Modeling longitudinal data with mixed

dropout mechanisms using extended pattern mixture model

By Jing Gong

A dissertation submitted to the

School of Public Health

University of Medicine and Dentistry of New Jersey

and the

Graduate School – New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

UMDNJ-School of Public Health

Awarded jointly by these institutions and

Written under the direction of

Professor Pamela Ohman-Strickland

And Approved by

_____

_____

_____

_____

Piscataway/New Brunswick, New Jersey

October, 2012

ABSTRACT OF THE DISSERTATION

Modeling longitudinal data with mixed

dropout mechanisms using extended pattern mixture model

By Jing Gong

Dissertation Director:

Pamela Ohman-Strickland

In physical and mental health research areas, longitudinal studies are popular tool for addressing outcome changes over time within and between individuals. However, monotone-type missing data caused by dropout is unavoidable in many longitudinal studies and may lead to biased inference and incorrect conclusions if the nature of dropout is ignored. The dropout process may cause three types missingness: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Most of existed statistical methods have treated the entire dropout the same, although this assumption may not be true in practice. For a real longitudinal study, based on observed dropout reasons, it may be more realistic to assume the nature of missingness to be a mixture of MCAR, MAR and MNAR.

In this thesis, two approaches are proposed to deal the mixed nature of dropout due to different longitudinal dataset's pattern and assumptions, and both of them are based on Pattern Mixture Model. If the outcome of interest can be assumed to follow a linear trend over time and the detailed dropout reasons for each subject are unknown, EM algorithm method will be added to Pattern Mixture Model to reflect a mixture of missing natures. If the outcome of interest is not linear with respect time and the detailed reasons

for each dropout are known, available-case missing value restriction and non-future-dependent missing value restriction will be used within Pattern Mixture Model to identify the distribution of unknown measurements caused by different dropout reasons. Multiple imputation method will be combined to impute multiple complete datasets to reflect the uncertainty caused by missing values.

Acknowledgement

I would like to express my great gratitude to my advisor, Dr. Pamela Ohamn-Strickland, for her encouragement, guidance and instruction in the past years. Without her constant support, the thesis would not have been completed in its current form. The mentorship provided by Dr. Ohman-Strickland has made significant impact and will continue to affect my life both professionally and personally.

I would also like to give my appreciation to my committee members Drs. Dirk Moore, Sinae Kim and Donald Hoover. Thank you all for your strong support and suggestion throughout my dissertation work. In addition, I want to thank Dr. Yong Wang, my supervisor in Forest Lab, for his valuable instruction and support in my research.

My husband Ji Xie and my lovely son Ryan Xie have given me enormous support during my Ph.D. study. My deepest gratitude and love are reserved for them. Without their effort, I would not have been able to finish the dissertation.

Table of Contents

List of Tables

List of Figures

List of Acronyms

ANCOVA	Analysis of covariance
BOCF	Baseline observation carried forward
EMA	European Medicines Agency
FDA	Food and Drug Administration
GEE	Generalized estimating equation
LOCF	Last observation carried forward
MAR	Missing at random
MCAR	Missing completely at random
MMRM	Mixed effects model for repeated measurements
MNAR	Missing not at random
MI	Multiple imputation
PMM	Pattern mixture model
SM	Select model
WGEE	Weighted generalized estimating equation

Glossary

Y_{it} = response for the i^{th} subject at time *t, i = 1,…,N, t = 1,…, T*

$Y_i = (Y_{i1},\ldots, Y_{iT})^{\mathrm{T}}$, vector of complete responses for i^{th} subject

$Y_{i.obs} = (Y_{i1},\ldots, Y_{iT_i})^{\mathrm{T}}$, vector of observed responses for ith subject

T_i = the last time point of measurements ($\leq$ T)

$X_{it} = (X_{it1}, \ldots, X_{itQ})$ Q-vector of fixed covariates for the i^{th} subject at time *t*

$X_i = (X_{i1}^T, \ldots, X_{iT_i}^T)^{\mathrm{T}}$, $J_i \times Q$ matrix of fixed covariates for i^{th} subject

$b_i = (b_{i1}^T, \ldots, b_{iT_i}^T)^{\mathrm{T}}$, $J_i \times K$ matrix of random covariates for i^{th} subject

U_i= K-vector of random effects, multi-normally distributed with mean 0 and covariance D

$\varepsilon_i = (\varepsilon_{i1},\ldots, \varepsilon_{iT_i})$ observed responses' error, normally distributed with mean 0 and covariance $\sigma^2 I_{J_i}$

$\beta^{(r)}$ = Q-vector of coefficients associated with fixed covariates for pattern r

$\pi^{(r)}$ = the probability of a subject falling in pattern r

$p^{(r)}$ = the probability of non-ignorable missingness for the rth dropout pattern

$R_i = r$ dropout pattern for ith subject

Z_i = indicator of the ignorable missingness or not for the ith subject, $Z_i = 1$ if subject i's missing responses are non-ignorable ($probability = p^{(r)}$) and $Z_i = 0$ if ignorable ($probability = 1 - p^{(r)}$).

1 Introduction

Longitudinal studies are widely used in clinical trial and public health to measure outcomes over time within individuals as well as differences in changes across individuals. Unfortunately, missing data are prevalent in longitudinal studies. The inference based solely upon the observed outcomes may be seriously flawed. So how to handle missing data is a very important problem in longitudinal studies. According to huge papers, it is clear that no universal method can be considered definitive for any missingness scenarios. Instead, different methods need to rely on the different patterns and types of missing data, the latter one is also known as missing data mechanisms.

There are two patterns of missing data. If at least some outcomes of interest are observed again after a missing value occurs, this is called intermittently missingness or non-monotone missingness. Alternatively, individuals may miss one visit and never return due to multiple reasons, such as relocation, death, side effect, no effect, and so on, which is called dropout or monotone missingness. In clinical trials, dropout usually causes much higher percentage of missing values than intermittently missingness. Moreover, dropout can't provide the important outcomes after the time of dropout, especially the last observation which is sometimes the researchers are mostly interested. Specifically, missingness caused by dropouts in clinical trials can destroy the benefit provided by randomization, which is typically used to make comparison between the treatment and the control groups as fair as possible. So dropout may cause more serious problem than intermittently missingness if they are ignored in studies' analysis. Properly handling and analyzing missing data caused by dropout is crucial to the validity of the conclusions

drawn from any longitudinal study. Thus in my thesis, I will focus on the methods for handling monotone missingness caused by dropout in longitudinal studies.

Many methods have been proposed for handling dropouts. The first key in selecting effective or meaningful methods is to make clear the dropout mechanisms of data. Traditionally the dropout mechanisms are divided into three types (Little, 1995; Little and Rubin, 2002; Diggle, 1994): (1) Missing Completely at Random (MCAR), in which the probability of dropout doesn't depend on the observed data and the missing data; (2) Missing at Random (MAR), in which the probability of dropout depends on the observed data, and (3) Missing Not at Random (MNAR), where the probability of dropout depends on the missing data and possibly the observed data.

A variety of literature focuses on regression-based methods for the longitudinal data with missingness caused by dropouts. For example, to MCAR data, standard methods include generalized estimating equations (GEE) (Liang and Zeger, 1986), complete case analysis (CC) (also known as listwise deletion), or simple imputation, such as regression imputation and hot deck imputation (Molenberghs, 2004). Among these methods, GEE is a better choice because it provides a more robust inferential procedure than other methods.

Compared to MCAR, MAR encompasses more dropout mechanisms thus lessening the assumptions for missingness. Mixed models, weighted generalized estimating equations(WGEE), and multiple imputation (MI) may be suitable methods used to handle

MAR missingness (Hogan, 2004; Panel on Handling Missing Data in Clinical Trials, 2010).

MNAR is the most complicated missing procedure, in which researchers need to make many unverifiable assumptions about the data distribution of the unobserved value. Actually, there are no existing methods that can be used to differentiate whether missing data are MAR or MNAR. So the MNAR methods usually are used as sensitivity analyses for MAR methods (Hogan, 2004). The majority of approaches for MNAR are based on models for the joint distribution of the outcome and the dropout mechanism. Most of them are likelihood-based methods and can be classified into three categories: selection models, pattern-mixture models and frailty models (Little 1995).

Last observation carried forward (LOCF) or baseline observation carried forward (BOCF) methods used to be considered providing valid and conservative results under MCAR or MAR assumptions. But actually, they do make MNAR assumption and not always conservative (details are discussed in the next chapter). Food and Drug Administration (FDA) recommends not using either of these approaches as the primary approach for handing missing data in clinical trials unless the assumptions that underlie them are scientifically justified (Panel on handling missing data in clinical trials, 2010).

Although there is an enormous literature on missing data methods for longitudinal studies, most of them have treated all observations with dropout as if they fall within the same dropout type. In practice, if we review the dropout reasons carefully, we would find

that different dropout reasons may be related to the outcomes in different ways, for example, in an implying different types of dropout HIV study, some patients may drop out because of “lost job”, or “moved from study area”, or “unrelated illness required treatment with contraindicated medicine”, which is not related to the outcome value and will cause MCAR. On the other hand some patients may drop out because they “did not get better”, which is clearly outcome related and will cause MAR or MNAR. It follows that in studies with either mixed dropout reasons or without clear dropout reasons, allowing mixed dropout mechanisms may be appropriate. In these situations, assuming one dropout mechanism may lead to biased inference. Methods of accounting for multiple dropout processes at the same time may lead to less bias when evaluating the longitudinal data. European Medicines Agency expressly pointed out that using different techniques for different dropout reasons is an attractive approach (Guideline on Missing Data in Confirmatory Clinical Trials, 2011).

Thus, my research is intended to investigate regression-based methods, specifically extensions of PMMs for handling monotone missingness caused by mixed dropout mechanisms and the potential impact on inference.

The remainder of this section presents two motivating examples with possibly mixed dropout mechanisms, followed by my thesis objectives. Chapter 2 discusses the effects of dropouts on data analyses and interpretation, describes the different types of dropout mechanisms in details, and introduces some regression-based statistical analysis methods for dealing with dropouts. Chapter 3 provides the details on one of my proposed

statistical approaches for handling mixed dropout mechanisms: EM algorithm used in Pattern Mixture - Within – Mixture model. Simulation results and a case study example are also presented in this section. Chapter 4 describes the other method I proposed – Pattern Mixture Models based multiple imputation with mixed missing values restrictions (non-future dependent and available-case) for handling mixed dropout problem, and a simulated chronic pain study is analyzed by different methods. Discussion and future work are provided in Chapter 5. And some methodological details and codes are given in an Appendix.

1.1 Motivating examples

Example 1: Schizophrenia study

Schizophrenia is a mental disorder that makes it difficult for a patient to tell the difference between real and unreal experiences, to think logically, to have normal emotional responses, and to behave normally in social situations. The National Institute of Mental Health Schizophrenia Collaborative Study collected longitudinal data on treatment related changes in overall severity (Hedeker and Gibbons,1997).

The main outcome of interest was drawn from the Inpatient Multidimensional Psychiatric Scale (IMPS), a series of questions which enables immediate stratification into psychotic types and the degree of mental illness. Specially, item #79 of IMPS, severity of illness can be used to evaluate the effect of treatment for Schizophrenia. It is scored as 1=normal, not at all ill; 2=borderline mentally ill; 3=mildly ill; 4=moderately ill; 5=markedly ill; 6=severely ill; and 7=among the most extremely ill. In this study,

437 patients were randomly assigned to receive one of four medications: placebo (108 patients), chlorpromazine, flupenzaine or thioridazine. Since Hedeker's previous analyses showed that the three drugs have similar effects on Schizophrenia, they were combined into one drug group (329 patients) for comparison with placebo here. Patients' IMPS were supposed to be measured every week. However since most of the measurements occurred at weeks 0, 1, 3, and 6 (see Table 1.1), we will focus on these weeks. If the completers are defined as those who were measured at week 6, the percentages of patients who dropped from the study were 34% and 19% for the placebo and drug groups respectively.

Hedeker and Gibbons used the PMMs with two patterns (one is for dropouts and one is for completers) to analyze the longitudinal IMPS #79 scores, assumed that all the dropouts are MNAR and the linear effects of square root of time and treatment by square root of time interaction to the IMPS #79. But here, I assume that the dropouts are under mixed dropout mechanisms.

Detailed dropout reasons for each patient are not publicized. However, based on the general nature of a longitudinal study for this kind of disease, it may be reasonable to assume that among all dropouts, some dropouts may cause MCAR or MAR, while some dropouts may cause MNAR. My first proposed method for handling mixed dropout will be used on analyzing this study's data in Chapter 3. Other popular methods, such Analysis of Covariance (ANCOVA) for complete case, Mixed Model for Repeated Measurements (MMRM), Generalized Estimating Equation (GEE), Weighted GEE, Last

Observation Carried Forward (LOCF), Baseline Observation Carried Forward (BOCF) and Pattern Mixture Model (PMM) will also be used to this data. And results from all methods will be compared.

Table 1.1. The experimental design and corresponding sample sizes:

	Sample size at Week							
	0	1	2	3	4	5	6	Completers
Placebo (n=108)	107	105	5	87	2	2	70	66%
Drug (n=329)	327	321	9	287	9	7	265	81%

Example 2: Simulated chronic pain study

Chronic pain is pain lasting more than 3 months, which may be caused by many reasons, such as Fibromyalgia, Diabetes, injury and so on. Due to confidential reason, a real data from a chronic pain clinical trial can't be used in my thesis. Thus, a similar study data mimicking is simulated from the real study through some transformation and randomization steps. This simulated study is to demonstrate the efficacy of test drug in the treatment of chronic pain. In detail, the simulated study mimics a multicenter, randomized, double-blind, placebo-controlled, monotherapy, and parallel-group phase III study for treating chronic pain. It has two treatment groups and 4 post-baseline-visits (weeks 4, 8, 12, 16). One thousand patients have been randomized in a 1:1 ratio to one of the two treatment groups - placebo and test drug. The population for the primary analysis consists of all patients with a baseline value and at least one post baseline measurement of the primary efficacy variable. The primary outcome measure is the

change from baseline in pain score. The range of pain scores is from 0 (no pain) to 10 (the highest level pain).

According to the mimic data, the baseline and measures at each post visit can be treated as normal distributions. And all the changes from baseline can be assumed as multivariate normal distribution. There are over 30% dropouts in this study, and slightly more dropouts occur in drug group (33.4%) than in placebo group (30.6%). Detailed dropout reasons for each patient are also simulated from the real study. Reasons includes: adverse events, lack of efficacy, lost contact, protocol violation, non-compliancy with respect to protocol requirements, and others. Some of these reasons may not be related to the patient's pain scores or may only be related to the observed pain scores, such as lost contact, protocol violation and so on. Dropouts due to these reasons can be assigned into MCAR and MAR. Alternatively, some reasons such as adverse events may have relationships with the missing measures of pain, causing MNAR. In this case, this study exhibits mixed dropout mechanisms with known reasons for dropout. Thus, my second proposed method will be used to analyze this data in Chapter 4, and the result will be also compared with the ones from other established analysis methods.

1.2 Objectives

Whether the detail dropout reasons for each subject are well known or unknown, longitudinal studies may actually included mixture of missing or dropout mechanisms. However, almost all the proposed regression-based methods for missing data have treated

the entire dropouts or the missingness the same. My thesis work is to fill in this gap between the practice and the existing statistical methods. The objectives of this thesis proposal are listed as follows:

1. Investigate proper regression-based methods for handling monotone missingness with multiply-caused dropout under the situations where
    a. The detail dropout reasons are not clear, and a function between the outcome of interest and time can be assumed.
    b. The detail dropout reasons are well known, although there is no function between the outcome of interest and time.
2. Compare my methods with other existing methods. See how the missingness or dropout mechanisms affect the results of analyses.

2 The effects of dropouts, dropout mechanisms and major regression-based statistical methods for handling dropouts in a longitudinal data

My thesis problem and objectives are generally described in the last Chapter. Now I will formalize the details about the effects of dropouts, dropout mechanisms, and some popular methods used for handling monotone missingness caused by dropouts.

2.1 The effects of dropout on longitudinal study

The most important problem resulting from dropout is the potential bias. In a clinical trial, randomization is applied to obtain balanced comparable treatment groups. However, if we ignore the dropouts and only analyze the completers, the treatment groups may be incomparable, and the estimation of the treatment effect may be biased. For example, if female patients with fewer efficacies are more likely to drop from study, then the gender rates in treatment groups will be unbalanced, and the analysis results will prefer to tested treatment. In addition, the completers may not be able to represent the target population, and then the external validity of this clinical trial is affected. For example, some patients drop out because they are not able to tolerate tested drug, and then cannot obtain benefits from this drug. If we simply ignore these patients from analysis, we may get biased results and target population.

Dropout may also lead to decreased power and variability. It's clearly that the sample size and the variability of the outcomes of interest affect the power of a longitudinal

study. If dropouts are simply excluded from the analysis, it will result in a reduction of statistical power. On the other hand, dropouts might be more likely to have extreme values, for example, due to lack of efficacy, some dropouts have much worse outcomes than completers'. Hence, excluding these dropouts from the analysis may lead to the underestimation of variability and then narrow the confidence interval for the treatment effect.

2.2 Dropout mechanisms

Under different dropout mechanisms, dropouts may lead to various effects in different statistical methods. To illustrate the various dropout mechanisms, we consider a longitudinal study with some dropouts. Suppose the full data consist of T repeated measurements taken on N subjects at a fixed set of time points $1, 2, \ldots, T$. The number and timing of measurements for all subjects is assumed to be equal. For i^{th} subject ($i=1,2,..,N$), his/her observed measurements can be represented by $Y_{i(obs)} = (Y_{i1},\ldots, Y_{iT_i})^T$, where T_i is the last time point of measurements ($T_i \leq T$). Subject i dropped out between time points T_i and T_i+1 if $T_i < T$, and his/her missed measurements is $Y_{i(mis)} = (Y_{i(T_i+1)},\ldots,Y_{iT})^T$. The full response vector for i^{th} subject combine $Y_{i(obs)}$ and $Y_{i(mis)}$ to get $Y_i = (Y_{i1},\ldots, Y_{iT})^T$. Let X_i be the vector of covariates for subject i, and Y_{it} the measurement for subject i at time point t. M_{it} is an indicator to express whether measurement Y_{it} is observed or not. M_{it} = 1 if Y_{it} is observed, and M_{it} = 0 if Y_{it} is missing. Define R_i as the indicator of dropout pattern, and $f(Y_i, R_i| X_i)$ as the full data density for subject i. The dropout mechanisms are presented in accordance with Little (1995).

Missing completely at random (MCAR):

The probability of dropout is independent of the observed data and the missing data. That is

$$f(R_i|Y_i, X_i) = f(R_i) \qquad (2.2\text{-}1)$$

A typical example is that a subject moved to a different location where the treatment can't be continued.

MCAR is simplest but most restrictive dropout mechanism. Under MCAR, the observed data can be treated as a random sample of all data. If a data set is MCAR, there is no impact on bias and most standard approaches of analysis are valid (e.g. complete case analysis, generalized estimating equation, etc.). Little's test (Little, 1988) can be used to test whether missingness is MCAR or not. This test compares the distribution of observed variables between dropouts and completers. If there is no significant difference found with respect to the variables, then MCAR can be assumed. Unfortunately, MCAR is often not plausible in longitudinal studies, especially in clinical trials.

Missing at random (MAR):

The probability of dropout is only dependent on the observed data (i.e. observed outcome of interest or covariates such as treatment) but not dependent on missing data. That is

$$f(R_i|Y_i, X_i) = f(R_i|Y_{i(obs)}, X_i) \qquad (2.2\text{-}2)$$

Here $Y_{i(obs)}$ is the observed dependent response vector, X_i is the observed covariate vector. MAR can be referred to outcome-dependent MAR and/or covariate-dependent MAR (DeSiyza 2009). Some researchers think that dropout due to the lack of efficacy will

cause MAR. For example, in a phase III study for controlling high blood pressure, some of patients drop out from the study because their diastolic blood pressures are still high after taking treatment. Hence the dropout process is related to the observed data.

MAR is a more realistic assumption than MCAR in a real study. The primary approach for clinical trial is usually based on this assumption. Under MAR, observed cases are no longer a random sample of full data, thus some standard approaches of analysis (e.g. complete case analysis, GEE) can't provide unbiased and efficient results for this situation any more. Mixed effects model for repeated measurements (MMRM), weighted generalized estimating equation (WGEE) and multiple imputation (MI) are three available and popular methods for MAR.

Likelihood and Bayesian approaches (e.g. MMRM and MI) are dependent on parametric models to catch the inference of interest. The joint distribution of measurement process and dropout process can be obtained from a product of two conditional distributions, one for full measurement data and one for dropout process. That is

$$f_{\theta,\varphi}(Y,R|X) = f_{\theta}(Y|X)f_{\varphi}(R|Y,X) \qquad (2.2\text{-}3)$$

θ presents the parameters used in model $f_{\theta}(Y|X)$ for the measurement process and φ is the parameters used in model $f_{\varphi}(R|Y,X)$ for dropout process. The likelihood function for (θ, φ) given Y_{obs}, Y_{mis}, X and R is obtained through averaging or integrating over all possible values of the missingness Y_{mis}.

$$L(\theta,\varphi|Y_{obs},X,R) = \int f_{\theta}(Y_{obs},Y_{mis}|X)f_{\varphi}(R|Y_{obs},Y_{mis},X)d\,Y_{mis} \qquad (2.2\text{-}4)$$

If the dropout process is MCAR or MAR, then the second term in the right side of equation 2.2-4 may be replaced by $f_\varphi(R|X)$ or $f_\varphi(R|Y_{obs}, X)$, respectively. If θ and φ are functionally independent, then the likelihood functions will be

$$L(\theta, \varphi|Y_{obs}, X, R) = f_\varphi(R|X) \int f_\theta(Y_{obs}, Y_{mis}|X) d\, Y_{mis} = f_\varphi(R|X) f_\theta(Y_{obs}|X)$$

for MCAR, and

$$L(\theta, \varphi|Y_{obs}, X, R) = f_\varphi(R|Y_{obs}, X) \int f_\theta(Y_{obs}, Y_{mis}|X) d\, Y_{mis} = f_\varphi(R|Y_{obs}, X) f_\theta(Y_{obs}|X)$$

for MAR.

According to above likelihood functions, we find that the parameters of interest, θ, for the measurement process are not dependent on the function for the dropout process under MCAR or MAR. Therefore, when either a likelihood-based method or a Bayesian approach are used, and when θ and φ are functionally independent, MCAR and MAR result in ignorable missingness since we don't need to model dropout process to get the estimates for θ.

However, if we use moment-based approaches (frequentist statistical procedures) (e.g. GEE), dropout process can't be ignored under MAR assumption. GEE doesn't specify a model for the whole multivariate distribution of a data vector, but only models the first moment mean response $E(Y_{it})$ at each visit t for the i^{th} subject. Thus, the dropout procedure needs to be considered in GEE to reduce the influence of dropout under MAR assumption. Weighted GEE (Robins, 1995) modifies the standard GEE through weights to address the missingness in the data under MAR assumption.

Missing not at random (MNAR):

Under MNAR, the probability of dropouts is dependent on the unobserved data and possibly the observed data. That is

$$f(R_i|Y_i, X_i) = f(R_i|Y_{i(mis)}, Y_{i(obs)}, X_i) \qquad (2.2\text{-}5)$$

For example, a subject drops out of the study because he/she doesn't feel well due to his/her unobserved health status (e.g. side effects of a treatment or the progression of the disease) after visit T_i.

Under MNAR, the dropout procedure is also dependent on the missing values given observed measures. According to the equation 2.2.-4, the parameters θ for the measurement process is not independent with the function for the dropout process. Thus MNAR is also called non-ignorable missing.

Since we don't know the exact value of the unobserved outcome, it's impossible to test whether the dropout is only dependent on the observed data (MAR) or dependent on missingness given observed data (MNAR). The analyses under MNAR need many strong untestable assumptions, so they are not supposed to be used as primary analysis in clinical trials. MAR is a more practical (compared to MCAR) and simpler (compared to MNAR) assumption for the dropout mechanism, so usually, a method under MAR assumption is set as primary analysis in a clinical trial, and then some analyses under MNAR (e.g. PMMs) and MCAR (e.g. ANCOVA) are used as sensitivity analyses for supporting the results obtained from primary analysis.

When data is MNAR, the simple and standard approaches will lead to biased and inefficient results. More complex methods based on likelihood such as Selection Models (SM), Pattern Mixture Models (PMM) and Shared-parameter models (or Frailty models) (Wu 1988, 1989, Follmann and Wu 1995, Albert and Follmann 2000) can be used for handling MNAR. All of them need to specify the joint distribution of the measurement process and dropout process. The former two methods SM and PMM are more popular used by researchers and will be discussed in the next section.

2.3 Major regression-based statistical methods

There are many methods have been proposed for handling longitudinal data with dropouts. Here, the most popular methods are introduced, including Complete Case Analysis (CC), Generalized Estimating Equation (GEE), Weighted GEE (WGEE), Mixed Model for Repeated Measures (MMRM), Last Observation Carried Forward (LOCF), Baseline Observation Carried Forward (BOCF), Multiple Imputation (MI), Selection Models (SM) and Pattern Mixture Models (PMM). Several ways can be used to classify these methods. For example, if we focus on which cases the method uses, we will find that some methods only use complete cases (e.g. CC), some methods use available cases (e.g. GEE, WGEE, MMRM, SM, PMM), and some methods based on imputed data (e.g. LOCF, BOCF, MI). If we classify the methods by from theoretical approach, we can separate them into likelihood –based (e.g. MMRM, SM, PMM), moment-based (e.g. GEE, WGEE), or Bayesian approaches (e.g. MI). However, the most important classification of these methods is based on dropout mechanisms. CC and GEE only can

be used under MCAR assumption; MMRM, WGEE, and MI are available under MAR assumption; and SM and PMM are popular methods for MNAR. LOCF and BOCF used to be thought are valid under MCAR or MAR, but they actually make strong MNAR assumptions. These methods will be introduced and discussed one by one.

2.3.1 Traditional Methods for dropout

2.3.1.1 Complete Case analysis (CC)

CC analysis used to be a popular method for handling missing data. This approach only includes the completers in the analysis. There are two scenarios in CC analysis: (1) Using those patients who have measurements at each scheduled time point in a longitudinal study. For example, repeated measures analysis of variance/covariance (ANOVA/ANCOVA) needs the subjects with all measurements to estimate the effect of covariates for a longitudinal study. (2) Using those patients who have the observed values at the time point that the researchers are interested (e.g. endpoint). For example, if the researchers want to know the efficacy of treatment at the end of study 3-month, then ANCOVA includes all cases with observed values at 3-months in model.

An obvious advantage of CC analysis is the simplicity of implementation. In addition, it can provide unbiased and valid results under MCAR. However, since it only includes completers in analysis, the smaller samples will reduce the statistical power and lead to inefficient estimates. If the data is not MCAR, then this analysis can produce biased results. For example, in a confirmatory clinical trial, the biggest concern with this

analysis may be a violation of the intention to treat principle (Guideline on missing data in confirmatory clinical trial 2010, Myers 2000). Imaging this scenario, treatment A has modest effect for reducing blood pressure no matter the baseline is severe or not, while treatment B provide a better effect for patients with less severe high blood pressure, but no effect for patients with severe high blood pressure. If the patients are like to drop out because of lack of efficacy, then CC analysis may produce bias and favor treatment B.

CC analysis has many disadvantages such as those mentioned above. Thus, it is not recommended as the primary analysis in clinical trials (Panel on handling missing data in clinical trials 2010, Guideline on missing data in confirmatory clinical trial 2010). However, this method still may be used in some situations. For example, it can be used as a sensitivity analysis to support the robustness of conclusions.

For population inferences, using the data from all subjects generally leads to more efficient and less biased results. The following methods use all observed cases for analysis.

2.3.1.2 Generalized Estimating Equations (GEE)

As a semi-parametric regression approach using moment-based inference, GEE was first introduced by Liang and Zeger in 1986. It is an extension of generalized linear models that account for correlated responses. Instead of attempting to specify a model for the whole multivariate distribution of a data vector, GEE only models the first moment,

specifically the mean response $E(Y_{it})$ at each visit t for the i^{th} subject. The part of the model that specifies the correlation is treated as a nuisance and not of scientific interest. In particular, GEE are specified via the following components:

- Mean response: $E(y_{it}) = \mu_{it}$
- Link function: $g(\mu_{it}) = X_{it}\beta$, β is a vector of fixed effects, X_{it} are covariates
- Variance of Y_{it}: $Var(Y_{it}) = \phi V(\mu_{it})$ such that $V(\mu_{it})$ is the usual variance predicted by the model for Y_{it} as a function of the mean and ϕ is an unknown dispersion parameter
- Variance matrix for repeated measurements on i^{th} subject: $\phi A_i = \phi diag[V(\mu_{i1}), \dots, V(\mu_{iT})]$
- Working covariance matrix for Y_i: $V_i(\alpha) = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}$, where $R_i(\alpha)$ is a "working correlation matrix" representing a guess at the true correlation structure for repeated measurements on i^{th} subject (e.g. independent, exchangeable, autoregressive, or unstructured).

Instead of a convenient likelihood, the following generalized estimating equation (GEE) is used to produce the estimates of β:

$$S(\beta, \alpha) = \sum_{i=1}^{n} \frac{\partial \mu_i'}{\partial \beta} [V_i(\alpha)]^{-1} \left(Y_i - \mu_i(\beta)\right) = 0 \qquad (2.3.1.2\text{-}1)$$

The estimators of β are consistent even if the working correlation matrix is not correct. However, a poor estimation of the correlation matrix may affect the efficiency of the estimators of β. So a sandwich estimator can be used to obtain a good estimate of $Cov(\hat{\beta})$ in large samples regardless of the true correlation model (White, 1982; Liang & Zegar, 1986). This is given as

$$cov(\hat{\beta}) = [X'\hat{V}X]^{-1}[\sum_i X_i'(Y_i - \widehat{\mu_i})(Y_i - \widehat{\mu_i})'X_i][X'\hat{V}X]^{-1} \quad (2.3.1.2\text{-}2)$$

While GEE relaxes the multivariate distribution assumptions about the data, it imposes stronger assumptions on the dropout mechanism. If dropouts are MCAR, both GEE and sandwich estimators are robust, regardless of whether the assumed working correlation matrix correctly specifies the true correlation structure. But if dropouts are MAR, GEE estimator is only consistent when using the true correlation structure for $V_i(\alpha)$ in equation 2.3.1.2-1, and sandwich estimator may not be robust even if working assumption is correct (Kenward 1998).

2.3.1.3 Weighted Generalized Estimating Equations (WGEE)

For adjusting the influence of dropout under MAR assumption, WGEE is proposed (Robins 1994, 1995) to modify standard GEE. It weights each subject's measurements *Y_i's* in the GEEs by the inverse of the probability which the subject drops at the observed dropout time, resulting in unbiased estimates of β.

Usually, the probability of dropout at any particular time point *t* for subject *i* is assumed to follow a logistic regression model with

$$logit\ P(M_{it} = 1 | X_{it}, Y_{i,t}) = \varphi_0 + \varphi_1 Y_{i,t} + \varphi_2 X_{it} \quad (2.3.1.3\text{-}1)$$

where M_{it} is the observed outcome indicator ($M_{ij}=1$ if observed, $M_{ij}=0$ if discarded), X_{it} are the covariates for i^{th} subject related to the probability of dropout (e.g. treatment), and $Y_{i,t}$ is the last observed response. Then for subject *i*, the probability he/she drops at time T_i, is given by

$$Prob(M_{iT_i} = 1|X_{it}, Y_{i,t}) = \pi_i$$

$= \{\prod_{t=1}^{T_i-1}[P(M_{i(t+1)} = 1|\, M_{it} = 1, X_{it}, Y_{i,t})]\} * [1 - P(M_{iT_i} = 1|M_{i(T_i-1)} =$

1,XiTi,Yi,Ti−1)]I(Ti≤T),

π_i is estimated by replacing φ_0, φ_1, and $\varphi_2$by their maximum likelihood estimates to yield $\hat{\pi}_i^{-1}$. The weights for the estimating equations are given by $\widehat{\omega}_i = \hat{\pi}_i^{-1}$.

Measurement model for outcomes of interest is the same as in standard GEE. The weights gotten from dropout model are combined with measurement model, and the estimate of β is the solution of the weighted GEE:

$$S(\beta, \alpha) = \sum_{i=1}^{n} \omega_i \frac{\partial \mu_i'}{\partial \beta} [V_i\,(\alpha)]^{-1}\big(Y_i - \mu_i(\beta)\big) = 0 \qquad (2.3.1.3\text{-}2)$$

In WGEE, we discard the missing observations and reweight the remaining (observed) to make them more representative. Under MAR assumption, WGEE yields consistent results when the dropout and mean models are correct and sample size is large.

2.3.1.4 Mixed Model for Repeated Measures (MMRM)

The MMRM model is also known as normal random-effects model, multilevel model, hierarchical linear model or Laird-Ware model (Laird and Ware 1982), intended for continuous and normally distributed outcomes. As a likelihood-based approach, a mixed model contains both fixed and random effects and offers a valid analysis for longitudinal

data under the MAR assumption (Cnaan 1997, Verbeke 2000). For a given individual *i* with T_i repeated measurements, the mixed model is

$$Y_i = X_i\beta + b_iU_i + \varepsilon_i \qquad (2.3.1.4\text{-}1)$$

where

- Y_i is T_i*1 repeated measures' vector;
- X_i is the matrix (T_i*p) of fixed effects covariates, where p is the number of predictors;
- β is the vector of unknown fixed effects;
- b_i is the matrix (T_i*q) of random effects covariates, where q is the number of random effects;
- $\varepsilon_i$is the vector of within-subject random errors follows multivariate normal distribution with mean vector 0 and covariance matrix σ^2V_i, $\varepsilon_i \sim MVN(0, \sigma^2V_i)$, often we use $V_i = I_{T_i}$the T_i* T_i identity matrix.
- U_i is the between-subject random components that assumed to follow a multivariate normal distribution given as $U_i \sim MVN(0, D)$, where D is the between-subject covariance matrix which is typically assumed to be unstructured.

Under these assumptions, the conditional distribution of Y_i given U_i is

$$f(Y_i|U_i; \Psi) = MVN(X_i\beta + U_ib_i, \sigma^2I_{T_i}) \qquad (2.3.1.4\text{-}2)$$

where Ψ is the complete set of unknown parameters $\{\beta, \sigma^2, D\}$.

By integrating over the random effects, the marginal distribution of Y_i can be demonstrated to be as follows:

$$f(Y_i|\Psi) = MVN(X_i\beta, b_i D b_i' + \sigma^2 I_{T_i}) \qquad (2.3.1.4\text{-}3)$$

As we discussed in section 2.2., in likelihood based or Bayesian method, the parameters of interest for measurement process doesn't depend on the function for the dropout process under MCAR or MAR assumption. Therefore, when the parameters for measurement process and for dropout process are functionally independent, the dropout process can be ignored under MCAR and MAR assumption. Since MMRM is a likelihood based method, it can provide valid results under MAR/MCAR assumption without any justification for the dropout procedure.

2.3.1.5 Last observation carried forward (LOCF)

Last observation carried forward (LOCF) is one widely used single imputation method for incomplete longitudinal data. It imputes all the missing values with the last observed value and assumes that the outcomes would not have changed from the last observed value (Molenberghs, 2004). After imputing all the missing values, usually analysis of (co)variance (AN(C)OVA) model is applied to analyze the imputed full data set. LOCF can be illustrated in the next two tables.

Table 2.1. Dropouts examples

ID	Baseline	Month 1	Month 2	Month3
02-001-1001	135	130		
02-001-1002	120	115	105	
02-001-1003	140	130	120	120

Table 2.2. LOCF to fill missing data

ID	Baseline	Month 1	Month 2	Month3
02-001-1001	135	130	→ 130	→ 130
02-001-1002	120	115	105	→ 105
02-001-1003	140	130	120	120

This method is sometimes wrongly considered to be valid under the assumptions of MCAR or MAR, but actually it assumes MNAR (Panel on Handling Missing Data in Clinical Trials, 2010). The reasons are simply explained in the following:

Assume there are *T* repeated measurements in a longitudinal study, and subjects drop out at last visit *T*. m_T is an indicator for the missingness at visit *J* (m_T =1 missing, and m_T =0 observed). Under MCAR, observed data form a random subsample of the full data. While under MAR assumption, the observed data form a random subsample of the data within a subclass defined by the observed data. Based on these two assumptions, the conditional density of missing Y_J given *X* and Y_{obs} is the same as the one for observed Y_T. In mathematical notation,

$$f(Y_T|X, Y_{obs}, m_T = 1) = f(Y_T|X, Y_{obs}, m_T = 0)$$

By contrast, LOCF assumes missing Y_J is equal to Y_{J-1} with the probability 1. Obviously,

$$f(Y_T = Y_{T-1}|X, Y_{obs}, m_T = 1) \neq f(Y_T|X, Y_{obs}, m_T = 0)$$

So LOCF doesn't make a MCAR or MAR assumption as some researchers used to think.

There is another popular and historical view about LOCF that is faulty: some have assumed that it is a conservative method for evaluating results of a clinical trial, understating differences in estimated time-trends between treatment groups. Actually, depending on the nature of the disease and tested treatment, LOCF may overestimate the

treatment effect in a clinical trial with dropouts (Molenberghs 2004, Shao 2003, Myers 2000). For example, if patients' disease is expected to grow worse over time (Guideline on missing data in confirmatory clinical trial 2010), and if patients in active treatment group are dropping out early due to adverse events, the LOCF analysis may produce anti-conservative results with respect to active treatment group. On the other hand, other clinical situations, LOCF may be conservative and underestimate the treatment effect. For example, patients' disease is expected to improve over time, and more patients drop out earlier in tested treatment group. In addition, LOCF will artificially reduce the standard error of the outcome of interest, especially with outcomes that have high variation within a subject, implying that this method is not necessarily conservative.

If the rates of dropout or time to dropout are different between treatment groups, and /or dropouts occur early but the outcome of interest is expected to change over time, then LOCF can produce biased results. Thus LOCF is not recommended as the primary analysis for clinical trial from a regulatory perspective, unless the assumptions that underlie it are scientifically justified or it clearly provides conservative results (Guideline on missing data in confirmatory clinical trial 2010). Although LOCF is not a good choice for primary analysis, it is a convenient candidate for sensitivity analysis, especially when accompanied by a discussion of appropriate assumptions.

2.3.1.6 Baseline Observation Carried Forward (LOCF)

Baseline observation carried forward (LOCF) is another single imputation method used for some incomplete longitudinal data. It imputes all the missing values with the baseline not the last observation (See Table 2.3).

Table 2.3. BOCF to fill missing data

ID	Baseline	Month 1	Month 2	Month3
02-001-1001	135	130	→ 135	→ 135
02-001-1002	120	115	105	→ 120
02-001-1003	140	130	120	120

This method has the similar problems as LOCF is that it is not valid under MCAR and MAR, not always conservative, and risks underestimation of the variability of outcome of interest. It is also not recommended as the primary analysis method for clinical trials, unless particular assumptions are reasonable.

There are some other single imputation methods, such as regression imputation, unconditional or conditional mean imputation, hot deck imputation (Molenberghs, 2004), worst and/or best case imputation (e.g. imputing the dropouts in active treatment group with the worst possible value of the outcome among the active treatment group, but assigning the best possible value among placebo group to dropouts in placebo group), and so on. Most of the single imputation methods ignore the uncertainty resulting from missingness, and thus create artificially smaller standard errors (Schafer, 1997). This risk of underestimating the variance of treatment effect can be reduced by multiple imputation method (Rubin 1987).

2.3.1.7 Multiple Imputation

Instead of filling in a single numerical value for each missing value, the multiple imputation technique replaces each missing with a set of plausible values that represent the uncertainty in predicting that missing value. Generally, multiple imputation involves three steps:

1. Imputation step: using an appropriate stochastic model to impute the missing values in a data set *m* times, resulting in *m* complete data sets
2. Analysis step: applying appropriate statistical procedures (e.g. ANCOVA, GEE, MMRM) for each data set
3. Combination step: Combining the results from these *m* complete data sets to obtain the overall statistical inferences

The combined result from the analyses for each complete data set reflects the extra variability. The total variability consists of both within and between imputation variability. Moreover, MI provides more efficient point estimates than single imputation because it reduces sampling variance through averaging over *m* data sets, even for very small *m*.

The imputation efficiency can be measure by the relative efficiency (RE) (Rubin, 1987). That is

$$RE = (1 + \frac{\lambda}{m})^{-1} \qquad (2.3.1.7\text{-}1)$$

Here λ is the rate of missing information. Table 2.4 shows relative efficiency RE of MI with different λ and *m.* Clearly, only a small number of imputations are necessary to get high RE. For example, if there is 50% of missingness, 10 times imputation can reach imputation efficiency 0.95.

Table 2.4. Relative Efficiency of MI

	λ			
m	10%	30%	50%	70%
3	0.97	0.91	0.86	0.81
5	0.98	0.94	0.91	0.88
10	0.99	0.97	0.95	0.93
20	1.00	0.99	0.98	0.97

There are many tools can be used to multiply impute missing values. Regression, predictive mean matching, propensity score, logistic regression, discriminant function and MCMC data augmentation are six methods available in SAS's multiple imputation procedure for generating multiple imputations for an incomplete data set. MCMC data augmentation and regression model are simple choice to impute continuous data with multivariate normal distribution assumption.

After obtaining *m* imputations of Y_{mis}, appropriate statistical method (e.g. MMRM, GEE, etc.) can be used to analyze the *m* completed data sets and the results are combined.

According to the discussion in section 2.2 dropout mechanisms, under likelihood-based method or Bayesian approach, MCAR and MAR can be treated as ignorable missingness, which means that the dropout process can be ignored in the analysis. MI is one type of Bayesian method, therefore, it can provide the valid results under MCAR/MAR

assumption. On the other hand, MI represents the uncertainty about imputed values, hence it is recommended by many researchers to instead of single imputation for handling missing data. However, if dropouts are MNAR, MI only is not suitable anymore.

2.3.2 Main methods as sensitivity analysis under MNAR assumption

MNAR is the most complicate d missing procedure. The majority of approaches for MNAR are based on models for the joint distribution of the measurement process and the dropout process. Most of them are likelihood-based methods and can be classified into three categories: selection models, pattern-mixture models and shared-parameter models (frailty models). Generally speaking, frailty models use latent frailties or random effects to capture the dependence between the outcome of interest process and dropout (Follmann and Wu 1995, Albert and Follmann 2000). The factorization is

$$f(Y, R|X) = \int f(Y|\gamma, X) f(R|\gamma, X) dF(\gamma|X)$$

where γ is the latent frailty term. Compared to frailty models, Selection Models and Pattern Mixture Models are more popular for handling MNAR.

2.3.2.1 Selection Model (SM)

Selection models (Rubin 1976, Little 1987, Diggle and Kenward 1994) specify the full data likelihood as the product of the marginal density of the measurement process and the

density of dropout process conditional on the outcomes. The full likelihood can be presented by

$$f(Y,R|X) = f(Y|X)f(R|\,Y,X) = f(Y_{obs},Y_{mis}|X)f(R|\,Y_{obs},Y_{mis},X) \qquad (2.3.2.1\text{-}1)$$

where Y is the outcome of interest which includes observed values (Y_{obs}) and missing values (Y_{mis}), X represents the covariates, and R is an indicator for dropout. The outcome variable in measurement process is analyzed the same as the outcome being analyzed in MAR analysis. For example, using a parametric model such as multivariate normal regression:

$$Y_{it} = X_{it}\beta + \varepsilon_{it} \quad t = 1,\dots,T \qquad (2.3.2.1\text{-}2)$$

The dropout process is often analyzed via a logistic regression, for example:

$$logit\{P(R|Y_{i1},..,Y_{iT},X)\} = \alpha_t + X_{it}\gamma + Y_{i,1}\theta_1 + \cdots + Y_{i,t+1}\theta_{t+1} \qquad (2.3.2.1\text{-}3)$$

This model shows that the probability of subject i dropping at time point t is dependent on some covariates, observed outcomes and the first missing outcome. If $\theta_{t+1} = 0$, the dropout is MAR, otherwise is MNAR.

There are two key assumptions in this parametric selection model: normality for the distribution of outcome and linear relationship between *logit(probability of dropout)* and observed/missing outcome *Y*. It's not possible to justify these two assumptions and distinguish between violations of them. Thus, extreme caution should be taken when parametric selection models are specified under the MNAR assumption (Kenward 1998, Hogan 2004).

2.3.2.2 Pattern Mixture Model (PMM)

Compared to the Selection Models, Pattern Mixture Models (Little 1993, 1994) are based on an alternative factorization of full data likelihood. They describe the full data likelihood as the product of the density of measurement process conditional on the dropout pattern and the density of dropout process. Conceptually, PMM treat the full data distribution as a mixture over dropout times or patterns and describe the observed outcomes within each pattern while assuming that the missingness is MAR within a pattern. The model can be fitted by

$$f(Y,R|X) = f(R|X)f(Y|R,X) = f(R|\,X)f(Y_{obs},Y_{mis}|R,X) \qquad (2.3.2.2\text{-}1)$$

where the former part indicates the distribution of dropout pattern, and the latter part gives the conditional distribution of outcome Y_{obs} and Y_{mis} given the dropout pattern.

Some researchers prefer Selection Models because they present the Little and Rubin missingness taxonomy (MCAR, MAR, MNAR) in a straight forward way. However, Pattern Mixture Models also explicitly present the nature of MNAR assumption in the model formulation, clearly specifying that dropouts and missing values are not independent.

For example, the conditional distribution *f(Y|R,X)* follows a multivariate normal distribution:

$$(Y_i|R_i = r) \sim N\left(\mu^{(r)}, \Sigma^{(r)}\right)$$

Here *r* denotes the dropout pattern for i^{th} subject; $\mu^{(r)}$ is the mean vector that can be estimated by a linear combination of covariates *X* ($X\beta^{(r)}$) for dropout pattern *r;* and $\Sigma^{(r)}$ is the error variance matrix for dropout pattern *r*. Using a Bayesian argument, the probability of dropout in Pattern Mixture Model can be proven to be dependent on missing outcomes, which represents MNAR mechanisms. As a special case, suppose there are two patterns, r=0, 1 (represents dropouts and completers respectively). Algebra shows that

$$logit\{pr(R_i = 0|Y_{i1,} \cdots, Y_{iT})\}$$

$$\propto det\Sigma^{(1)} - det\Sigma^{(0)} + \{\mathrm{Y_i} - \mu^{(0)}\}\{\Sigma^{(0)}\}^{-1}\{\mathrm{Y_i} - \mu^{(0)}\}$$

$$-\{\mathrm{Y_i} - \mu^{(1)}\}^{\mathrm{T}}\{\Sigma^{(1)}\}^{-1}\{Y_i - \mu^{(1)}\}$$

Thus, the probability of dropout is related to Y_{it}, Y_{it}^2, and all first-order cross-products $Y_{it}Y_{is}$ for all *t* and *s* and $t \neq s$. If $\Sigma^{(0)} = \Sigma^{(1)}$, then the logit of the dropout probability is linear with respect to Y_{it} for all *t* (Hogan 2004). That means the probability of dropout is related to all the observed and unobserved measurements, reflecting the nature of MNAR.

If the marginal distribution of Y_i (e.g. blood pressure) is of interest, it can be obtained through a mixture of normal distributions with mean and variance given by

$$E(Y) = \Psi\mu^{(0)} + (1 - \Psi)\mu^{(1)}$$

$$var(Y) = \Psi(1 - \Psi)(\mu^{(0)} - \mu^{(1)})^2 + \Psi\Sigma^{(0)} + (1 - \Psi)\Sigma^{(1)}$$

where $\Psi = pr(R_i = 0)$.

Under certain conditions, PMMs are under-identified due to the missing values. That is not all parameters are estimable from incomplete pattern data. Extra assumptions about

the distributions of missing values are needed to produce additional restrictions that result in identifiable models. Depending on the nature of the study and data patterns, there is a variety of parameter constraints can be used to fix under-identification of the model parameters. For example, assuming the outcome of interest for each pattern is a linear function of time and other covariates resolves the under-identification problem (Hogan, 2004). This restriction method will be used in this dissertation and is explained in my first proposed approach for handling mixed dropout mechanisms. While for my second proposed method, available-case missing value restriction (Thijs, 2002) and non-future-dependent (NFD) missing value restriction (Molenberghs et al. 2003) within PMMs are used to deal with the under-identification problem.

2.4 Summary

Missing data has received increasing attention in healthcare field in recent years, especially from regulatory agencies (e.g. U.S.Food and Drug Administration (FDA) or European Medicines Agency (EMA)). Further, more researches and policy makers have recognized the need for analyses to reflect the nature of missingness when it is not MCAR or MAR. Therefore, statistical methods such as Pattern Mixture Models (PMMs) or Selection Models (SMs) for handling MNAR data have obtained more attention recently. However, both PMM and SM assume the missing mechanism is unique – MNAR, which may be a questionable assumption in many longitudinal studies, especially in confirmatory clinical trials. Usually, dropouts or missingness are occurs for multiple reasons, and may include MCAR, MAR and MNAR mechanisms within the

same study. Unfortunately, little research has focused on mixed dropout mechanisms. Ofer Harel and Joseph Schafer gave a general introduction about some methods that can be used in dealing with partial and latent ignorable missingness problems, but didn't provide the theoretical details or simulations (Harel O., 2009).

Therefore I will provide additional choices for handling mixed dropout to satisfy these increasing needs. Although all the methods for handling MNAR missingness including the ones I will propose need some unverifiable assumptions about the distribution of unobserved values, they provide important sensitivity analyses that support results from any primary analysis.

My proposed two approaches for handling mixed dropout mechanisms are motivated from two different studies' scenarios. Both approaches are based on Pattern Mixture Models, because relatively PMMs are more flexible, computationally less expensive, and lead to simpler exploration of the sensitivity of results to assumptions. The first proposed method is pattern mixture - within – mixture model with EM algorithm. This approach would be appropriate when the detailed dropout reasons are not clear and are suspected to result from mixed dropout mechanisms, as well as when a function between the outcome of interest and time can be assumed. The second approach is PMMs with mixed missing value restrictions combined with multiple imputation. This approach is appropriate when the detailed dropout reason for each subject is known and can result from either a MCAR/MAR or a MNAR dropout.

3 EM algorithm used in Pattern mixture-within-mixture model for handling longitudinal data with mixed dropouts mechanisms

In this chapter, I present an approach to handle continuous data with mixed dropout mechanisms that leverages both the PMMs and the EM algorithm. It is assumed that the true reason for dropout is unknown. More specifically, I assume that do not know whether the dropout is MCAR/MAR or MNAR.

To express the mechanism of MNAR, Pattern Mixture Models factor the joint distribution as the marginal distribution of the dropout and the conditional distribution of the response given the dropout patterns. The dropout patterns are usually dependent on the dropout times. For example, in a longitudinal study with four visits, some subjects drop at last visit. Then we have two dropout patterns R =1, 2. The second dropout pattern (R=2) represents the completers.

O	O	O	M	R=1
O	O	O	O	R=2

Here O represents an observed value; the capital M indicates the missing values caused by dropouts.

Under the MNAR assumption, PMMs treat the full data distribution as a mixture over dropout patterns, but assume that outcomes from two individuals with the same dropout pattern follow the same distribution. For example, the trend of treatment effect for the dropouts is different with the ones for completers, which will result in the different distribution of the response. This scenario is represented by Figure 3.1.

Alternatively, if all dropouts are MCAR/MAR, the data distribution is considered homogeneous over patterns and an Expectation Maximization algorithm can be applied which is theoretically equivalent to considering a likelihood of only the observed data (after integrating over the missing values). This alternative scenario is represented by Figure 3.2.

However, if the dropout mechanisms are mixed with both ignorable missingness (MCAR/MAR) and non-ignorable missingness (MNAR), subjects with the same dropout time may have different response distributions if these subjects have different dropout mechanisms. Furthermore, subjects with different dropout patterns might or might not follow different response distributions. Therefore, under mixed dropout, some combination of these approaches may be needed to handle mixed dropout.

Thus with this chapter, I use mixture models within each level of dropout pattern to allow for the possibility of different missingness mechanisms. I fit mixture models within Pattern Mixture Models to handle mixed dropout problem. Specifically, within each dropout pattern, a mixture distribution takes into account the possibility of different missingness mechanisms. Suppose the outcome of interest is a continuous measurement (e.g. blood pressure) that can be assumed as normality. The model for my method is laid out pictorially in Figure 3.3.

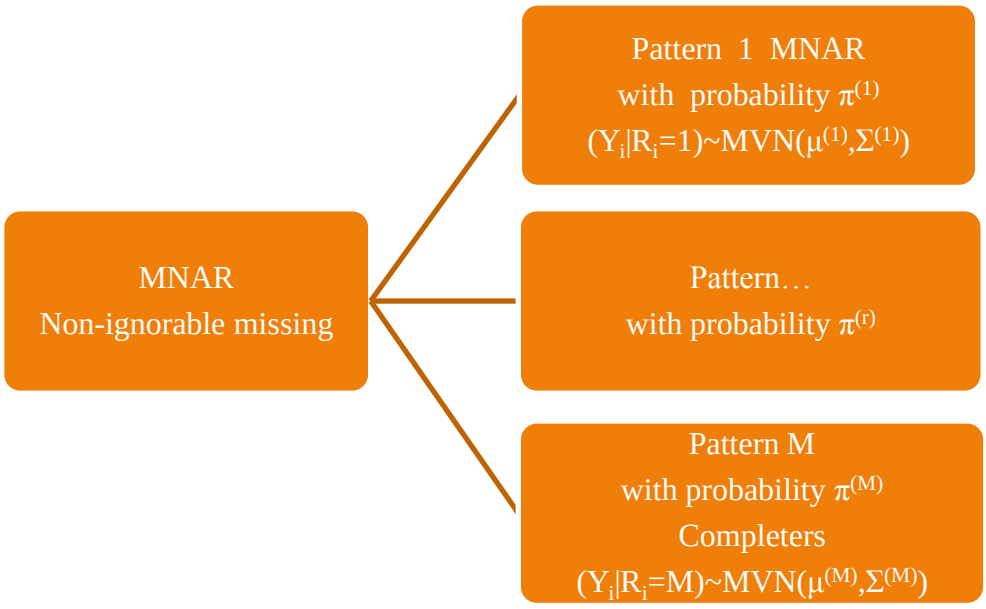

Figure 3.1. Assuming all missingness are MNAR

Where $\pi^{(r)}$ is the known probability of a subject following pattern *r*.

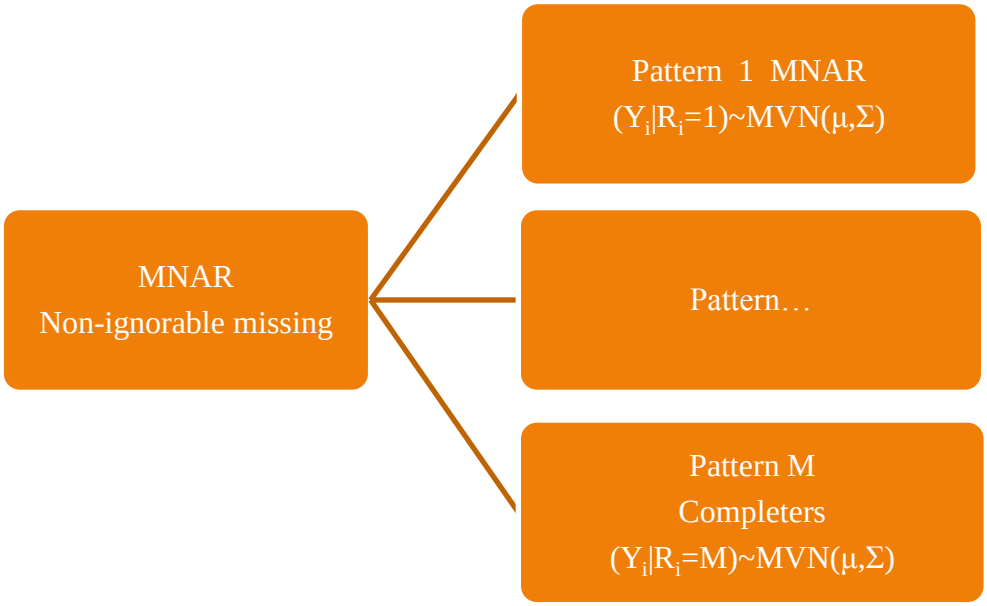

Figure 3.2. Assuming all missingness are MCAR/MAR

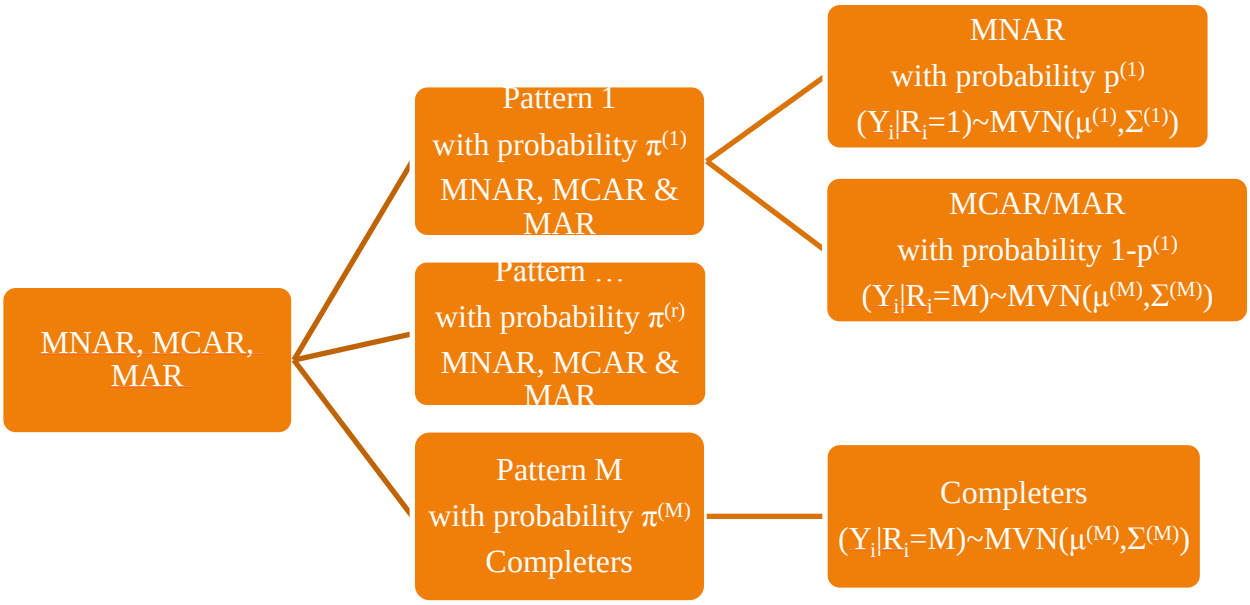

Figure 3.3. Proposed approaches based on Pattern Mixture Model, assuming missingness mechanisms are a mixture of MNAR, MCAR and MAR

$p^{(1)}$ denotes the probability of non-ignorable missingness given pattern *1*. Given a subject drops out at a particular time prior to completing the study, the response may follow one of two distributions.

The distribution of responses from subjects with ignorable missingness (MCAR, MAR) is assumed to be the same as the distribution of responses from completers. However, for subjects with non-ignorable missingness (MNAR), the distributions of responses are assumed different for each dropout pattern.

In order to fit mixture models within Pattern Mixture Models, there are two key issues need to be addressed. First, how does one separate mixed dropouts into ignorable missingness (MCAR and MAR) and non-ignorable missingness (MNAR). Second, how does one solve the under-identified problem for Pattern Mixture Models.

For the first key issue, if the dropout reasons are recorded clearly for each dropout (e.g. in confirmatory clinical trials), then we can simply class the dropouts into different types according to those recorded dropout reasons. For example, the dropouts caused by adverse events usually produce non-ignorable missingness, while ignorable missingness is usually caused by reasons like move to other state, etc.. The case where dropout reasons may be categized as ignorable or non-ignorable based on known information will be handed in the next chapter. In this chapter, handling the setting in which the dropout reasons are not clearly recorded in a study (e.g. some historical studies or studies without recording dropout reasons). In this case, we cannot classify dropout reasons as ignorable or non-ignorable with certainty and this categorization may be treated as an unobserved variable. EM algorithm will be used to solve this issue.

The second key issue centers on the under-identification of PMMs. That is not all parameters are estimable from incomplete pattern data. Thus, some extra assumptions about the distributions of missing values are needed to identify the models for the subjects with dropouts. In my first proposed method, I'll assume that the response for each pattern is a defined function of time and other covariates, and then the under-indentified problem can be solved naturally.

3.1 Model specification and estimates

Suppose the complete data consist of T repeated observations taken on N subjects at a fixed set of time points *1, 2, …, T*. The number and timing of measurements for all

subjects is assumed to be equal. For i^{th} subject ($i=1,2,..,N$), the full data response vector is $Y_i = (Y_{i1},..., Y_{iT})^T$. Since subject i may drop out from the study, we use $Y_{i\,(obs)} = (Y_{i1},..., Y_{iT_i})^T$ to denote the observed response vector, where T_i is the last time point of measurement ($T_i \leq T$).

If there are M dropout patterns, let R_i be a random variable representing the last observed value for subject i, which may take on values $r=1,..,M$. we would say "subject i has missing pattern r). Given $R_i = r\ (<M)$, Y_i is either from $f_{IGM}^{(r)}(Y)$ (the density of Y for ignorable missingness) with probability $(1-p^{(r)})$ or $f_{NIGM}^{(r)}(Y)$ (the density of Y for non-ignorable missingness) with probability $p^{(r)}$. Thus, the marginal distribution of Y_i given pattern r is

$$f_{Y_i}(Y_i)^{(r)} = p^{(r)} f_{NIGM}^{(r)}(Y_i|X_i) + (1-p^{(r)}) f_{IGM}^{(r)}(Y_i|X_i). \qquad (3.1\text{-}1)$$

We have $p^{(r)}$ denotes the probability of non-ignorable missingness given pattern r, and X_i represents fixed, potentially time varying covariates.

A linear mixed model is fitted for non-ignorable missing Y_i given pattern r ($=1,...,M-1$) with

$$[Y_i; X_i, U_i|r]_{NIGM} = X_i\beta^{(r)} + b_i U_i + \varepsilon_i \qquad (3.1\text{-}2)$$

X_i is a $T_i \times Q$ matrix of fixed covariates such that $X_i = (X_{i1}^T, \dots, X_{iT_i}^T)^{\mathrm{T}}$. X_{it} is a vector of length Q the number of covariate measured at each time t. The random covariates are denoted by b_i, a $T_i \times K$ matrix. This matrix can be written as $b_i = (b_{i1}^T, \dots, b_{iT_i}^T)^{\mathrm{T}}$, where the length of each vector b_{it} K is the number of random covariates included in the model

at each time *t*. For example, if the random covariates only include the intercept, the *K* would equal 1. The *K*-vector of random effects U_i is assumed to follow a multivariate normally distributed with mean 0 and covariance *D*. The error vector $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ_i})^{\mathrm{T}}$, is independent of b_i and is assumed to be normally distributed with mean 0 and covariance matrix $\sigma^2 I$, where *I* is the identity matrix. $\beta^{(r)}$ is the *Q*-vector of coefficients associated with fixed covariates specific to informative dropout in pattern *r*. Thus, Y_i with non-ignorable missing with dropout pattern *r* is from multivariate normal distribution with mean vector $X_i\beta^{(r)} + U_i b_i$ and covariance matrix $\sigma^2 I$.

$$f_{NIGM}^{(r)}(Y_i; X_i, U_i) \sim N(X_i\beta^{(r)} + U_i b_i, \sigma^2 I)$$

A linear mixed model is also fitted for the completers in the last pattern $r = M$, which includes only completers. This model is

$$[Y_i; X_i, U_i | r = M] = X_i\beta^{(M)} + U_i b_i + \varepsilon_i \qquad (3.1\text{-}3)$$

Under ignorable missingness, effect of treatment should be the same, independent of dropout time, thus we assume the parameters for ignorable missingness are the equal across patterns with dropouts, as well as equal to the parameters for the last pattern (completers). That is,

$$[Y_i; X_i, U_i | r = 1]_{IGM} = X_i\beta^{(M)} + U_i b_i + \varepsilon_i$$

…

$$[Y_i; X_i, U_i | r = M - 1]_{IGM} = X_i\beta^{(M)} + U_i b_i + \varepsilon_i$$

further, as well as $[Y_i; X_i, U_i | r = M] = X_i\beta^{(M)} + U_i b_i + \varepsilon_i$, assume variances are constant across patterns. This assumption can easily be relaxed when appropriate.

For any pattern r (<M), we define a Bernoulli random variable Z_i for subject i with $Z_i = 1$, subject i's dropout is considered non-ignorable missingness ($probability = p^{(r)}$); and $Z_i = 0$ if it is ignorable missingness ($probability = 1 - p^{(r)}$). $p^{(r)}$ may vary by pattern when the probabilities of MNAR dropout for different patterns are not the same. Since Z_i is unobserved, that is we don't know whether Y_i is from $f_{NIGM}^{(r)}(Y_i; X_i, U_i)$ or $f_{IGM}^{(r)}(Y_i; X_i, U_i)$, the EM algorithm is naturally used for estimation in this model.

3.2 EM algorithm

The EM algorithm is an iterative procedure for maximizing observed data likelihood. Let Y be the random vector corresponding to the observed data. Ψ is the vector containing the unknown parameters in the postulated form for the distribution of Y. If using M as the missing data vector, the complete-data log likelihood function is given by

$$logL_c(\Psi; Y, M) = logf_c(Y, M; \Psi).$$

The EM algorithm maximizes the observed data log likelihood function $logL(\Psi; Y)$ by proceeding iteratively in terms of the complete-data log likelihood function $logL_c(\Psi; Y, M)$ through E-step and M-step. E-step calculates the conditional expectation of the complete data log-likelihood given the observed data such that parameters are replaced by their previous estimates $\psi^{(t)}$, and M-step updates the estimates to $\psi^{(t+1)}$ using this expected log-likelihood. Here, $\psi^{(t)}$ and $\psi^{(t+1)}$ represents the estimated complete set of unknown parameters Ψ at t^{th} and $(t+1)^{th}$ iteration respectively. On the $(t+1)^{th}$ iteration, the E-step and M-step are defined as follows:

E-step: Calculate $Q(\Psi;\Psi^{(t)})$ where

$$Q(\Psi;\Psi^{(t)}) = E_{\Psi^{(t)}}\{logL_c(\Psi;Y,M)|Y\}$$

M-step: Choose $\psi^{(t+1)}$ to maximize $Q(\Psi;\Psi^{(t)})$; that is

$$Q(\Psi^{(t+1)};\Psi^{(t)}) \geq Q(\Psi;\Psi^{(t)})$$

The E-step and M-step are alternated repeatedly until the difference $L\left(\Psi^{(t+1)};Y\right) - L\left(\Psi^{(t)};Y\right)$ changes by a very small amount to show the convergence of the sequence of observed likelihood values $\{L\left(\Psi^{(t)};Y\right)\}$.

To simplify computation, we suppose there is only one random effect – intercept, which is from normal distribution with mean 0 and variance δ^2. Random covariates are $T_i \times 1$ matrix $b_i = (1,\dots,1)^{\mathrm{T}}$.

The complete data likelihood function for this Pattern Mixture-Within-Mixture Model can be written as

$$L(\Psi;Y,X,U,R,Z) = f(Y,X,U,R,Z;\Psi)$$

$$= \prod_{i=1}^{N}\left[\prod_{t=1}^{T_i}[f(Y_{it}|X_{it},U_i,R_i,Z_i;\beta)]\, f(U_i;\delta^2) f(Z_i|R_i,p) f(R_i)\right]$$

$$= \prod_{r=1}^{M-1} \prod_{i \in pattern\, r} \left\{ \begin{array}{l} \left\{ \pi^{(r)} p^{(r)} [\prod_{t=1}^{T_i} f\left(Y_{it} \middle| X_{it}, U_i, R_i = r, Z_i; \beta^{(r)}\right)] f(U_i; \delta^2) \right\}^{Z_i} \\ \left\{ \pi^{(r)} (1-p)^{(r)} [\prod_{t=1}^{T_i} f\left(Y_{it} \middle| X_{it}, U_i, R_i = r, Z_i; \beta^{(M)}\right)] f(U_i; \delta^2) \right\}^{1-Z_i} \end{array} \right\}$$

$$\times \prod_{i \in pattern\, M} \left\{ \pi^{(M)} \left[\prod_{t=1}^{T_i} f\left(Y_{it} \middle| X_{it}, U_i, R_i = M; \beta^{(M)}\right) \right] f(U_i; \delta^2) \right\}$$

(3.2-1)

where $\pi^{(r)}$ is the fixed probability of a subject following pattern *r* to estimated from the data. For example, if there are 20% of subjects drop at pattern *r*, then $\pi^{(r)} = 0.2$. Ψ represents the complete set of unknown parameters $\{\beta, \sigma^2, \delta^2, p\}$, where β refers to $\{\beta^{(1)}, \dots, \beta^{(M)}$ and $p = (p^{(1)}, \dots, p^{(M)})$.

The log-likelihood function is

$$\log L(\Psi; Y, X, U, R, Z)$$

$$= \prod_{r=1}^{M-1} \prod_{i \in pattern\, r} \left\{ \begin{array}{l} Z_i \left[\log \pi^{(r)} + \log p^{(r)} + \sum_{t=1}^{T_i} log\, f\left(Y_{it} \middle| X_{it}, U_i, R_i = r, Z_i = 1; \beta^{(r)}\right) + logf(U_i; \delta^2) \right] + \\ (1 - Z_i) \left[\log \pi^{(r)} + \log(1 - p^{(r)}) + \sum_{t=1}^{T_i} logf\left(Y_{it} \middle| X_{it}, U_i, R_i = r, Z_i = 0; \beta^{(r)}\right) + logf(U_i; \delta^2) \right] \end{array} \right\}$$

$$+ \prod_{i \in pattern\, M} \left\{ \log \pi^{(M)} + \sum_{t=1}^{T_i} logf\left(Y_{it} \middle| X_{it}, U_i, R_i = M; \beta^{(M)}\right) + logf(U_i; \delta^2) \right\}$$

(3.2-2)

The unknown parameters can be estimated via maximization of the conditional expectation of the log-likelihood of the complete data given the observed data $E[\log f(\ Y, X, U, R, Z; \Psi)|Y, X, \Psi]$. Here, all the random effect and indicators of non-ignorable missingness are considered as missing data.

3.2.1 E-step

At the $(t+1)^{th}$ interation, calculate

$$Q\left(\Psi\middle|\Psi^{(t)}\right) = \mathrm{E}[\log L(\ \Psi; Y, X, U, R, Z)|Y, X, \Psi^{(t)}], \qquad (3.2.1\text{-}1)$$

the expectation with respect to $f(U, R, Z|Y, X, \Psi^{(t)})$, where $\Psi^{(t)}$ represents the values of parameters estimated at the t[th] iteration.

$$Q\left(\Psi|\Psi^{(t)}\right) =$$

$$\sum_{r=1}^{M-1}\sum_{i\in pattern\ r}\{\log \pi^{(r)} +$$

$$EZi|Ri=r\log Pr+EZit=1T\log fYitXit,Ui,Ri=r,Zi=1;\beta r+EZ\log fUi;\delta 2+E1-Zi|Ri=r\log 1-P$$

$$r+EZit=1T\log fYitXit,Ui,Ri=r,Zi=0;\beta M+E1-Z\log fUi;\delta 2+i\in pattern$$

$$M\log\pi(M)+Et=1T\log fYitXit,Ui,Ri=M;\beta M+E\log fUi;\delta 2$$

Where

$$f(U_i; \delta^2) = \frac{1}{\sqrt{2\pi}\delta} exp\left(-\frac{U_i^2}{2\delta^2}\right)$$

$$f\left(Y_{it}|X_{it},U_i,Z_i=1,R_i=r;\beta^{(r)}\right)=\frac{1}{\sqrt{2\pi}\sigma}exp\left(-\frac{\left(Y_{it}-X_{it}\beta^{(r)}-U_i\right)^2}{2\sigma^2}\right)$$

$$f\left(Y_{it}|X_{it},U_i,Z_i=0,R_i=r;\beta^{(M)}\right)=f\left(Y_{it}|X_{it},U_i,,R_i=M;\beta^{(M)}\right)$$

$$=\frac{1}{\sqrt{2\pi}\sigma}exp\left(-\frac{\left(Y_{it}-X_{it}\beta^{(M)}-U_i\right)^2}{2\sigma^2}\right)$$

The following conditional expectations are needed:

$$E(U_i|Y_i,X_i,R_i=r,Z_i=1)=\frac{\sum_{t=1}^{T_i}\left(Y_{it}-X_{it}\beta^{(r)}\right)\delta^2}{\sum_{t=1}^{T_i}\delta^2+\sigma^2}$$

$$E(U_i^2\big|Y_i,X_i,R_i=r,Z_i=1)=\frac{\sigma^2\delta^2}{\sum_{t=1}^{T_i}\delta^2+\sigma^2}+[E(U_i|Y_i,X_i,R_i=r,Z_i=1)]^2$$

$$E(U_i|Y_i,X_i,R_i=r,Z_i=0)=\frac{\sum_{t=1}^{T_i}\left(Y_{it}-X_{it}\beta^{(M)}\right)\delta^2}{\sum_{t=1}^{T_i}\delta^2+\sigma^2}$$

$$E(U_i^2\big|Y_i,X_i,R_i=r,Z_i=0)=\frac{\sigma^2\delta^2}{\sum_{t=1}^{T_i}\delta^2+\sigma^2}+[E(U_i|Y_i,X_i,R_i=r,Z_i=0)]^2$$

$$E(U_i|Y_i,X_i,R_i=M)=\frac{\sum_{t=1}^{T_i}\left(Y_{it}-X_{it}\beta^{(M)}\right)\delta^2}{\sum_{t=1}^{T_i}\delta^2+\sigma^2}$$

$$E(U_i^2\big|Y_i,X_i,Z=1,R_i=M)=\frac{\sigma^2\delta^2}{\sum_{t=1}^{T_i}\delta^2+\sigma^2}+[E(U_i|Y_i,X_i,R_i=M)]^2$$

$$E(Z_iU_i|Y_i,X_i,R_i=r,Z_i=1)=E(Z_i|Y_i,X_i,R_i=r)E(U_i|Y_i,X_i,R_i=r,Z_i=1)$$

$$E(Z_iU_i^2|Y_i,X_i,R_i=r,Z_i=1)=E(Z_i|Y_i,X_i,R_i=r)E(U_i^2\big|Y_i,X_i,R_i=r,Z_i=1)$$

$$E[(1-Z_i)U_i|Y_i,X_i,R_i=r,Z_i=0]$$

$$=E[(1-Z_i)|Y_i,X_i,R_i=r]E(U_i|Y_i,X_i,R_i=r,Z_i=0)$$

$$E[(1-Z_i)U_i^2|Y_i,X_i,R_i=r,Z_i=0]$$

$$=E[(1-Z_i)|Y_i,X_i,R_i=r]E(U_i^2\big|Y_i,X_i,R_i=r,Z_i=0)$$

$$E(Z_i|Y_i, X_i, R_i = r)$$

$$= \frac{P^{(r)} \sum_{t=1}^{T_i} \left\{ \frac{1}{\sqrt{2\pi}\sqrt{\delta^2+\sigma^2}} exp\left(-\frac{\left(Y_{it} - X_{it}\beta^{(r)}\right)^2}{2(\delta^2+\sigma^2)} \right) \right\}}{P^{(r)} \sum_{t=1}^{T_i} \left\{ \frac{1}{\sqrt{2\pi}\sqrt{\delta^2+\sigma^2}} exp\left(-\frac{(Y_{it} - X_{it}\beta^{(r)})^2}{2(\delta^2+\sigma^2)} \right) \right\} + (1-P^{(r)}) \sum_{t=1}^{T_i} \left\{ \frac{1}{\sqrt{2\pi}\sqrt{\delta^2+\sigma^2}} exp\left(-\frac{(Y_{it} - X_{it}\beta^{(M)})^2}{2(\delta^2+\sigma^2)} \right) \right\}}$$

$$E[(1-Z_i)|Y_i, X_i, R_i = r] = 1 - E[Z_i|Y_i, X_i, R_i = r]$$

3.2.2 M-step

New estimates $\Psi^{(t+1)}$ are obtained at each iteration via maximizing equation (3.2.1-1). This is done by setting the first derivatives equal to 0 and solving these equations. For example, to obtain $\beta^{(t+1)}$, we need to solve

$$\frac{\partial \hat{Q}(\Psi|\Psi^{(t)})}{\partial \beta} = 0$$

In detail,

$$\hat{\beta}_q^{(r)(t+1)}$$

$$= \frac{\sum_{i \in pattern\, r} \left\{ E(Z_i|R_i = r)^{(t)} \sum_{t=1}^{T_i} \left(Y_{it} - X_{it}\beta^{(r)(t)} + X_{itq}\beta_q^{(r)(t)} - E(U_i|Z_i = 1, R_i = r)^{(t)} \right) \cdot X_{itq} \right\}}{\sum_{i \in pattern\, r} \left\{ E(Z_i|R_i = r)^{(r)} \sum_{t=1}^{T_i} X_{itq}^2 \right\}}$$

$$\hat{\beta}_q^{(M)(t+1)}$$

$$= \frac{\sum_{r=1}^{M-1} \sum_{i \in pattern\, r} \left\{ \begin{matrix} E((1-Z_i)|R_i = r)^{(t)} \sum_{t=1}^{T_i} \left(Y_{it} - X_{it}\beta^{(M)(t)} + X_{itq}\beta_q^{(M)(t)} - E(U_i|Z_i = 0, R_i = r)^{(t)} \right) \cdot X_{itq} \\ + \sum_{i \in pattern\, M} \left\{ \sum_{t=1}^{T_i} \left(Y_{it} - X_{it}\beta^{(M)(t)} + X_{itq}\beta_q^{(M)(t)} - E(U_i|R_i = M)^{(t)} \right) \cdot X_{itq} \right\} \end{matrix} \right\}}{\sum_{r=1}^{M-1} \sum_{i \in pattern\, r} \left\{ E((1-Z_i)|R_i = r)^{(r)} \sum_{t=1}^{T_i} X_{itq}^2 \right\} + \sum_{i \in pattern M} \sum_{t=1}^{T_i} X_{itq}^2}$$

where r=1,…,M-1; β_q denotes the coefficient for q^{th} covariate X_{itq}, $q = 1, \ldots, Q$.

$$\hat{p}^{(r)(t+1)} = \frac{\sum_{i \in pattern\, r} E(Z_i | R_i = r)^{(t)}}{n^{(r)}}$$

$$\hat{\sigma}_q^{2(t+1)}$$

$$= \frac{\left\{ \sum_{r=1}^{m} \sum_{i \in pattern\, r} \left\{ \begin{array}{c} E(Z_i|R_i=r)^{(t)} \left[\sum_{t=1}^{T_i} \left(Y_{it} - X_{it}\beta^{(r)}\right)^t - 2\sum_{t=1}^{T_i} \left(Y_{it} - X_{it}\beta^{(r)}\right)^t E(U_i|Z_i=1, R_i=r)^{(t)} \right] \\ + \sum_{t=1}^{T_i} E\left(U_i^2 | Z_i = 1, R_i = r\right)^t + \\ E\left((1-Z_i)|R_i=r\right)^{(t)} \left[\sum_{t=1}^{T_i} \left(Y_{it} - X_{it}\beta^{(M)(t)}\right)^2 - 2\sum_{t=1}^{T_i} \left(Y_{it} - X_{it}\beta^{(M)(t)}\right) E(U_i|Z_i=0, R_i=r)^{(t)} \right] \\ + \sum_{t=1}^{T_i} E\left(U_i^2 | Z_i = 0, R_i = r\right)^t \end{array} \right\} \right\}}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} 1}$$

$$+ \sum_{i \in pattern\, M} \left\{ \sum_{t=1}^{T_i} \left(Y_{it} - X_{it}\beta^{(M)(t)}\right)^2 - 2\sum_{t=1}^{T_i} \left(Y_{it} - X_{it}\beta^{(M)(t)}\right) E(U_i|R_i=M)^{(t)} \right\} + \sum_{t=1}^{T_i} E\left(U_i^2 | R_i = M\right)^t$$

$$\hat{\delta}^{2(t+1)}$$

$$= \frac{\sum_{r=1}^{m} \sum_{i \in pattern\, r} \left\{ E(Z_i|R_i=r)^{(t)} E\left(U_i^2|Z_i=1, R_i=r\right)^t + E\left((1-Z_i)|R_i=r\right)^{(t)} E\left(U_i^2|Z_i=0, R_i=r\right)^t \right\} + \sum_{i \in pattern\, M} E\left(U_i^2|R_i=M\right)^t}{\sum_{i=1}^{N} 1}$$

E- and M-steps are alternately repeated until the sequence of maximum likelihood estimates is convergent (when $(\Psi^{(t+1)} - \Psi^{(t)})(\Psi^{(t+1)} - \Psi^{(t)})^T \leq 0.000001$).

3.2.3 Standard errors

Standard errors are estimated by the inverse of the Empirical Fisher Information (Meilijson 1989). Empirical Fisher Information matrix is given by

$M(\Psi) = \sum_{i=1}^{N} s(Y_i; \Psi) s^T(Y_i; \Psi) - \frac{1}{N} S(Y; \Psi) S^T(Y; \Psi)$,

where $s(Y_i; \Psi)$ is the score function, and $S(Y; \Psi) = \sum_{i=1}^{N} s(Y_i; \Psi)$.

At t^{th} iteration, it is given by

$$s\left(Y_i;\Psi^{(t)}\right)=\frac{\partial Q_i(\Psi|\Psi^{(t)})}{\partial\Psi}$$

and

$$S\left(Y;\Psi^{(t)}\right)=\sum_{i=1}^{N}s\left(Y_i;\Psi^{(t)}\right)=\frac{\partial Q(\Psi|\Psi^{(t)})}{\partial\Psi}$$

At the convergence of the EM algorithm, $S\left(Y;\widehat{\Psi}\right)=0$. Thus the empirical fisher information is

$$M(\widehat{\Psi})=\sum_{i=1}^{N}s\left(Y_i;\widehat{\Psi}\right)s^T\left(Y_i;\widehat{\Psi}\right) \qquad (3.2.3\text{-}1)$$

where $s\left(Y_i;\widehat{\Psi}\right)=\frac{\partial Q_i(\Psi|\widehat{\Psi})}{\partial\Psi}$

In detail, to $i\in Pattern\ r$, r=1,…,M-1, the elements in $s\left(Y_i;\widehat{\Psi}\right)$ are given by:

$$\frac{\partial Q_i}{\partial\beta_q^{(r)}}=\frac{E(Z_i|R_i=r)\sum_{t=1}^{T_i}\left(Y_{it}-X_{it}\beta^{(r)}-E(U_i|Z_i=1,R_i=r)\right)X_{itq}}{\sigma^2}$$

$$\frac{\partial Q_i}{\partial\beta_q^{(M)}}=\frac{E\left((1-Z_i)|R_i=r\right)\sum_{t=1}^{T_i}\left(Y_{it}-X_{it}\beta^{(M)}-E(U_i|Z_i=0,R_i=r)\right)X_{itq}}{\sigma^2}$$

$$\frac{\partial Q_i}{\partial P^{(r)}}=\frac{E(Z_i|R_i=r)-P}{P(1-P)}$$

$$\frac{\partial Q_i}{\partial\sigma}=-\sum_{j=1}^{J_i}\frac{1}{\sigma}+\frac{\left\{\begin{array}{c}E(Z_i|R_i=r)\left[\sum_{t=1}^{T_i}\left(Y_{it}-X_{it}\beta^{(r)}-E(U_i|Z_i=1,R_i=1)\right)^2+\sum_{t=1}^{T_i}Var(U_i|Z_i=1,R_i=1)\right]+\\E\left((1-Z_i)|R_i=r\right)\left[\sum_{t=1}^{T_i}\left(Y_{it}-X_{it}\beta^{(M)}-E(U_i|Z_i=0,R_i=r)\right)^2+\sum_{t=1}^{T_i}Var(U_i|Z_i=0,R_i=r)\right]\end{array}\right\}}{\sigma^3}$$

$$\frac{\partial Q_i}{\partial \delta} = -\frac{1}{\delta} + \frac{E(Z_i U_i^2 | Z_i = 1, R_i = r) + E\left((1 - Z_i) U_i^2 | Z_i = 0, R_i = r\right)}{\delta^3}$$

To $i \in Pattern\ M$, the elements in $s\left(Y_i; \widehat{\Psi}\right)$ are given by:

$$\frac{\partial Q_i}{\partial \beta_q^{(M)}} = \frac{\sum_{t=1}^{T_i} \left(Y_{it} - X_{it}\beta^{(M)} - E(U_i | R_i = M)\right) \cdot X_{itq}}{\sigma^2}$$

$$\frac{\partial Q_i}{\partial \sigma} = \frac{\sum_{t=1}^{T_i} \left(Y_{it} - X_{it}\beta^{(M)} - E(U_i | R_i = M)\right)^2 + \sum_{t=1}^{T_i} Var(U_i | R_i = M)}{\sigma^3} - \sum_{t=1}^{T_i} \frac{1}{\sigma}$$

$$\frac{\partial Q_i}{\partial \delta} = -\frac{1}{\delta} + \frac{E(U_i^2 | R_i = M)}{\delta^3}$$

3.2.4 Marginal results

I focus on estimating the treatment effect in a longitudinal clinical trial with a continuous outcome of interest, such that incomplete cases are due to mixed dropout mechanisms. The treatment effect is expressed by comparing the means of outcome of interest (e.g. blood pressure) at the end of study between two treatment groups (e.g. placebo vs. tested drug). The proposed method only provides results from specific dropout pattern and type. The marginal mean of outcomes of interest at the last visit can be obtained by averaging over dropout patterns and types.

$$E(Y) = \sum_r \sum_Z E(Y | r, Z) \widehat{P(Z | r)} P(r) \qquad (3.2.4\text{-}1)$$

where $P(r) = \pi^{(r)}$ is the known probability of a subject drop at pattern r. $\widehat{P(Z \mid r)} =$ $p(r)$ is the probability of non-ignorable missingness ($Z=1$) at pattern *r*, while$\widehat{P(Z \mid r)} = 1 - p^{(r)}$ is the probability of ignorable missingness ($Z = 0$) at pattern *r*.

$E(Y|r, Z = 1) = X\beta^{(r)}$,

$E(Y|r, Z = 0) = X\beta^{(M)}$,

and $E(Y|r = M) = X\beta^{(M)}$.

Then

$E(Y) =$

$\sum_{r=1}^{M-1}(\pi^{(r)}p^{(r)}E(Y|r, Z = 1) + \pi^{(r)}\left(1 - p^{(r)}\right)E(Y|r, Z = 0) + \pi^{(M)}E(Y|r = M)$

$= \sum_{r=1}^{M-1}(\pi^{(r)}p^{(r)}E(Y|r, Z = 1)) + (1 - \sum_{r=1}^{M-1}\pi^{(r)}p^{(r)})E(Y|r = M)$

(3.2.4-2)

Here $\pi^{(M)} = (1 - \sum_{r=1}^{M-1}(\pi^{(r)}))$

3.2.5 Delta Method

Delta method is used to derive the estimate of variance of a function of parameters, e.g. the weighted average parameters. In detail, Taylor series expansions are used to approximate variances for functions of random variables. Assume an estimate Ψ converges in probability to its true values ψ, then $\sqrt{n}(\Psi - \psi)$ is convergence in distribution to $N(0, \Sigma)$, which is represented by

$$\sqrt{n}(\Psi - \psi) \rightarrow N(0, \Sigma)$$

where n is the number of observations and Σ is a covariance matrix. Using Taylor series expansions, a function g of the estimator Ψ may be approximated by

$$g(\Psi) \approx g(\psi) + \nabla g(\psi)^T(\Psi - \psi)$$

The variance of $g(\Psi)$ is approximately

$$var(g(\Psi)) \approx var(g(\psi) + \nabla g(\psi)^T(\Psi - \psi)) = \nabla g(\psi)^T(\Sigma/n)\nabla g(\psi)$$

In our case, the variance for the marginal result estimate can be calculated by

$$\text{var}(g(\Psi)) =$$

$$\left(\frac{\partial g}{\partial \beta^{(1)}}, \cdots, \frac{\partial g}{\partial \beta^{(M)}}, \frac{\partial g}{\partial p^{(1)}}, \cdots, \frac{\partial g}{\partial p^{(M-1)}}\right) \begin{pmatrix} \sigma^2_{\beta^{(1)}\beta^{(1)}} & \cdots & \sigma^2_{\beta^{(1)}\beta^{(M)}} & \sigma^2_{\beta^{(1)}P^{(1)}} & \cdots & \sigma^2_{\beta^{(1)}P^{(M-1)}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{\beta^{(1)}\beta^{(M)}} & \cdots & \sigma^2_{\beta^{(M)}\beta^{(M)}} & \sigma^2_{\beta^{(M)}P^{(1)}} & \cdots & \sigma^2_{\beta^{(M)}P^{(M-1)}} \\ \sigma^2_{\beta^{(1)}P^{(1)}} & \cdots & \sigma^2_{\beta^{(M)}P^{(1)}} & \sigma^2_{P^{(1)}P^{(1)}} & \cdots & \sigma^2_{P^{(1)}P^{(M-1)}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{\beta^{(1)}P^{(M-1)}} & \cdots & \sigma^2_{\beta^{(M)}P^{(M-1)}} & \sigma^2_{P^{(1)}P^{(M-1)}} & \cdots & \sigma^2_{P^{(M-1)}P^{(M-1)}} \end{pmatrix} \begin{pmatrix} \frac{\partial g}{\partial \beta^{(1)}} \\ \vdots \\ \frac{\partial g}{\partial \beta^{(M)}} \\ \frac{\partial g}{\partial p^{(1)}} \\ \vdots \\ \frac{\partial g}{\partial p^{(M-1)}} \end{pmatrix}$$

In practice, the unknown parameters are replaced by their estimates.

Here, the function of interest is

$$g(\hat{\Psi}) = E(Y) = \sum_{r=1}^{M-1}(\pi^{(r)}\hat{p}^{(r)}E(Y|r, Z = 1)) + (1 - \sum_{r=1}^{M-1}\pi^{(r)}\hat{p}^{(r)})E(Y|r = M),$$

where $E(Y|r, Z = 1)$ and $E(Y|r = M)$ are functions of $\hat{\beta}^{(r)}$ and $\hat{\beta}^{(M)}$ respectively.

$\hat{\Psi} = \begin{pmatrix} \hat{\beta}^{(1)} \\ \cdots \\ \hat{\beta}^{(M)} \\ \hat{p}^{(1)} \\ \cdots \\ \hat{p}^{(M-1)} \end{pmatrix}$, $\pi^{(r)}$is the observed proportion of subjects with pattern *r*, *r=1,…,M-1*.

$$\frac{\partial g}{\partial \hat{\beta}^{(r)}} = \pi^{(r)}\hat{p}^{(r)}X$$

$$\frac{\partial g}{\partial \hat{\beta}^{(M)}} = (1 - \pi^{(r)}\hat{p}^{(r)})X$$

$$\frac{\partial g}{\partial \hat{P}^{(r)}} = \pi^{(r)}(X\hat{\beta}^{(r)} - X\hat{\beta}^{(M)})$$

The variance of $g(\Psi)$ can be obtained through the above approximations.

PMWMM and a standard Pattern Mixture Model are compared in Appendix through likelihood functions and inferences.

3.3 Simulation

3.3.1 Data simulation

A simple study is simulated to examine the performance of proposed pattern mixture-within-mixture models. This study has 400 subjects (200 subjects in placebo group (treatment=0), 200 subjects in drug group (treatment=1)), and 4 equally spaced time points (0, 1, 2, 3). It was assumed that no subjects withdraw from this study at time points 1 and 2 and 40% of subjects drop out at time point 3. Therefore, we will have two patterns (*r=1,2*) in which pattern 1 represents dropouts and pattern 2 represents completers; as depicted in the following diagam:

O O O M *r*=1

O O O O *r*=2

The data are generated from a linear mixed model that includes the fixed covariates of time (baseline, 1 month, 2 month, 3 month), treatment (0 and 1), and a treatment - by - time interaction. Time is a continuous covariate with values of 0, 1, 2 and 3. Here, the "treatment effect" measures the difference between treatment groups at baseline while the treatment by time interaction measures the effect of treatment in changes in outcome (treatment - placebo) over time. The model error ε_{ij} is from $N(0,0.2^2)$ and the random effect -intercept U_i is from $N(0, 0.5^2)$. Four scenarios are considered:

(1) All missing values caused by dropouts are missing completely at random (MCAR).

All responses are generated from the model $Y_i=X_i\beta + U_i + \varepsilon_i$, where

$\beta = (\beta_{intercept}, \beta_{time}, \beta_{treatment}, \beta_{time*treatment}) = (10, -0.4, 0, -0.4)$

Randomly select 40% of subjects (0.4*200=80) from each treatment group, then set the values of these subjects at time point 3 as missing.

Here $f(R_i|Y_i,X_i)=f(R_i)=0.4$

(2) All missing values caused by dropouts missing at random (MAR).

All responses are generated from the model $Y_i=X_i\beta + U_i + \varepsilon_i$, where

$\beta = (\beta_{intercept}, \beta_{time}, \beta_{treatment}, \beta_{time*treatment}) = (10, -0.4, 0, -0.4)$

For each treatment group, order the values at time point 2 and then set the top 40% of values (0.4*200=80) as missing.

Here the dropout is dependent on the value of the last observation, so $f(R_i|Y_i,X_i)=f(R_i|Y_{i2})=0.4$, where Y_{i2} is the last observed value.

(3) All missing values caused by dropouts are missing not at random (MNAR).

The distribution of responses for dropouts with the missing value at time point 3 is set using dropout pattern 1, while the distribution for completers uses dropout pattern 2. The outcomes for the dropouts (80 (=0.4*200) in placebo group, 80 in treatment group) are generated from the model I: $Y_i = X_i\beta^{(1)} + U_i + \varepsilon_i$ for dropout pattern 1 (*r=1*), where $\beta^{(1)}$ is a set of fixed covariates' parameters corresponding to the dropouts. For the completers (400-80*2=240), data are simulated from the model II: $Y_i = X_i\beta^{(2)} + U_i + \varepsilon_i$ for dropout pattern 2 (*r=2*), where fixed covariates' parameters $\beta^{(2)}$ is for the completers. Here

$$\beta^{(1)} = \left(\beta^{,(1)}_{intercept}, \beta^{,(1)}_{time,}\, \beta^{,(1)}_{treat}, \beta^{,(1)}_{time*treat}\right) = (12, -0.2, 0, -0.2),$$

$$\beta^{(2)} = \left(\beta^{,(2)}_{intercept}, \beta^{,(2)}_{time,}\, \beta^{,(2)}_{treat}, \beta^{,(2)}_{time*treat}\right) = (10, -0.4, 0, -0.4).$$

The conditional distributions *f(Y|R,X)* follows two normal distributions:

$$(Y_i|R_i = r) \sim N\left(\mu^{(r)}, \Sigma^{(r)}\right)$$

r=1 represents the dropouts and *r=2* represents the completers. Algebra shows that

$$logit\{pr(R_i = 1|Y_{i1,} \cdots, Y_{iT})\}$$

$$\propto det\Sigma^{(2)} - det\Sigma^{(1)} + \{Y_i - \mu^{(1)}\}\{\Sigma^{(1)}\}^{-1}\{Y_i - \mu^{(1)}\}$$

$$- \{Y_i - \mu^{(2)}\}^{T}\{\Sigma^{(2)}\}^{-1}\{Y_i - \mu^{(2)}\}$$

Thus, the probability of dropout is related to Y_{it}, Y_{it}^2, and all first-order cross-products $Y_{it}Y_{is}$ for all *t* and *s* and $t \neq s$, which includes all the observed and missing measurements, reflecting the nature of MNAR.

(4) For mixed dropouts, some are missing completely at random (5%), some are missing at random (15%), and some are missing not at random (20%).

For MNAR data, the outcomes of interest for 80 subjects (0.2*400=80, 40 from placebo group, 40 from drug group) are generated from the model I: $Y_i = X_i\beta^{(1)} + U_i + \varepsilon_i$ for MNAR in dropout pattern 1 (*r=1*), where $\beta^{(1)}$ is the set of fixed covariates' parameters corresponding to the MNAR in pattern 1. For the other 320 subjects (400-80=320, 160 in placebo group, 160 in drug group), data are simulated from the model II: $Y_i = X_i\beta^{(2)} + U_i + \varepsilon_i$ for completers in pattern 2 (*r=2*) and subjects with ignorable missingness (MCAR & MAR) in pattern 1 (*r=1*). To generate ignorable missingness for each treatment group, we randomly select 10 subjects (0.05*200=10) from 160 subjects in each group and make them drop at time point 3 to get MCAR. Then order the values at time point 3 for the remaining completers and let the top 30 subjects (0.15*200=30) drop at time point 3 to generate MAR. We use

$$\beta^{(1)} = \left(\beta^{,(1)}_{intercept}, \beta^{,(1)}_{time,}\ \beta^{,(1)}_{treat}, \beta^{,(1)}_{time*treat}\right) = (12, -0.2, 0, -0.2)$$

and

$$\beta^{(2)} = \left(\beta^{,(2)}_{intercept}, \beta^{,(2)}_{time,}\ \beta^{,(2)}_{treat}, \beta^{,(2)}_{time*treat}\right) = (10, -0.4, 0, -0.4).$$

In this scenario, 50% of dropouts are MCAR /MAR, and 50% of dropouts are MNAR. The main simulation steps are shown in figure 8.

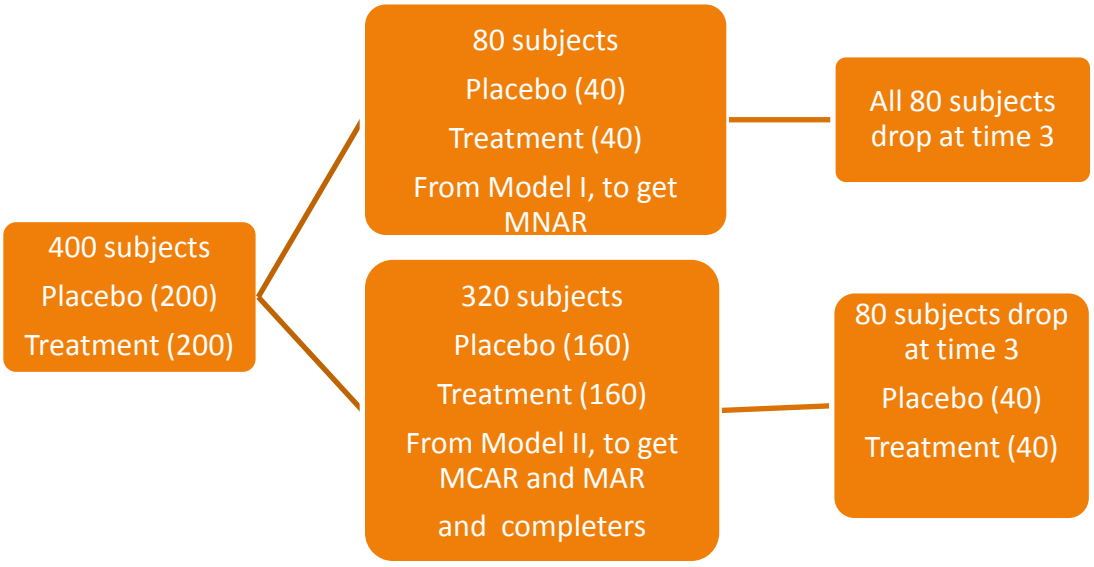

Figure 3.4. Mixed dropouts' data simulation steps (p=0.5)

1000 simulated datasets are used for each scenario.

3.3.2 Simulation results

For both dropout patterns, regression coefficients for both the fixed effects time and treatment - by - time interaction are assumed to be negative. If we assume a smaller outcome is better, this model reflects that the subjects' outcomes are expected to improve in both the drug and placebo groups over time with more improvement among the drug group. It is also noted that the same dropout proportions and time for placebo and tested drug groups are simulated for the entire four dropout scenario. The parameter of interest is the difference in outcome between placebo group and tested drug group at time point 3 (the end of study).

The popular methods ANCOVA for complete case, Mixed Model for Repeated Measures (MMRM), Generalized Estimation Equation (GEE) with exchangeable or compound symmetric (CS) working correlation matrix, Last Observation Carried Forward (LOCF) followed by ANCOVA, Baseline Observation Carried Forward (BOCF) followed by ANCOVA, Weighted GEE (WGEE), Pattern Mixture Models and proposed PMWMM are applied and compared. All the models I used here are linear models with treatment, time and the treatment - by - time interaction as covariates.

3.3.2.1 Results for MCAR scenario

Table 3.1 shows the models' results for the first scenario – all missingness is completely at random (MCAR). ANCOVA for complete case, MMRM, GEE, WGEE, and PMM all provide the unbiased results in least square means and difference in outcome between two treatment groups at the end of study. PMWMM cannot obtain convergent results for data with MCAR dropout. In comparison, LOCF and BOCF produce conservative estimates with estimates biased towards the null hypothesis – no statistical significant difference in outcome of interest between placebo group and tested drug group at the end of study. These conservative results are predictable since the subjects' outcome of interest is expected to be better over time and the dropout situation are balanced between two treatment groups, so more conservative outcomes in tested drug group than in placebo group would be obtained under LOCF or BOCF method. Therefore, these two methods are not valid under MCAR and this scenario.

Table 3.1. Means of outcome of interest at the end of study (time point 3) from the simulated study with 400 subjects and 4 equally spaced time points. All missing caused by dropout are MCAR.

Pain Score	Placebo	Drug	LSMD (95% CI)	P-value
	Mean (SEM)	Mean (SEM)		
True Value	8.80	7.60	-1.20	
BOCF[a]	9.14 (0.04)	8.27 (0.04)	-0.86 (-0.99, -0.74)	<0.0001
LOCF[b]	8.91 (0.03)	7.82 (0.03)	-1.09 (-1.18, -1.00)	<0.0001
CC[c]	8.80 (0.04)	7.60 (0.04)	-1.20 (-1.30, -1.10)	<0.0001
MMRM[d]	8.80 (0.04)	7.60 (0.04)	-1.20 (-1.31, -1.09)	<0.0001
WGEE[e]	8.80 (0.04)	7.60 (0.04)	-1.20 (-1.30, -1.10)	<0.0001
GEE[f]	8.80 (0.04)	7.60 (0.04)	-1.20 (-1.31, -1.10)	<0.0001
PMM[g]	8.80 (0.04)	7.60 (0.04)	-1.20 (-1.31,-1.09)	<0.0001
PMWMM[h]	-	-	-	-

a: Using baseline observation carried forward method to impute missing values, then use ANCOVA to analyze this imputed full data set.
b: Using last observation carried forward method to impute missing values, then use ANCOVA to analyze this imputed full data set.
c: Using ANCOVA to analyze complete case.
d: Using Linear Mixed Model with random intercept to analyze observed data.
e: Using logistic model with explanatory variables treatment group and last observed change to get the weights, then use GEE model with weight to analyze observed data.
f: Using GEE model with compound symmetry correlation structure to analyze observed data.
g: Using PMMs with two dropout patterns, one is for dropouts and one is for completers
h: Using proposed PMWMM

3.3.2.2 Results for MAR scenario

Table 3.2 shows the models' results for data with MAR observations. MMRM, GEE and WGEE all produce unbiased estimates of group means and treatment effects. Here, the

GEE estimate is still consistent because it uses the correct working correlation matrix (Compound Symmetry) which matches with the structure of our simulated data set. When including only complete cases, the ANCOVA model, correctly estimates the difference between placebo and drug groups; however, the means of interested outcomes for each group are underestimated. The robust estimate of the difference here may occur because the numbers of dropouts between two treatment groups are balanced in this scenario. Hence complete case is not suitable under MAR assumption. LOCF and BOCF still generate the conservative estimates out the same levels as when missing data are MCAR, so they are not valid under MAR assumption either. PMM only provides correct estimates for the difference between two treatment groups, but wrong means for each treatment group, and larger standard errors for all estimates. PMWMM cannot obtain convergent results for data with MAR dropout again.

Table 3.2. Means of outcome of interest at the end of study (time point 3) from the simulated study with 400 subjects and 4 equally spaced time points. All missing caused by dropout are MAR.

Pain Score	Placebo	Drug	LSMD (95% CI)	P-value
	Mean (SEM)	Mean (SEM)		
True Value	8.80	7.60	-1.20	
BOCF[a]	9.14 (0.05)	8.27 (0.05)	-0.87 (-1.00, -0.73)	<0.0001
LOCF[b]	8.93 (0.04)	7.84 (0.04)	-1.09 (-1.19, -0.99)	<0.0001
CC[c]	8.63 (0.04)	7.43 (0.04)	-1.20 (-1.30, -1.10)	<0.0001
MMRM[d]	8.80 (0.04)	7.60 (0.04)	-1.20 (-1.31, -1.10)	<0.0001
WGEE[e]	8.80 (0.04)	7.60 (0.04)	-1.20 (-1.30, -1.10)	<0.0001
GEE[f]	8.80 (0.04)	7.60 (0.04)	-1.20 (-1.31, -1.10)	<0.0001
PMM[g]	9.68 (0.07)	8.48 (0.07)	-1.20	<0.0001

			(-1.40, -1.00)	
PMWMM[h]	-	-	-	-

a: Using baseline observation carried forward method to impute missing values, then use ANCOVA to analyze this imputed full data set.
b: Using last observation carried forward method to impute missing values, then use ANCOVA to analyze this imputed full data set.
c: Using ANCOVA to analyze complete case.
d: Using Linear Mixed Model with random intercept to analyze observed data.
e: Using logistic model with explanatory variables treatment group and last observed change to get the weights, then use GEE model with weight to analyze observed data.
f: Using GEE model with compound symmetry correlation structure to analyze observed data.
g: Using PMMs with two dropout patterns, one is for dropouts and one is for completers
h: Using proposed PMWMM

3.3.2.3 Results for completely MNAR scenario

Table 3.3 shows the models' results for completely MNAR. Except PMM and PMWMM, all the other approaches obtain biased results, especially WGEE. The extreme estimates of WGEE may be caused by the incorrect mean structure (measurement model) and the wrong weights obtained from the logistic regression model for the probability of dropout. The real distribution of outcomes for our simulated data is the mixture of two multivariate normal distributions rather than the marginal distribution used in the WGEE measurement model. Except for the Pattern Mixture Model, LOCF produces closed estimates to the true values. BOCF keeps giving the conservative estimates under this scenario. Complete case ANCOVA, MMRM and GEE result in biased treatment effects with this effects that are greater than the true value under the alternative hypothesis – the tested drug has significant positive effect, which bias is the most concern and tried to be avoided by FDA.

Table 3.3. Means of outcome of interest at the end of study (time point 3) from the simulated study with 400 subjects and 4 equally spaced time points. All missing caused by dropout are MNAR.

Pain Score	Placebo	Drug	LSMD	P-value
	Mean (SEM)	Mean (SEM)	(95% CI)	
True Value	9.84	8.88	-0.96	
BOCF[a]	10.01 (0.09)	9.22 (0.09)	-0.79 (-1.03, -0.55)	<0.0001
LOCF[b]	9.89 (0.08)	8.99 (0.08)	-0.90 (-1.13, -0.68)	<0.0001
CC[c]	9.23 (0.08)	8.13 (0.08)	-1.10 (-1.33, -0.86)	<0.0001
MMRM[d]	9.74 (0.09)	8.70 (0.09)	-1.04 (-1.30, -0.80)	<0.0001
WGEE[e]	12.61 (0.13)	12.48 (0.12)	-0.13 (-0.48, 0.20)	0.3919
GEE[f]	9.73 (0.09)	8.68 (0.09)	-1.05 (-1.30, -0.80)	<0.0001
PMM[g]	9.84 (0.04)	8.88 (0.04)	-0.96 (-1.07, -0.85)	<0.0001
PMWMM[h]	9.85 (0.04)	8.88 (0.04)	-0.97 (-1.08, -0.86)	<0.0001

a: Using baseline observation carried forward method to impute missing values, then use ANCOVA to analyze this imputed full data set.
b: Using last observation carried forward method to impute missing values, then use ANCOVA to analyze this imputed full data set.
c: Using ANCOVA to analyze complete case.
d: Using Linear Mixed Model with random intercept to analyze observed data.
e: Using logistic model with explanatory variables treatment group and last observed change to get the weights, then use GEE model with weight to analyze observed data.
f: Using GEE model with compound symmetry correlation structure to analyze observed data.
g: Using PMMs with two dropout patterns, one is for dropouts and one is for completers
h: Using proposed PMWMM

3.3.2.4 Results for Mixed dropouts

The results for the mixed missingness are shown in Table 3.4. The proposed Pattern Mixture-Within-Mixture Model produces the unbiased estimates. Comparing to the

proposed approach, Pattern Mixture Model also provides very close estimates to the true values but relatively bigger standard errors and confidence interval under our simulation conditions. When the data distribution and dropout structure are different with our simulated scenario, the difference between these two methods may be more obviously. The results of MMRM and GEE are a little far from the true ones and with larger effect. BOCF and LOCF underestimate the true effect for the simulated data with mixed dropouts.

Table 3.4. **Means of outcome of interest at the end of study (time point 3) from the simulated study with 400 subjects and 4 equally spaced time points. Missingness mechanisms caused by dropout are mixed: 20% MNAR, 15% MAR and 5% MCAR (the probability of MNAR dropout is 0.5)**

Pain Score	Placebo	Drug	LSMD	P-value
	Mean (SEM)	Mean (SEM)	(95% CI)	
True Value	9.32	8.24	-1.08	
BOCF[a]	9.52 (0.07)	8.72 (0.07)	-0.80 (-1.00, -0.60)	<0.0001
LOCF[b]	9.37 (0.07)	8.40 (0.07)	-0.97 (-1.16, -0.79)	<0.0001
CC[c]	8.90 (0.07)	7.78 (0.07)	-1.12 (-1.31, -0.94)	<0.0001
MMRM[d]	9.27 (0.08)	8.15 (0.08)	-1.12 (-1.33, -0.91)	<0.0001
WGEE[e]	9.45 (0.10)	8.47 (0.12)	-0.98 (-1.28, -0.68)	0.0067
GEE[f]	9.25 (0.08)	8.12 (0.09)	-1.13 (-1.35, -0.90)	<0.0001
PMM[g]	9.30 (0.05)	8.24 (0.05)	-1.06 (-1.21,-0.91)	<0.0001
PMWMM[h]	9.31 (0.04)	8.23 (0.04)	-1.08 (-1.19,-0.97)	<0.0001
	p (the probability of MNAR) =0.50			

a: Using baseline observation carried forward method to impute missing values, then use ANCOVA to analyze this imputed full data set.
b: Using last observation carried forward method to impute missing values, then use ANCOVA to analyze this imputed full data set.
c: Using ANCOVA to analyze complete case.
d: Using Linear Mixed Model with random intercept to analyze observed data.

e: Using logistic model with explanatory variables treatment group and last observed change to get the weights, then use GEE model with weight to analyze observed data.
f: Using GEE model with compound symmetry correlation structure to analyze observed data.
g: Using PMMs with two dropout patterns, one is for dropouts and one is for completers
h: Using proposed PMWMM

3.4 Worked Example: Schizophrenia study

The National Institute of Mental Health Schizophrenia Collaborative Study collected longitudinal data on treatment related changes in overall severity (Hedeker and Gibbons,1997). The main outcome of interest was drawn from the Inpatient Multidimensional Psychiatric Scale (IMPS), a series of questions which enables immediate stratification into psychotic types and the degree of mental illness. Specially, item #79 of IMPS, severity of illness can be used to evaluate the effect of treatment for Schizophrenia. It is scored as 1 = normal, not at all ill; 2 = borderline mentally ill; 3 = mildly ill; 4 = moderately ill; 5 = markedly ill; 6 = severely ill; and 7 = among the most extremely ill.

In this study, 437 patients were randomly assigned to receive one of four medications: placebo (108 patients), chlorpromazine, flupenzaine or thioridazine. Since Hedeker's previous analyses showed that the three drugs have similar effects on Schizophrenia, they were combined into one drug group (329 patients) for comparison with placebo here. We want to compare the tested drug group (329 patients) with the placebo group (108 patients). Patients' IMPS were focused measured on weeks 0, 1, 3, and 6. The completers are defined as those who were measured at week 6, and the percentages of

patients who dropped from the study were 34% and 19% for the placebo and drug groups respectively. Detailed dropout reasons for each patient are not publicized. However from other similar studies, common dropout reasons can be predicted, such as adverse events, lack of treatment effect and withdrawal for unspecified reasons, etc.. Reasons such as lack of treatment effect, adverse events may be related to the outcome of interest, but others may not be. The nature of dropout or missingness for this study may be a mixture of MCAR, MAR and MNAR

Dropout numbers and cumulative information at each mainly visit are represented at Table 3.5 and Figure 3.5.

Table 3.5. Dropout number by week and treatment group

	Placebo (N=108)			Drug (N=329)		
week	Count	Cum count	Cum Percent (%)	Count	Cum count	Cum Percent (%)
1	2	2	2	6	6	2
3	18	20	19	34	40	12
6	17	37	34	22	62	19

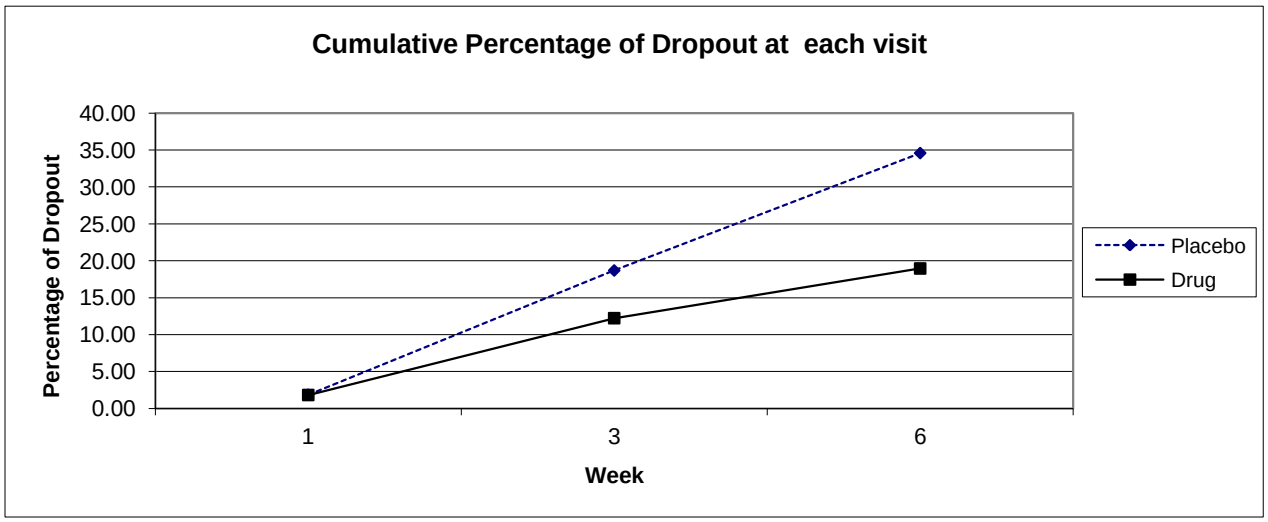

Figure 3.5. Cumulative percentage of dropout by week

From above table and figure, we can easily find that there are many more percentages of dropouts occur in placebo group (total 34%) than in tested drugs group (total 19%). It indicates that the dropout mechanism is not MCAR, but related to at least treatment.

The primary outcome of interest indicating efficacy for treatment is the IMPS #79 score, severity of illness. Table 3.5 showed that the size of dropouts at each time point is small, so all the dropouts are grouped together. The means of IMPS #79 score at each visit in placebo and drug groups for dropouts, completers and all patients are presented in Table 3.6 and Figure 3.6. The means of IMPS #79 score decrease *less* in placebo dropouts' group (from 5.58 to 5.31 to 5.23) than in placebo completers' group (from 5.23 to 4.82 to 4.59 to 4.25). In comprison, IMPS #79 scores decrease *more* in drug dropouts' group (from 5.32 to 4.19 to 2.74) than in drug completers' group (from 5.39 to 4.49 to 3.95 to 3.06). Figure 3.6 suggests that dropout is not MCAR again because the mean IMPS #79 scores have different trends for completers and dropouts. Although the detailed dropout reasons for each patient were unknown, based on these results, one might hypothesize more patients dropped out in the placebo group due to lack of efficacy, while in drug group more patients may dropped out because of good efficacy. Figure 3.6 also demostrates that the means of IMPS #79 score are approximately linear with square root of week, and the linear trends for dropouts and completers are not the same.

Table 3.6. Mean of IMPS #79 score at each visit by treatment group for dropouts and completers

		All		Dropouts		Completers	
week	Sqrt. of week	Placebo	Drug	Placebo	Drug	Placebo	Drug
0	0	5.35	5.38	5.58	5.32	5.23	5.39
1	1	4.99	4.43	5.31	4.19	4.82	4.49
3	1.73	4.74	3.80	5.23	2.74	4.59	3.95
6	2.45	4.25	3.06			4.25	3.06

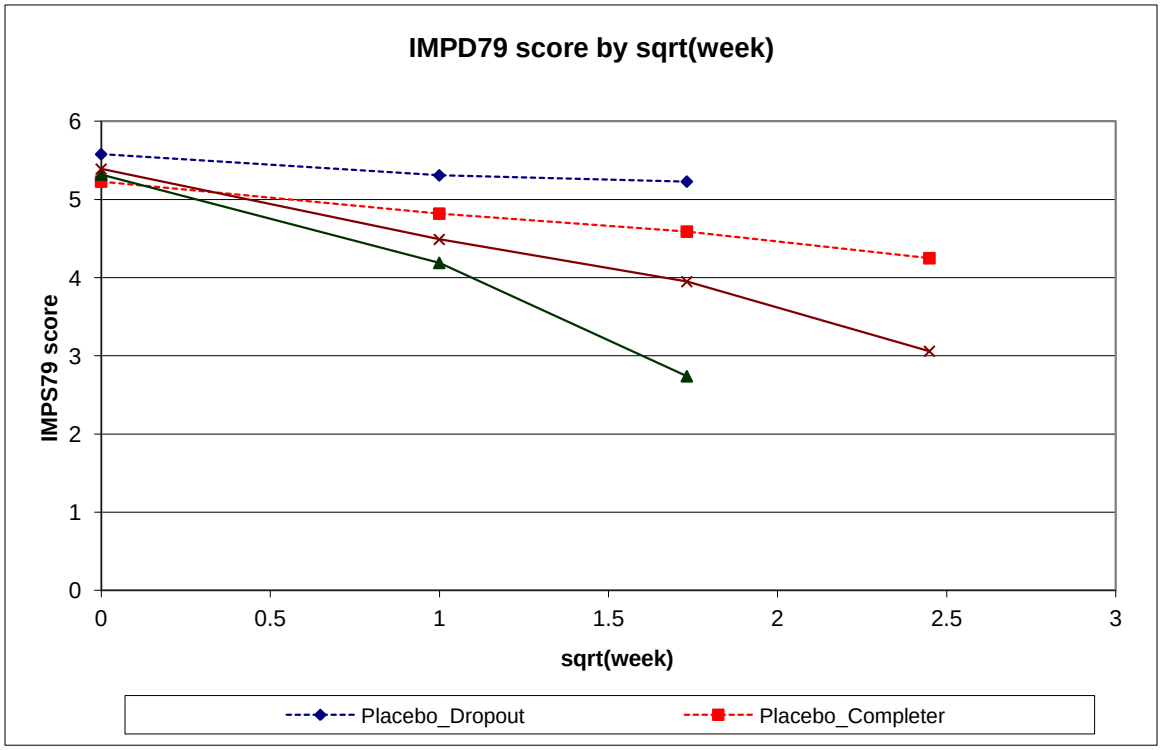

Figure 3.6. Mean of IMPS79 score at each visit by treatment group for dropouts and completers

The Pattern Mixture-Within-Mixture Model, which uses an EM algorithm within a Pattern Mixture Model was motivated by this study, illustrating the need of accounting for mixed dropout mechanisms. In particular, for this study, dropout reasons are not clear and a linear relationship between the outcome of interest and time can be assumed. In analyzing this longitudinal study, both my proposed method Pattern Mixture-Within-

Mixture Model and Pattern Mixture Model assume that the data is separated into two patterns – dropouts and completers. While the Pattern Mixture Model assumes missing values of all patients with dropout follow the same pattern represented by a single multivariate normal distribution. The PMWMM assumes that dropouts responses may follow two patterns, as modeled by a mixture of two multivariate normal distributions, which one is for MNAR and the other is for MAR and MCAR. The latter distribution is also assumed to be the same as completers'.

Last Observation Carried Forward (LOCF), Baseline Observation Carried Forward (BOCF), ANCOVA for complete case, Mixed Models for Repeated Measurements (MMRM), Generalized Estimation Equation (GEE), and Weighted GEE (WGEE) are also applied to the data from this study. Description of these approaches can be found in section 2.3. All the linear models use treatment group, time (square root of week) and the treatment - by - time interaction as covariates:
Imps #79 score = time +treat + time*treat.

The results are compared in Table 3.7. We can find that all the approaches give significant results for the difference in mean IMPS#79 score between placebo and drug groups at the end of study (week 6). IMPS #79 score in drug group is significantly lower than in the placebo group, showing a robust positive effect of drug. Among these methods, our proposed method give the largest difference of IMPS #79 score (-1.62) between two treatment groups. Pattern Mixture Model also provides big difference (-

1.57) compare to the remaining methods. MMRM, GEE and WGEE methods obtain estimates that are similar to one another (-1.38, -1.38 and -1.36 respectively).

In this study, complete case ANCOVA induces the smallest difference (-1.25). From Figure 3.6, we can find that the difference between treatment groups is obviously smaller in completers than in dropouts. If we excluded all the dropouts, the results must be biased towards null hypothesis direction – no difference between placebo and tested drug groups. BOCF method also produces similar small difference (-1.27) as complete case analysis does.

The simulation study suggested that LOCF can be a conservative approach, yielding estimates that are closed to the null value. However, as evidenced here, compared to MMRM and WGEE methods under MAR assumption, LOCF may not always be a conservative statistical technique for longitudinal data analysis which is preferred by regulation agency. Since the outcome of interest (IMPS #79 score) decreases on average over time, the biased estimates toward bigger IMPS #79 score are obtained for both treatment groups under the LOCF method. However, the difference between two groups may not be biased towards null hypothesis, because there are more dropouts in placebo group than in drug group (34% vs. 19%). Thus, the influence of LOCF is more serious in placebo group than in drug group.

Table 3.7. Evaluation of IMPS #79 score at 6-weeks using different methods.

IMPS #79 score at end of study				
	Placebo (N=450)	Drug (N=450)	LSMD (95%CI)	P-value

	Mean (SEM)	Mean (SEM)		
LOCF[a]	4.65 (0.10)	3.15 (0.06)	-1.50 (-1.72,-1.27)	<.0001
BOCF[b]	4.70 (0.10)	3.44 (0.06)	-1.27 (-1.49,-1.04)	<.0001
CC[c]	4.34 (0.11)	3.09 (0.06)	-1.25 (-1.50, -1.01)	<.0001
MMRM[d]	4.43 (0.12)	3.05 (0.06)	-1.38 (-1.64, -1.11)	<.0001
GEE[e]	4.43 (0.14)	3.05 (0.08)	-1.38 (-1.70, -1.05)	<.0001
WGEE[f]	4.44 (0.15)	3.08 (0.08)	-1.36 (-1.70, -1.03)	<.0001
PMM[g]	4.45 (0.12)	2.88 (0.07)	-1.57 (-1.85, -1.30)	<.0001
PMWMM[h]	4.45 (0.12)	2.83 (0.06)	-1.62 (-1.89, -1.35)	<.0001
	p (the probability of MNAR) =0.69			

a: Using last observation carried forward method to impute missing values, then use ANCOVA to analyze this imputed full data set.
b: Using baseline observation carried forward method to impute missing values, then use ANCOVA to analyze this imputed full data set.
c: Using ANCOVA to analyze complete case.
d: Using Linear Mixed Model with random intercept to analyze observed data.
e: Using GEE model with compound symmetry correlation structure to analyze observed data.
f: Using logistic model with explanatory variables treatment group and last observed change to get the weights, then use GEE model with weight to analyze observed data.
g: Using Pattern Mixture Model with two dropout patterns to analyze observed data.
h: Using proposed Pattern Mixture-Within-Mixture Model to analyze observed data.

3.5 Summary

It is clearly that all methods for handling data with MNAR must make some unverifiable assumptions concerning the unknown missing values. Therefore, no single MNAR approach can be used definitively. In particular with unknown dropout reasons, it is difficult to tell which sensitive method (PMM or PMWMM) is more correct or efficient since they are dependent on different assumptions. Rather, in any study with dropout,

several methods based on different reasonable assumptions regarding to the missingness mechanisms should be considered as a sensitivity analyses family. Obtaining similar results from these methods will point to robust conclusions. On the other hand contrasting results may aid investigators to understand their study results more deeply.

The results from simulation and Schizophrenia study showed that there may be no big different marginal outcome of interest between PMWMM and PMM for some studies. If the distinction between two interested groups (e.g. placebo and tested drug) is very significant or very non-significant, we don't need to pay much attention on explaining these two analyses methods and their results. But if the difference is very close to the critical value, then these two methods may produce distinct test results – one significant and one non-significant. Under this situation, more discussions will be needed to explain the different MNAR methods' results and the assumptions about the missing values.

Both PMM and PMWMM discussed in this Chapter assume that the response for each pattern is a function of time and other covariates. For example, in Schizophrenia study, we assume the linear relationship between IMPS #79 score and the time (square root of week) and time-treatment interaction. However, this assumption may not be correct to many other studies. So other approaches need to be developed under mixed dropout assumptions if a specific function cannot be specified for the relationship between response and time. My next proposed method for handling mixed dropout will combine Pattern Mixture Model and Multiple Imputation, and use mixed missing values restriction

to resolve the under-identification of the PMM when the response is not necessary a predefined function of time and other covariates.

4 Pattern Mixture Models based Multiple Imputation with mixed missing values restriction

Dropout is an important and unavoidable problem in longitudinal clinical trials. The dropout mechanisms need to be considered very carefully in the data analyses in order to reduce as much as possible the possibility that the dropouts will bias the study results. Much literature has focused on statistical methods for handling dropouts. For continuous data, regression methods range from using simple ANCOVA with only complete cases to more complicated approaches such as Pattern Mixture Models, which treats all the dropouts as if they fall in the same dropout mechanism. In this chapter, I focus on estimating the treatment effect in a longitudinal clinical trial with a continuous outcome of interest, such that incomplete cases are due to mixed dropout mechanisms (ignorable missingness (MCAR/MAR), and non-ignorable missingness(MNAR)). The treatment effect is expressed by comparing the changes from baseline at the end of study between two treatment groups (e.g. placebo vs. tested drug).

For analyzing the repeated continuous data with monotone missingness caused by dropouts (described in more detail in Chapter two), ANCOVA and GEE are only suitable under MCAR assumption, whereas MMRM and MI are recommended under the MAR/MCAR assumption. Selection Models, shared-parameter models and Pattern Mixture Models are used as sensitivity analyses under MNAR assumption. Although LOCF and BOCF used to be popular methods for longitudinal clinical trials with dropouts, they do make strong MNAR assumption and are no longer recommended by FDA as primary analyses.

In the last chapter, I proposed Pattern Mixture-Within-Mixture Models, integrating an EM algorithm within PMMs to handle the mixed dropout problem when the reasons for dropout are unknown. However, in contemporary clinical trials, the reasons of dropout are often recorded as clearly as the investigators can, so that the dropout reasons are known rather than unknown. The differential dropouts may result in missing values that can be classified as ignorable missingness or non-ignorable missingness, depending on whether the dropout is related to the observed or/and unobserved data. Further, to overcome the under-identified problem of PMMs, the previous proposed method assumes the linear effects of time and treatment by time interaction to the outcome of interest, a very strong and restrictive assumption which may be untrue in many clinical trials with longitudinal follow-up.

As another approach, Lu (2011) proposed imposing a non-future missing value restriction when using PMMs combined with multiple imputation for handling dropouts under the MNAR assumption. I extended this approach to deal with mixed dropout mechanisms, a potentially more realistic scenario for some longitudinal studies.

The proposed imputation strategy handles longitudinal studies with continuous responses and mixed dropout mechanisms – ignorable missingness and non-ignorable missingness. I apply two missing value restrictions (available-case and non-future dependence) within PMMs, combined with multiple imputation to fill in the missing values caused by different dropout categories. And then the multiple imputed complete data sets are analyzed by a selected method, e.g. GEE, to catch the treatment effects – the difference

of changes from baseline at the end of study between two treatment groups. Finally, the results from each imputed complete data set are combined based on standard multiple imputation methodology (see section 2.3.1.7). The proposed imputation step uses the dropout information from the trials and relaxes the assumption of linear effects of time and treatment by time interaction, flexibly allowing the user to choose the identifying restrictions or assumptions regarding the inestimable parameter that best fit the problem at hand.

The pattern mixture models (PMMs) based multiple imputation (MI) with mixed missing value restrictions (available-case and non-future dependence) approach is described in the next two sections, along with details for implementation. This method is applied to a simulated chronic pain clinical trial, demonstrating the effect of a drug for chronic pain. Results are compared to more traditional methods for longitudinal study with dropouts. Finally, procedures for sample size and power calculations through simulation are provided for the proposed method and these results are compared to the ones from other approaches.

4.1 Two missing data structures within PMMs

Assume there are *T* designed visits in a longitudinal study and let $y_t(t = 1,2, \dots, T)$ represent the value of a patient's outcome measure (continuous variable) at visit *t*. Before introducing imputation strategy, I will fully describe the data structure and the assumptions for modeling. The structure of the data is laid out in the following figure.

O	M	M	M	*R*=1
O	O	M	M	*R*=2
O	O	O	M	*R*=3
O	O	O	O	*R*=4

Here, *R* denotes the Dropout Pattern, *R=1,…,T*. *R=t* indicates the number of observed measurements in this pattern, such that subjects with this pattern drop out of the study between visit *t* and visit *t+1* (the measurements from visit *t+1* to *T* are missing); O represents an observed value; the gray capital M indicates the first missing value or current missing value; and black capital M denotes a future missing value. Throughout discussion that follows, the "future" for any particular subject refers to any time point beyond the first one with a missing value, i.e., *T*>*t*+1.

Recall, a PMM describes the full data likelihood as the product of the density of measurement process given dropout pattern and the density of dropout process. Specifically, it can be decomposed as follows:

$$f(y_1, \dots, y_T, R = t|X) = f(y_1, \dots, y_T|R = t, X)f(R = t|X)$$

$$= f(Y_{obs}|R, X)f(Y_{mis}|Y_{obs}, R, X)f(R = t|X), \quad (4.1\text{-}1)$$

where $Y_{obs} = y_1, \dots, y_t$, and $Y_{mis} = y_{t+1}, \dots, y_T$. $f(Y_{obs}|R, X)$ represents the conditional distribution for observed outcomes given each dropout pattern and covariates (e.g. treatment), and $f(Y_{mis}|Y_{obs}, R, X)$ represents conditional distribution for missing values given observed data and covariates within each dropout pattern.

The distribution function of PMMs in equation (4.1-1) can be further decomposed as

$$f(y_1, \dots, y_T, R = t|X)$$

$$= f(y_1, \dots, y_t|R = t, X) f(y_{t+1}|y_1, \dots, y_t, R = t, X) \prod_{s=t+2}^{T} f(y_s|y_1, \dots, y_{s-1}, R = t, X) f(R = t|X) \quad (4.1\text{-}2)$$

In this function, only the first and the last factors can be identified from the data, while other factors – the distributions of first and future missingness are under-identifiable, because the missing values are unobserved and the true relationship between missingness and observations are unknown. The under-identification of PMMs actually increases the flexibility of these models, allowing the researchers to use different reasonable identifying restrictions for the missing values.

To handle the mixed dropout mechanisms, the model will be based on the following two missing data restrictions– available-case missing value restriction for ignorable missingness and non-future dependent missing value restriction for non-ignorable missingness.

4.1.1 Restrictions for ignorable missingness

Only one restriction, available-case missing value restriction, is needed for the first and future missing values. This restriction assumes that the conditional density of a missing observation is equal to the density of that observation for the available cases. This is represented by

$$f(y_t|y_1,\dots,y_{t-1},R=j,\mathrm{X}) = f(y_t|y_1,\dots,y_{t-1},R\geq t,\mathrm{X})$$

for all $t \geq 2$ and $j < t$.

Molenberges gave a detailed proof in his paper (1998) to show that for longitudinal data with dropouts, available-case missing value restriction is equivalent to MAR, as expressed by

$$f(y_t|y_1,\dots,y_{t-1},R=j,\mathrm{X}) = f(y_t|y_1,\dots,y_{t-1},R\geq t,\mathrm{X})$$

$$\Leftrightarrow$$

$$f(R=t|y_1,\dots,y_T,\mathrm{X}) = f(R=t|y_1,\dots,y_t,\mathrm{X}),$$

which means that the MAR assumption holds if the predictions of missing values for those who drop at time *j* are equivalent in distribution to the predictions for the available cases (observations). Thus, under MAR, missing values at time *j* and beyond can be predicted sequentially from the subjects who have the observations at those times.

Available-case missing value restriction method will be used to impute the missing values caused by MAR, which is ignorable missingness. For example, if the repeated measures of interest $Y=(y_1,\dots,y_T)$ follows a multivariate normal distribution, given as

$$Y \sim MVN(\mu, \Sigma).$$

For the dropouts at pattern *t*, $Y=(Y_{obs}, Y_{mis})$, where $Y_{obs}=(y_1,\dots,y_t)$ and $Y_{mis}=(y_{t+1},\dots,y_T)$. The conditional distribution of Y_{mis} given Y_{obs} follows a multivariate normal distribution too, given by

$$Y_{mis|obs} \sim MVN\left(\mu_{mis|obs}, \Sigma_{mis|obs}\right)$$

MI MCMC data augmentation will be used to impute these missing values. The details of imputation method and steps will be introduced in section 4.2.2.3.

4.1.2 Restrictions for Non-ignorable missingness

Missing data from participants with non-ignorable missingness are separated into two types: first missing value and future missing values. Different assumptions or "restrictions" are used for each type.

For the first missing value y_{t+1}, National Research Council in "*The prevention and treatment of missing data in clinical trials*" (2010) provides a way to identify the distribution of missing value by introducing a sensitivity parameter Δ, expressed as

$$f(y_{t+1}|y_1, \dots, y_t, R = t, X) = f(y_{t+1} + \Delta|y_1, \dots, y_t, R \geq t + 1, X)$$

Through modifying the values of Δ, the analysis can be conducted so that the results favor the null hypothesis, an approach preferred by regulation agencies. Hence, adding shift for the first missing values will be used for treating non-ignorable missingness dropout in my proposed method.

For example, supposing the repeated measures of interest $Y=(y_1, \dots, y_{t+1})$ follows a multivariate normal distribution, y_{t+1} is the first missing. We draw a value Y_{t+1}^* from the conditional distribution of the first missing value $Y_{t+1|obs} \sim MVN(\mu_{t+1|obs}, \Sigma_{t+1|obs})$ for Y_{t+1} through MCMC data augmentation firstly, then impute the first missing value by $Y_{t+1}^* + \Delta$.

Although the exact shift value is unknown, a reasonable approximated range based on the minimum and maximum values of the primary efficacy variable can be set (Lu, 2011).

According to the relevant historical data, shift value can be selected as a series of proportions of the observed treatment efficacy (eg. 10%, 20%, and 30%, etc.).

For future missing values y_j where $j>t+1$, Kenward et al. proposed a non-future-dependent (NFD) missing value restriction (Kenward 2003) for PMMs. This restriction assumes that the probability of dropout is only related to the observations and the first missing value, but not related to the future missing values. Within the PMMs framework, the non-future dependent missing value restriction can be defined as follows:

$$f(y_s|y_1,\dots,y_{s-1},R=t,X)=f(y_s|y_1,\dots,y_{s-1},R\geq s-1,X) \qquad (4.1.2\text{-}2)$$

for all $T\geq s\geq t+2$, and all $t\geq 1$. For example, at pattern R=1such that T=4, we can demonstrate using the NFD missing value restriction that

$$f(y_3|y_1,,y_2,R=1,X)=f(y_3|y_1,y_2,R\geq 2,X)$$

and

$$f(y_4|y_1,\dots,y_3,R=1,X)=f(y_4|y_1,\dots,y_3,R\geq 3,X)$$

For the remainder of this section, we demonstrate the connection under this restriction with Selection Models and justify its use when data are assumed to be non-ignorable (MNAR). Recall, Selection Models can be factorized as the product of the marginal density of the measurement process and the density of dropout process conditional on the outcomes

$$f(y_1,\dots,y_T,R=t|X)=f(y_1,\dots,y_T|X)f(R=t|y_1,\dots,y_T,X)$$

When selection models are used, researchers usually assume that the probability of dropout at pattern *t* is only related to the covariates, observations $y_1, \ldots, y_t$, and the first missingness y_{t+1}, which is called "missing non-future dependent". It can be expressed by

$$f(R = t|y_1, \ldots, y_T, X) = f(R = t|y_1, \ldots, y_{t+1}, X)$$

For example, at pattern *R=1*, the subjects drop out between visit *1* and *2*. The probability of dropout at this pattern is dependent on the covariates, observation y_1, and the first missing value y_2,

$$f(R = 1|y_1, \ldots, y_T, X) = f(R = 1|y_1, y_2, X).$$

Kenward et al. (2003) proved that the "non-future dependent missing value restriction" used in PMMs is equivalent to the "missing non-future dependent" assumption used in SMs. Specifically

$$f(y_s|y_1, \ldots, y_{s-1}, R = t, X) = f(y_s|y_1, \ldots, y_{s-1}, R \geq s - 1, X)$$

$$\Leftrightarrow$$

$$f(R = t|y_1, \ldots, y_T, X) = f(R = t|y_1, \ldots, y_{t+1}, X)$$

In addition, we can demonstrate that the non-future dependent missing value restriction for PMMs implies a MNAR mechanism. Note that this restriction within PMMs is recommended by National Research Council in "*The prevention and treatment of missing data in clinical trials*" (2010) to handle missingness under MNAR assumption.

Non-future dependent missing value restriction within PMMs will be only used to impute missing values caused by non-ignorable missingness dropout; while for the missing

values from ignorable missingness dropout, available-case missing value restriction is applied as discussed next.

4.2 PMMs based Multiple Imputation with mixed missing value restrictions

4.2.1 Classification of dropout categories

A key issue to mixed dropout problem is how to assign the dropouts to ignorable missingness or non-ignorable missingness category. Usually, the assignment is based on the dropout information collected from a clinical trial termination form that includes a list of potential dropout reasons. The main reasons may be: (1) didn't get better, give up; (2) death; (3) side effects; (4) move from study area (5) study requirement is too onerous, give up (6) unrelated illness required treatment with contraindicated medicine, and so on. Some of dropout reasons may not be related to the missing measurements, which can be assigned into ignorable missingness. Alternatively, some reasons may be related to the missing measures of interest, implying non-ignorable missingness, e.g. side effects or death. Due to the study design, the termination form may vary from one study to another, so that these details concerning dropout classification will need to be determined within each study.

4.2.2 Imputation strategy

Based on dropout reasons, the dropouts can be separated into ignorable and non-ignorable. A Bernoulli variable *G* is defined for the differential dropout categories. *G=1* indicates non-ignorable missingness dropout, while *G=0* indicates ignorable missingness dropout. Treatment is set as the covariate *X* (*0* – placebo, *1* – tested drug).

Let $f_{G=g}(y_i, \dots, y_j | R = t, X = x)$ express the conditional density of y_i ,... y_j, given the last observed measurement is at visit *t* ,the participant is receiving treatment *x* and the dropout category is *g.* Then the complete data density for pattern *t* and dropout type *g* can be written as

$$f_{G=g}(y_1, \dots, y_T | R = t, X = x)$$

$$= f_{G=g}(y_1, \dots, y_t | R = t, X = x) f_{G=g}(y_{t+1} | y_1, \dots, y_t, R = t, X = x) \prod_{s=t+2}^{T} f_{G=g}(y_s | y_1, \dots, y_{s-1}, R = t, X = x)$$

(4.2.2-1)

The first factor on the right side of (4.2.2-1) is clearly identifiable from the observed values. The second factor $f_{G=g}(y_{t+1} | y_1, \dots, y_t, R = t, X = x)$ represents the conditional distribution for the first missingness. The third and beyond factor $f_{G=g}(y_s | y_1, \dots, y_{s-1}, R = t, X = x)$ (with all $T \geq s \geq t + 2$) represent the conditional distribution for the future missing values. All the missing values can be identified by various assumptions due to the dropout categories - ignorable or non-ignorable.

4.2.2.1 Imputation for the first missing

4.2.2.1.1 First missing from ignorable missingness dropout

Available-case missing value restriction is used for the first missing value at pattern *t* in treatment *x* from ignorable missingness dropout. The conditional densities can be imputed by

$$f_{G=0}(y_{t+1}|y_1,\dots,y_t,R=t,X=x)=f(y_{t+1}|y_1,\dots,y_t,R\geq t+1,X=x), \quad (4.2.2.1\text{-}1)$$

which means the predictions of missing values given previous measures under ignorable missingness assumption for those who drop at time *t* in treatment *x* are equivalent in distribution to the predictions for the observations at that time given previous measures in treatment *x*.

4.2.2.1.2 First missing from non-random dropout

The conditional densities for the first missing at pattern *t* in treatment *x* from non-random dropout are selected as

$$f_{G=1}(y_{t+1}|y_1,\dots,y_t,R=t,X=x)=f(y_{t+1}+\Delta|y_1,\dots,y_t,R\geq t+1,X=x) \quad (4.2.2.1\text{-}2)$$

The difference between these two distributions in (4.2.2.1-2) is a shift Δ parameter. Compared to the distribution of all observed y_{t+1}, the distribution of the missing y_{t+1} caused by non-random dropout at pattern *t* will shift Δ units. When $\Delta \neq 0$, the conditional distribution of y_{t+1} in pattern *t* is different from that of observed measurements, representing a scenario of missing not at random.

To make sure the sensitivity analysis under mixed dropouts assumption is conservative, with estimates biased towards the null hypothesis, a different shift parameter Δ may be

preferred for various treatment groups. For example, for the placebo or control group, we may set Δ=0 for the non-ignorable missingness dropouts; while for the drug group, we may try a series of non-zero shifts for these dropouts.

4.2.2.2 Imputation for the future missing

4.2.2.2.1 Future missing from ignorable missingness dropout

Under random dropout, available-case missing value restriction is also used for the future missing value at pattern *t* in treatment *x*. The conditional densities of the missing values can be obtained by assuming

$$f_{G=0}(y_s|y_1,\dots,y_{s-1},R=t,X=x)=f(y_s|y_1,\dots,y_{s-1},R\geq s,X=x)$$

(4.2.2.2-1)

for all $T\geq s\geq t+2$, and all $t\geq 1$.

4.2.2.2.2 Future missing from non-ignorable missingness

Non-future dependent missing value restriction is applied to the future missing values from non-random dropout. It assumes that the dropout is only related to the observed and first missing measurements, but not related to future missing measurements. The unidentifiable conditional distributions of future missing y_s at pattern *t* in treatment *x* satisfy

$$f_{G=1}(y_s|y_1,\dots,y_{s-1},R=t,X=x)=f(y_s|y_1,\dots,y_{s-1},R\geq s-1,X=x)$$

(4.2.2.2-2)

for all $T\geq s\geq t+2$, and all $t\geq 1$.

This means that all the future missing values in non-random dropouts are imputed based on the observed and imputed first missing values.

The right hand side of (4.2.2.2-2) can further be partitioned into

$$\mathrm{f}(\mathrm{y_s}|\mathrm{y_1},\dots,\mathrm{y_{s-1}},\mathrm{R}\geq \mathrm{s}-1,\mathrm{X}=\mathrm{x})$$

$$=\frac{f(y_1,\dots,y_s,R\geq s-1|\mathrm{X}=\mathrm{x})}{f(y_1,\dots,y_{s-1},R\geq s-1|\mathrm{X}=\mathrm{x})}$$

$$=\frac{f(y_1,\dots,y_s,R=s-1|\mathrm{X}=\mathrm{x})+f(y_1,\dots,y_s,R\geq s|\mathrm{X}=\mathrm{x})}{f(y_1,\dots,y_{s-1},R\geq s-1|\mathrm{X}=\mathrm{x})}$$

$$=\frac{f(y_1,\dots,y_{s-1},R=s-1|\mathrm{X}=\mathrm{x})}{f(y_1,\dots,y_{s-1},R\geq s-1|\mathrm{X}=\mathrm{x})}\mathrm{f}(\mathrm{y_s}|\mathrm{y_1},\dots,\mathrm{y_{s-1}},\mathrm{R}=\mathrm{s}-1,\mathrm{X}=\mathrm{x})$$

$$+\frac{f(y_1,\dots,y_{s-1},R\geq s|\mathrm{X}=\mathrm{x})}{f(y_1,\dots,y_{s-1},R\geq s-1|\mathrm{X}=\mathrm{x})}\mathrm{f}(\mathrm{y_s}|\mathrm{y_1},\dots,\mathrm{y_{s-1}},\mathrm{R}\geq \mathrm{s},\mathrm{X}=\mathrm{x})$$

$$=\delta_{\mathrm{s}-1}\mathrm{f}(\mathrm{y_s}|\mathrm{y_1},\dots,\mathrm{y_{s-1}},\mathrm{R}=\mathrm{s}-1,\mathrm{X}=\mathrm{x})+(1-\delta_{\mathrm{s}-1})\,\mathrm{f}(\mathrm{y_s}|\mathrm{y_1},\dots,\mathrm{y_{s-1}},\mathrm{R}\geq \mathrm{s},\mathrm{X}=\mathrm{x})$$

(4.2.2.2-3)

where

$$\delta_{\mathrm{s}-1}=\frac{\mathrm{f}(\mathrm{y_1},\dots,\mathrm{y_{s-1}},\mathrm{R}=\mathrm{s}-1|\mathrm{X}=\mathrm{x})}{\mathrm{f}(\mathrm{y_1},\dots,\mathrm{y_{s-1}},\mathrm{R}\geq \mathrm{s}-1|\mathrm{X}=\mathrm{x})}$$

$$=1-\frac{\mathrm{f}(\mathrm{y_1},\dots,\mathrm{y_{s-1}}|\mathrm{R}\geq \mathrm{s},\mathrm{X}=\mathrm{x})\mathrm{P}(\mathrm{R}\geq \mathrm{s}|\mathrm{X}=\mathrm{x})}{\mathrm{f}(\mathrm{y_1},\dots,\mathrm{y_{s-1}}|\mathrm{R}\geq \mathrm{s}-1,\mathrm{X}=\mathrm{x})\mathrm{P}(\mathrm{R}\geq \mathrm{s}-1|\mathrm{X}=\mathrm{x})}$$

Note on the right side of (4.2.2.2-3), the conditional density for the first missing $\mathrm{f}(\mathrm{y_s}|\mathrm{y_1},\dots,\mathrm{y_{s-1}},\mathrm{R}=\mathrm{s}-1,\mathrm{X}=\mathrm{x})$ is a mixture distribution of non-ignorable missingness dropout and ignorable missingness dropout:

$$\mathrm{f}(\mathrm{y_s}|\mathrm{y_1},\dots,\mathrm{y_{s-1}},\mathrm{R}=\mathrm{s}-1,\mathrm{X}=\mathrm{x})$$

$$= \alpha_{s-1,x} f_{G=0}(y_s|y_1, \dots, y_{s-1}, R = s-1, X = x)$$

$$+ \left(1 - \alpha_{s-1,x}\right) f_{G=1}(y_s|y_1, \dots, y_{s-1}, R = s-1, X = x)$$

$$=$$

$$\alpha_{s-1,x} f(y_s|y_1, \dots, y_{s-1}, R \geq s, X = x) + \left(1 - \alpha_{s-1,x}\right) f(y_s + \Delta|y_1, \dots, y_{s-1}, R \geq s, X = x)$$

(4.2.2.2-4)

where $\alpha_{s-1,x}$ denotes the fraction of patients from ignorable missingness dropout group at pattern *s-1* in treatment *x*.

Then we can formulate the right side of (4.2.2.2-3) as the following:

$$f(y_s|y_1, \dots, y_{s-1}, R \geq s-1, X = x)$$

$$= \delta_{s-1}\left(1 - \alpha_{s-1,x}\right) f(y_s + \Delta|y_1, \dots, y_{s-1}, R \geq s, X = x)$$

$$+ \left(1 - \delta_{s-1} + \alpha_{s-1,x}\delta_{s-1}\right) f(y_s|y_1, \dots, y_{s-1}, R \geq s, X = x)$$

$$= \lambda_{s-1}\, f(y_s + \Delta|y_1, \dots, y_{s-1}, R \geq s, X = x) + (1 - \lambda_{s-1}) f(y_s|y_1, \dots, y_{s-1}, R \geq s, X = x)$$

(4.2.2.2-5)

Here,

$$\lambda_{s-1} = \left(1 - \frac{f(y_1,\dots,y_{s-1}|R \geq s, X=x) P(R \geq s|X=x)}{f(y_1,\dots,y_{s-1}|R \geq s-1, X=x) P(R \geq s-1|X=x)}\right)(1 - \alpha_{s-1,x})$$

(4.2.2.2-6)

Therefore, under non-future dependent missing value restriction, the unidentifiable conditional densities for visit *s* in pattern *t* ($s \geq t+2$) can be expressed as a mixture distribution of $f(y_s|y_1, \dots, y_{s-1}, R \geq s, X = x)$, the conditional distributions of y_s from pooled patterns with observed measurement at visit *s* and beyond in treatment *x*, and $f(y_s + \Delta|y_1, \dots, y_{s-1}, R \geq s, X = x)$, the previous conditional distributions of y_s adding a shift. The weight λ_{s-1} is based on (a) identifiable densities $f(y_1, \dots, y_{s-1}|R \geq$

$s, X = x)$ and $f(y_1, \dots, y_{s-1} | R \geq s - 1, X = x)$ - both pooled patterns including completers in treatment *x*, (b) $P(R \geq s | X = x)$ and $P(R \geq s - 1 | X = x)$, which can be estimated using the fractions of patients from the pooled patterns in treatment *x*, and (c) $\alpha_{s-1,x}$, the fraction of random dropout at Pattern *s-1* in treatment *x*.

If Δ equals to 0, the right side of (4.2.2.1-2) for the first missing value at pattern *t* in treatment *x* becomes

$$f(y_{t+1} + \Delta | y_1, \dots, y_t, R \geq t + 1, X = x) = f(y_{t+1} | y_1, \dots, y_t, R \geq t + 1, X = x)$$

For future missing measures at pattern *t*, based on equation (4.2.2.2-5), the right side of (4.2.2.2-2) can be expressed as

$$f(y_s | y_1, \dots, y_{s-1}, R \geq s - 1, X = x)$$

$$= \lambda_{s-1} f(y_s + \Delta | y_1, \dots, y_{s-1}, R \geq s, X = x) + (1 - \lambda_{s-1}) f(y_s | y_1, \dots, y_{s-1}, R \geq s, X = x)$$

$$= \lambda_{s-1} f(y_s | y_1, \dots, y_{s-1}, R \geq s, X = x) + (1 - \lambda_{s-1}) f(y_s | y_1, \dots, y_{s-1}, R \geq s, X = x)$$

$$= f(y_s | y_1, \dots, y_{s-1}, R \geq s, X = x)$$

for all $T \geq s \geq t + 2$, and all $t \geq 1$.

Hence, under 0 shift, NFD missing value restriction is equivalent to available-case missing value restriction, which leads to MAR.

4.2.2.3 Multiple Imputation (MI) steps

4.2.2.3.1 Markov Chain Monte Carlo (MCMC)

When the continuous outcome of interest can be treated as multivariate normal distribution, MCMC data augmentation is one available Bayesian MI method in SAS to impute the missing values.

MCMC is a class of algorithms that produce a chain of simulated draws from probability distribution via Markov Chains. A Markov Chain is a sequence of random variables in which the distribution of each element depends only on the previous set of imputed values. As a Bayesian method, MCMC iteratively generates imputed values for the missing elements using reasonable starting values. After a sufficient number of iterations (complete sets of updated imputations for all missing measures), the Markov Chain converges to a stable stationary distribution, which is the distribution of interest. A complete imputed sample is then drawn via the next iteration of the Markov Chain.

Consider, the repeated measures of interest $Y=(y_1,\ldots,y_T)$ are from a multivariate normal distribution, given by

$$Y \sim MVN(\mu, \Sigma).$$

$Y=(Y_{obs}, Y_{mis})$, where $Y_{obs}=(y_1,\ldots,y_t)$ are fully observed, and $Y_{mis}=(y_{t+1},\ldots,y_T)$ have missing values at visits from $t+1$ to T. The conditional distribution of Y_{mis} given Y_{obs} follows a multivariate normal distribution too, given by

$$Y_{mis|obs} \sim MVN\left(\mu_{mis|obs}, \Sigma_{mis|obs}\right)$$

The detail formulas of $\mu_{mis|obs}$ and $\Sigma_{mis|obs}$ are given in the following imputation I - step.

The MCMC data augmentation conducts the imputation I-step and the posterior P-step iteratively to achieve the reliable multiply imputed data set (Schafer 1997). At m^{th} iteration:

1. The imputation I-step:

Given an estimated mean vector ($\mu^{(m)}$) and covariance matrix($\Sigma^{(m)}$), the I-step draws the missing values ${Y_{mis}}^{(m)}$ from a conditional distribution for Y_{mis} given Y_{obs} for each subject independently.

Suppose $\mu^{(m)} = \left(\mu_{obs}^{(m)\prime}, \mu_{mis}^{(m)\prime}\right)'$ is the partitioned predicted mean vector for Y_{obs} and Y_{mis}, where $\mu_{obs}^{(m)}$is for Y_{obs} and $\mu_{mis}^{(m)}$is for Y_{mis}. And also define

$$\Sigma^{(m)} = \begin{bmatrix} \Sigma_{11}^{(m)} & \Sigma_{12}^{(m)} \\ \Sigma_{12}^{(m)\prime} & \Sigma_{22}^{(m)} \end{bmatrix}$$

a partitioned covariance matrix, where $\Sigma_{11}^{(m)}$ is the covariance matrix for observations Y_{obs}, $\Sigma_{22}^{(m)}$ is the covariance matrix for missing values Y_{mis}, and $\Sigma_{12}^{(m)}$ is the covariance matrix between Y_{obs} and Y_{mis}. The conditional distribution of Y_{mis} given Y_{obs} is a multivariate normal distribution with the mean vector

$$\mu_{mis|obs}^{(m)} = \mu_{mis}^{(m)} + \Sigma_{12}^{(m)\prime}{\Sigma_{11}^{(m)}}^{-1}(Y_{obs} - \mu_{obs}^{(m)}),$$

and the covariance matrix

$$\Sigma_{mis|obs}^{(m)} = \Sigma_{22}^{(m)} - \Sigma_{12}^{(m)\prime}{\Sigma_{11}^{(m)}}^{-1}\Sigma_{12}^{(m)}$$

2. The posterior P-step:

   Given a complete data $({Y_{mis}}^{(m)}, Y_{obs})$, the P-step draws the new posterior population mean vector ($\mu^{(m+1)}$) and covariance matrix($\Sigma^{(m+1)}$) from $f(\mu, \Sigma|Y_{obs}, {Y_{mis}}^{(m)})$.

   Jeffreys noninformative prior distribution is used in this step. The posterior distributions of μ and Σ are given by:

   - $\Sigma^{(m+1)}|Y^{(m)} \sim W^{-1}(N-1, (N-1)S)$ inverted Wishart distribution

- $\mu^{(m+1)}|\left(\Sigma^{(m+1)}, Y^{(m)}\right) \sim N(\overline{Y^{(m)}}, \frac{1}{N}\Sigma^{(m+1)})$

Where $Y^{(m)} = (Y_{obs}, {Y_{mis}}^{(m)})$, $S = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i^{(m)} - \overline{Y^{(m)}})(Y_i^{(m)} - \overline{Y^{(m)}})'$, and $i=1,\ldots,N$ denotes i^{th} subject.

These new estimates will be used in the next I-step.

A Markov chain $({Y_{mis}}^{(1)}, \mu^{(1)}, \Sigma^{(1)})$, $({Y_{mis}}^{(2)}, \mu^{(2)}, \Sigma^{(2)})$, … is created, which converges in distribution to $f(Y_{mis}, \mu, \Sigma|Y_{obs})$. When the iterates converge to a stationary distribution, an approximately independent draw of the missing values can be simulated from this distribution.

4.2.2.3.2 Detailed multiple imputation steps for PMMs with mixed missing value restrictions

m sets of plausible values are imputed for each missing measure to yield *m* imputed complete data sets. Then GEE procedure is applied for each imputed data set. The results from these *m* complete data sets are combined to obtain the overall statistical inferences.

The details of imputation steps at pattern *t* in treatment *x* are presented as the following:

Step 1. **Impute the first or current missing value y_{t+1} for each subject at pattern *R*=*t* (*t* = 1, ..., *T*-1).**

Using the MCMC method described in previous section, a single value of y_{t+1}^{*} can be imputed firstly based on the observed values $y_1, \ldots, y_t$ and the parameters of multivariate

normal distribution from pooled patterns *t+1* and above in treatment *x* (placebo or tested drug). Initial parameter estimates for the mean vector and covariance matrix to be used in the MCMC are derived using an EM algorithm. 200 burn-in iterations are conducted before a stationary distribution has been achieved and each of the next iteration are used for the imputed values.

Based on formulas 4.2.2.1-1 and 4.2.2.1-2, the first missing values can be imputed by the following three types:

- If the patient's dropout is from placebo group, the missing y_{t+1} is imputed as $\tilde{y}_{t+1} = y^*_{t+1}$.
- If the patient's dropout is from tested drug group and is ignorable (*G=0*), then the missing y_{t+1} in this pattern is also imputed as $\tilde{y}_{t+1} = y^*_{t+1}$.
- If the patient's dropout is from tested drug group and is non-ignorable (*G=1*), then the missing y_{t+1} in this data pattern is imputed as $\tilde{y}_{t+1} = y^*_{t+1} + \Delta$.

***Step 2*. Impute the future missing values of $y_{t+2}, y_{t+3}, ..., y_T$ for patients at pattern *R=t* (*t* = 1, ..., *T*-1).**

Starting with imputation for y_{t+2}, first, similar to Step 1, draw y^*_{t+2} through MCMC method. According to formulas (4.2.2.2-1) and (4.2.2.2-5), the future missing y_{t+2} is imputed as:

- $\tilde{y}_{t+2} = y^*_{t+2}$, if the dropout is from placebo group.
- $\tilde{y}_{t+2} = y^*_{t+2}$, if the dropout is ignorable missingness dropout from tested drug group.

- $\tilde{y}_{t+2} = y^{*}_{t+2} + \Delta$ with probability λ_{t+1}, if the dropout is non-ignorable missingness dropout from tested drug group where λ_{t+1} is calculated through formula 4.2.2.2-6.
- $\tilde{y}_{t+2} = y^{*}_{t+2}$ with probability $1 - \lambda_{t+1}$, if the dropout is non-ignorable missingness dropout from tested drug group.

Missing values of y_{t+3} through y_T can be imputed similar to the missing values for y_{t+2}.

The imputations of y_{t+1} through y_T are created recursively within each treatment group, pattern (for all $t = 1, \ldots, T-1$) and dropout category *G* (*G=0, 1*) to create a complete dataset. This is repeated *m* times to create *m* complete data sets. Small *m* can provides efficient point estimates through averaging over *m* data sets. Based on the relative efficiency of MI table in section 2.3.1.7, if there is 50% of missingness, 10 times imputation can reach imputation efficiency 0.95. Thus 10 imputation is enough in my method.

The shift adding for non-ignorable missingness makes the imputed complete data ruined the assumption of multivariate normal distribution. Since GEE procedure doesn't make any assumption about the distribution of repeated measures, it is performed to each imputed complete data set.

Finally, the results from these *m* complete data sets are combined as follows to obtain the overall statistical inferences. Suppose *Q* is a generic scalar quantity to be estimated (e.g. a mean, or regression coefficient). The overall estimate of Q is

$$\bar{Q} = m^{-1} \sum_{l=1}^{m} \hat{Q}^{(l)},$$

where $\hat{Q}^{(l)}$ is the estimate of Q based on the l^{th} imputed data set.

and estimated total variance is

$$T = (1 + m^{-1})B + \bar{U}$$

B is the between imputation variance

$$B = (m-1)^{-1} \sum_{l=1}^{m} (\hat{Q}^{(l)} - \bar{Q})^2$$

$\bar{U}$ is the within imputation variance

$$\bar{U} = m^{-1} \sum_{l=1}^{m} U^{(l)}$$

Here $\hat{Q}^{(l)} = \hat{Q}(Y_{obs}, Y_{mis}^{(l)})$, and $U^{(l)} = U(Y_{obs}, Y_{mis}^{(l)})$, the variance for $\hat{Q}^{(l)}$.

Tests are based on Student's t-approximation

$$(\bar{Q} - Q)/\sqrt{T} \sim t_v$$

4.3 Worked Example: Simulated Chronic Pain Study

A chronic pain study is mimicked here to demonstrate the efficacy of tested drug in the treatment of chronic pain. In detail, the simulated study mimics a multicenter, randomized, double-blind, placebo-controlled, monotherapy, and parallel-group phase III study for treating chronic pain. It has two treatment groups, one baseline visit and four post-baseline-visits (weeks 4, 8, 12, 16). One thousand patients have been randomized in a 1:1 ratio to one of the two treatment groups - placebo and tested drug. The population for the primary analysis consists of all patients with baseline value and at least one post baseline measurement of the primary efficacy variable. The range of

pain scores is from 0 (no pain) to 10 (the highest level pain). The primary outcome measure is the change from baseline in pain score. The treatment effect is expressed through the difference of primary outcome between two treatment groups (placebo vs. tested drug) at the end of study.

Dropout numbers and cumulative information at each visit are presented at Table 4.1 and Figure 4.1. There are over 30% dropouts in this study, and slightly more dropouts occur in drug group (33.4%) than in placebo group (30.6%).

Table 4.1. Dropout number by visit and treatment group

anvis	Pl_count	Pl_cum	Pl_Percent(%)	Dr_count	Dr_cum	Dr_Percent(%)
1	50	50	10	50	50	10
2	39	89	17.8	47	97	19.4
3	29	118	23.6	31	128	25.6
4	35	153	30.6	39	167	33.4

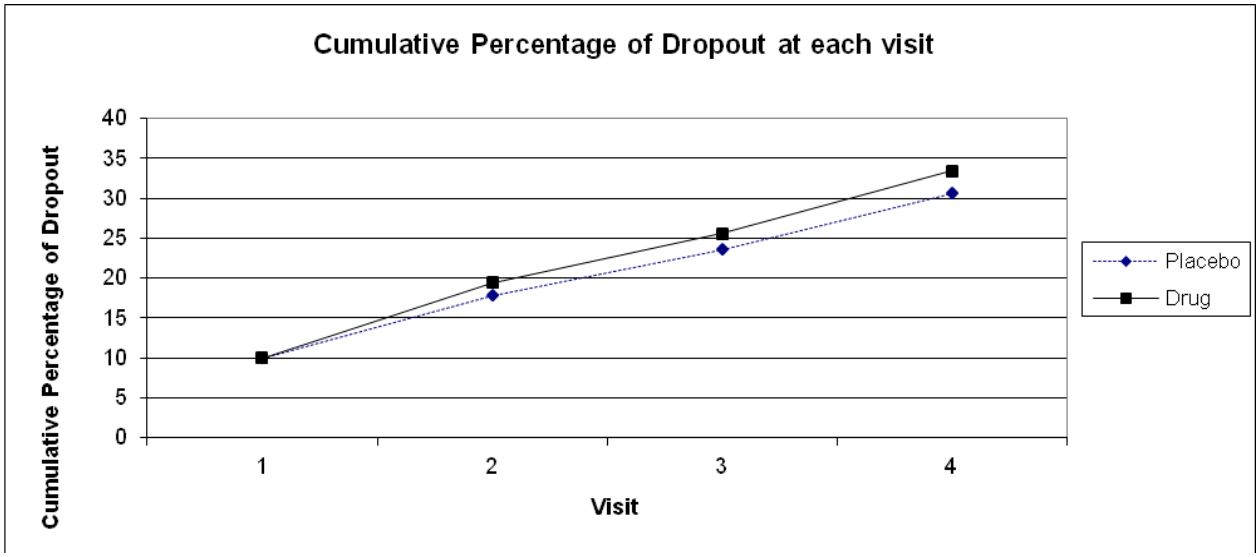

Figure 4.1. Cumulative Percentage of Dropout at each visit

The means of change from baseline in pain score by treatment group and dropout pattern are shown in Table 4.2 and Figure 4.2.

Table 4.2. Mean of change from baseline at each post-baseline visit.

visit	Pl_I	Pl_II	Pl_III	Pl_IV	Dr_I	Dr_II	Dr_III	Dr_IV
1	-0.91	-1.32	-1.07	-1.22	-0.91	-1.20	-2.15	-1.73
2		-0.87	-0.92	-1.55		-1.53	-2.17	-2.15
3			-1.01	-1.46			-2.00	-2.27
4				-1.45				-2.14

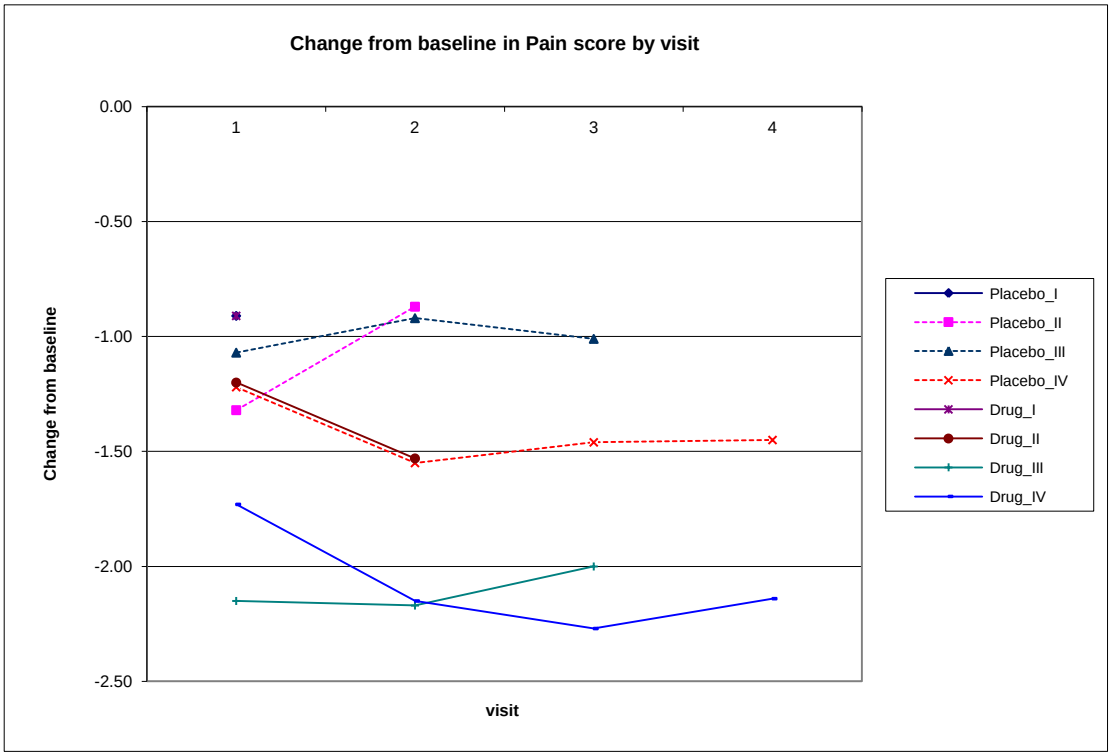

Figure 4.2. Change from baseline in Pain score by visit.

Figure 4.2 shows that in both treatment groups, the completers changed more than dropouts. However, there is no obvious relationship between change and visit. So the linear function of visit for the changes from baseline can't be assumed.

Detailed dropout reasons for each patient are also simulated from the real study. Reasons includes: adverse events, lack of efficacy, lost contact, protocol violation, non-compliance with respect to protocol requirements, and others. Dropout due to an Adverse Event (AE) in tested treatment group is of most concern to the FDA. AE may lead to non-ignorable missingness dropout since an AE is likely related to the patient's missing chronic pain score. Comparable to AE, "Lack of Efficacy" (LOE) is another major dropout reason. However, it is debatable whether a dropout caused by LOE is MAR or MNAR. Except AE and LOE, all the other dropout reasons may be assumed as ignorable missingness dropout. Thus two scenarios are considered, as laid out into Table 4.3 and Table 4.4 respectively. According to the first dropout classification system, "Lack of Efficacy" is assigned as non-ignorable, but in the second dropout classification, it is assumed as ignorable.

Table 4.3. Dropout Category I

Dropout Reasons	**Classification of Dropout Category**
1. Adverse Event	Non-ignorable missingness
2. Lack of Efficacy	Non-ignorable missingness
3. lost contact	Ignorable missingness
4. protocol violation	Ignorable missingness
5. became pregnant	Ignorable missingness
6. non-compliant with protocol requirements	Ignorable missingness
7. Other	Ignorable missingness

Table 4.4. Dropout Category II

Dropout Reasons	**Classification of Dropout Category**
1. Adverse Event	Non-ignorable missingness
2. Lack of Efficacy	Ignorable missingness
3. lost contact	Ignorable missingness
4. protocol violation	Ignorable missingness
5. became pregnant	Ignorable missingness

6. non-compliant with protocol requirements	Ignorable missingness
7. Other	Ignorable missingness

The dropout information is presented in Table 4.5 and Figure 4.3. Generally speaking, more dropouts in drug group are caused by Adverse Events, while in placebo group more dropouts are caused by Lack of Efficacy.

Table 4.5. Dropout number by visit, treatment group and dropout reasons.

visit	Treatment	AE[a]	LOE[b]	Others[c]
1	Placebo	25	10	15
	Drug	34	5	11
2	Placebo	19	9	11
	Drug	33	8	6
3	Placebo	14	9	6
	Drug	13	12	6
4	Placebo	6	2	27
	Drug	10	2	27

a: Dropout caused by adverse events
b: Dropout caused by lack of efficacy
c: Dropout caused by other reasons except AE and LOE

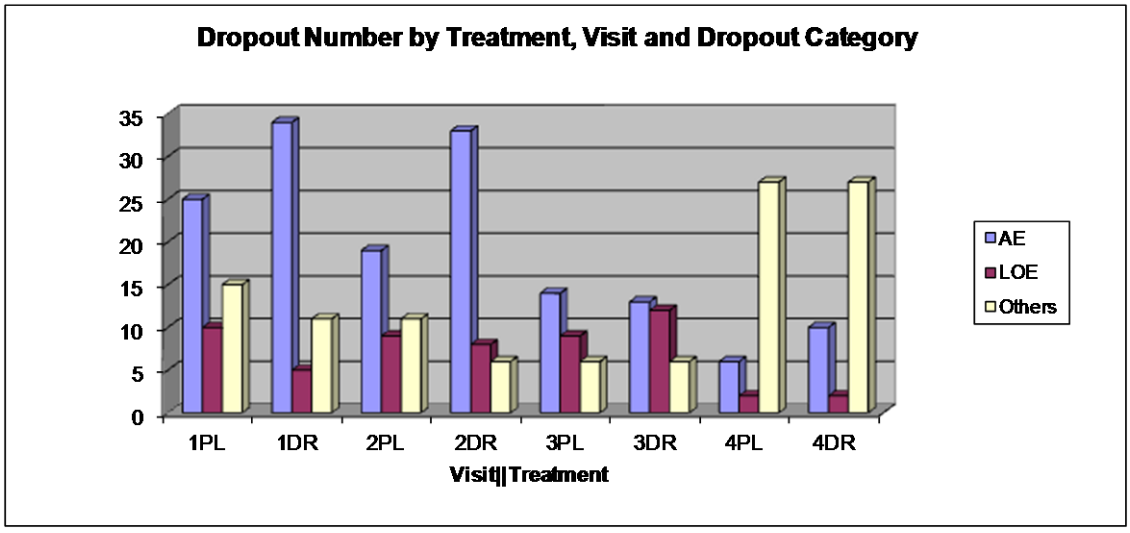

Figure 4.3. Dropout Number by Treatment, Visit and Dropout Category

In statistical analyses, the changes from baseline in pain score are treated as continuous variables and assumed to follow a multivariate normal distribution. The treatment effect is defined as the difference of changes from baseline between placebo and drug groups at the end of study.

PMM-based MI with mixed missing value restrictions (available-case and NFD) is applied to this simulated study in order to handle mixed dropout mechanisms based on the dropout categories expressed in Table 4.6 and Table 4.7 respectively. According to the observed treatment efficacy, shift series are selected as 0, 0.2, 0.4, 0.6, 0.8 and 1.0 for drug group and 0 for placebo group.

These results may be compared to results from analyses that do not rely on MI, including BOCF, LOCF, Modified BOCF (impute missing values caused by Adverse Events or Lack of Efficacy by BOCF, but impute other missing values by LOCF), and MMRM (Mixed Model for Repeated Measures), four main approaches used in clinical trials of chronic pain, either as primary analysis or sensitivity analysis. The results from these four methods are shown in Table 4.8 with the results from GEE, WGEE and ANCOVA model. Except for the ANCOVA model, all the other models analyze the change from baseline in pain score from all post visits with treatment, visit (as a factor), treatment by visit interaction, baseline pain score, and baseline pain score by visit interaction as explanatory variables. In ANCOVA model, change from baseline in pain score at the end of study (visit=4) is analyzed, and only treatment is set as a factor and baseline pain score as a covariate.

MMRM assumes the outcome of interest is multivariate normal distribution, so if the imputed data can't be assumed as multivariate normal distribution, MMRM is not suitable anymore. GEE doesn't make any assumption about the distribution of outcome of interest, and also it can provide unbiased estimators regardless of whether the correlation structure is correct or not. Here BOCF, LOCF, mBOCF, and the proposed method produce imputed values that do not follow a multivariate normal distribution. Hence GEE is a better choice to apply to these imputed complete longitudinal data sets. GEE procedure with an autoregressive correlation structure (expressed as AR(1) in SAS) is fit to analyze each imputed complete data set, with treatment group, visit (as a factor), treatment by visit interaction, baseline pain score, and baseline pain score by visit interaction as explanatory variables. Based on this estimation procedure, least squares means with standard errors for each treatment group and the difference of least squares means between two groups with 95% confidence interval are obtained. Results are shown in Table 4.6 , Table 4.7 , and Table 4.8.

Table 4.6. Change from baseline at the end of study (using proposed PMMs based MI with mixed missing value restrictions method, based on dropout category I, treat dropout caused by AE and LOE as non-ignorable), GEE model with autoregressive correlation structure for the imputed complete datasets.

Change from baseline at end of study					
shift		Placebo (N=450)	Drug (N=450)	LSMD* (95%CI)	P-value
Placebo	Drug	Mean (SEM)	Mean (SEM)		
0	0	-1.330 (0.104)	-2.030 (0.109)	-0.701 (-0.993, -0.409)	<0.0001
0	0.2	-1.332 (0.101)	-1.986 (0.114)	-0.655 (-0.955, -0.354)	<0.0001
0	0.4	-1.331 (0.104)	-1.970 (0.108)	-0.639 (-0.936, -0.342)	<0.0001
0	0.6	-1.323 (0.102)	-1.929 (0.106)	-0.606 (-0.894, -0.317)	<0.0001

0	0.8	-1.329 (0.102)	-1.918 (0.115)	-0.589 (-0.891, -0.287)	0.0001
0	1.0	-1.323 (0.104)	-1.873 (0.114)	-0.550 (-0.855, -0.245)	0.0004

*: The difference between the least square means for placebo and drug groups.

Table 4.7. Change from baseline at the end of study (using proposed PMMs based MI with mixed missing value restrictions method, based on dropout category II, only treat dropout caused by AE as non-ignorable). GEE model with autoregressive correlation structure for the imputed complete datasets.

Change from baseline at end of study					
shift		Placebo (N=450)	Drug (N=450)	LSMD* (95%CI)	P-value
Placebo	Drug	Mean (SEM)	Mean (SEM)		
0	0	-1.330 (0.104)	-2.030 (0.109)	-0.701 (-0.993, -0.409)	<0.0001
0	0.2	-1.332 (0.101)	-1.995 (0.114)	-0.663 (-0.964, -0.363)	<0.0001
0	0.4	-1.332 (0.104)	-1.988 (0.108)	-0.656 (-0.952, -0.360)	<0.0001
0	0.6	-1.324 (0.102)	-1.956 (0.108)	-0.632 (-0.919, -0.344)	<0.0001
0	0.8	-1.330 (0.102)	-1.953 (0.114)	-0.623 (-0.923, -0.323)	<0.0001
0	1.0	-1.325 (0.104)	-1.917 (0.112)	-0.593 (-0.894, -0.291)	0.0001

*: The difference between the least square means for placebo and drug group.

Table 4.8. Change from baseline at the end of study (using different methods).

Change from baseline at end of study				
Pain Score	Placebo (N=450)	Drug (N=450)	LSMD* (95%CI)	P-value
	Mean (SEM)	Mean (SEM)		
LOCF[a]	-1.30 (0.10)	-1.92 (0.10)	-0.62 (-0.90, -0.34)	<0.0001
BOCF[b]	-1.10 (0.10)	-1.60 (0.10)	-0.50 (-0.76, -0.23)	0.0003
MBOCF[c]	-1.23 (0.10)	-1.77 (0.10)	-0.54 (-0.82, -0.26)	0.0001
MMRM[d]	-1.33 (0.11)	-2.04 (0.11)	-0.70 (-1.00, -0.40)	<0.0001
GEE[e]	-1.33 (0.10)	-2.03 (0.11)	-0.70	<0.0001

			(-1.00, -0.40)	
WGEE[f]	-1.34 (0.10)	-2.05 (0.11)	0.71 (-1.01, -0.41)	<0.0001
ANCOVA[g]	-1.42 (0.12)	-2.17 (0.12)	-0.75 (-1.07, -0.42)	<0.0001

a: Using Last Observation Carried Forward method to impute missing values, then use GEE (autoregressive correlation structure) to analyze this imputed full data set.
b: Using Baseline Observation Carried Forward method to impute missing values, then use GEE (autoregressive correlation structure) to analyze this imputed full data set.
c: Using modified BOCF method (impute missing values caused by Adverse Events or Lack of Efficacy by BOCF, while impute other missing values by LOCF), then use GEE (autoregressive correlation structure) to analyze this imputed full data set.
d: Using Mixed Model for repeated measurements with unstructured correlation to analyze observed data. Model with fixed terms of treatment group, visit, treatment by visit interaction, baseline pain score, and baseline pain score by visit interaction as explanatory variables
e: Using GEE (autoregressive correlation structure) model to analyze observed data.
f: Using logistic model with explanatory variables treatment group and last observed change to get the weights, then use GEE (autoregressive correlation structure) model with weight to analyze observed data.
g: Using ANCOVA to analyze observed case at the end of study.
*: The difference between the least square means for placebo and drug group.

Three tables show that all the approaches obtain a significant difference in change from baseline between the placebo group and drug group, in favor of the drug. When there is no shift ($\Delta=0$), the Least Square Means and Least Square Means Difference (LSDM) from the proposed PMM-based MI with mixed missing value restriction approach are the same as the results from original MMRM analysis for observed data which assumes the missingness is MAR. Comparing Table 4.6 and Table 4.7, whether dropouts caused by "Lack of Efficacy" are set as ignorable or non-ignorable, the LSDM decreases as the shift value increases. Also, when LOE dropouts are treated as non-ignorable, the difference reduces more quickly than when LOE dropouts are treated as ignorable. These results can be easily explained by the value of shift and the percentage of measurements adding shift.

If the shift and/or the percentage of non-random dropouts in drug group become larger, then the analysis becomes more conservative, in favor of placebo treatment.

4.4 Simulation for sample size and power calculation

Many software programs have been developed to calculate samples size or powers (e.g. PS, nQuery, etc.). However, most of them are based on formulas of sample size (e.g. two samples t-test, one sample t-test, etc.), which are only suitable for simple studies. If the studies are complicated, for example, longitudinal studies with many dropouts, then these software routines may not produce the correct estimates. Hence, in this section, simulation is conducted to produce the power based on the pre-specified sample size according to the proposed approach PMM-based MI with mixed missing value restrictions and other main methods under differential dropout assumptions.

It is well known that the dropout problem is unavoidable to many clinical trials and may interfere with the ability of the study to detect an effect, so it is necessary to pay attention to the influence of dropouts when designing a study. To account for potential dropouts, researchers have traditionally increased the sample size through simple mathematical calculations. For example, if the estimated sample size is n and the dropout rate is d, then the final sample size would be $n/(1-d)$. However, this estimate is actually based on the assumption that observed and missing values will follow the same distribution as would be the case if dropout is MCAR. If the dropout is not MCAR, this estimate may not accurately reflect the sample sizes necessary to achieve the nominal power.

In recent years for longitudinal studies with many dropouts, the Food and Drug Administration (FDA) and European Medicines Agency (EMA) have asked pharmaceutical companies to provide sensitivity analyses under the MNAR assumption. They want consistent results to support the effect of drug as obtained from the primary analysis. If the influence of the dropout mechanisms is ignored, contradictory results may be obtained due to unsuitable sample size and power. Then the effect of the studied drug may be questioned by agency, and a new trial may also be requested. To ensure an efficient study, it may be better to calculate the sample sizes or powers for the study up front considering different dropout mechanisms. Then, armed with more information about the sample size and power in the presence of various types of dropout, researchers have a better chance of determining a suitable size for the study, resulting in consistent positive results across primary and sensitivity analyses.

There are no straightforward formulations that can be used to calculate the sample size or power for longitudinal study with dropouts under various dropout mechanisms. Instead, simulation procedures are needed to provide the power based on pre-specified sample size. A SAS macro (SAS 9.2) is developed here to do this simulation. This simulation procedure includes the power calculation for methods MMRM, ANCOVA (ANCOVA model for the observed data at the end of study), BOCF – GEE (use BOCF to impute the missing values, then use GEE to analysis imputed complete data), LOCF - GEE, mBOCF – GEE, MI – GEE, and my proposed approach PMM-based MI with mixed missing value restrictions – GEE.

I give an example to show the different powers obtained from various dropout assumptions, based on the same pre-specified sample size. The following simulation conditions are mimicked from one existed real clinical trial. Because the relationship between the response and time cannot be specified (e.g. linear), a model with covariate time is not convenient to be used for simulating the longitudinal data. Thus, multivariate normal distributions are used to generate the longitudinal continuous variable for two treatment groups (placebo vs. drug), with 200 subjects per group, and 1 baseline (visit=0) and 4 post visits (visit=1, 2, 3, 4) per subject. The overall dropout rate is assumed 30% for placebo group and 40% for drug group. For both groups, there are 40% of the dropouts are assumed to occur between visits 1 and 2, 30% between visits 2 and 3, and 30% after visit 3. The mean vectors are (0, -13, -16, -18, -20) for placebo group and (0, -17.55, -21.6, -24.3, -27) for drug group. The standard deviation vectors are (12, 18, 20, 20, 20) for placebo group and (12, 19, 21, 21, 21) for drug group. The correlation matrix is

$$\begin{bmatrix} 1 & -0.2 & -0.2 & -0.2 & -0.2 \\ -0.2 & 1 & 0.9 & 0.8 & 0.65 \\ -0.2 & 0.9 & 1 & 0.9 & 0.8 \\ -0.2 & 0.8 & 0.9 & 1 & 0.9 \\ -0.2 & 0.65 & 0.8 & 0.9 & 1 \end{bmatrix}$$

The correlation structure for the changes from baseline at four post visits can be treated as autoregressive. The primary efficacy parameter is the change from baseline in the outcome of interest. The treatment effect is also measured as the difference of the changes from baseline between placebo and drug groups at the end of study (visit 4).

As mentioned before, adverse event (AE) is one of most popular and important dropout reasons, and the dropout caused by AE in tested treatment group is of the most concern

by regulation agency, so it's better to provide sensitive analysis that treats these dropouts under MNAR assumption. Since usually more AE occur within drug group, I set 60% of dropouts in drug group are caused by AE, while only 30% of dropouts in placebo group are assumed due to AE in our example.

MMRM model assumes that all the missingness caused by dropout is MAR. It uses fixed terms of treatment group (as a factor), visit (as a factor), treatment by visit interaction, baseline score, and baseline score by visit interaction as explanatory variables to analyze all the observed change from baseline. ANCOVA model (under MCAR assumption) is fitted to analyze the observed change from baseline at the end of study (visit 4), with treatment as a factor and baseline as a covariate. The imputed methods BOCF, LOCF, mBOCF, and MI have been introduced in previous sections already. GEE models with autoregressive correlation structure are used to analyze these imputed complete data sets. Explanatory variables include treatment (as a factor), visit (as a factor), treatment by visit interaction, baseline score, and baseline score by visit interaction.

In the PMM-based MI with mixed missing value restrictions approach, the dropouts caused by AE in drug group are set as non-ignorable, and the remaining dropouts are assumed to be ignorable. Then the imputation follows the steps introduced in 4.3 section, and the shift parameter is varied from 0 to 10 by 2 for the non-ignorable dropouts in drug group, and 0 for all the other dropouts. 1000 data sets are simulated and analyzed. The mean of simulation results and power are shown in Table 4.9 and Table 4.10.

Table 4.9. Power calculation (1000 simulations), using main methods.

Change from baseline at visit 4				
Pain Score	Placebo (N=200)	Drug (N=200)	LSMD (SEM) (95%CI)	Power (%)
	Mean (SEM)	Mean (SEM)		
MMRM	-20.08 (1.53)	-27.03 (1.58)	-6.95 (2.20) (-11.26, -2.63)	88.30
ANCOVA	-20.03 (1.70)	-27.02 (1.83)	-6.99 (2.50) (-11.91, -2.07)	78.40
LOCF	-18.69 (1.37)	-24.53 (1.45)	-5.84 (2.00) (-9.76,-1.92)	83.50
BOCF	-14.02 (1.42)	-16.22 (1.56)	-2.20 (2.12) (-6.34, 1.95)	17.00
mBOCF	-17.31 (1.41)	-19.55 (1.56)	-2.24 (2.11) (-6.37, 1.89)	18.40
MI	-20.08 (1.55)	-27.02 (1.70)	-6.94 (2.31) (-11.46, -2.42)	86.00

Table 4.10. Power calculation (1000 simulations), using proposed PMMs based MI with mixed missing value restrictions method, GEE used for the imputed complete datasets.

Change from baseline at end of study					
shift		Placebo (N=200)	Drug (N=200)	LSMD (SEM) (95%CI)	Power
Placebo	Drug	Mean (SEM)	Mean (SEM)		
0	0	-20.07 (1.55)	-27.04 (1.70)	-6.97 (2.30) (-11.54, -2.40)	86.20
0	2	-20.07 (1.55)	-26.53 (1.70)	-6.46 (2.30) (-11.03, -1.89)	82.00
0	4	-20.07 (1.55)	-26.02 (1.71)	-5.95 (2.30) (-10.53, -1.38)	73.30
0	6	-20.07 (1.55)	-25.52 (1.71)	-5.45 (2.31) (-10.03, -0.86)	65.30
0	8	-20.07 (1.55)	-25.01 (1.72)	-4.94 (2.32) (-9.54, -0.34)	56.40
0	10	-20.07 (1.55)	-24.50 (1.74)	-4.43 (2.33) (-9.05, 0.19)	46.30

The results in above two tables show that under ignorable missingness dropout assumption (the distribution of the missing values is the same as that of observations), MMRM, MI and proposed method with shift=0 produce unbiased point estimates for the

least square means (-20 for placebo & -27 for tested treatment) and the least square means difference (-7) at the end of study. The powers obtained from these three methods are also similar (88.3%, 86% & 86.2% respectively). ANCOVA model ignores the correlation among repeated measures, and only use the observed changes at the end of study. Although it also gives unbiased point estimators here, it does produce larger standard errors and make the results less efficient (power is 78.4%).

This simulated study assumes that the smaller measure is better, so the evolution of disease is improved in both placebo and tested treatment group. Moreover, the dropout rate in tested treatment group is larger than the ones in placebo group. Therefore, LOCF, BOCF, and mBOCF all give smaller point estimators (LSMD -5.84, -2.20 & -2.24 respectively) than MMRM, MI and proposed method with shift=0. However, the standard error from LOCF is obviously less than other methods, so we still get 83.5% power in this simulation. In some circumstances, although the point estimators of LOCF are smaller, LOCF still may produce higher power than the ones from MMRM or other methods under MAR assumption. Therefore, treating LOCF as a conservative method, prefer to placebo or control treatment is a mistake to many studies. BOCF and mBOCF provide much smaller estimators and powers (17.0% & 18.4% respectively) in this simulated scenario. So if the process of disease is improved over time, and the dropout rate in tested treatment group is obviously higher than the ones in placebo group, then BOCF or mBOCF may produce too conservative results.

Thus, under this simulation scenario, if the investigator wants to use proposed method with some shifts, BOCF, mBOCF, or ANCOVA as the sensitivity analysis for supporting the results from primary analysis under MAR assumption (e.g. MMAR, MI), larger sample size may be needed to achieve the higher power.

4.5 Summary

This chapter proposed a multiple imputation strategy for handling mixed dropout mechanisms for clinical trials with continuous outcomes measured longitudinally. Within PMMs, NFD missing value restriction is used for imputing missing values caused by non-ignorable missingness dropout, and available-case missing value restriction is applied to ignorable missingness dropout. In addition, in contrast to single imputation, MIs overcome the problem of underestimating the variance to researchers. Through changing the value of shift, the level of conservativeness can be easily controlled by investigators. Further, the sample size and power calculation based on this proposed method may help investigators finding a suitable sample size for clinical trial, and make the trial achieves the enough power not only for primary analysis, but also for sensitivity analyses.

5 Discussion and Future Works

5.1 Discussion

With longitudinal data, dropout can impact both the bias of treatment effects as well as decreased power and narrowed confidence interval for the treatment effects. Although there are extensive literatures on methods for handling dropout, no single method is flexible enough to handle all types of dropout. Rather, for different dropout mechanism assumptions (under MCAR, MAR and MNAR assumptions), different methods are recommended as presented in Chapter Two. Therefore, my thesis explores two regression-based approaches for the continuous outcomes measured longitudinally with mixed dropout mechanisms. These approaches, presented in Chapters 3 and 4, differ according to whether the reasons for dropout are unknown or known, respectively.

The first proposed method, the pattern mixture-within-mixture model, is developed for use when the reasons for dropout are unknown. It assumes linear trends in the outcome over time within each treatment group. It also assumed that within each dropout pattern, a mixture distribution takes into account the possibility of different missingness mechanisms. Overall, in simulation studies, this approach provided results with estimates close to the true values and standard errors. In comparison to this approach, PMMs and GEE also provides very close estimates to the true values but relatively bigger standard errors under our simulation conditions. BOCF and LOCF underestimate the true effect for the simulated data sets with mixed dropouts. When applied to the Schizophrenia

study, however, the proposed approach produced somewhat different results than those from alternative methods, may caused by the unbalanced dropout rates.

Potential limitations include the computational effort for EM convergence and the added assumptions for the pre-specification of linear trends of time. For many studies, the assumption of a linear trend in the outcome over time may be reasonable. For example, some HIV studies during short time period (e.g. 3 months). However, it may be not true for other longitudinal studies. For example, the chronic pain studies. In addition, estimates of interest cannot be obtained through this method under MCAR/MAR assumption, since there are no mixed distributions can be separated by EM algorithm.

The second approach supposes that the reasons for dropouts are well known and relaxes the assumption that the outcome of interest is linearly associated with time. In fact, no particular structure is assumed for the trend in outcome over time. This approach multiply imputes missing values assuming a PMM with multiple missing value restrictions dependent on (1) whether dropout is ignorable or non-ignorable and (2) whether the missing value is the first missing value or a future missing value. Standard methods are used to generate and combine the estimates from the multiply imputed data. The imputation procedure makes this method flexible for handling different dropout types ignorable or non-ignorable. Thus this approach is more practical to contemporary longitudinal clinical trials with many dropouts.

This proposed method can be implemented under SAS macro, no heavy or slow computation problem. Compared to the MMRM (under MAR assumption), this proposed method provide the same result in simulation studies when shift is set as zero, and conservative results when shift is larger than 0. The conservative level of analysis can readily be controlled by modifying the values of shift in this method. The potential problem of this method may be the explanation of the value of shift and the non-future dependent missing value restriction to clinician.

My proposed methods provide reasonable sensitivity analyses under mixed dropout mechanisms (MCAR/MAR and MNAR) for supporting the primary analysis. The second proposed method is also response to the suggestion by European Medicines Agency at Guideline on Missing Data in Confirmatory Clinical Trials (2011), using different techniques for different dropout reasons. It flexibly allows users to impute the missing values under MAR assumption (set shift=0) or same MNAR assumption (use same shift for all dropouts) or mixed dropout assumption. Future works are included in the following section.

5.2 Future Works

A specific imputation model may be reasonable for some studies, but not suitable to others, depending on the nature of the disease and/or drug. Therefore, additional models may need to be crafted to suit each future study of interest. For example, in some studies, it might be reasonable to assume that dropped subjects from the treatment group will

exhibit the same future evolution of the disease as the subjects from the placebo group. In such a case, missing values for both the placebo and treatment groups should be imputed based only on the parameters from the placebo group. This may be just one sensitivity analysis under MNAR assumption to support the primary analysis. It can also be combined with my proposed method. In detail, for the MNAR dropout, the first missing value will be imputed by the placebo based imputation approach, and then followed by non-future dependent missing value restriction to impute the future missing values.

A more complex problem arises for longitudinal clinical trials that have high rates of dropouts, e.g. over 50%. Clinicians may feel uncomfortable about the analysis results based on the imputed complete data. My proposed method also can be extended to treat this situation. The first missing value under MNAR assumption will be imputed based on my proposed imputation method of adding shift. Through non-future dependent missing value restriction, the future missing values will be imputed or un-imputed due to a calculated probability. The entire MCAR/MAR dropouts will be left un-imputed. The imputed incomplete data cannot be assumed as multivariate normal distribution, so MMRM is not suitable. Moreover, the left dropouts may be caused by reasons for MAR dropout, and thus GEE is questionable too. WGEE may be a good choice to this imputed incomplete data, but this needs to be confirmed by future research.

Additionally, in my proposed method, the dropout caused by death is treated as MNAR, assuming that the missing values after death can be imputed. This assumption may be

implausible. How to handle death as a reason of dropout is a perplexing problem that needs to be solved in the future.

Appendix

Details of the EM and Standard Error computation for PMMs

To simplify computation, we suppose there is only one random effect – intercept, which is from normal distribution with mean 0 and variance δ^2. Random covariates are $T_i \times 1$ matrix $b_i = (1, \dots, 1)^{\mathrm{T}}$.

The complete data likelihood function for PMM can be written as

$$L(\Psi; Y, X, U, R) = f(Y, X, U, R; \Psi) = \prod_{i=1}^{N}\left[\prod_{t=1}^{T_i}[f(Y_{it}|X_{it}, U_i, R_i; \beta)]\, f\left(U_i; \delta^2\right) f(R_i)\right]$$

$$= \prod_{r=1}^{M} \prod_{i \in pattern\ r} \left\{\pi^{(r)}[\prod_{t=1}^{T_i} f\left(Y_{it}\middle|X_{it}, U_i, R_i = r; \beta^{(r)}\right)] f\left(U_i; \delta^2\right)\right\}$$

EM algorithm

1) In E-step, at the $(t+1)^{th}$ iteration, the conditional expectation of the log-likelihood of the complete data given the observed data is

$$Q\left(\psi|\psi^{(t)}\right) = \sum_{r=1}^{M} \sum_{i \in pattern\ r} \left\{\log \pi^{(r)} + E\left(\sum_{t=1}^{T_i}\left[\log f\left(Y_{it}\middle|X_{it}, U_i, R_i = r; \beta^{(r)}\right)\right] + E(\log f(U_i; \delta^2))\right)\right\}$$

Where

$$f(U_i; \delta^2) = \frac{1}{\sqrt{2\pi}\delta} exp\left(-\frac{U_i^2}{2\delta^2}\right)$$

$$f\left(Y_{it}|X_{ij},U_i,R_i=r;\beta^{(r)}\right)=\frac{1}{\sqrt{2\pi}\delta}exp\left(-\frac{(Y_{it}-X_{it}\beta^{(r)}-U_i)^2}{2\delta^2}\right)$$

Then we can get

$$Q\left(\psi|\psi^{(t)}\right)=$$

$$\sum_{r=1}^{M}\sum_{i\in pattern\ r}\left\{\log\pi^{(r)}-\sum_{t=1}^{T_i}\frac{(Y_{it}-X_{it}\beta^{(r)})^2}{2\sigma^2}+\sum_{t=1}^{T_i}\frac{(Y_{it}-X_{it}\beta^{(r)})E(U_i|R_i=r)}{2\sigma^2}-\right.$$

$$t=1Ti12\sigma2EUi2Ri=r-EUi2Ri=r2\delta2-t=1Ti\log2\pi+\log\sigma-\log2\pi+\log\delta$$

The following conditional expectations are needed:

$$E(U_i|Y_i,X_i,R_i=r)=\frac{\sum_{t=1}^{T_i}\left(Y_{it}-X_{it}\beta^{(r)}\right)\delta^2}{\sum_{t=1}^{T_i}\delta^2+\sigma^2}$$

$$E(U_i^2|Y_i,X_i,R_i=r)=\frac{\sigma^2\delta^2}{\sum_{t=1}^{T_i}\delta^2+\sigma^2}+[E(U_i|Y_i,X_i,R_i=r)]^2$$

2) In M-step, the updated parameters are

$$\hat{\beta}_q^{(r)(t+1)}$$

$$=\frac{\sum_{i\in pattern\ r}\left\{\sum_{t=1}^{T_i}\left(Y_{it}-X_{it}\beta^{(r)(t)}+X_{itq}\beta_q^{(r)(t)}-E(U_i|R_i=r)^{(t)}\right)\cdot X_{itq}\right\}}{\sum_{i\in pattern\ r}\left\{\sum_{t=1}^{T_i}X_{itq}^2\right\}}$$

where r=1,…,M; β_q denotes the coefficient for q^{th} covariate X_{itq}, $q=1,\dots,Q$.

$$\hat{\sigma}_q^{(r)(t+1)}$$

$$=\frac{\sum_{r=1}^{m}\sum_{i\in pattern\ r}\left\{\sum_{t=1}^{T_i}\left(Y_{it}-X_{it}\beta^{(r)}\right)^t-2\sum_{t=1}^{T_i}\left(Y_{it}-X_{it}\beta^{(r)}\right)^t E(U_i|R_i=r)^{(t)}+E(U_i^2|R_i=r)^t\right\}}{\sum_{i=1}^{N}\sum_{t=1}^{T_i}1}$$

$$\hat{\delta}^{2(t+1)} = \frac{\sum_{r=1}^{M}\sum_{i\in pattern\ r}\left\{E(U_i^2|R_i=r)^t\right\}}{\sum_{i=1}^{N}1}$$

Standard Errors

To $i \in Pattern\ r$, r=1,…,M, the elements in $s(Y_i;\widehat{\Psi})$ are given by:

$$\frac{\partial Q_i}{\partial \beta_q^{(r)}} = \frac{\sum_{t=1}^{T_i}\left(Y_{it} - X_{it}\beta^{(r)} - E(U_i|R_i=r)\right)X_{itq}}{\sigma^2}$$

$$\frac{\partial Q_i}{\partial \sigma} = -\sum_{t=1}^{T_i}\frac{1}{\sigma} + \frac{\left[\sum_{t=1}^{T_i}\left(Y_{it} - X_{it}\beta^{(r)} - E(U_i|R_i=1)\right)^2 + \sum_{t=1}^{T_i}Var(U_i|R_i=1)\right]}{\sigma^3}$$

$$\frac{\partial Q_i}{\partial \delta} = -\frac{1}{\delta} + \frac{E(U_i^2|R_i=r)}{\delta^3}$$

From these expressions for updating fixed effects coefficients, we can find that the formulas for calculating pattern and dropout specific effects derived from PMM are different with ones derived from my proposed method PMWMM for mixed dropout mechanisms, but the exactly quantity of difference between two methods are intractable.

Reference

Albert PS, Follmann DA. Modeling repeated count data subject to informative dropout. *Biometrics* 2000; 56: 667-677

Committee for Medicinal Products for Human Use (CHMP). Guideline on missing data in confirmatory clinical trials. European medicines agency, 2011

Cnaan A, Laird MN, Slasor P. Using the general linear mixed model to analyses unbalanced repeated measures and longitudinal data. *Statistics in Medicine* 1997; 16:2349-2380

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of theroyal Statistical Society,* Series B 1977; 39: 1-38

DeSouza CM, Legedza ATR. An overview of practical approaches for handling missing data in clinical trials, Journal of Biopharmaceutical Statistics 2009; 19: 1055-1073

Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis. *Applied Statistics* 1994; 43: 49-93

Follmann DA, Wu M. An approximate generalized linear model with random effects for informative missing data. *Biometrics* 1995; 51: 151-168

Harel O., Schafer JL. Partial and latent ignorability in missing-data problems. *Biometrika* 2009; 96, 1: 37-50

Hedeker D and Gibbons RD. Application of Random-Effects Pattern-Mixture Models for Missing Data in Longitudinal Studies, *Psychological Methods* 1997; 2, 1: 64-78

Hogan JW, Laird NM. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* 1997; 16: 239-257

Hogan JW, Roy J., and Korkontzelou C. Tutorial in Biostatistics Handling Drop-out in longitudinal studies, *Statistics in Medicine* 2004; 23: 1455-1497

Kenward MG, Molenberghs G. Likelihood based frequentist inference when data are missing at random. *Statistical Science* 1998; 13: 236-247

Kenward MG, Molenberghs G, Thijs H. Pattern-mixture models with proper time dependence. *Biometrika* 2003; 90(1): 53-71.

Laird NM. Missing data in longitudinal studies. *Statistics in Medicine 1988*; 7:305-315

Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*, 1982; 38:963-974

Liang KY and Zeger S. Longitudinal data analysis using generalized linear models". *Biometrika* 1986; 73 (1): 13-22

Lin H, McCulloch CE, and Rosenheck RA. Latent pattern mixture models for informative intermittent missing data in longitudinal studies. *Biometrics* 2004; 60:295-305

Little RJA, Rubin D. Statistical Analysis With Missing Data. New York: John Willey; 1987

Little RJA. A test of missing completely at random for multivariate data with missing values. *Journal of American statistical Association* 1988; 83: 1198-1202

Little RJA. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 1993; 88: 125-134

Little RJA. A class of pattern-mixture model for normal incomplete data. *Biometrika*, 1994; 81: 471-483

Little RJA. Modeling the drop-out mechanism in repeated measures studies. *Journal of American statistical Association* 1995; 90: 1112-21

Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley; 2002
Meilijson I, A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Satistical Society. Series B (Methodological)*1989;51:127-138

Lu K., Chen C. Li D., A parametric pattern-mixture model with non-future dependent missing value restrictions (2011). Submitted to *Statistics in Medicine*

Molenberghs G, Thijs H, Jansen I and Beunckens C. Analyzing incomplete longitudinal clinical trial data. *Biostatistics* (2004), 5, 3: 445-464

Mallinckrodt CH, Clark WS and Stacy R. Type I error rates from mixed-effects model repeated measures versus fixed effects analysis of variance with missing values impurted via last observation carried forward. *Drug Information Journal* 2001a; 35: 1215-1225

Mallinckrodt CH, Clark WS, Carroll RJ, and Molenberghs G., Accounting for dropout bias using mixed-effects models. *Journal of Biopharmaceutical statistics* 2001b; 11: 9-21
Myers WR. Handling missing data in clinical trials: An overview. *Drug information Journal*, 2000; 34:525-533

Panel on Handling Missing Data in Clinical Trials; National Research Council. The Prevention and Treatment of Missing Data in Clinical Trials. http://www.nap.edu/catalog/12955.html; 2010

Robins JM, Rotnitzky A, Zhao LP.Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association* 1994; 89:846-866

Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of American Statistical Association* 1995; 90:106-121

Rotnitzky A, Robins JM, Scharfstein DO. Semi-parametric regression for repeated outcomes with nonignorable nonresponse. *Journal of American Statistical Association* 1998; 93: 1321-1339

Roy J. Modeling longitudinal data with non-ignorable dropouts using a latent dropout class model. *Biometrics* 2003; 59: 829-836

Rubin DB. Multiple imputation after 18+ years. *Journal of the American Association* 1996; 91:473-489

Rubin DB. Multiple imputation for Nonresponse in Surveys, 1987, New York: John Wiley & Sons, Inc.

Rubin DB. Inference and missing data. *Biometrika*. 1976; 63:581-592

Schafer JL. *Analysis of Incomplete Multivariate Data*, 1997, New York: Chapman & Hall/CRC.

Thijs H, Molenberghs G, Michiels B., Verberke G & Curran D. Strategies to fit pattern-mixture models. *Biostatistics* 3; 245-65.

Shao J, Zhong B. Last observation carry-forward and last observation analysis. *Statistics in Medicine* 2003; 22:2429-2441

Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. New York:Springer; 2000

Wu MC, Bailey KR. Analysis changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine 1988*; 7:337-346

Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 1989; 44:175-188

Wu MC, Bailey KR. Estimation and comparison of changes in the presence of informative right censoring:conditional linear model. *Biometrics* 1989; 45:939-955
White H. Maximum likelihood estimation of misspecified models. *Econometrica* 1982; 50: 1-25

Yuan YC. Multiple Imputation for Missing Data: Concepts and New Development (Version 9.0)

Vita

Jing Gong

Education

2012 Ph. D. in Biostatistics, University of Medicine and Dentistry of New Jersey, Piscataway, NJ

2003 M.S. in Statistics, Rutgers University, NJ

2000 M.S. in Analytical Chemistry, Fudan University, Shanghai, China

1997 B.S. in Chemistry, Fudan University, Shanghai, China

Employment

2011-2012 Consultant Statistician in Forest Lab, Jersey City, NJ

2009-2010 Research Assistant, Department of Biostatistics, University of Medicine and Dentistry of New Jersey, Piscataway, NJ

2009 Intern, SHERING-PLOUGH Corporation, Kenilworth, NJ

2006-2009 Research Assistant (Biostatistician), Clinical Research and Occupational Medicine Division, University of Medicine and Dentistry of New Jersey, Piscataway, NJ

Publications

1. Laumbach RJ, Kipen HM, Kelly-McNeil K, Zhang J, Zhang L, Lioy PJ, Ohman-Strickland P, Gong J, Kusnecov A. Fiedler N. Sickness response symptoms among healthy volunteers after controlled exposures to diesel exhaust and psychological stress. Environmental Health Perspectives 2011; 119(7): 945-950. PMID 21330231.

2. Laumbach RJ, Kipen HM, Kelly-McNeil K, Gandhi S, Pham P, Gong J, Ohman-Strickland P, Effects of diesel exhaust and stress on systemic inflammation and the acute phase response in humans. American Thoracic Society (ATS) 2006 International Conference, San Francisco, CA. May 19-24, 2007.