Treatment of Missing Data in Randomized Clinical Trials

Math 654 Design and Analysis of Clinical Trails

Design and Analysis of Clinical Trials Project

Victor Moya

Yat Fan WONG

Introduction

Missing Data refers to an outcome that is meaningful for analysis that was not collected, it may includes numeric and character data. The followings are several reasons on missing data in clinical trial. Patient refusal to continue, withdrawal of consent because they moved away and left no contact. There is an adverse event that resulted in discontinuation of study treatment. Also, poor record keeping is another main reason for missing data.

In December 2000, 66% of 519 clinical trials published did not report how they handled missing data. Moreover, between July and December in 2001, 89% of in 71 trials published by the BMJ, JAMA, Lancet or New England Journal of Medicine had missing outcome data in their clinical trial experiments. It seems like most researchers do not pay attention to missing data and might not think it is a serious problem. According to 'ICH Guidance on Efficacy Evaluations in Clinical Trials, Guidelines E9: Statistical Principles for Clinical Trials', we should avoid missing data if it is possible because missing data will introduce a potential source of biases in our experiments. A trial with missing data may be valid as long as sensible methods are used for dealing with missing data. However, the guideline does not recommended any applicable methods of handling missing data. In fact, there are many methods and procedures for handling missing data were introduced in these years but none of them was approved by the agency the any major country.

In this project, we will introduce three traditional and two modern methods on handling missing data in clinical trials. Before that, we will talk about the nature of missing of missing data because we need to do assumptions for the missing data during using any method or procedure is applied. Also, we will do a comparison on each methods with a real data set. Of course, the best solution to missing data in clinical trials is to set up some preventions before it happens. So, we will recommend some prevention at the end of our report.

The nature of missing data

Although we cannot tell from the data on hand whether the missing observations are MCAR, MAR or MNAR, assumption of missingness mechanism will be made when a specific analysis

method is being used. For example, in the case of likelihood-based estimation, the unbiased parameter estimates can be obtained from the observed data. The missingness mechanism is ignorable only if it arises from an MCAR or MAR.

Missing Completely At Random (MCAR)

$Pr(r \mid y_o, y_m) = Pr(r)$

The probability of a subject dropping out is independent of outcomes, visible or invisible, or any other variables in the analysis. It means the missingness does not depend on observed or unobserved data, or any analysis valid for the whole dataset is valid for the observed data. For example, participant's data were missing because he was stopped for a traffic violation and missed the data collection session. The other example is a laboratory sample is dropped, so the resulting observation is missing.

Missing At Random (MAR)

$Pr(r \mid y_o, y_m) = Pr(r \mid y_o)$

The probability of a subject dropping out is conditionally *independent* of future (current) observations, given the observed data. It means the missingness depends on the observed data but does not depend on the unobserved data. For example, women are less likely to disclose weight. Therefore, the probability of missing depends on gender and does not depend of weight itself. We can make more realistic assumption under MAR and the estimations can be biased. Also, the likelihood based analyses of the outcome will be valid under MAR. For example, in the following table, subjects 1 and 2 have the same values where both are observed. Under the assumption of MAR, variables 5 and 6 from subject 2 have the same distribution (not the same value!) as variables 5 and 6 from subject 1

	Variables					
Subject	1	2	3	4	5	6
1	1	3	4.3	3.5	1	4.6
2	1	3	4.3	3.5	?	?

Missing Not At Random (MNAR)

The probability of a subject dropping out is conditionally *dependent* on future unseen observations, given the observed history. Equivalently, the future statistical behavior of subjects is not the same for those who drop out and those who don't, even if their history is identical. It only depends on the unobserved outcomes of the variable being analyzed. So, the pattern will not be random nor ignorable. For example, we are studying mental health and people who have been diagnosed as depressed but they are less likely than others to report their mental status. The other example is the overweight people are not likely to report their weight. In result, the missingness of weight will depends on weight only.

Treatments for missing data

Basically, we can divided the methods into two approaches, one is traditional approach and the other is modern approach. The traditional approach uses mainly deletion and imputation to handle the missing data. Because of the convenience of using software with calculation device like computers, more complicated statistical theory and calculation can be applied and done in a short period of time. Handling missing data with modern approach becomes more and more popular today.

A. Traditional Approaches

1. Listwise Deletion

This method is to omit those cases with missing data and to run our analyses on what remains. So, all analyses are conducted with the same number of cases at the end and we will have a complete case to do the analysis. For example, we are working on a clinical trial study on depression and we divide our patients into two groups, Group 1: Control (No drug) and Group 2: Treatment (with drugs). At the end of the study, we have the following data set.

Subject	Group	Baseline	Follow-up Week		
			1st month	3th month	6th month
1	1	296	175	187	192
2	1	376	329	236	76
3	1	150	?	?	?
4	2	282	186	225	134
5	2	317	31	85	120
6	2	362	104	?	?

By using listwise deletion method, we just simply remove the all the data from patient 3 and 6 from the sample before performing any further analysis.

Subject	Group	Baseline	Follow-up Week		
			1st month	3th month	6th month
1	1	296	175	187	192
2	1	376	329	236	76
3	X	X	X	X	X
4	2	282	186	225	134
5	2	317	31	85	120
6	X	X	X	X	X

The only advantage of listwise deletion is it is easy to approach because we don't have to both any missing data and we will have all subject with the same number of cases. However, it will affect the statistical power because the loss of too much valuable data and reduction in the sample size. We may have a biased result at the end.

2. Simple Imputation Method - Last Observation Carried Forward (LOCF)

By using LOCF method, subject's missing responses is equal to their last observed response and it is developed under Missing Completely At Random (MCAR) framework. So, we just

impute values to the missing data from the patient's last observation if there is any missing data. This method is usually used in longitudinal (repeated measures) studies of continuous outcomes. Also, this is the most popular method for handling missing data nowadays. For example, in our clinical trial study on depression, patient 3 dropped out from the study after the baseline data was recorded. Patient 6 dropped out after the 1st month follow up.

Subject	Group	Baseline	Follow-up Week		
			1st month	3th month	6th month
1	1	296	175	187	192
2	1	376	329	236	76
3	1	150	?	?	?
4	2	282	186	225	134
5	2	317	31	85	120
6	2	362	104	?	?

For patient 3, we assume the remain missing visit value will be the same his last visit record. i.e. we fill in the missing values for 1st, 3th, and 6th month follow up records with his last visit (baseline) data value, 150. For the patient 6, we just fill in his 3th and 6th month follow up missing data with his last observed 1st month record, 104.

Subject	Group	Baseline	Follow-up Week		
			1st month	3th month	6th month
1	1	296	175	187	192
2	1	376	329	236	76
3	1	150	150	150	150
4	2	282	186	225	134
5	2	317	31	85	120
6	2	362	104	104	104

3. Simple Imputation Method – Baseline Observation Carried Forward (BOCF)

Like LOCF, it is developed under Missing Completely At Random (MCAR) framework and it is usually used in longitudinal (repeated measures) studies of continuous outcomes. But this time we assume a patient's missing responses is equal to their baseline observed response. It means we impute values to the missing data from the patient's baseline observation. For example,

in our clinical trial study on depression, patient 3 dropped out from the study after the baseline data was recorded. Patient 6 dropped out after the 1st month follow up.

Subject	Group	Baseline	Follow-up Week		
			1st month	3th month	6th month
1	1	296	175	187	192
2	1	376	329	236	76
3	1	150	?	?	?
4	2	282	186	225	134
5	2	317	31	85	120
6	2	362	104	?	?

For patient 3, we assume the remain missing visit value will be the same his baseline record. i.e. we fill in the missing values for 1st, 3th, and 6th month follow up records with his last visit (baseline) data value, 150. For patient 6, we just fill in his 3th and 6th month follow up missing data with his baseline record, 362.

Subject	Group	Baseline	Follow-up Week		
			1st month	3th month	6th month
1	1	296	175	187	192
2	1	376	329	236	76
3	1	150	150	150	150
4	2	282	186	225	134
5	2	317	31	85	120
6	2	362	104	362	362

Basically, the simple imputation method (LOCF and BOCF) are really popular because it is easy to understand and approach. Also, it minimizes the number of the subjects who are eliminated from the analysis. In result, it provides conservative results with respect to active treatment if placebo patients drop out early because of lack of efficacy. However, patients who are given treatments should get better, the score for any patient should not be remain unchanged. In our example, we expect the score for patient 3 from the Control group will increase and the score for patient 6 from the Treatment group decrease.

Subject	Group	Baseline	Follow-up Week		
			1st month	3th month	6th month
1	1	296	175	187	192
2	1	376	329	236	76
3	1	150	155	160	165
4	2	282	186	225	134
5	2	317	31	85	120
6	2	362	104	95	90

If simple imputation method (LOCF and BOCF) is applied, the record data will remain unchanged. It means there is no improvement on the treatment group and no difference between both groups.

Subject	Group	Baseline	Follow-up Week		
			1st month	3th month	6th month
1	1	296	175	187	192
2	1	376	329	236	76
3	1	150	150	150	150
4	2	282	186	225	134
5	2	317	31	85	120
6	2	362	104	104	104

So, LOCF or BOCF is not an analytic approach but a method for imputing missing values. It will tend to underestimates or overestimates its variance and the biased estimates of the treatment effects as well.

B. Modern Approaches - Likelihood Base

1. EM algorithm

Full Information Maximum Likelihood

Probably the most pragmatic missing data estimation approach for structural equation modeling is full information maximum likelihood (FIML), which has been shown to produce unbiased parameter estimates and standard errors under MAR and MCAR. FIML, sometimes called “direct maximum likelihood,” "raw maximum likelihood" or just "ML," is currently

available in all major SEM packages. FIML requires that data be at least MAR (i.e., either MAR or MCAR are ok). The process works by estimating a likelihood function for each individual based on the variables that are present so that all the available data are used. For example, there may be some variables with data for all 389 cases but some variables may have data for only 320 of the cases. Model fit information in is derived from a summation across fit functions for individual cases, and, thus, model fit information is based on all 389 cases. Full Information Maximum Likelihood uses the Expectation-Maximization (EM) algorithm. A general approach to iterative computation of maximum-likelihood estimates when the observations can be viewed as incomplete data. Since each iteration of the algorithm consists of an expectation step followed by a maximization step we call it the EM algorithm.

Maximum-likelihood

Recall the definition of the maximum-likelihood estimation problem. We have a density function $p(x|\theta)$ that is governed by the set of parameters θ (e.g., p might be a set of Gaussians and θ could be the means and covariances). We also have a data set of size N, supposedly drawn from this distribution, i.e., $X = \{x_1 \ldots x_n\}$. That is, we assume that these data vectors are independent and identically distributed (i.i.d.) with distribution p. Therefore, the resulting density for the samples is

$$P(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = L(\theta|X).$$

This function $L(\theta|X)$ is called the likelihood of the parameters given the data, or just the likelihood function. The likelihood is thought of as a function of the parameters θ where the data X is fixed. In the maximum likelihood problem, our goal is to find the θ that maximizes L. That is, we wish to find θ^* where

$$\theta^* = \operatorname*{argmax}_{\theta} L(\theta|X)$$

Often we maximize $\log(L(\theta|X))$ instead because it is analytically easier. Depending on the form of $p(x|\theta)$ this problem can be easy or hard.

For example, if p (x|θ) is simply a single Gaussian distribution where $\theta = (\mu,\delta^2)$, then we can set the derivative of log (L (θ|X)) to zero, and solve directly for μ and δ^2(this, in fact, results in the standard formulas for the mean and variance of a data set). For many problems, however, it is not possible to find such analytical expressions, and we must resort to more elaborate techniques. Suppose that we had the sample data 1, 4, 7, 9 and wanted to estimate the population mean. You probably already know that our best estimate of the population mean is the sample mean, but forget that bit of knowledge for the moment. Suppose that we were willing to assume that the population was normally distributed, simply because this makes the argument easier. Let the population mean be represented by the symbol μ, although in most discussions of maximum likelihood we use a more generic symbol, θ, because it could stand for any parameter we wish to estimate. We could calculate the probability of obtaining a 1, 4, 7, and 9 for a specific value of μ. This would be the product p(1)*p(4)*p(7)*p(9). You would probably guess that this probability would be very small if the true value of μ = 10, but would be considerably higher if the true value of μ were 4 or 5. (In fact, it would be at its maximum for μ = 5.25.) For each different value of μ we could calculate p(1), etc. and thus the product. For some value of μ this product will be larger than for any other value of μ. We call this the maximum likelihood estimate of μ. It turns out that the maximum likelihood estimator of the population mean is the sample mean, because we are more likely to obtain a 1, 4, 7, and 9 if μ = the sample mean than if it equals any other value.

Overview of the EM Algorithm

1.Maximum likelihood estimation is ubiquitous in statistics
2. EM is a special case that relies on the notion of missing information.
3. The surrogate function is created by calculating a certain conditional expectation.
4. Convexity enters through Jensen's inequality.
5. Many examples were known before the general principle was enunciated.

Ingredients of the EM Algorithm

1. The observed data y with likelihood f(y |θ). Here θ is a parameter vector.
2. The complete data x with likelihood g(x |θ).

3. The conditional expectation, $Q(\theta|\theta^{\mathbf{n}}) = E[\ln g(x \mid \theta) \mid y, \theta^{\mathbf{n}}]$ furnishes the minimizing function up to a constant. Here $\theta^{\mathbf{n}}$ is the value of θ at iteration n of the EM algorithm.
4. Calculation of $Q(\theta \mid \theta^{\mathbf{n}})$ constitutes the E step; maximization of $Q(\theta \mid \theta^{\mathbf{n}})$ with respect to θ constitutes the M step.

Minimization Property of the EM Algorithm

1. The proof depends on Jensen's inequality $E[h(Z)] \geq h[E(Z)]$ for a random variable Z and convex function h(z).
2. If p(z) and q(z) are probability densities with respect to a measure μ, then the convexity of $-\ln z$ implies the information inequality
$E_{\mathbf{p}}[\ln p] - E_{\mathbf{p}}[\ln q] = E_{\mathbf{p}}[-\ln(q/p)] \geq -\ln E_{\mathbf{p}}(q/p) = -\ln\int(q/p)pd\mu = 0$, with equality when p = q.
3. In the E step minimization, we apply the information inequality to the conditional densities $p(x) = f(x \mid \theta^{\mathbf{n}})/g(y \mid \theta^{\mathbf{n}})$ and $q(x) = f(x \mid \theta)/g(y|\theta)$ of the complete data x given the observed data y.
4. The information inequality $E_{\mathbf{p}}[\ln p] \geq E_{\mathbf{p}}[\ln q]$ now yields

$$Q(\theta \mid \theta^{\mathbf{n}}) - \ln g(y \mid \theta) = E[\ln\{f(x \mid \theta)/g(y \mid \theta)\} \mid y, \theta^{\mathbf{n}}]$$
$$\leq E[\ln\{f(x \mid \theta^{\mathbf{n}})/g(y \mid \theta^{\mathbf{n}})\} \mid y, \mid \theta^{\mathbf{n}}) = Q(\theta^{\mathbf{n}} \mid \theta^{\mathbf{n}}) - \ln g(y \mid \theta^{\mathbf{n}}), \text{ with equality when } = n.$$

5. Thus, $Q(\theta \mid \theta^{\mathbf{n}}) - Q(\theta^{\mathbf{n}} \mid \theta^{\mathbf{n}}) + \ln g(y \mid \theta^{\mathbf{n}})$ minorizes $\ln g(y \mid \theta)$.
6. In the M step it suffices to maximize $Q(\theta \mid \theta^{\mathbf{n}})$ since the other two terms of the minimizing function do not depend on θ.

Schafer (1999) phrased the problem well when he noted "If we knew the missing values, then estimating the model parameters would be straightforward. Similarly, if we knew the parameters of the data model, then it would be possible to obtained unbiased predictions for the missing values." Here we are going to do both.

We will first estimate the parameters on the basis of the data we do have. Then we will estimate the missing data on the basis of those parameters. Then we will re-estimate the parameters based on the filled-in data, and so on. We would first take estimates of the variances, covariances and means, perhaps from listwise deletion. We would then use those estimates to solve for the regression coefficients, and then estimate missing data based on those regression

coefficients. (For example, we would use whatever data we have to estimate the regression $\hat{Y} = bX + a$, and then use X to estimate Y wherever it is missing.) This is the estimation step of the algorithm.

Having filled in missing data with these estimates, we would then use the complete data (including estimated values) to recalculate the regression coefficients. But recall that we have been worried about underestimating error in choosing our estimates. The EM algorithm gets around this by adding a bit of error to the variances it estimates, and then uses those new estimates to impute data, and so on until the solution stabilizes. At that point we have maximum likelihood estimates of the parameters, and we can use those to make the final maximum likelihood estimates of the regression coefficients.

There are alternative maximum likelihood estimators that will be better than the ones obtained by the EM algorithm, but they assume that we have an underlying model (usually the multivariate normal distribution) for the distribution of variables with missing data.

SAS Example:

32 students take six tests. These six tests are indicator measures of two ability factors: verbal and math.

Suppose now due to sickness or unexpected events, some students cannot take part in one of these tests. Now, the data test contains missing values at various locations, as indicated by the following DATA step:

```
data missing;
input x1 x2 x3 y1 y2 y3;
datalines;
23  .  16 15 14 16
29 26 23 22 18 19
14 21   . 15 16 18
20 18 17 18 21 19
25 26 22   . 21 26
26 19 15 16 17 17
  . 17 19  4  6  7
12 17 18 14 16  .
25 19 22 22 20 20
 7 12 15 10 11  8
29 24  . 14 13 16
28 24 29 19 19 21
12  9 10 18 19  .
11  . 12 15 16 16
20 14 15 24 23 16
26 25  . 24 23 24
20 16 19 22 21 20
14  . 15 17 19 23
14 20 13 24  .  .
29 24 24 21 20 18
26  . 26 28 26 23
20 23 24 22 23 22
23 24 20 23 22 18
14  . 17  . 16 14
28 34 27 25 21 21
17 12 10 14 12 16
 .  1 13 14 15 14
22 19 19 13 11 14
18 21  . 15 18 19
12 12 10 13 13 16
22 14 20 20 18 19
29 21 22 13 17  .
;
```

The maximum likelihood method, as implemented in PROC CALIS, deletes all observations with at least one missing value in the estimation. In a sense, the partially available information of

these deleted observations is wasted. This greatly reduces the efficiency of the estimation, which results in higher standard error estimates.

To fully utilize all available information from the data set with the presence of missing values, you can use the full information maximum likelihood (FIML) method in PROC CALIS, as shown in the following statements:

```
proc calis method=fiml data=missing;
 factor
 verbal ---> x1-x3,
 math ---> y1-y3;
 pvar verbal = 1., math = 1.;
 run;
```

In the PROC CALIS statement, you use METHOD=FIML to request the full-information maximum likelihood method. Instead of deleting observations with missing values, the full-information maximum likelihood method uses all available information in all observations.

Output shows some modeling information of the FIML estimation of the confirmatory factor model on the missing data.

Confirmatory Factor Model With Missing Data: FIML

FACTOR Model Specification

The CALIS Procedure

Mean and Covariance Structures: Model and Initial Values

Modeling Information

Data Set	WORK.MISSING
N Records Read	32
N Complete Records	16
N Incomplete Records	16
N Complete Obs	16
N Incomplete Obs	16
Model Type	FACTOR
Analysis	Means and Covariances

PROC CALIS shows you that the number of complete observations is 16 and the number of incomplete observations is 16 in the data set. All these observations are included in the estimation. The analysis type is 'Means and Covariances' because with full information maximum likelihood, the sample means have to be analyzed during the estimation.

Output shows the parameter estimates.

Factor Loading Matrix: Estimate / StdErr / t-value

	verbal	math
x1	5.5003	
	1.0025	
	5.4867	
	[_Parm1]	0
x2	5.7134	
	0.9956	
	5.7385	
	[_Parm2]	0
x3	4.4417	
	0.7669	
	5.7918	
	[_Parm3]	0
y1	0	4.9277
		0.6798
		7.2491
		[_Parm4]
y2	0	4.1215
		0.5716
		7.2100
		[_Parm5]
y3	0	3.3834
		0.6145
		5.5058
		[_Parm6]

Factor Covariance Matrix: Estimate/StdErr/t-value

	verbal	math
verbal	1.0000	0.5014
		0.1473
		3.4029
		[_Add01]
math	0.5014	
	0.1473	
	3.4029	
	[_Add01]	1.0000

Error Variances

Variable	Parameter	Estimate	StandardError	t Value
x1	_Add08	12.72770	4.77627	2.66478
x2	_Add09	9.35994	4.48806	2.08552
x3	_Add10	5.67393	2.69872	2.10246
y1	_Add11	1.86768	1.36676	1.36650
y2	_Add12	1.49942	0.97322	1.54067
y3	_Add13	5.24973	1.54121	3.40623

2. Mixed-Effect Model Repeated Measure (MMRM) model

By using MMRM model method, we can use all of the data we have. Missing data are no explicitly imputed. It has no effect on other scores from that same patient. It applies with a Restricted Maximum Likelihood (REML) solution to study longitudinal (repeated measures) analyses under the Missing At Random (MAR) assumption. The REML is able to give an unbiased estimate of the covariate structure where the MLE's estimation is biased. Linear Mixed Model is defined as: $\mathbf{Y_i = X_i\beta + Z_ib_i + \varepsilon_i}$

Y, **X**, and **β** are as in Simple Linear Model

$\mathbf{Z}_i$ - $\mathbf{n_i}$ **x q** known design matrix for the **random effects**

$\mathbf{b}_i$ - **q x 1** vector of **unknown random effects** parameters

ε - $\mathbf{n_i}$ **x 1** vector of **unobserved random errors**

$\mathbf{X}_i\boldsymbol{\beta}$ - denotes fixed effects

$\mathbf{Z}_i\mathbf{b}_i$ - denotes the random effects, was selected at random from the population of interest.

$\boldsymbol{\varepsilon}_i$ - denotes repeated measures effects

b ~ N_p(**0, G**) i.e., multivariate normal with mean vector **0** and covariance matrix **G**

ε ~ N_p(0, R) i.e., multivariate normal with mean vector **0** and covariance matrix **R** (repeated measures structure)

b, ε are uncorrelated

$$E\begin{bmatrix} b \\ \varepsilon \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad Var\begin{bmatrix} b \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}$$

For example, we are working on a clinical trial study on depression and we divide our patients into two groups, Group 1: Control (No drug) and Group 2: Treatment (with drugs). The fixed effect will be the treatment group (between-subjects factor) and the random effect will be time (within-subjects factor), and time*group. During the first stage, a linear regression portion of a model is introduced and describes a fixed portion. The fixed segment of the model is applicable to every individual with varying parameters respectively. The distributions of these random effects constitute the second stage of the model and construct a special covariate structure. Basically, our goal is to determine which covariance structure best fits the random variances and covariance of data.

Structure	Description	# of Parameters	{i,j}th element
AR(1)	Autoregressive(1)	2	$\sigma_{ij} = \sigma^2 \rho^{\lvert i-j \rvert}$
CS	Compound Symmetry	2	$\sigma_{ij} = \sigma_1 + \sigma^2 1(i=j)$
UN	Unstructured	t(t+1)/2	$\sigma_{ij} = \sigma_{ij}$
TOEP	Toeplitz	t	$\sigma_{ij} = \sigma_{\lvert i-j \rvert+1}$
VC	Variance Components	q	$\sigma_{ij} = \sigma_k^2 1(i=j)$ and i corresponds to the kth effect
ARH(1)	Heterogeneous AR(1)	t+1	$\sigma_{ij} = \sigma_i \sigma_j \rho^{\lvert i-j \rvert}$
CSH	Heterogeneous CS	t+1	$\sigma_{ij} = \sigma_i \sigma_j \left[\rho 1(i \neq j) + 1(i=j)\right]$
TOEPH	Heterogeneous TOEP	2t-1	$\sigma_{ij} = \sigma_i \sigma_j \rho_{\lvert i-j \rvert}$

Using SAS, we selected the best covariance structure from the results of the AIC (Akaike's information Criteria) and BIC (Schwarz's Bayesian Criteria). When using these measures, the number closest to zero is the better fit. If SAS gives conflicting results, the simpler model is probably better. In addition, we also use the F-test, but it is only approximate, and is not very accurate when there are a large number of missing data. Usually the unstructured (model) will be

the best model and it approaches to the model that fits both the treatment-by-time means E [Y] = X β and the (co)variances Var [Y] = ZGZ' + R.

$$\begin{matrix} Time\_1 \\ Time\_2 \\ Time\_3 \\ Time\_4 \end{matrix}\begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{12} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{43} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{bmatrix}$$

Covariance structures: Unstructured

The mixed model is more efficient and reliable because it is able to implement the missing data. Also, it is a greater analysis of variation both within and between individuals is available. In addition, the study of background effects is offered. However, it is more complex for software analysis which is not always easy to set up. Moreover, it take times to find out which covariance structure best fits the random variances and covariance of data.

Analysis data with different Methods

The way we test the efficiency of each methods is simple. We find a complete data set and remove some data to make an assumption some patients dropped out during the process of the study. For example, we are working on a clinical trial study on depression and we divide our patients into two groups, Group 1: Control (No drug) and Group 2: Treatment (with drugs).

Group	Subj	Base	1st	3th	6th
1	1	296	175	187	192
1	2	376	329	236	76
1	3	309	238	150	123
1	4	222	60	82	85
1	5	150	271	250	216
1	6	316	291	238	144
1	7	321	364	270	308
1	8	447	402	294	216
1	9	220	70	95	87
1	10	375	335	334	79
1	11	310	300	253	140
1	12	310	245	200	120
2	13	282	186	225	134
2	14	317	31	85	120
2	15	362	104	144	114
2	16	338	132	91	77
2	17	263	94	141	142
2	18	138	38	16	95
2	19	329	62	62	6
2	20	292	139	104	184
2	21	275	94	135	137
2	22	150	48	20	85
2	23	319	68	67	12
2	24	300	138	114	174

Group	Subj	Base	1st	3th	6th
1	1	296	175	187	192
1	2	376	329	236	76
1	3	309	238	150	123
1	4	222	60	82	85
1	5	150	.	.	.
1	6	316	291	238	144
1	7	321	364	270	308
1	8	447	402	.	.
1	9	220	70	95	87
1	10	375	335	334	79
1	11	310	300	253	.
1	12	310	245	200	120
2	13	282	186	225	134
2	14	317	31	85	120
2	15	362	104	.	.
2	16	338	132	91	77
2	17	263	94	141	142
2	18	138	38	16	95
2	19	329	.	.	.
2	20	292	139	104	.
2	21	275	94	135	137
2	22	150	48	20	85
2	23	319	68	67	.
2	24	300	138	114	174

We assume those missing data are missing completely at random (MCAR) or missing at random (MAR). Also, F-test is being used this time.

MMRM (AR(1)) is the best method comparing to the others since it has the closer result to the original full data set. So, it. The interaction time*group: the drug treatment is having a differential effect on the 2 groups.

Method	SAS Code	Pr > F group	Pr > F time*group
Full Data Set	Proc GLM	0.0012	<0.0001
Listwise Deletion	Proc GLM	0.0216	0.0037
LOCF	Proc GLM	0.0302	0.0013
BOCF	Proc GLM	0.1194	0.0019
EM Algorithm	Proc Calis	0.0213	0.0014
MMRM (UN)	Proc Mixed	0.0111	0.0015
MMRM (AR (1))	Proc Mixed	0.0068	<0.0001

- Type = CS (Compound Symmetry)
- AIC = 856.6 (smaller)
- BIC = 859.0

Fit Statistics

-2 Res Log Likelihood	852.6
AIC (smaller is better)	856.6
AICC (smaller is better)	856.8
BIC (smaller is better)	859.0

- Type = UN (Unconstructed)
- AIC = 854.1 (smaller)
- BIC = 865.9

Fit Statistics

-2 Res Log Likelihood	834.1
AIC (smaller is better)	854.1
AICC (smaller is better)	857.5
BIC (smaller is better)	865.9

- Type = AR (1) (Autoregressive (1))
- AIC = 852.2 (smallest)
- BIC = 854.6 (smallest)

Fit Statistics

-2 Res Log Likelihood	848.2
AIC (smaller is better)	852.2
AICC (smaller is better)	852.4
BIC (smaller is better)	854.6

Prevention

It is smart to avoid missing data in clinical trials before it happens. For the study design, we should select an easily realizably endpoints and do a better adjustment on sample size. Besides, we should avoid complicated and messy record keeping, adopt a flexible appointment schedule, minimize the waiting time during appointment and select a convenient location for the

participants. Before each follow up time, we can remind patients about appointments and follow-up immediately after he missed an appointment. If the participant didn't show up, we may remind him by phone or home visit him.

Conclusion

We introduced three traditional (Listwise deletion, simple imputation method – LOCF and BOCF), and two likelihood-base modern methods (EM algorithm and MMRM) on handling missing data in clinical trials. Also, we did a comparison on each methods with a real data set. Assumptions (MCAR, MAR or MNAR) were made for the missing data during using any method or procedure is applied. We found that MMRM is the best method comparing to the others since it has the closer result to the original full data set under the F-test. If we have a chance, we should find couple more data sets to verify our conclusion under the same procedure. Of course, the best solution to missing data in clinical trials is to set up some preventions before it happens. Because missing data can lead to biased estimates of treatment differences and reduces the benefit provided by randomization. We should pay more attention to focus on preventing missing data during our clinical trial study.

Reference

David C. Howell,
http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html

David C. Howell, http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Mixed-Models-Repeated/Mixed-Models-for-Repeated-Measures1.html

http://en.wikipedia.org/wiki/Analysis_of_clinical_trials

Craig H. Mallinckrodt, Peter W. Lane, Dan Schnell, Yahong Peng, James P. Mancuso, Recommendations for the Primary Analysis of Continuous Endpoints in Longitudinal Clinical Trials

James R. Carpenter, Michael G. Kenward, Missing data in randomized controlled trials – a practical guide

Guidance for Industry E9 Statistical Principles for Clinical Trials

European Medicines Agency, London, 23 April 2009, Guidline on Missing Data in Confirmatory Clinical Trials

Lancet 2005 365: 1159-1162 and Clin Trials 2004; 1:368-376.