjected MI and IPW methods gives less biased results compared to CC analysis. [3,4,5]. [6,7] also investigated MI and found out that it gives less biased results compared to CC analysis. Despite of all these, little has been done to compare these methods when fitting joint models of longitudinal and survival data in situations of missing data.

Andersson et al. [8] used a joint model to model long-term trends in breast cancer while accounting for drop-out as well as for measurement error. They argue that ignoring drop-outs when fitting a joint model may lead to biased results. Sweeting [9], also used a joint model to associate longitudinal measurements of aortic diameter with the risk of aneurysm rupture. He argues that a well fitted joint model can be used to make predictions that can be utilized within a fuller decision modelling framework, to allow planning of future interventions for patients under a 'watchful waiting' care pathway.

Crowther MJ et al. [10] looks at application of joint model of longitudinal and survival data while incorporating delayed entry (missing data), which has received relatively little attention. They argue that, incorporating the missing entries requires introduction of methods of handling missing data.

This paper seeks to examine the performance of CC and IPW methods when applied to missing at random (MAR) data in the joint models of longitudinal and survival data.

1.1 Definitions and Notations

1.1.1 Rubin's classification of missing data mechanisms

The missing data terminology used in this paper were developed by Rubin [11]. To describe these mechanisms, notations for missing cases are introduced.

Let R_{ij} denote indicator variable that takes on value 1 if subject i is observed at time j, and 0 if the subject was not observed at this time point. Here we consider whether the dependent variable y was observed or not. For measurement at n time points, then the $n\,x\,1$ complete dependent variable vector is

$$y_i^{'} = (y_{i1}, y_{i2}, \dots, y_{in}).$$

The $n\,x\,1$ missing data indicator vector for a subject is then

$$R_i^{'} = (R_{i1}, R_{i2}, \dots, R_{in}).$$

Where the specific R_{ij} values equal 1 or 0 depending on whether y_{ij} is observed or not

(*i.e.*, $R_{ij} = 1$ if subject i is observed at time j, or $R_{ij} = 0$ if subject i is missing at time j).

Based on R_i the complete dependent variable vector y_i can be partitioned into its observed y_i^O and unobserved y_i^M components for a given subject i. Here, y_i is a potential dependent variable vector for subject i, which differs notation-wise from usual treatment of this vector, and y_i^O as the actually observed dependent variable vector for subject i. Analogously, y_i^M is the component of the dependent variable vector that is missing.

1.1.2 Missing completely at random (MCAR)

This is the case where a patient is missing for completely random reasons at a particular time point that is, the missing responses probability is not related to the longitudinal outcome. For example, when a patient relocates to a different place or forgets to attend a scheduled appointment. This means that the indicators for missing data, R_i do not dependent on both y_i^O and y_i^M; they are independent of the values that were observed and those that were not.

1.1.3 Missing at random (MAR)

This occurs in cases where the probability of missing cases depends on both the covariates X_i of a fully observed model and vector y_i^O of observed dependent variable but is unrelated to unobserved dependent variable vector y_i^M . Example, when the doctor advises the patient to stop attending treatment due to measurements previously observed.

1.1.4 Missing not at random (MNAR)

This is where the missingness is related to vector y_i^M of unobserved dependent variable taking into account the observed variables $X_i \; and \; y_i^O$. This implies that there is an association between what would have been observed, y_i^M and the missingness R_{ij} . A good example is when a patient stops treatment due to death which is related to his or her CD4 Count level, including those that would have been observed if he continued with the treatment.

2. METHODOLOGY

2.1 Description of the ART Data

A completely fully observed data (data with no missing cases) has been considered as baseline data. The randomized AIDS clinical trial data containing both longitudinal and survival data used in this paper is publicly available for free to aid in research in HIV/AIDS in the website [12]. The dataset has a total of 1405 observations and 467 patients who were followed in the clinical trial. This population consists of HIV+ patients who are at least 18 years old and are on Antiretroviral drugs (ARVs). The longitudinal variable was CD4 counts measured until the patient was lost to follow-up. The survival variable was time to death. The other explanatory variables considered in this paper are as in Table 1.

Table 1. Explanatory variables used joint modeling of the ART data

No	Variable	Description
1	Drug	ddI, ddC
2	Gender	Male, female
3	Age	Years
4	Weight	Kilograms
5	Marital status	Not married, married, others
6	Education level	None, primary, secondary, tertiary
7	Employment	Yes, no
8	Clinic	Urban or rural

2.2 Joint Model Application to Fully Observed Data

A joint model comprising longitudinal and survival sub-models has been applied to fully observed data. This is set as a baseline model.

2.2.1 Longitudinal data modeling

First, we consider a linear mixed effects model, where longitudinal measurements, $y_{i1}, \dots, y_{ini}$ for the i^{th} subject at times S_{i1},… S_{ini} is given as:

$$\begin{aligned} y_i &= \mu_i(s) + M_{1i}(\mathrm{s}) + \varepsilon_i \\ &= X_{1i}^T(s)\beta_1 + Z_{1i}^T(s)b_i + \varepsilon_i \qquad (1) \\ b_i &\sim N(0,\varphi), \quad \varepsilon_i \sim (0,\delta_\varepsilon^2 I) \end{aligned}$$

In this case, y is a vector of responses that were observed with a dimension n_i, β_1 is a vector of fixed effects of dimension p, b_i is a random effects vector of dimension q, $X_{1i}^T(s)$ is a matrix of fixed effects covariates that are time-varying of (size $n \, x \, p$), $Z_{1i}^T(s)$ is ($n \, x \, q$) dimensional matrix of random effects covariates and ε_i represents a within-group error vector of dimension n_i with a Gaussian distribution.

Here, $\mu_i(s)$ = $X_{1i}^T(s\,)\beta_1$ is mean response and M_{1i}(s) = $Z_{1i}^T(s)b_i$ contains random effects. M_{1i}(s) are the adjusted CD4 trajectories. Z_{1i} are random effect covariates while X_{1i} are fixed effects covariates.

2.2.2 Survival data modeling

The survival Cox proportional hazard model is given by;

$$\begin{aligned} h_i(t) &= h_0(t)\mu_i(t) \\ &= h_0(t)\exp\,(X_{2i}^T(s)\beta_2) \end{aligned} \qquad (2)$$

Where $h_0(t)$ represents baseline hazard function, $X_{2i}^T(s)$ is a time-dependent covariates vector and β_2 is a vector of fixed effects parameters.

2.2.3 The joint models structure

2.2.3.1 The longitudinal sub model specification

The longitudinal sub-model as described by [13], is given as below;

$$\begin{aligned} y_i &= \mu_i(s) + W_{1i}(\mathrm{s}) + \varepsilon_i \\ &= X_{1i}^T(s)\beta_1 + Z_{1i}^T(s)b_i + \varepsilon_i \end{aligned}$$
$$b_i \sim N(0,\varphi),\ \varepsilon_i \sim (0, V_i I),\ \log\,(V_i) \sim N(\mu_v, \delta_v^2) \qquad (3)$$

Where V_i is the within-subject variability which follows a log-normal distribution with mean μ_v and variance δ_v^2

2.2.3.2 The survival sub model specification

Specification of survival sub-model is given as:

$$h_i(t) = h_0(t)\exp\,(X_{2i}^T(s)\beta_2 + W_{2i}(t) \qquad (4)$$

In this cases, the association function, $W_{2i}(t)$, in (4) is similar to $W_{1i}(s)$ in longitudinal sub-model (3) which is given as;

$$W_{2i}(t) = \gamma_0 b_{0i} + \gamma_1 b_{1i} + \gamma_2 b_{2i} + \gamma_3 \log(V_i) + b_{3i} \qquad (5)$$

Where

γ_0, γ_1, γ_2 and γ_3, shows the association between the two sub-models, that is, the random intercepts, linear slope, quadratic slope and the CD4 variability, respectively.

2.2.3.3 Bayesian estimation and inference

The two sun models are then linked together using Bayesian estimation approach as follows:

$$f(y,T,\delta|\Theta_1,\Theta_2) = \prod_{i=1}^{N} \int (y|\Theta_1, n_i) f(T,\delta|y,\Theta_2,n_i) f(n_i) dn_i$$

Having a likelihood function given as

$$L(y,T,\delta|\Theta_1,\Theta_2) = \prod_{i=1}^{N} \int (y|\Theta_1, n_i) f(T,\delta|y,\Theta_2,n_i)^{\delta i} (1 - F(T,\delta|y,n_i,\Theta_2)^{(1-\delta i)} f(n_i) dn_i$$

Where $n_i = \{b_i,\ V_i I\}$ represents the shared underlying process, $\Theta_1 = \{\beta_1, \varphi, \mu_v, \delta_v^2\}$ and $\Theta_2 = \{\beta_2, \gamma, \delta_3^2\}$ represents population parameters specified in the mixed and survival models respectively. $f(.)$ and $F(.)$ denotes density and distribution functions, respectively.

2.3 Application of Joint Model to Fully Observed Data

A joint model of longitudinal and survival data has been fitted to the HIV/AIDS data that has all observation, that is, data without any missing case.

2.4 Application of the Joint Model to Complete Case Data

About 30% missing at random (MAR) observations have been simulated from the fully observed data. We then discard the missing observations from the data. This forms our complete case data. The joint model of longitudinal and time to event data has been then applied to the complete case data.

2.5 Application of Joint Model to the Inverse Probability Weighting (IPW) Repeated Measures Data

We assume that the expected outcome is Y, given the design variables X, and auxiliary variables V. We also assume that, $M = 1$ depict that Y is missing. IPW can be described by introducing a response indicator, $R = 1 - M$, such that $R = 0$ when Y is missing and $R = 1$ when it is observed.

In this case, the auxiliary information must be included in the history of the observed data, $Z_j^- = (Y_1, \ldots, Y_{j-1},\ V_1, \ldots, V_{j-1})$ also in the model for $\pi(X,V,\theta)$. The procedure for performing IPW on the incomplete data is as follows:

1. Identify the regression model to be used if all the intended data were observed.
2. Let $\emptyset_j(X, Z_j^-;\theta) = P(R_j = 1|R_{j-1} = 1, X, Z_j^-;\theta)$ denote the probability that Y_j was observed.

3. Fit the model for $\emptyset_j$ and represent the estimated parameters by $\hat{\theta}$.
4. Let $\pi_j(X, Z_j^-; \theta) = \prod_{k=1}^{j} \emptyset_j (X, Z_k^-; \theta)$ to represent the probability that the patient was retained in the study to time j.
5. By using independence correlation structure, fit the regression in Step 1, weighting the individual contributions to the model by $\{\pi_j(X, Z_j^-; \hat{\theta})\}$.
6. Estimate the error by use of the bootstrap technique.

2.6 Comparison of Joint Models Obtained from CC and IPW Data with that of Fully Observed Data

A joint model of longitudinal and survival data applied to fully observed data has been set as baseline model which is assumed to be the correctly specified model. The joint models fitted from CC and IPW datasets are compared with joint model fitted from the fully observed dataset. The precise nature of the joint models is selected using the DIC (Deviance Information Criterion). The joint model with DIC close to that of a full data model indicates preferred model and the method used to estimate missing data in the said model is considered the best.

3. RESULTS

3.1 Descriptive Statistics of the Explanatory Variables in ART Data

Table 2 shows descriptive statistics of the explanatory variables present in the Anti-retroviral therapy (ART) data that was used to evaluate missing data estimation methods. A total of 467 HIV-positive patients on ART treatment were considered for the study. Out of these, 422 (90.36%) were males and 45 (9.64%) were females. The majority (49.89%) of the infected patients were below 25 years with mean age of 31.34 (SD=10.50), Most of the patients 187 (40.04%) were married while 154 (32.98%) were not married. 126 (26.98%) were either in the divorced or widowed group. Patients who had no formal education were 13(2.78%) while those who had completed at least primary education were 87(97.22%). The majority (53.32%) of the patients were not employed. The mean CD4 count at baseline was 7.13 (SD=4.71) cells/mm3. Finally, the mean weight at baseline was 51.30 kg (SD=7.03).

Table 2. Baseline characteristics of patients on ART. (N=467)

Characteristics	Descriptive statistics
Gender	
Female	45(9.64%)
Male	422(90.36%)
Age group	
Mean age(SD)	31.34(10.50)
<25	233(49.89%)
26-30	45(9.64%)
31-35	58(12.42%)
36-40	41(8.78%)
41-45	41(8.78%)
46+	23(4.93%)
Education	
None	13(2.78%)
Primary	159(34.05%)
Secondary	116(24.84%)
Tertiary	110(23.55%)
Marital Status	
Not married	154(32.98%)
Married	187(40.04%)
Others	126(26.98%)
Employed	
Yes	218(46.68%)
No	249(53.32%)
Mean CD4+ at diagnosis (SD)	7.13(4.71)
Mean weight(SD)	51.30(7.03)

3.2 Fitting the Separate and Joint Models for the Fully Observed Dataset

Table 3 summarizes the estimates, standard errors and p-values of the parameters of the separate and joint models run with full data (all cases observed). In longitudinal sub-model, the predictors, time and age, were statistically significant at 5% level of significance. The time effect coefficient has a negative sign indicating that CD4 cell count decreases on average with time. The estimate for gender in the longitudinal regression model has a negative sign though not significantly different from zero, suggesting that, throughout the follow-up, the male patients had lower CD4 cell counts than females. The coefficient for age is -0.057 (P-value 0.001), suggesting that as the age of patients increases there is a significant decrease in CD4 count cells/mm3 level over the time. Similarly, the predictor coefficient for weight (-0.009) indicates that CD4 count decreases over the period of study for a unit decrease in weight of the patients.

Table 3. Separate and joint model of longitudinal and time to event fully observed ART data

	Parameters	Separate analysis			Joint analysis		
		Estimate	Std. error	P-value	Estimate	Std. error	P-value
Longitudinal sub-model (CD4 Count)	Intercept	9.001	1.569	<0.001	8.894	0.024	<0.001
	Obstime	-0.147	0.012	<0.001	-0.161	0.002	<0.001
	Drug(ddl)	0.605	0.433	0.163	0.582	0.010	0.182
	gender(male)	-0.563	0.740	0.447	-0.471	0.018	0.509
	Age	-0.057	0.018	0.001	-0.052	0.004	<0.001
	Marital status	0.278	0.244	0.254	0.307	0.058	0.214
	Weight	-0.009	0.011	0.427	-0.009	0.002	0.361
	Education	0.203	0.177	0.252	0.187	0.041	0.271
	Clinic	-0.089	0.408	0.827	-0.125	0.121	0.857
	Employment	0.043	0.417	0.918	-0.003	0.059	0.930
Survival sub-model (time to death)	Intercept						
	Drug(ddl)	0.250	0.148	0.092	0.396	0.032	0.011
	Gender(male)	-0.156	0.245	0.523	-0.273	0.062	0.266
	Age	0.000	0.007	0.947	-0.012	0.002	0.086
	Marital status	-0.062	0.098	0.529	0.018	0.028	0.919
	Weight	0.019	0.010	0.071	0.023	0.002	0.032
	Education	-0.049	0.070	0.484	-0.009	0.016	0.906
	Clinic	-0.166	0.153	0.279	-0.233	0.040	0.187
	Employment	0.175	0.156	0.263	0.252	0.031	0.095
	γ_1				-0.276	0.002	<0.001
	γ_2				363.501	43.627	<0.001
	DIC					**10603.94**	

In survival sub-model, sex and age are not found to be significant predictors for death event of HIV patients. The risk of an educated patient being lost to follow-up is [exp(-0.049) = 0.952] times higher compared to an uneducated patient. Patients attending rural clinic have the higher hazard of loss to follow-up from the treatment as compared to the patients who belong to the urban area though not statistically significant. Body weight is associated positively with CD4 count trajectory. Also, after initiation of ART, patients with higher body weight are associated with lower hazard for loss to follow-up.

The parameter estimates for the two models i.e separate and joint models are quite similar to each other though not identical. The posterior estimates of the association parameters for the joint model are statistically significant, implying that there is an association between the two sub-models. The parameter estimate of association due to the trend of CD4 is negative ($\gamma_1 = -0.276$). This implies that there is a negative association between the slope of the CD4 count and the hazard of HIV patients who die while in ART treatment. This implies that there is a significantly reduced risk of dying in patients undergoing ART treatment with an increasing trend in the CD4 count. Also, the association parameter estimate due to CD4 count variability is positive ($\gamma_2 = 363.501$). This indicates that the higher CD4 fluctuation in HIV-positive patients is significantly associated with the higher hazard of dying.

The full data joint model has DIC of 10603.94. This Deviance Information Criterion is used as the reference for the other models.

3.3 Application of Joint Model to Complete Case Data

Table 4 shows the summary of the full data, observed data obtained after deleting the missing cases created from fully observed data and the missing cases on the longitudinal variable.

Table 4. Summary of full, observed and missing values simulated on the longitudinal variable CD4 counts

Observations	Frequency	Percentage
N (Full data)	1405	100.00
Observed	995	70.82
Missing	410	29.18

Separate and joint models are then fitted from the complete case data. As shown in Table 5, the separate and joint model parameter estimates for the complete case data are not similar to those of full data model. The Deviance Information Criteria (DIC) of the complete case joint model is 8410.332 which greatly varies from that of full data joint model.

A comparison is made on a few selected parameter estimates for full data and complete case joint models. As shown in Table 6, in the longitudinal sub-model, the predictor age is statistically significant at 5% level of significance in full data joint model with P-value <0.001 but is not a statistically significant predictor in the complete case joint model (P-value 0.08). The coefficient for education is a statistically significant predictor in the complete case data joint model (P-value 0.02) but not in full data joint model (P-value 0.27). The coefficient for the employment has a negative impact on the full data joint model (-0.003) but positive impact on the complete case joint model (0.517). In the survival sub-model, age is statistically significant in the complete case data joint model (P-value 0.04) but insignificant in full data joint model (P-value 0.09). Body weight is statistically significant in full data joint model (P-value 0.03) but not significant in the complete case joint model (P-value 0.15).

Table 5. Separate and joint model of longitudinal and time to event complete case ART data

	Parameters	**Separate analysis**			**Joint analysis**		
		Estimate	**Std. error**	**P-value**	**Estimate**	**Std. error**	**P-value**
Longitudinal sub-model (CD4 Count)	Intercept	8.432	1.750	0.000	7.618	0.023	<0.001
	obstime	-0.152	0.016	0.000	-0.159	0.003	<0.001
	Drug(ddl)	0.705	0.453	0.121	0.353	0.010	0.413
	gender(male)	-1.167	0.783	0.137	-0.618	0.016	0.390
	Age	-0.052	0.020	0.009	-0.037	0.005	0.081
	Marital status	0.294	0.285	0.304	0.394	0.042	0.070
	Weight	-0.001	0.015	0.956	-0.011	0.001	0.196
	Education	0.349	0.196	0.075	0.437	0.032	0.020
	Clinic	-0.246	0.442	0.579	-0.332	0.071	0.422
	Employment	0.384	0.450	0.393	0.517	0.077	0.153
Survival sub-model (time to death)	Intercept						
	Drug(ddl)	0.168	0.161	0.297	0.520	0.036	<0.001
	Gender(male)	0.012	0.282	0.966	-0.446	0.051	0.120
	Age	0.001	0.008	0.873	-0.017	0.002	0.042
	Marital status	-0.096	0.106	0.364	0.010	0.026	0.933
	Weight	0.011	0.012	0.329	0.017	0.003	0.148
	Education	-0.055	0.075	0.459	-0.004	0.023	0.983
	Clinic	-0.219	0.164	0.182	-0.231	0.040	0.164
	Employment	0.175	0.168	0.297	0.233	0.035	0.151
	γ_1				-0.293	0.002	<0.001
	γ_2				414.392	48.016	<0.001
	DIC					**8410.332**	

Table 6. Comparison of the selected parameter estimates in the full data and complete case joint model

	Parameters	**Full data joint model**			**Complete case joint model**		
		Estimate	**Std. error**	**P-value**	**Estimate**	**Std. error**	**P-value**
Longitudinal sub-model	Age	-0.052	0.004	<0.001	-0.037	0.005	0.081
	Education	0.187	0.041	0.271	0.437	0.032	0.020
	Employment	-0.003	0.059	0.930	0.517	0.077	0.153
Survival sub-model	Age	-0.012	0.002	0.086	-0.017	0.002	0.042
	Weight	0.023	0.002	0.032	0.017	0.003	0.148
	DIC		**10603.94**			**8410.332**	

3.4 Application of Joint Model to Inverse Probability Weighting Data

Table 7 shows results of the joint model fitted from IPW data. From the results, the separate and joint models parameter estimates are almost similar to those from full data model though not identical. The DIC for the inverse probability weighting joint model is 10600.95 which is almost equal to that of the full data joint model.

3.5 Comparisons of CC and IPW Data Joint Models with Full Data Joint Model

Table 8, shows results of joint models applied to CC and IPW data, compared to full data joint model using the magnitude of DIC. From the results, Inverse Probability Weighting is found to be the best method of handling missing data. Inverse Probability Weighting model has Deviance Information Criterion very close to that of full data joint model which is set as a baseline model. Complete Case analysis shows a significant variation in its DIC compared to the full data model.

4. DISCUSSION

In everyday activities, it is important to address issues dealing with missing data since it occurs in almost all investigations in the real-world. It is more important to account for missing data especially on studies dealing with clinical data. In this paper, around 30% missing observations are imputed randomly on the longitudinal variable of the treatment and follow-up data used. Discarding all these missing cases and only working with complete data will result in a biased model. This paper addresses the performance of using the Complete Case analysis (CC) and Inverse Probability Weighting (IPW) methods for handling missing covariate values and applies them to joint models of longitudinal and survival data using the AIDS dataset.

Table 7. Separate and joint model of longitudinal and time to event Inverse Probability Weighting ART data

	Parameters	**Separate analysis**			**Joint analysis**		
		Estimate	**Std. error**	**P-value**	**Estimate**	**Std. error**	**P-value**
Longitudinal sub-model (CD4 Count)	Intercept	8.964	1.569	<0.001	8.887	0.022	<0.001
	Obstime	-0.146	0.012	<0.001	-0.166	0.003	<0.001
	Drug(ddl)	0.608	0.433	0.161	0.586	0.010	0.191
	gender(male)	-0.557	0.740	0.452	-0.442	0.018	0.549
	Age	-0.056	0.018	0.002	-0.058	0.003	<0.001
	Marital status	0.283	0.243	0.245	0.301	0.037	0.143
	Weight	-0.008	0.011	0.460	-0.010	0.003	0.371
	Education	0.204	0.177	0.251	0.268	0.040	0.086
	Clinic	-0.102	0.409	0.804	-0.125	0.086	0.785
	Employment	0.039	0.418	0.926	0.043	0.100	0.864
Survival sub-model (time to death)	Intercept						
	Drug(ddl)	0.252	0.148	0.089	0.304	0.027	0.003
	Gender(male)	-0.159	0.244	0.514	-0.178	0.096	0.549
	Age	0.000	0.007	0.946	-0.008	0.001	0.191
	Marital status	-0.065	0.098	0.510	0.020	0.027	0.906
	Weight	0.019	0.010	0.067	0.022	0.004	0.039
	Education	-0.050	0.070	0.479	0.016	0.015	0.712
	Clinic	-0.165	0.153	0.281	-0.247	0.029	0.075
	Employment	0.173	0.157	0.268	0.279	0.034	0.080
	γ_1				-0.273	0.001	<0.001
	γ_2				501.507	32.577	<0.001
	DIC					**10600.95**	

Table 8. Summary of deviance information criteria for CC and IPW joint models compared to full data model

	Model		
	Full data	**Complete case**	**IPW**
DIC	10603.94	8410.33	10600.95
Difference from full data	0.00	2193.61	2.99

From the results obtained, the parameter estimates and standard errors of the joint model obtained from IPW are quite similar to those of full data joint model though not identical. The parameter estimates for CC joint model shows a huge variation compared to that of the full data model. This implies that under MAR mechanism, using CC analysis will give biased results as compared to IPW analysis for joint models. IPW joint model has DIC of 10600.95 which is too close to 10603.94 of the full data joint model. The CC analysis joint model has a DIC of 8410.33 which was far from the full data model DIC. This huge variation further shows the biasedness of assuming missing data when fitting a joint model.

5. CONCLUSIONS

The results, in general, reveals that IPW is likely to be the best method of handling missing data under the MAR mechanism. In this paper, IPW joint model outperforms CC in terms of DIC when compared to full data joint model. This advantage of the IPW is well documented in terms of the MAR mechanism [14,15].

The findings further suggests the inappropriateness of CC analysis. The study shows that CC analysis can lead to the loss of power of the covariates and imprecise parameter estimates. To avoid this, an application of IPW can be utilized. This study supports [16] recommendation to avoid CC analysis where possible.

Missingness mechanism is simulated to be MAR, indicating that the CC performance is unsatisfactory under this assumption. This is in line with previous studies which shows that CC is more widely used under MCAR than under MAR [16]. Therefore, the preferred method of dealing with missing covariate values under MAR in joint models of longitudinal and survival data is IPW.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Little RJA, Rubin DB. Statistical Analysis with Missing Data. New York: John Wiley & Sons; 2002.
2. Höfler M, Pfister H, Lieb R, Wittchen H. The use of weights to account for non-response and drop-out. Social Psychiatry and Psychiatric Epidemiology. 2005;40: 291–299.
3. Karangwa I. Using the Markov Chain Monte Carlo Method to Make Inferences on Items of data contaminated by missing values. American Journal of Theoretical and Applied Statistics. 2013;2(3):48.
4. Kropko J, Goodrich B, Gelman A, Hill J. Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. Political Analysis. 2014;1-23.
5. Raghunathan T, Lepkowski J, VanHoewyk, M, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology. 2001;27(1):85-95.
6. Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. American Journal of Epidemiology. 2010; 171(5):624-32.
7. Finch WH. Imputation methods for missing categorical questionnaire data. A Comparison of Approaches. Journal of Data Science. 2010;8:361-378.
8. Andersson TM, Crowther MJ, Czene K, Hall P, Humphreys K. Mammographic density reduction as a prognostic marker for postmenopausal breast cancer; results using a joint longitudinal-survival modelling approach. American Journal of Epidemiology; 2017.
9. Sweeting MJ. Using predictions from a joint model for longitudinal and survival data to inform the optimal time of intervention in an abdominal aortic aneurysm screening programme. Biometrical Journal; 2017.

10. Crowther MJ, Andersson TML, Lambert P C, Abrams KR, Humphreys K. Joint modelling of longitudinal and survival data: incorporating delayed entry and an assessment of model misspecification. Statistics in medicine. 2016;35(7):1193-1209.
11. Rubin DB. Inference and missing data. Biometrika. 1976;63:581-592.
12. http://www.biostat.umn.edu/~brad/software.html.
13. Lyles RH, Munz A, Xu J, Taylor JMG, Chmiel JS. Adjusting for measurement error to assess health effects of variability in biomarkers. Stat.Med. 1999;18:1069-1086.
14. Little R, Rubin DB. Statistical analysis with missing data. New York: John Wiley; 1987.
15. Schafer JL. Analysis of Incomplete Multivariate Data, (1st Edn). Chapman & Hall; 1997.
16. Kenward M, Molenberghs G. Last observation carried forward: A crystal ball. Journal of Biopharmaceutical Statistics. 2009;19(5):872–888.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle3.com/review-history/34913