UHASSELT

KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics

Master's thesis

The performance of different sample size calculation approaches for repeated measures designs

John Andrew

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

dr. Francesca SOLMI

dr. Robin BRUYNDONCKX

2019

2020

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics

Master's thesis

The performance of different sample size calculation approaches for repeated measures designs

John Andrew

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

dr. Francesca SOLMI

dr. Robin BRUYNDONCKX

Acknowledgements

I am deeply grateful to my supervisors Dr. Francesca Solmi and Dr. Robin Bruyndonckx for their encouragement, patience, dedication, and uncountable support during my thesis period. The meetings we had and the advice you provided have contributed much to make this thesis successful. Working with you was a great pleasure and I real appreciate.

I would also like to express my gratitude to all my lecturers at Hasselt University from whom I acquired enormous knowledge in Biostatistics. Also, I wish to express my gratitude to my beloved father, mother, brothers and sisters for their support and encouragement. Finally, I thank all my friends who provided advice and support. I am grateful for you all.

Abstract

Many researchers prefer repeated measure study because they allow to study a within subject evolution. Moreover, they have higher statistical power compared to a cross-sectional study. However, the plentiful inputs needed for sample size calculation for repeated measure studies makes it difficult to perform. This force researchers to prefer simplified methods when estimating the sample size. Assuming simplified versions of the study during sample size calculation can lead to either too large or too small sample, in which a too small study can fail to produce useful results, while a too large study is costly and unnecessary.

Using a simulation study, we compare the performance, in terms of power, between a Linear Mixed-effect Model (LMM) approach, which is a repeated measure design approach, and an independent t-test approach, a simplified approach preferred in repeated measure context. The comparison is made in two scenarios: the number of measurements per subject and the correlation between the residuals. We find that for a fixed sample size in both scenarios, the LMM is more powerful than the independent t-test. This implies that, for a fixed power, fewer subjects are required when the LMM approach is used than when the independent t-test is used. We also find that, for a fixed sample size, using the LMM approach, the power decreases with an increase in correlation between the residuals. Also, with a fixed sample size, taking more measurements beyond 5 does not add much power.

Therefore, LMM is recommended as the appropriate method to use when estimating the sample size for repeated measure studies with a moderate to large number of measurements. Also, the study design should neither be simplified nor ignored during the sample size calculation to ensure the appropriate sample size is obtained for the success of the study, ethical adherence, and cost effectiveness.

Keywords: *Linear Mixed-effects Model; Independent t-test; Power of the test; Sample size; Simulation*

Contents

List of Tables

List of Figures

1 Introduction

1.1 Background

Selecting an appropriate sample size is one of the most important step to perform when designing a study. This is because, a too small study can waste resources by failing to produce useful results, while a too large study uses more resources than required. Ethical issues always arise in studies involving human or animal subjects. A too small study involving human exposes the subjects to potentially harmful treatments without advancing knowledge as it will produce questionable results. On the other hand, a too large study exposes an unnecessary number of subjects to potentially harmful treatment. Thus, the number of subjects (sample size) in a study should be large enough to provide a reliable answer to the research objective (EMA, 2012). Also, choosing the appropriate sample size increases the probability of the study to detect the hypothesized effect, ensures cost-effectiveness, and ethical considerations (Krzywinski and Altman, 2013; Lenth, 2001).

Different factors affect sample size calculation, for example, the nature of the response of interest, effect size, the response variability, power of test, type I error, study design, and statistical method for data analysis (Ahn et al., 2014). The nature of the response influence the choice of the statistical analysis method and when other factors are fixed, different statistical methods can give a different sample size. The effect size is the hypothesized difference between the groups; however, when all factors are constant, an increase in the effect size reduces the sample size and vice versa. Assuming that all factors are fixed, when variability of the response increases, the sample size increases, the same goes with an increase in power. Contrary, the sample size decreases, when type I error increases.

The design of the study should be considered when estimating sample size since it also influences the choice of the statistical model to be used. In this study, we focus on the study design, specifically a repeated measure design. Studies with repeated measures are widely used in medical, behavioural and epidemiological fields.

Repeated measure studies are mostly preferred because they have more advantages than cross sectional studies. One of the advantages is that they allow researchers to study the within subject evolution over time. Also, collecting repeated measurements increases statistical power to detect the hypothesized difference, consequently reducing the cost of the study as fewer number of subjects will be required compared to cross sectional study. In a repeated measure study, subjects are randomized into two (or more) groups, and every subject is evaluated repeatedly over the follow up time at well defined time points,(Caruana et al., 2015). On the other hand, in a cross-sectional design, subjects are evaluated at one point only, usually at the end of the study,(Sedgwick, 2014).

Another advantage of repeated measure study is its ability to test several hypotheses in the same study. One can test for the main treatment effect, which evaluates treatment difference averaged across the repeated measures (Rochon, 1991). Also, a test for the difference at specific time points can be done. Moreover, the treatment difference across time can also be tested.

Despite the advantages of the repeated measure study over cross-sectional study, sample size calculation for such design tend to be complex. The complexity is due to the correlation that exists between the measurements taken from the same subject (Guo et al., 2013). Therefore, this existing correlation between the repeated measurements must be taken into account when calculating the sample size, (Muller et al., 1992).

There are different approaches used for sample size calculation including power analysis approach (Ahn et al., 2014), Bayesian approach (Adcock, 1988; Wang et al., 2005; Zhang et al., 2011), and precision analysis approach (Ahn et al., 2014).

In this study, we focused on the power analysis approach, which focuses on testing the null hypothesis a prespecified level of significance. The approach is characterized by many factors that can be estimated, for example, power, sample size, and effect size. Therefore, it is upon a researcher to decide what to estimate, depending on the primary objective. Using the power analysis approach, one can arrive at an appropriate sample size from two perspectives; either by fixing the power and estimate the corresponding sample size or by fixing the sample size and estimate the power that can be achieved. For example, one can directly compute the required sample size for a 90% power or power that can be achieved by a sample size of 100 subjects. In this study, we used the second perspective, that is estimating the power that can be achieved with the pre-specified sample.

For some statistical tests and models, power analysis calculations have the exact mathematical formula. These formulas express power as a function of other components, for instance, effect size and sample size. When there is no mathematical formula, approximations can be adopted. In the absence of reasonable approximations, the best alternative is the use of simulation. The process of simulation involves specifying the values for the model parameters and use them to generate randomly, a large number of data sets. Following that, one can either fit a model or perform a statistical test depending on the hypothesis of interest to each data set obtained. The power is then estimated from the proportion of the number of times the null hypothesis gets rejected. Power analysis based on simulation is always a valid alternative and gives accurate estimates when a large number of data sets are generated, (Castelloe, 2000).

Many researchers tend to use the available, and well established simplified (approximation) approaches to estimate the sample size for a repeated measure study hoping that they extrapolate accurately to the situation at hand. For example, in randomized studies with a continuous response comparing the mean response between groups, the mostly used approach is an independent t-test method which compares the mean response of different groups at the point(s) of interest (Dawson, 1998; Everitt, 1995; Frison and Pocock, 1992; Overall and Doyle, 1994; Senn et al., 2000).

Using an independent t-test involves estimating sample size on the outcome expected at a single time point. This approach ignores the study design and may lead to a too small or too large sample. The overestimation or underestimation of the sample size may lead to giving a right answer to the wrong question (Kimball, 1957). Also, since t-test assumes independence, the conclusions based on independence assumption for correlated data may be invalid (DeRouen et al., 1991; Fleiss et al., 1987). This approach is more preferred because the sample size calculation procedures are well established. For example, Ahn et al. (2014) discusses sample size calculation using univariate t-test for the continuous outcome. Also there are enough software packages for this purpose including Power and Precision (Borenstein et al., 1997), PASS (Hintze, 2008), nQuery Advisor (Elashoff, 2000), SAS/IML (Institute, 2013), and UnifyPow (O'Brien, 1998).

When the response variable is measured on a continuous scale in repeated measurement study, the plausible model is Linear Mixed-effects Model (LMM). However, sample size estimation formula associated with a test procedure in LMM situations is not mathematically tractable hence making it difficult to calculate the sample size analytically. The way forward in LMM situation is always through a computer simulation. Computer simulation was once limited by slow processing speed, but recently, computer speeds have increased, enabling simulation studies to be done in a reasonable period of time.

There are few softwares available for sample size calculation for a limited repeated measure design including PASS (Hintze, 2008). PASS software support LMM with a scalar covariate only which implies that only compound symmetry structure can be assumed. This covariance structure assumes a constant variance and a constant correlation between the repeated measurements over time. This is not always the case for measurements recorded over time since the correlation between the near measurements from the same subject can not be the same as that for measurements further apart.

In this study, we compare the performance of the two sample size calculation methods under the power analysis approach; Independent t-test and LMM through a computer simulation study. The performance of the two methods was measured by the power to reject a specified null hypothesis under a certain scenario. Different scenarios that were considered in this study were number of measurements recorded per subject, and correlation between error terms. The main aim was to show, in terms of power and sample size, what a researcher gains by using the appropriate approach and what would be the expected loss when an inappropriate approach is used.

The outline of this work is organised as follows; first in the background, we began by introducing the problem at hand and some concepts on sample size calculation for repeated measure study. We also reviewed the two methods that researchers use in practice. Following, we presented the data used in this study and laid a methodology which covers all the methods applied to answer research objectives. We proceeded to the comparison of the two methods used for sample size calculation in repeated measure design which are LMM and independent t-test. The comparison was made under two scenarios, that is, number of measurements and residual correlation through a simulation study by focusing on the empirical power to reject null hypothesis of no difference between the treatment groups at specified time points. The simulation study was used as it allowed us to compare the performance of many tests even under too small sample size scenarios. The purpose of the comparison was to demonstrate which method performs better in terms of power under the mentioned scenarios. In the results section, we presented sample sizes for some arbitrary chosen power for the two methods. Finally, we closed with a discussion of the findings, conclusion and some recommendations.

1.2 Objectives of the study

This study focused on the following objectives;

- Comparing the performance of independent t-test and LMM approaches in sample size calculation under two scenarios: number of measurements recorded per subject and correlation between the error terms (residual).
- Quantifying the difference between the methods in terms of sample size for different scenarios.
- Recommendation of the method to be used in practice for different scenarios under a pragmatic view.

1.3 Data description

This analysis used the data from a pilot study where the primary aim was to investigate the effect of the administered treatment over time. The response used to evaluate the effect of the treatment over time was a continuous variable which was recorded over time and the lower values of the response variable denoted a better condition (desired treatment effect). The experiment involved 30 subjects randomized into two treatment groups (group 1 and 2), each group having 15 subjects. Group 1 received placebo and group 2 received active treatment. Subjects received the respective treatment in month zero (starting time of the study) and further followed for 36 months. The first measurement was taken at month zero and the second measurement at the 12^{th} month since the effect of the treatment was expected to start being observable after 12 months. Other measurements were recorded after every 2 months afterwards, that is month 14, 16, 18,..., 34, 36. This resulted into 14 measurements per subject. All subject characteristics were controlled by design to ensure all subjects are comparable at month zero, hence none of them were recorded. For exploratory analysis, all the data were used; however, during modelling, the observation at month zero were discarded since no significant difference is expected at the beginning of a properly randomised study. .

2 Methodology

This section gives an overview of exploratory data analysis for longitudinal data, statistical methods used, and closes by presenting how the comparison between the performance of the two methods was carried out.

2.1 Exploratory data analysis

For longitudinal data, the preliminary analysis are done by exploring the subject profile, the mean structure, variance structure, and correlation structure in order to understand the underlying structure of the data. These analysis are intended to provide an insight into the data structure and credible implications to be considered during model fitting. As argued by Ilk (2002), visualizing data should always be the essential primary part of the analysis because it often provides information that sometimes may differ from those given by numeric summaries. Also, it provides an easy way to digest complex summaries as it requires neither model fitting nor assumptions.

2.1.1 Subject profile

The subject profile is plotted to visualize how each subject evolved over the entire study period. It is the easy way to visualize the variability present in the data. We plottted the subject profile because we want to see if there are between subject and within subjects variability which can give us an insight on whether to choose the model that model these variability or not. If there is variability observed between and within subjects, a plausible model can be a Linear Mixed-effects Model since it accommodates the between and within variability by including random effects (Gałecki and Burzykowski, 2013). Whereas, if no variability observed between subjects, a fixed effect model with correlated error terms will be a reasonable choice. This model accommodates the correlation between the repeated measurements in the residual variance-covariance by using the appropriate covariance structure.

2.1.2 Mean structure

The plot of average evolution describes how the profile for each treatment group and also for the whole population evolved over time. We used this plot to get a clue of how the fixed effects should be included in the model. If the average evolution seems linear over time, linear time effects in the model may be plausible. Moreover, if non linear, appropriate forms of time effect such as quadratic or cubic form is chosen depending on how the profile appear (Verbeke and Molenberghs, 2000).

2.1.3 Variance structure

Repeated measurements may exhibit different variability pattern. To have a picture of the variance structure between the repeated measurements over time within a subject, we plotted the average squared ordinary residuals against time. These residuals are obtained by fitting a linear regression of Y (the response) on Time. The evolution of these residuals show how the variance changes over time and hence the choice is made on which random effects to include in the model (Verbeke and Molenberghs, 2000). In general, variability can either be stable over time, increasing over time, decreasing over time or unstructured (Guo et al., 2013). Stable variability pattern is the simplest variance pattern that assumes equal variance among repeated measurements over time. Increasing variability pattern assumes the variance between repeated measurements to increase over time, whereas decreasing variability pattern assumes a decrease in variance over time among the repeated measurements. Unstructured variability pattern is the complex variance structure that assumes the variance between any pair of the measurements to be unique and with no specific form. With a stable (constant) variability over time, only a random intercept is plausible (Verbeke and Molenberghs, 2000). For other patterns, a guideline on the possible choice of additional random effects to random intercept is explained by Verbeke and Molenberghs (2000).

2.1.4 Correlation structure

Measurements from the same subject recorded over time exhibit a correlation. One way we used to visualise the existing correlation in the data was by using a scatter plot matrix. The evolution of the scatters over time shows how correlation changes over time. From the scatter plot the following can be concluded;

1. zero correlation which is concluded when the scatters show no existence of any kind of pattern between the repeated measurements at time one and over time. This implies the independence between repeated measurements.

2. constant correlation which is concluded when the scatters show the presence of a pattern at time one, with the pattern remaining constant over time. This indicates the presence of a constant correlation between the repeated measurements with time. A plausible choice for the covariance structure is compound symmetry structure which is the common correlation structure for clustered data.

3. structured correlation is concluded when the scatters exhibit a certain correlation structure over time which will imply the presence of a specific correlation pattern between the repeated measurements over time. A typical example is the first order autoregressive (AR(1)), which assumes the correlation between repeated measurements from the same subject to decline exponentially with an increase in distance between the two measurements (Guo et al., 2013). This is common for time series data and longitudinal data.

4. unstructured correlation which is concluded when the scatters shows the presence of a correlation between repeated measurements which changes over time with no specific pattern. Under unstructured correlation pattern, any two repeated measures have unique correlation, that is, if there are q repeated measures, there will be $q \times (q-1)/2$ distinct correlations to be estimated (Guo et al., 2013). Other ways to visualise correlation is by plotting standardized residual against time as explained by Verbeke and Molenberghs (2000). In all patterns, it is assumed that all subjects exhibit the same correlation pattern.

2.2 Statistical methods used

In this section we present the overview of the two methods used in this study. The method are the LMM and Independent t-test. We also present the information needed to estimate the sample size for a repeated measure study.

2.2.1 Linear Mixed effects Model

According to Stiratelli et al. (1984), an extension on models for independent observations has been proposed for repeated measurements. These models account for the correlation between repeated measurements from one subject so that correct inference can be made about the effects of covariates on the outcome variable. They also allow us to study the mean evolution of individuals observed over some period of time, and helps us to evaluate the individual deviations from this mean profile (Singer et al., 2003).

Assume that the total number of subjects is N and that there are n_i repeated measurements on each subject. The responses for the subject i are collected in the response vector $\mathbf{Y}_i$. Hence, the general linear mixed-effects model for Gaussian outcome can be written as:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon_i}, \qquad i = 1, 2, \cdots, N \tag{1}$$

$$\mathbf{b}_i \sim N(\mathbf{0}, D),$$
$$\boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \Sigma_i)$$
$$\mathrm{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}$$

where $\mathbf{X}_i$ is the $n_i \times p$ fixed between-subject design matrix, $\boldsymbol{\beta}$ is the p dimensional vector of fixed effects assumed to be common for all subjects. $\mathbf{Z}_i$ is the $n_i \times q$ random within-subject design matrix, $\mathbf{b}_i$ is the q dimensional vector of random effects, and $\boldsymbol{\varepsilon_i}$ is the vector of the measurement errors associated with the response vector $\mathbf{Y}_i$. Linear mixed models are often preferred over more traditional approaches such as repeated measures ANOVA because of their advantage in dealing with missing values.

For this study, the formulation of the LMM used for the analysis is presented below;

$$y_{ij} = \beta_0 + b_{0i} + \beta_1 Group + (\beta_2 Group_{1i} + \beta_3 Group_{2i} + b_{1i})Time_j + \epsilon_{ij} \tag{2}$$

$$\epsilon_{ij} = \rho\epsilon_{ij-1} + s_{ij}$$

$$(b_{0i}, b_{1i})' \sim N(\mathbf{0}, D)$$
$$s_{ij} \sim N(0, (1-\rho^2)\sigma^2)$$
$$\mathrm{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}$$

where $i = 1, \cdots, N$, $j = 12, 14, \cdots, 36$, y_{ij} is the response measurement for the i^{th} subject at the j^{th} month. Group is the dummy variable where group=1 is the placebo group, group=2 is the active treatment group, and group 1 is the reference group. β_0 is the mean response for group 1 at month zero and β_1 the difference in the mean response between the two treatment groups at month zero. β_2 is the effect of placebo treatment over time, β_3 is the effect of active treatment over time, and $Time$ is the continuous variable which is the month of measurement. b_{0i} is the random intercept and b_{1i} is the random slope associated with subject i. ρ is the lag 2 correlation between two adjacent error terms from one subject and s_{ij} is an independent term in the ϵ_{ij}. In this model ϵ_{ij} is split into two parts, one which depend on ϵ_{ij-1} and another which is independent from all, s_{ij}.

We first fitted the LMM (2) to the pilot data to check whether there was a difference in evolution over time between the placebo and active treatment group. We then compared the mean response between the two groups at month 12, 24 and 36 since these three time point were of interest. Following that, we tested for the need of the random slope (random time effect).

This corresponds to the hypothesis: $H_0 : d_{12} = d_{22} = 0$, which is clearly on the boundary of the parameter space. Therefore, the classical likelihood-based inference cannot be applied since one of its regularity condition is that, chi-square approximation is valid when H_0 is not on the boundary of parameter space. Thus, a mixture of chi-square distribution, $0.5\chi_2^2 + 0.5\chi_1^2$ was used. If we found significant results, we will retain the random slope in the model, if not, the model with random intercept only will be considered. Therefore, the final model will be used to evaluate the performance of LMM through simulation. The motivation for the final model used for further analysis is explained in detail in Section 3.1.

2.2.2 Independent t-test

The independent t-test is the common used method for comparison of means and in sample size calculation for the study involving independent observations. This method can also be used in repeated measure studies where the independence assumption is used. In a repeated measure study, the independent t-test involves comparing the treatment groups at each time point or for some time points of interest by just adjusting for the multiple testing. We used Bonferroni correction to determine the adjusted significance level to conclude a significant difference between the two groups. This method is a preferred due to its simplicity. However the downside follows when the number of measurements increases, the number of significance tests will increase, but the increased information about the difference between the treatment groups may be very small (Everitt, 1995). The t-test for independent observations prescribed below;

$$T = \frac{\bar{Y_1} - \bar{Y_2}}{\sqrt{\frac{1}{2}(S_1^2 + S_2^2)}} \sim t_{n_1+n_2-2} \tag{3}$$

Where $\bar{Y_1}$ and $\bar{Y_2}$ are the sample means corresponding to group 1 and group 2, S_1^2, S_2^2 are the sample variance, and n_1, n_2 are number of subjects in each group, respectively. Equal group sample sizes were considered and the variances for the two groups were assumed to be equal.

2.2.3 Sample size calculation for a repeated measure study

In this section, we described the information needed to estimate the sample size for a repeated measure study. These information should be specified prior to sample size estimation. Knowing the estimate of the number of subjects required for the study will enable a researcher to estimate the total cost of the study.

i. Choosing the hypothesis of interest

   When designing a study, a researcher must decide on the hypothesis to be tested. This always depends on the primary objective of the study (Ahn et al., 2014). Thus, the sample size calculation should be based on the stated hypothesis. In repeated measures, there are many choices of hypotheses that can be tested, but the interaction between treatment and time hypothesis is usually of interest. This hypothesis shows if the trend of the response variable across the whole follow up period is the same between the two groups. In our case, the main hypothesis was to test whether the mean response was the same between active treatment group (Group 2) and placebo group (group 1) at months 12, 24 and 36.

ii. Specifying variance and correlation patterns

Specifying variance and correlation between the repeated measurements when planning for a study with repeated measures is the most challenge since the repeated measurements from the same subject are correlated and their variance may be constant or not. If incorrect variance and or correlation are used, sample size and the final analysis will be questionable (Muller et al., 1992). In repeated measures, four variability patterns can be assumed, namely; stable variability, decreasing variability, increasing variability and unstructured variability. These variability patterns are as explained in Section 2.1.3.

For correlation patterns, four types can be thought of; zero correlation, equal correlation, structured correlation, and unstructured correlation. These are as explained in Section 2.1.4. For the repeated measures measured over time with equally spaced time intervals like the data we have at hand, the measurements taken closer in time are highly correlated than distant in time measurement. Such correlation is a structured correlation which is AR(1). The choice for the variance and correlation structures used in the analysis for this study are clearly explained and motivated in Section 3.1.

iii. Selecting a power analysis method

The method for power analysis should be chosen depending on the study design, the nature of the response variable and the planned analysis (Ilk and Cook, 2004). For example, our study planned to test whether the mean response at month 12, 24, and 36 differs between the active treatment group and placebo group. In this case, a sample size calculation based on a two group t-test would be inappropriate because the planned analysis is not a t-test. The mismatch between the study design and the planned data analysis may lead to either a too small sample size or too large sample (Muller et al., 1992). A large sample is costly, and for a clinical trial may be unethical since more subjects will be exposed to unnecessary risk. A small sample will have small power to detect the hypothesized effect and the inference may be questionable.

Our study involved repeated measurements. Therefore, a method that corrects for the within subject correlation is appropriate for the sample size calculations and analysis of the data. In practice, models for correlated data have become the most used method for analyzing repeated measures and longitudinal data but there are no general sample size calculation formula available for such models. However, there are few and limited validated power analysis packages available for some class of these models. One example is PASS (Hintze, 2008) which support a random intercept model.

The choice used by many researchers in practice is the simple methods which always assume simple assumptions about the study design. For complex study designs, the power calculations are not straight forward as they require more programming skills and theoretic knowledge of statistics. Hence the appropriate method to find appropriate sample size is through a simulation study. In this study, simulation was used since the since sample calculation formula corresponding to the model used is not mathematical tractable.

iv. Finding inputs for sample size calculation

The inputs needed to compute a sample size are type I error rate, predictor variables, primary hypothesis being tested, difference in the means response, variances between repeated measurements, and the correlation among the repeated measurements. When designing a study, inputs which are known to a researcher are the type I error rate, the predictor variable(s), and the target hypothesis. The type I error rate, chosen is the probability of rejecting the correct null hypothesis, (Guo et al., 2013). The commonly used significance levels are 0.01 or 0.05, and for this study, we considered a significance level of 0.05. Predictor variables depend on the design of the study, and for our study, the predictor was the group variable with two levels (1=placebo and 2 =active treatment).

The mean difference, variance and correlation between repeated measurements can be obtained from, either the previous studies,or the pilot study, or from the specialist experience. For this study, the inputs were obtained from the pilot data. Model 2 was fitted to the pilot data and the estimates for variance, correlation, and mean difference were obtained and used in the power calculation process.

v. Choice of the appropriate software

There are software packages and internet-based programs which are available that compute sample size for t-tests, regression models, and ANOVA. A few cover a limited range of repeated measures designs. The available software can not handle complex designs, (Castelloe, 2000). Thus, the best option for the researcher is to use computer simulation study to perform power analysis. In this report, we used simulation study in SAS to perform power analysis.

2.3 Comparing the performance of LMM and independent t-test

We measured the performance of a method by its power, which is the number of times the method detect a significant group difference at month 12, 24 and 36. This was done through a simulation study and for the two scenario as explained in the following subsections below. In each of the scenarios, the power achieved by each of the two methods under specified conditions were compared in order to identify which method is more powerful than the other.

2.3.1 Simulation study

Simulation is the way to assess the performance of a method in a situation where there is no theoretical background or for a situation where there is no closed mathematical formula. It is used to give the empirical estimation of sampling distributions, studying the assumed statistical assumptions in statistical procedures, and determining power in the hypothesis testing (Burton et al., 2006). In our case, we used simulation for power estimation in order to compare the performance of the two approaches. Simulation in this study was done as described below;

1. Model 2 was fitted to the pilot data to get the estimates for different parameters which were used as inputs to generate the data.
2. Random effects b_{0i}, b_{1i}, and s_{ij} for each subject were generated from their assumed distributions using the estimates from step 1.
3. The error term, ϵ_{ij} was obtained from s_{ij} using $\epsilon_{ij} = \rho\epsilon_{ij-1} + s_{ij}$.
4. Assignment of group indicator was done by assigning the first half of the subjects to group 1 and the remaining half to group 2.

5. The response values for each subject for the 13 time points per group were obtained as follows; for group 1, $y_{ij} = \beta_0 + b_{0i} + (\beta_2 + b_{1i}) * Time_j + \epsilon_{ij}$, and for group 2, $y_{ij} = \beta_0 + b_{0i} + \beta_1 + (\beta_3 + b_{1i}) * Time_j + \epsilon_{ij}$.

6. For each scenario, 1000 data sets were generated.

7. For each of the 1000 data sets, LMM and independent t-test were applied to test for the difference in the treatment effect between the two groups at month 12, 24, and 36. For LMM, Model 2 was fitted and for independent t-test, we used formula 3.

8. The power, which is the number of times the test concluded a significant result at $\alpha = 0.0167$ for each of the three time points was compared for the two approaches.

2.3.2 Scenarios considered

i. Correlation between error terms

The effect of the correlation between error terms, here referred to as rho on power was examined by varying its value from the observed and assess how it influences the change in power. The simulation was done as explained in the aforementioned simulation section. The parameter estimates were replaced by the values obtained after fitting Model 2 to the pilot data except for the rho which was the parameter of interest. Five arbitrary chosen values for the rho were considered which were 0.1, 0.3, 0.5, 0.7 and 0.9. For each of the rho value, a sample size of 5, 6, 7, ..., 25 subjects per group were considered. For each combination of rho value and number of subjects, 1000 data sets were simulated and analysed using Model 2. The test of the difference in mean response between the two treatment groups was done at month 12, 24 and 36. The power to conclude a significant difference at each of the three time point were examined. The correction for multiple testing was done by dividing alpha=0.05 by 3, giving alpha=0.0167, thus a p value less than 0.0167 was considered significant. The power corresponding to each value of rho were compared to see how power changed as rho value changed.

To assess the difference in the power obtained under the LMM and independent t-test, the simulated data for rho value of 0.7 were used. Other parameters and number of subjects were fixed as explained above. The value 0.7 was arbitrary chosen. For each dataset generated an independent t-test and LMM were used to test for the difference in the mean response between the two treatment groups. For every number of subjects per group a proportion of significant results at month 12, 24 and 36 was calculated and a comparison was made to see which method gave higher power than the other.

ii. Number of measurements per subject

To assess the effects of the increase in number of measurements recorded per subject on the power, a simulation was done using Model 2 as described above. The parameter estimates used in the simulation process were obtained after fitting Model 2 to the pilot data. The number of subjects per group was fixed to 15 and the number of measurements considered were 3, 5, 7, and 13. The number of subjects used here (15) was arbitrary chosen. By 3 measurements, every subject was measured at month 12, 24 and 36, and by 5 measurements a subject was measured at month 12, 18, 24, 30 and 36. For 7 measurements, every subject was measured after every 4 months from month 12 to month 36, and for 13 measurements, every subject was measured after every two months starting from 12^{th} month to the end of the study. For each number of measurements, 1000 data sets were generated and analysed using Model 2. The power to detect a significant difference between the two groups at month 12, 24 and 36 were calculated and compared to see how it changed when the number of measurements were increased.

Also a comparison of LMM and independent t-test was done under this scenario where by for every simulated data set independent t-test was applied to test the mean difference between the two groups at the same time points. Their power to detect the significant difference between the groups at the points of interest (month 12, 24 and 36) were compared.

3 Results

The methods described in Section 2 were implemented to answer the research objectives of this study. This section is organised as follows; exploratory data analysis which gave the insight of the data and motivation for the selected model, model results which presented the results of LMM result fitted when fitted to the pilot data. These estimates were used as starting values for the simulation study. Lastly, a comparison between LMM and independent t-test.

3.1 Exploratory data analysis

Under this sub-section, we take a look into the data in details by looking at subject profile, mean structure, variance structure and correlation structure. These gave the structure of the data and hence we get the evidence for the possible plausible model that fitted the data at hand.

3.1.1 Subject profile

Figure 1 shows the evolution of the response over time for 30 subjects randomized into two treatment groups, Placebo group (group 1), and the active treatment group (group 2). As seen in the plot, there is variability within and between subjects. This suggests difference in the measurements at the beginning of the treatment and also a different evolution over time between the subjects. Also, some subject profile cross depicting that subjects which start with high response value do not remain high, and those which start with low response value do not remain low. This gives the indication to fit a Linear mixed-effects Model with random effects (both random slope and random intercept). As stated by Verbeke and Molenberghs, (2000), random effects are subject specific corrections to the overall mean since they capture variability that is not captured by covariates included in the model. That is, the i^{th} random intercept tells how the i^{th} subject's intercept differ from the overall intercept. Similarly, the i^{th} random slope shows the deviation of the i^{th} subject slope from the overall slope.

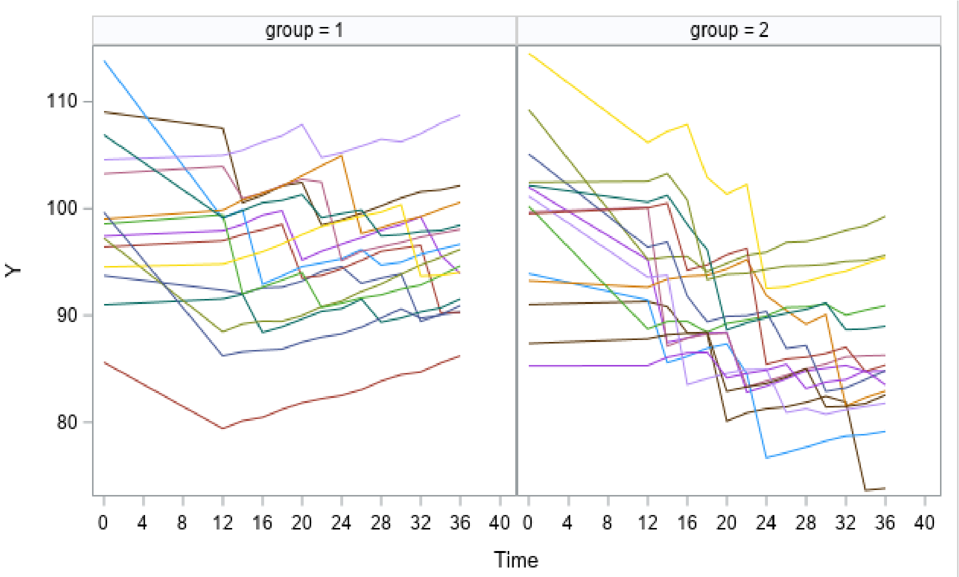
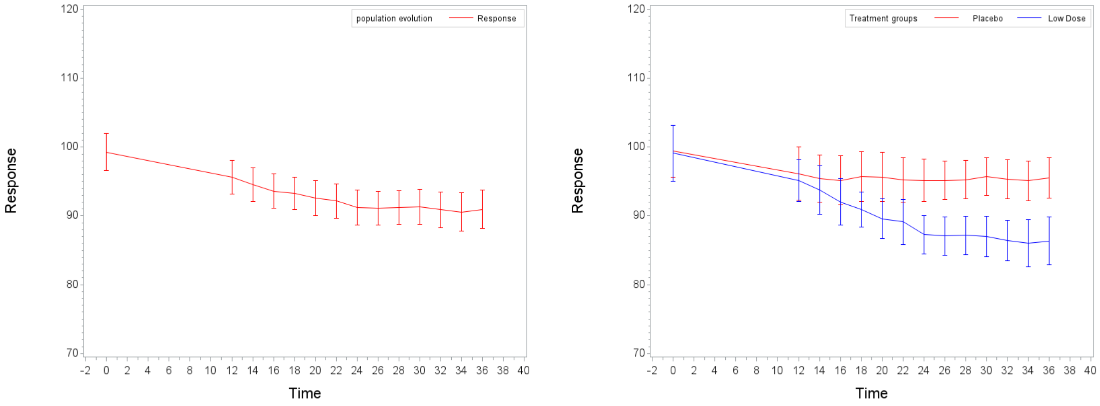


Figure 1: Individual profile plot for placebo (group 1) and active treatment (group 2). Y is the response variable

3.1.2 Mean structure

To visualize how the profile for each of the two groups evolved over time, the average trend was plotted (Figure 2). Generally, the average response for both groups seems to start at a similar point. This indicated that proper randomization of the study was done to ensure that all subjects had acceptable differences at the beginning of the study. Thus, any observed changes after time zero signal the effect of the drug being used by the subjects. The average trend showed that the mean response decreased over time indicating the possibility of having a treatment effect over time. The significance of the treatment effect will be checked in the model which will be fitted to the data. From Figure 2a, the overall time effect, that is the relationship between the response and time seems to be linear. Therefore, a linear mean model may be appropriate to model the mean structure.



(a) Population evolution (b) Evolution by group

Figure 2: Average evolution of the response variable over time.

3.1.3 Variance structure

The variability structure of the observed data is important since it acts as a guideline in selecting the variance-covariance structure of the model that can be used to describe the variability present in the data. An average trend of the squared ordinary least squares residuals was used to explore the variance structure. Figure 3 shows the evolution of variance function over time. The variability seems to be non constant over time, suggesting that the model with random intercept and random slope may be plausible for the data at hand. But the need for the random slope will be formally tested. However, it is always better to include many random effects to ensure that the remaining variability is not due to any missing random effect (Verbeke and Molenberghs, 2000). The unstructured covariance structure for the random effects was assumed since in the presence of only random slope and random intercept, only unstructured covariance structure or variance components structures are meaningful. Unstructured covariance structure assumes the random effects to be dependent while variance component structure assumes random effects to be independent. The independent assumption between random effects may sometimes be irrelevant especially for longitudinal data.

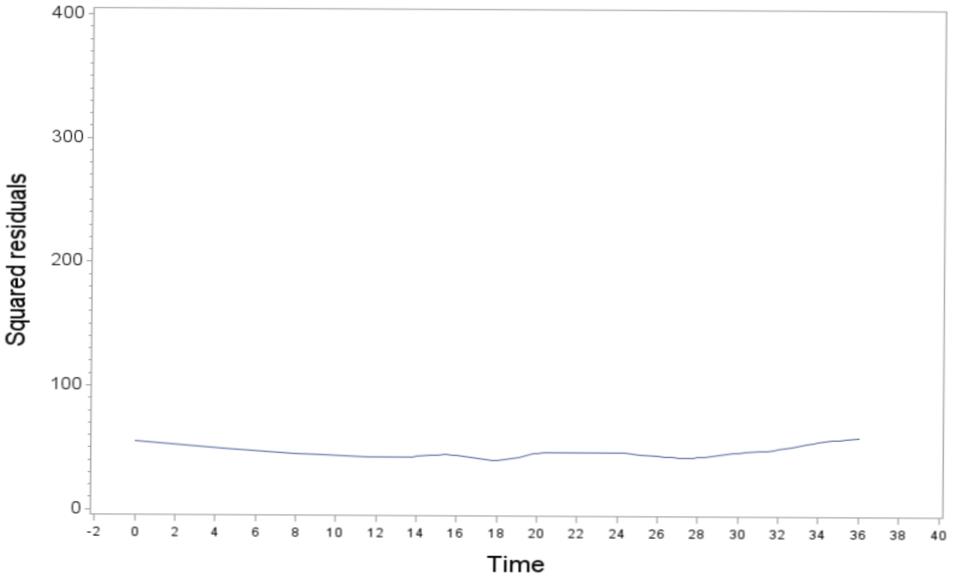


Figure 3: Evolution of variance over time

3.1.4 Correlation structure

The correlation structure describes how measurements within a subject are correlated. As it can be viewed from the scatter plot matrix (Figure 13, in the Appendix), observations from any two time points were positively correlated. This was also seen from the numerical result of the correlation matrix, A, given below. From Matrix A, it can be observed that the correlation between two closer measurements from the same subject is higher as compared to distant measurements. Thus correlation seems to decrease with an increase in the time gap between the two measurements. This suggests any correlation structure with a decaying correlation to be one of the candidate for the model to be fitted. Diggle et al.(1994) suggests the use of simple covariance structure for the residual covaiance structure in the presence of random effects other than random intercept with an argument that all dependence present in the data are assumed to be accounted for by the random effects included in the model. However one can still use appropriate covariance structure for residual. Thus, in this study AR(1) covariance structure for the residual was used in the analysis on top of unstructured covariance for random intercept and random slope.

$$A = \left(\begin{array}{cccccccccccccc}
 & \mathrm{T12} & \mathrm{T14} & \mathrm{T16} & \mathrm{T18} & \mathrm{T20} & \mathrm{T22} & \mathrm{T24} & \mathrm{T26} & \mathrm{T28} & \mathrm{T30} & \mathrm{T32} & \mathrm{T34} & \mathrm{T36} \\
\mathrm{T12} & 1 & 0.86401 & 0.8335 & 0.80996 & 0.77338 & 0.68299 & 0.56167 & 0.55831 & 0.57501 & 0.55998 & 0.59882 & 0.5639 & 0.56131 \\
\mathrm{T14} & & 1 & 0.91379 & 0.85898 & 0.78869 & 0.82321 & 0.70885 & 0.68377 & 0.68292 & 0.65182 & 0.65731 & 0.60992 & 0.60854 \\
\mathrm{T16} & & & 1 & 0.96311 & 0.8795 & 0.88242 & 0.78514 & 0.79303 & 0.80373 & 0.78675 & 0.7695 & 0.72352 & 0.71636 \\
\mathrm{T18} & & & & 1 & 0.92405 & 0.90839 & 0.82097 & 0.82888 & 0.83861 & 0.82855 & 0.80711 & 0.75089 & 0.74089 \\
\mathrm{T20} & & & & & 1 & 0.9612 & 0.85636 & 0.84837 & 0.84931 & 0.84447 & 0.82144 & 0.83643 & 0.83717 \\
\mathrm{T22} & & & & & & 1 & 0.88492 & 0.8614 & 0.85722 & 0.84987 & 0.80294 & 0.80855 & 0.80951 \\
\mathrm{T24} & & & & & & & 1 & 0.96481 & 0.95766 & 0.93645 & 0.88325 & 0.88254 & 0.87865 \\
\mathrm{T26} & & & & & & & & 1 & 0.98991 & 0.97912 & 0.93286 & 0.91894 & 0.91234 \\
\mathrm{T28} & & & & & & & & & 1 & 0.98671 & 0.94475 & 0.92259 & 0.91721 \\
\mathrm{T30} & & & & & & & & & & 1 & 0.9502 & 0.92151 & 0.91333 \\
\mathrm{T32} & & & & & & & & & & & 1 & 0.96005 & 0.95042 \\
\mathrm{T34} & & & & & & & & & & & & 1 & 0.99717 \\
\mathrm{T36} & & & & & & & & & & & & & 1
\end{array}\right)$$

3.2 Model results

This section presents the results for Model 2 when fitted to the pilot data.

3.2.1 Testing for the need of the random slope in the model

From the subject profile in Section 3.1, it was suggested that the model need to have random slope but this need to be tested formally. The result for the test is given in Table 1. From Table 1, Model 2 was fitted with b_{0i} and b_{1i} (Model 1), and with only b_{0i} (Model 2) and compared. The result was statistically significant at 5% level of significance. Hence, evidently, null hypothesis as stated in Section 2.2.1 was rejected and concluded that the model with random intercept and random time effects(slope) was appropriate. This conclusion supports the exploratory analysis explained in Section 3.1. Further analysis were done using the model with both random intercept and random slope.

Table 1: Test for the need of the random slope

Model	Random effects	REML
1	Intercept + slope	1895.8
2	Intercept	1990.3
	-2 $\ln(\lambda_N)$	94.5
	Asymptotic null distribution	$0.5\chi_2^2 + 0.5\chi_1^2$ = 4.915
	p-value	<0.0001

3.2.2 Difference between placebo group and active treatment group

The interest here was to check if the treatment has effect on the response of interest over time, and if so, to check if this difference differs between experimental (group 2) and placebo (group 1) at month 12, 24, and 36. To study these differences, the response was assessed on how it evolved over time.

Table 2 shows the placebo effect over time to be insignificant (p = 0.6306), and significant active treatment effect (p <0.0001). This indicated that, the subjects in active treatment group benefited from the drug they received unlike to subjects in placebo group. The difference between the slopes (evolution rate of the response) of the two treatment groups was statistically significant (p <0.0001). For the subjects on the active treatment group (group 2), the average response value decreased by 0.3660 units for every one month, which was 0.3440 units more than the placebo group. This was also seen on the subject profile plot, Figure 1, where the average response value seems to decline at high rate in the active treatment group (group 2) than in placebo group (group 1).

Table 2: Solution for fixed effects

Effect		Estimate	Standard Error	DF	t Value	Pr >|t|
Intercept		96.1018	2.2028	28	43.63	<.0001
Group	2	2.3623	2.8226	330	0.84	0.4032
T*group	1	-0.0220	0.0458	330	-0.48	0.6306
T*group	2	-0.3660	0.063	330	-5.81	<.0001
T*group1 - T*group2		0.3440	0.0779	330	4.42	<.0001

As seen in Table 3, the variability among subjects intercept (d_{11}) was 38.2287 and that among slopes (d_{22}) was 0.0121. This indicated that subjects' response was highly different at baseline. Also, subjects had different evolution over time. The random intercept and random slopes were negatively correlated (Correlation = - 0.57), indicating that subject that started with large response value had a low evolution rate and those started with small response value had a high evolution rate.

Table 3: Covariance parameter estimates

Cov Parm	Estimate
d_{11}	38.2287
d_{22}	0.0121
$d_{12} = d_{21}$	-0.3870
Correlation (d_{11}, d_{22})	-0.5700
ρ (rho)	0.7844
Residual variance	11.4549

ρ (rho) is a lag 2 correlation between two adjacent error terms

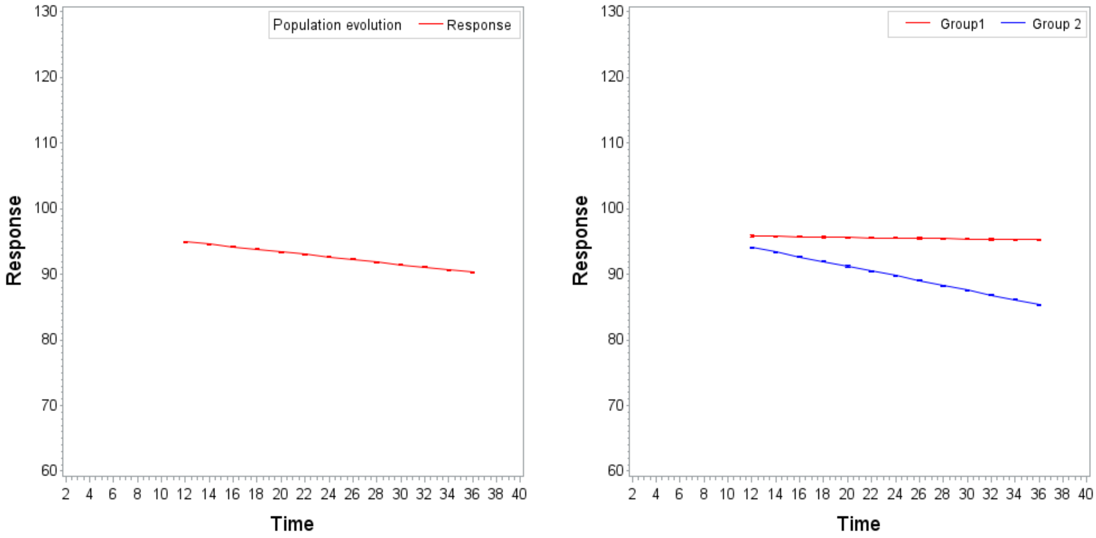
To assess if the two groups differ at month 12, 24 and 36, a comparison of the two groups was done at month 12, 24 and 36. Since there were three tests to perform, Bonferroni correction was done by dividing alpha = 5% by three, giving an adjusted alpha = 0.0167. The p value corresponding to each of the three tests was compared to alpha = 0.0167 for significance check. Table 4 showed the two groups to be significantly different at month 24 and 36, and insignificant different at month 12. The difference kept increasing with an increase in time, as shown in Table 4 where the difference increases from 1.7654 units after 12 months to 10.0209 units after 36 months. This shows that the subjects benefited more in the long run than in the short run.

Table 4: Difference between group 1 and group 2 at month 12, 24 and 36

Parameter (group 1 - group 2)	Estimate	Standard Error	DF	t Value	Pr >|t|
month 12	1.7654	2.2487	330	0.79	0.4330
month 24	5.8931	1.9728	330	2.99	0.0030
month 36	10.0209	2.1151	330	4.74	<0.0001

3.3 Simulation results

To test the performance of our simulation study, the average evolution for the two groups and a population evolution averaged over all 1000 simulated datasets using the estimates given in Table 2 and Table 3 were plotted as seen in Figure 4. The average evolution of the response from simulated data seemed to be similar to that of the original data given in Figure 2.



(a) Population evolution

(b) Evolution by group

Figure 4: Average evolution of the response variable for 1000 simulated data sets

Table 5 summarizes the estimates from the simulation study when Model 2 was fitted to the simulated data under pilot data characteristics. The results were similar to those presented from the pilot data in Table 2. This indicated that the simulation was correctly done.

Table 5: Average parameter estimates from simulation done under pilot data characteristics

Parameter (group 1 - group 2)	Estimate	Standard Error	95% LCL	95% UCL
month 12	1.6800	2.2251	-2.6812	6.0412
month 24	5.7664	1.9559	1.9328	9.6000
month 36	9.8529	2.1000	5.7369	13.9689

3.3.1 Effect of residual correlation (rho) on power

From the simulation study, the chance to reject the null hypothesis of the equality between the two groups at month 12 was small as seen in Figure 5. The power ranged between 5.4% and 11.8% for different rho values. This indicates that there is a small chance to observe a significant difference between the placebo and active treatment effect in early months. For a fixed number of subjects per group, a small change in the power due to the changes in the rho value was observed. These difference can not be clearly distinguished when increasing the number of subjects since the power curves are very close and crossing. Thus, the difference in the power for the considered rho values (0.1, 0.3, 0.5, 0.7, and 0.9) at month 12 can not be easily distinguished.

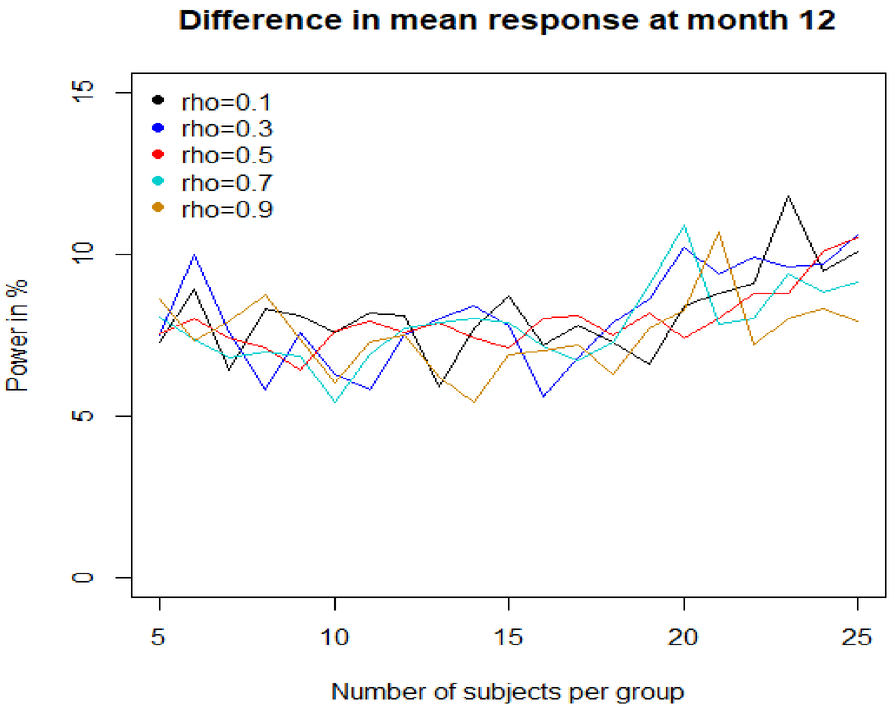


Figure 5: Power curves for different rho values for an increase in the number of subjects per group at month 12

Figure 6 shows the power of the test to give a significant difference between the two treatment groups at month 24. From the figure, the power seems to increase with an increase in number of subjects per group. The power achieved under rho values 0.1, 0.3 and 0.5 are more similar since the power curves crosses for some number of subjects. Thus, the difference in the power achieved for rho values 0.1, 0.3 and 0.5 can not be distinguished. A clear difference in the power achieved for rho value 0.7 and 0.9 was observed. The minimum power for rho=0.9 was 33.2% whereas for rho=0.7 was 35.3% and the maximum power was 89.2% for rho=0.9 and 91.1% for rho=0.7. The power decreased on average by 4% for any number of subjects by increasing the rho value from 0.1 to 0.9. This indicated that the power of the test decreases with an increase in the strength of the correlation between the residuals and vice versa. An increase in correlation implies an increase in variability between the error terms which in turn results into small power. This is because by definition, correlation is the ratio of covariance and the product of standard deviations. Hence, high correlation corresponds to high covariance which is the variability between error terms.

Figure 7 shows the power of the test to reject the null hypothesis of equality between the two treatment groups at month 36. From the figure, the power seems to decrease with an increase in the rho value. For example, with 5 subjects per group, the power decreased from 75.2% for rho=0.1 to 66% for rho=0.9, the trend continues but with a shrinkage in the difference between the power achieved for an increasing in number of subjects. For rho=0.1, 100% power was achieved with 15 subjects per group whereas 100% power was achieved with 18 subjects per group for rho=0.9. Thus, for a fixed number of subjects per group, more power is achieved when the residuals are weakly correlated than when they are strongly correlated.This is true for the opposite.

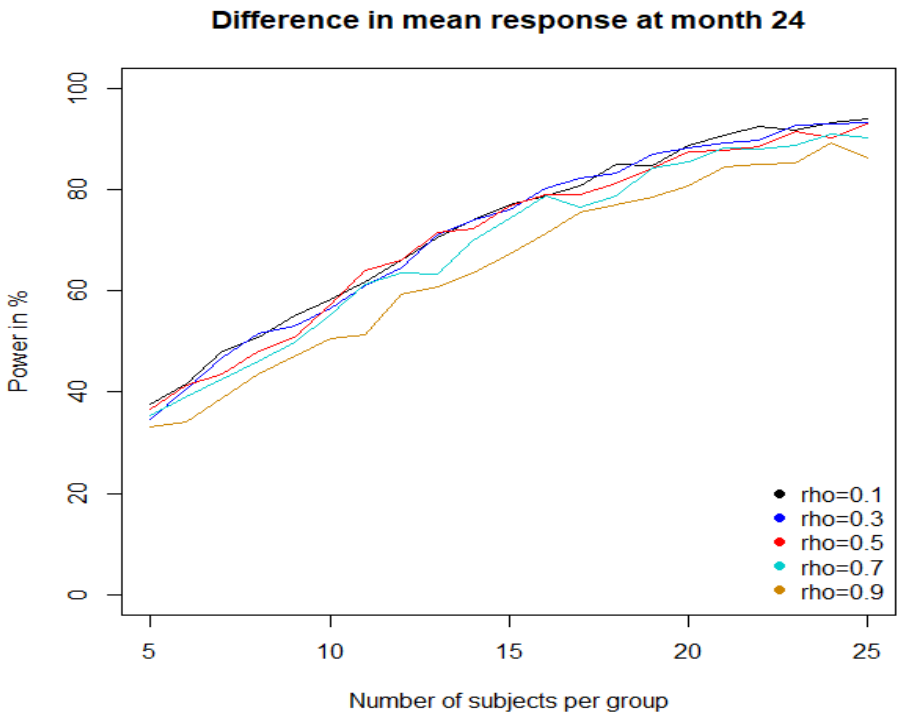


Figure 6: Power curves for different rho values for an increase in the number of subjects per group at month 24

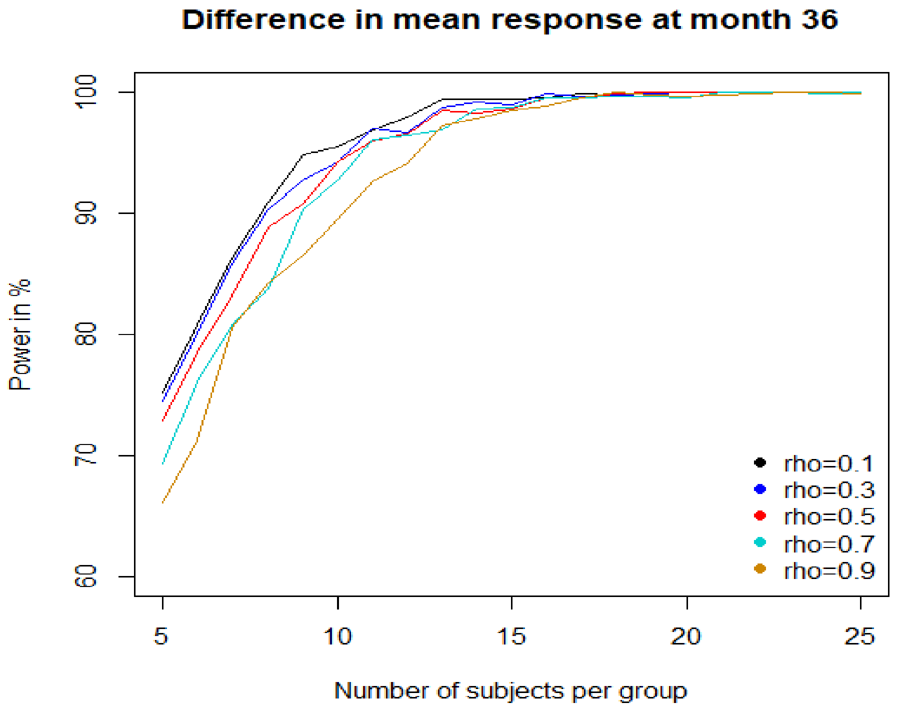


Figure 7: Power curves for different rho values for an increase in the number of subjects per group at month 36

3.3.2 The number of measurements per subject

The results are summarized in Figure 8. There was a low rate of convergence (49.2%) for the 3 measurements. This can be due to lack of enough data to estimate the covariance structures specified, or there may be no much variability in the data to be modeled with the specified model. Therefore, the 3 measurements were not considered for comparison with the other cases. The comparison was done for 5 to 13 measurements which had high convergence rate (more than 91%). As seen in Figure 8, the power remains constant with an increase in the number measurements. So, for a fixed number of subjects in a study, measuring 5 measurements per subject give the same power as measuring 13 measurements per subject. Therefore, with a fixed number of subjects per group, 5 measurements are enough for the study to have the desired power to detect the difference of interest. Taking more than 5 measurements does not give the study extra power, rather it adds costs. These costs may include cost such as clinical procedure costs, staff costs, and laboratory costs when a study is a clinical trial.

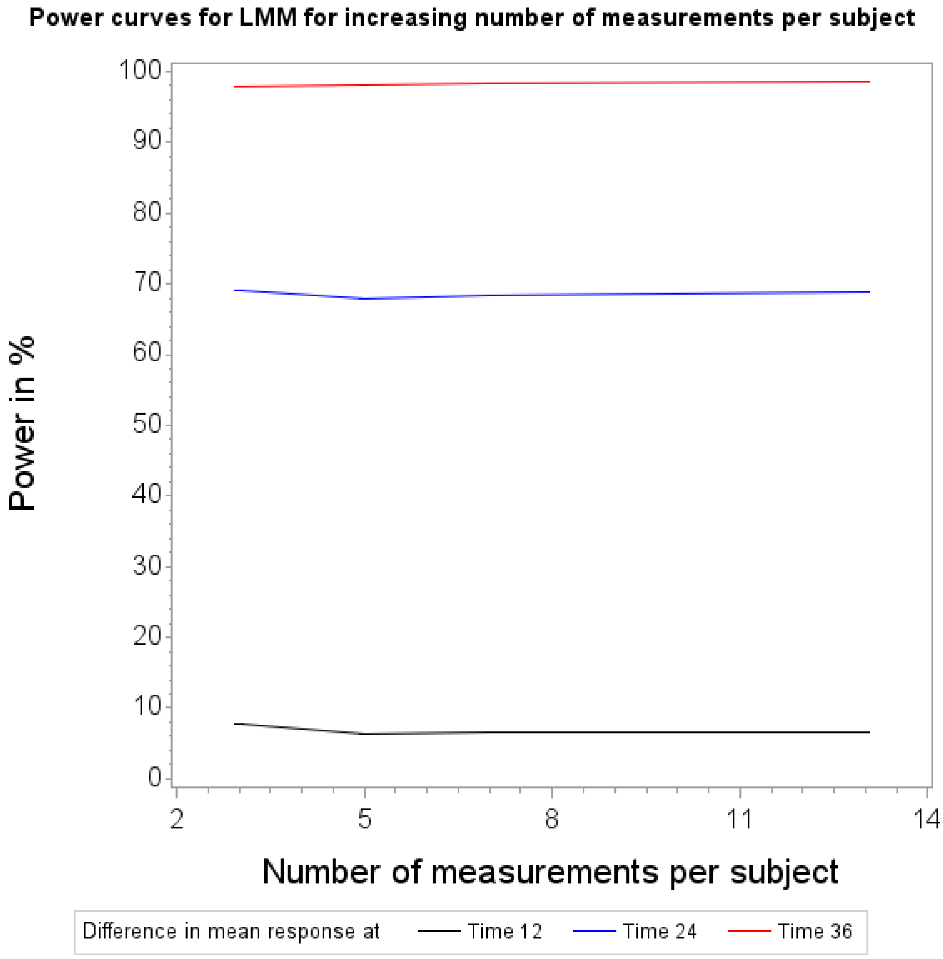


Figure 8: Power curve when 3, 5, 7 and 13 measurements are recorded per subject.

3.4 Comparison of LMM and independent t-test

The power for LMM and the independent t-test when testing for the difference in the mean response between the groups at month 12 is presented in Figure 9. We notice that, increasing the number of subjects per group from 5 to 25, the LMM power increased from 5.4% to 10.9% as compared to 2.1% to 7.3% for independent t-test. Therefore, the LMM seems to be more powerful than independent t-test by more than 3% for any number of subjects per group.

The same trend of an increase in power for an increase in number of subjects was observed at month 24 (See Figure 10). As seen in Figure 10, LMM is more powerful than independent t-test. The power increased from 35.3% to 91.1% when LMM is used as compared to 11.7% to 80% when independent t-test is used. For example, with only 5 subjects per group, LMM is more than 20% powerful than independent t-test. The difference can also be seen in number of subjects for some arbitrary selected power as given in Table 6 .As presented in the Table 6, to achieve a power of 80% we need 9 more subjects per group when using independent t-test as compared to when LMM is used.

From Figure 11, similar result was observed as in previous Figures. With 5 subjects per group, LMM is more powerful than independent t-test for more than 25%. This also was seen in terms of number of subjects per group, in Table 6. For example, with 7 subjects per group, LMM can achieve 80% power while with independent t-test, 3 more subjects are needed to achieve the same power. Thus, using LMM is more cost effective than using independent t-test since it uses less subjects to achieve the same power.

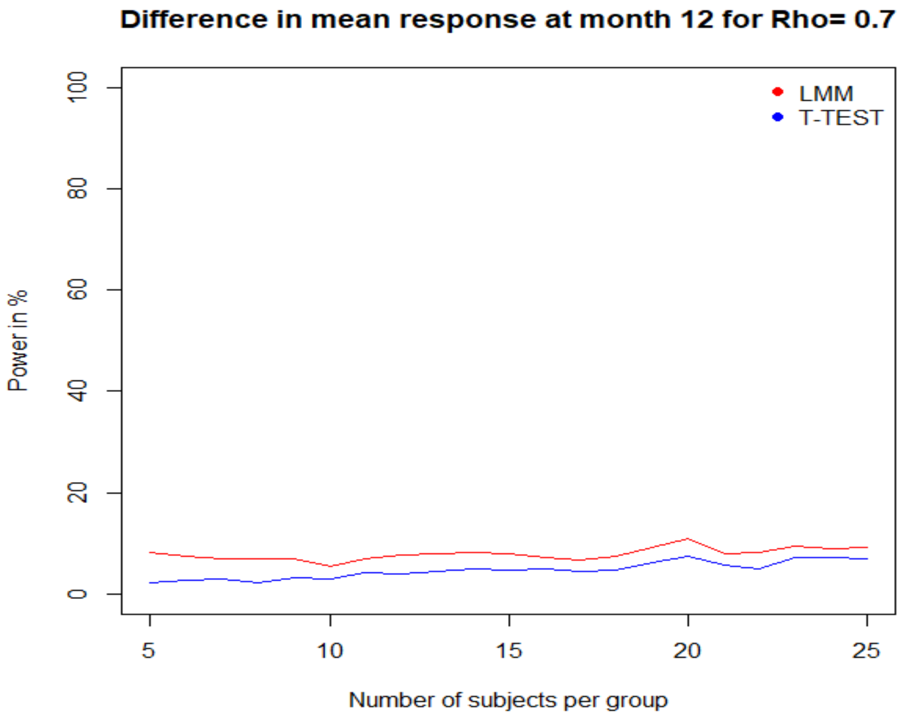


Figure 9: Power curves comparing LMM and t-test at month 12 for rho = 0.7

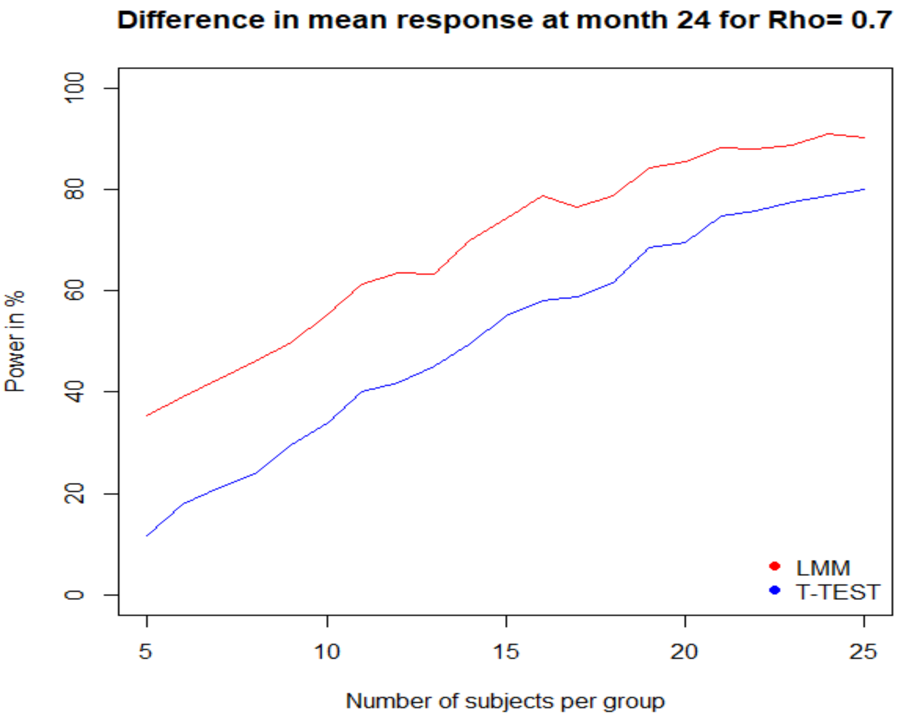


Figure 10: Power curves comparing LMM and t-test at month 24 for rho = 0.7

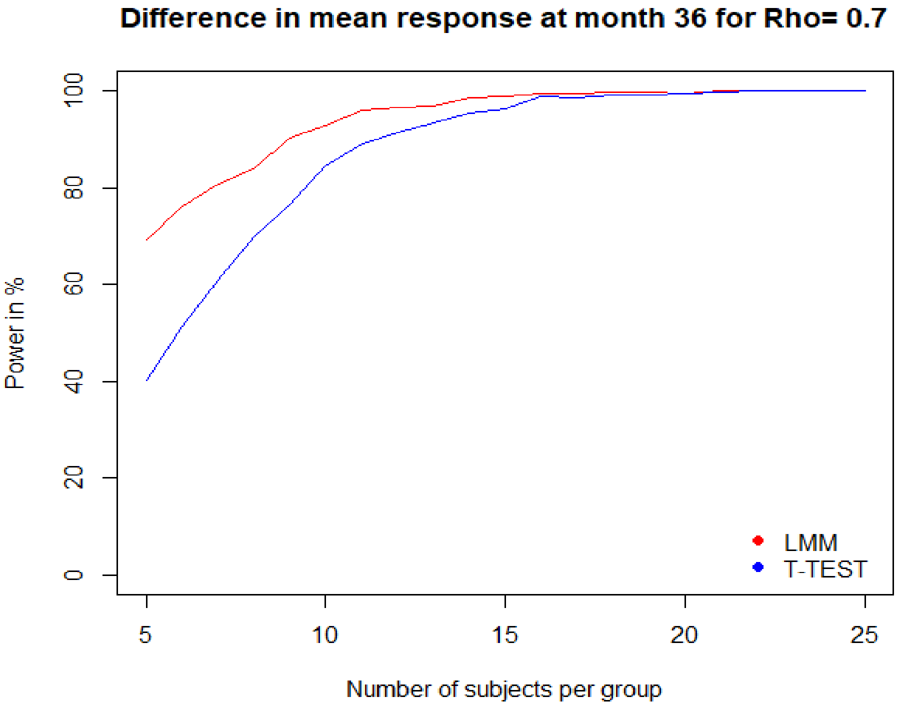


Figure 11: Power curves comparing LMM and t-test at month 36 for rho = 0.7

Table 6: Number of subjects per group for LMM and t-test for rho=0.7

Power (%)	24^{th} month		36^{th} month	
	LMM	t-test	LMM	t-test
80	16	25	7	10
85	19	26	8	11
90	23	30	9	12

The difference in power between the LMM and an independent t-test when the measurements taken from each subject are increasing is shown in Figure 12. The power from the independent t-test seems to be less than that from LMM for all the three time points compared. The difference is large for the month 24. The power from LMM is higher than that from independent t-test the test ignores the dependence present among the observations from the same subject. However, the power for all the two methods remain constant for 5 to 13 measurements. The power corresponding to 3 measurements per subject was not compared to other number of measurements due to the sane reason explained in Section 3.3.2. Therefore, since power does not increase with an increase in number of measurements taken, 5 measurements are sufficient enough to achieve the required power. Taking more measurements will increase cost for the study with no much gain in power.

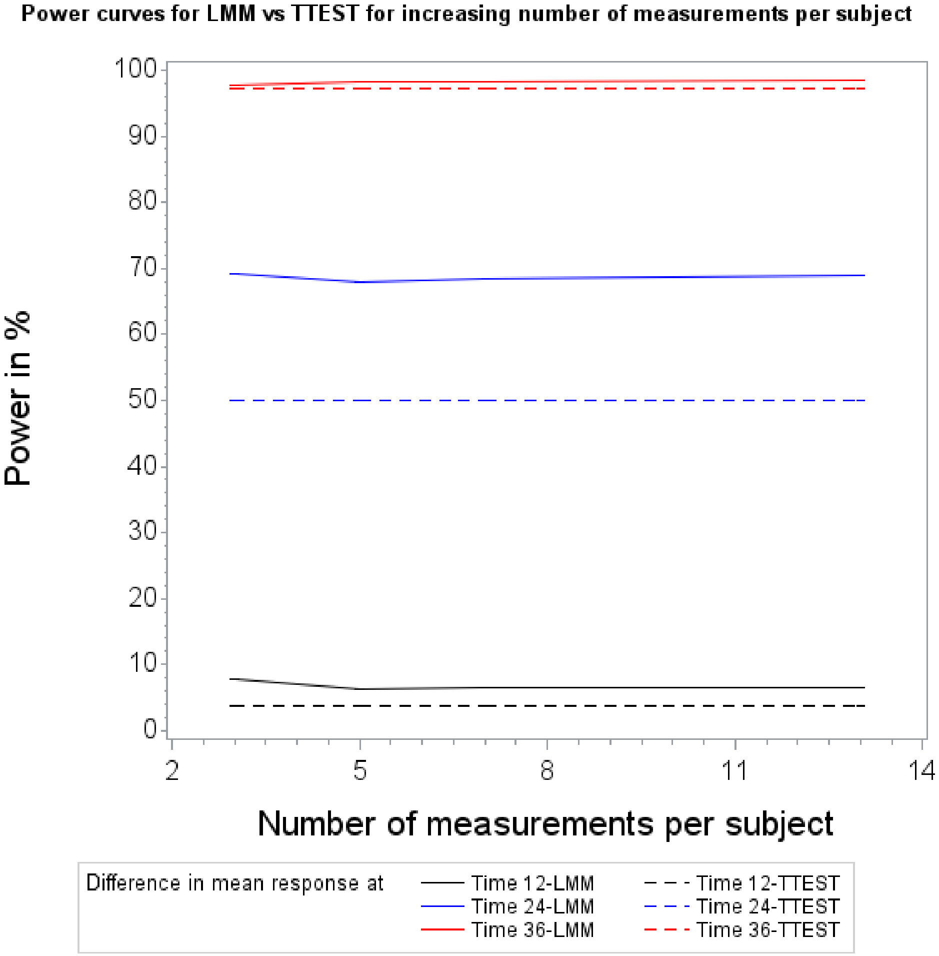


Figure 12: LMM and t-test power for increasing number of measurements per subject

4 Discussion and conclusion

The main objective of this study was to compare the performance of the approaches used for sample size calculation in the repeated measure design, and to recommend the approach to be used in practice. Two approaches, LMM and independent t-test were compared in two scenarios, namely, number of measurements per subjects,and residual correlation. The comparison aimed at showing the gain in terms of power when the appropriate approach is used and the loss when inappropriate approach is used. Simulation study was used to assess the performance of each approach and then later compared. The results showed LMM to have more power than independent t-test in both scenarios.

The LMM used for the analysis was built after exploring the pilot data using subject specific profile, mean structure, variance and correlation structures. These preliminary analyses were done to ensure we get the best fitting LMM to the data. This is because LMM may lead to misleading inference if the covariance structures for the error and or random effects are misspecified (Verbeke and Molenberghs, 2000; Vossoughi et al., 2012).

As outlined in Section 2.3.1, the pilot data was analysed first to estimate the model parameters, and these parameters were used to generate the data. LMM and independent t-test were then applied to each of the generated data set and the results (power) from the two methods were compared.

The results indicated that when other parameters are fixed, changing the residual correlation (rho) has impact on the power under both approaches. The power decreased with an increase in the strength of error correlation. This was contrary to Zhang and Ahn (2011) findings where large rho values were associated with higher power.

Also, fixing other parameters, the increase in number of measurements beyond 5 did not add much power of the test, although in general increasing number of measurements taken per subject adds power, implying that for a fixed power less subjects are required. This result was similar to Overall and Doyle (1994) and Zhang and Ahn (2011) findings as they showed that any additional measurements taken from the subject beyond four did not add power. Therefore, from our results, taking additional measurements beyond 5 will adds no power but cost.

Comparing the two methods, the result showed a higher power when LMM was used than when independent t-test is used. For different values of rho (correlation between the error terms), the power of the independent t-test remained constant since the test uses the information at a particular point only. Unlike the independent t-test, the power of LMM decreased with an increase in the correlation between the error terms. Therefore, the difference in power between the two method decreases with an increase in the strength of correlation.

When increasing the number of measurements taken per subject, LMM is powerful than independent t-test on all the time points of interest. For any additional measurement taken beyond 5, the power remained constant for both LMM and t-test. This indicates that the difference in the power achieved by the two methods does not change by increasing in number of measurements.

A difference between the two methods was also observed in the number of subject needed to detect the difference between the two groups at month 24 and 36. To detect the difference at month 24, independent t-test requires 7 more subjects per group (14 in total for the two groups) as compared to LMM to achieve 80% or 90% power. Similarly, at month 36, independent t-test requires 3 more subjects per group to achieve the same power as compared to LMM. Therefore, using LMM for sample size calculation results into using fewer subjects without losing power. This implies cost effectiveness and ethical adherence.

This study did not consider a situation with missing observations and subject characteristics like age and gender. However LMM is not highly affected by the presence of missing observations since it uses available measurements of each subject (Gałecki and Burzykowski, 2013; Verbeke and Molenberghs, 2000). This is not the case when independent t-test is used since the mean response estimate which is used under t-test method is always affected with the missing observation(s) and may lead to estimation of inappropriate sample size.

Sample size calculation in repeated measure study is more challenging as it involves solving the problem with many answers especially when it comes to specifying the variance and correlation between the repeated measurements. For example, with only 3 measurements per subject, one possibility is 3 variance values and 3 correlation which will be needed to be specified when estimating the sample size. To obtain accurate sample size, the specified values should be closer or similar to those expected from the data. The easy way to get the estimates for the variance and correlation values is to use the estimates from the previous studies or from a pilot study. The simplicity assumed by some researchers like assuming a constant variance or correlation or both over the follow up time is not recommended because it does not reflect the reality.

In conclusion, this study compared the performance of two approaches, LMM and independent t-test for sample size calculation in a repeated measure study (longitudinal) with a continuous response. LMM can handle complex design structures in repeated measure experiment unlike independent t-test. The power to detect the difference between the groups at month 12, 24 and 36 were compared. The results obtained for the two scenarios showed that the power for LMM is higher compared to that of the independent t-test. Also, for a fixed power say 80%, LMM requires fewer subjects than independent t-test implying that LMM is cost effective and ethical. LMM is also easy to implement since there are many softwares available which support LMM including SAS and R. Therefore, LMM is recommended as the first choice to use when estimating the sample size for repeated measure studies with a moderate to large number of measurements. We therefore advice the researchers to conduct appropriate and affordable power analysis to ensure they use appropriate sample size for the success of the study. We also recommend researchers to remember the following:

- Aligning the study design and hypothesis of the power analysis with the planned data analysis because misalignment can lead to overestimating or underestimating the sample size and hence lead to wastage of resources or less powerful study, respectively.
- Both study design, hypothesis, research objectives and final data analysis should be considered during sample size calculation because any change in any of these may have impact on sample size.
- Ethics and monetary cost should be considered.

With what we have demonstrated in the results between the two approaches, we hope to encourage the researchers to use the appropriate sample size calculation methods, and also to consult the statisticians from the designing stage of the study. Sample size calculation allows a researcher to evaluate the trade-off among choice of analysis, choice of the test and type I and II errors.

Also, as it is well established in literature that, the power of the study can be increased through either taking additional measurement from the existing subjects or by increasing the number of subjects (Ahn et al., 2014). Therefore, a researcher has to make a choice between the two options. The choice should depend on the differences between the cost of recruiting/getting a new subject and the cost of obtaining additional measurement from the subject. If obtaining additional measurement is cheaper than recruiting/getting a new subject, then additional measurement should be taken to achieve the desired power and vice versa.

If taking additional measurement is the desired option, the process to determine the number of measurements to be taken should be as follows; a researcher should specify the minimum number of the repeated measurements to be taken and then apply the statistical methods to obtain the optimal number of measurement to be taken. If it is in a clinical trial, it will be important to discuss with the regulatory agency about the number of measurements that satisfy the regulatory requirements.

For further study, we recommend the investigation of the performance of the two methods when there are missing observations, and considering the baseline characteristics like age and gender. As we know in longitudinal studies involving human the missing data always happen due to many reason such as drop out. This would generalize our simulation results.

References

Adcock, C. (1988), 'A bayesian approach to calculating sample sizes', *Journal of the Royal Statistical Society: Series D (The Statistician)* **37**(4-5), 433–439.

Ahn, C., Heo, M. and Zhang, S. (2014), *Sample size calculations for clustered and longitudinal outcomes in clinical research*, Chapman and Hall/CRC.

Borenstein, M., Rothstein, H. and Cohen, J. (1997), 'Power and precision, biostat'.

Burton, A., Altman, D. G., Royston, P. and Holder, R. L. (2006), 'The design of simulation studies in medical statistics', *Statistics in medicine* **25**(24), 4279–4292.

Caruana, E. J., Roman, M., Hernández-Sánchez, J. and Solli, P. (2015), 'Longitudinal studies', *Journal of thoracic disease* **7**(11), E537.

Castelloe, J. M. (2000), Sample size computations and power analysis with the sas system, *in* 'Proceedings of the Twenty-Fifth Annual SAS User's Group International Conference', Citeseer, pp. 265–25.

Dawson, J. D. (1998), 'Sample size calculations based on slopes and other summary statistics', *Biometrics* pp. 323–330.

DeRouen, T. A., Mancl, L. and Hujoel, P. (1991), 'Measurement of associations in periodontal diseases using statistical methods for dependent data', *Journal of periodontal research* **26**(3), 218–229.

Diggle, P., Liang, K.-Y. and Zeger, S. L. (1994), 'Longitudinal data analysis', *New York: Oxford University Press* **5**, 13.

Elashoff, J. (2000), 'nquery advisor release 4.0', *Statistical Solutions, Cork, Ireland, Software for MS-DOS systems* .

EMA (2012), 'Note for guidance on statistical principles for clinical trials (cpmp/ich/363/96): European medicine agency; 1998.'.

Everitt, B. (1995), 'The analysis of repeated measures: a practical review with examples', *Journal of the Royal Statistical Society: Series D (The Statistician)* **44**(1), 113–135.

Fleiss, J. L., Park, M. H. and Chilton, N. W. (1987), 'Within-mouth correlations and reliabilities for probing depth and attachment level', *Journal of periodontology* **58**(7), 460–463.

Frison, L. and Pocock, S. J. (1992), 'Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design', *Statistics in medicine* **11**(13), 1685–1704.

Gałecki, A. and Burzykowski, T. (2013), *Linear mixed-effects models using R: A step-by-step approach*, Springer Science & Business Media.

Guo, Y., Logan, H. L., Glueck, D. H. and Muller, K. E. (2013), 'Selecting a sample size for studies with repeated measures', *BMC medical research methodology* **13**(1), 100.

Hintze, J. (2008), 'Pass (power analysis and sample size)', *NCSS, LLC. Kaysville, UT: NCSS, LLC* .

Ilk, O. (2002), Exploring mean structure in longitudinal data with graphics : Linked brushing approach.

Ilk, O. and Cook, D. (2004), 'Graphical methods for exploratory multivariate longitudinal data analysis'.

Institute, S. (2013), 'Sas/iml user's guide, version 9.4'.

Kimball, A. (1957), 'Errors of the third kind in statistical consulting', *Journal of the American Statistical Association* **52**(278), 133–142.

Krzywinski, M. and Altman, N. (2013), 'Points of significance: Power and sample size'.

Lenth, R. V. (2001), 'Some practical guidelines for effective sample size determination', *The American Statistician* **55**(3), 187–193.

Muller, K. E., Lavange, L. M., Ramey, S. L. and Ramey, C. T. (1992), 'Power calculations for general linear multivariate models including repeated measures applications', *Journal of the American Statistical Association* **87**(420), 1209–1226.

Overall, J. E. and Doyle, S. R. (1994), 'Estimating sample sizes for repeated measurement designs', *Controlled clinical trials* **15**(2), 100–123.

O'Brien, R. G. (1998), A tour of unifypow: a sas module/macro for sample-size analysis, *in* 'Proceedings of the 23rd SAS Users Group International Conference', SAS Institute Cary (NC), pp. 1346–1355.

Rochon, J. (1991), 'Sample size calculations for two-group repeated-measures experiments', *Biometrics* pp. 1383–1398.

Sedgwick, P. (2014), 'Cross sectional studies: advantages and disadvantages', *Bmj* **348**, g2276.

Senn, S., Stevens, L. and Chaturvedi, N. (2000), 'Repeated measures in clinical trials: simple strategies for analysis using summary measures', *Statistics in medicine* **19**(6), 861–877.

Singer, J. D., Willett, J. B., Willett, J. B. et al. (2003), *Applied longitudinal data analysis: Modeling change and event occurrence*, Oxford university press.

Stiratelli, R., Laird, N. and Ware, J. H. (1984), 'Random-effects models for serial observations with binary response', *Biometrics* pp. 961–971.

Verbeke, G. and Molenberghs, G. (2000), *Linear mixed models for longitudinal data*, Springer-Verlag.

Vossoughi, M., Ayatollahi, S., Towhidi, M. and Ketabchi, F. (2012), 'On summary measure analysis of linear trend repeated measures data: performance comparison with two competing methods', *BMC medical research methodology* **12**(1), 33.

Wang, H., Chow, S.-C. and Chen, M. (2005), 'A bayesian approach on sample size calculation for comparing means', *Journal of Biopharmaceutical Statistics* **15**(5), 799–807.

Zhang, S. and Ahn, C. (2011), 'How many measurements for time-averaged differences in repeated measurement studies?', *Contemporary clinical trials* **32**(3), 412–417.

Zhang, X., Cutter, G. and Belin, T. (2011), 'Bayesian sample size determination under hypothesis tests', *Contemporary clinical trials* **32**(3), 393–398.

A Appendix

A.1 Figures

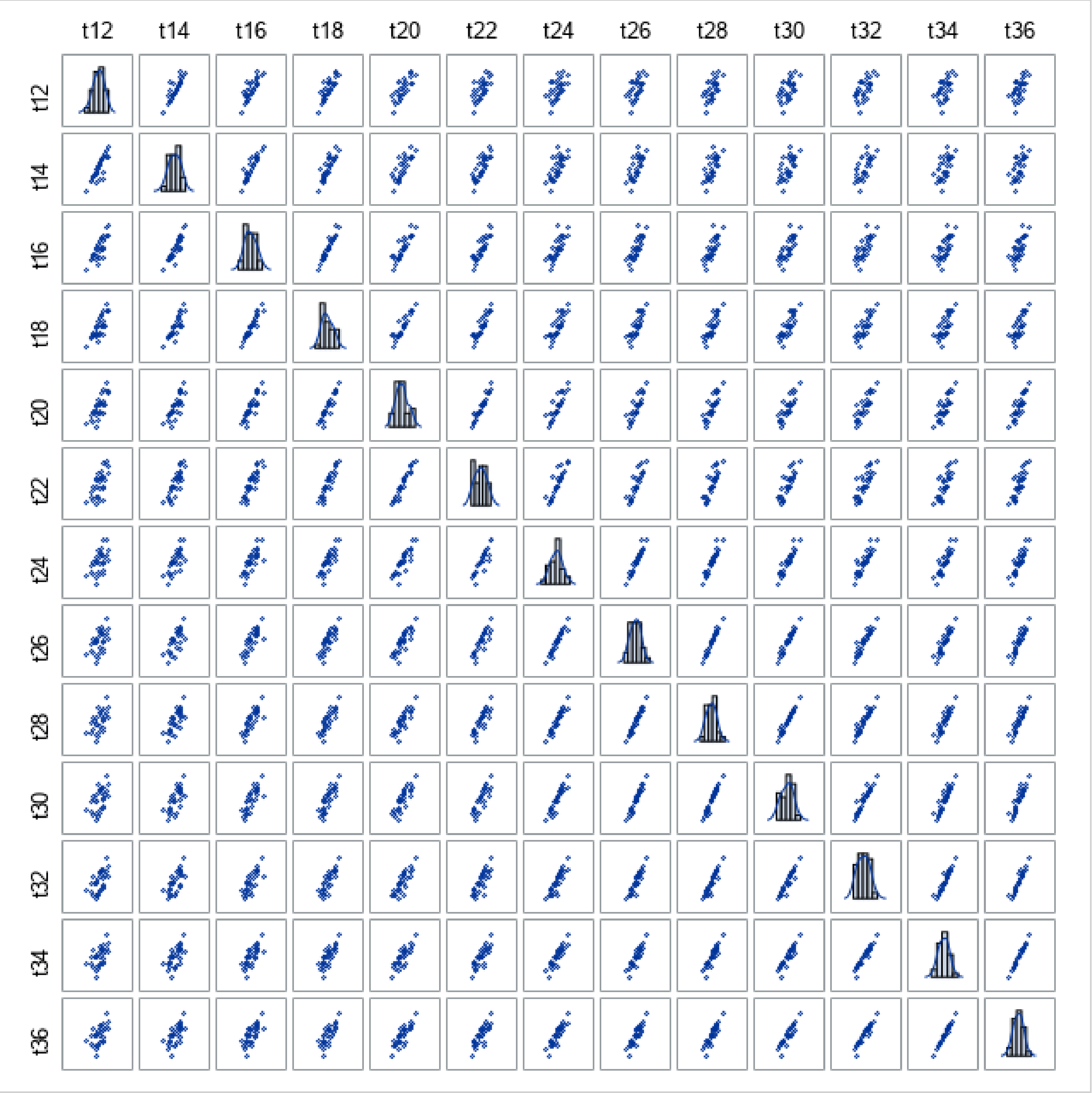


Figure 13: Scatter plot matrix.

A.2 SAS Code

```
*....Fitting LMM to pilot data.............;
proc mixed data=jo.pilotab1 method=reml empirical;
class subject group(ref="1") TC;
model Y =group T*group  / solution;
random intercept T / type=un subject=subject v vcorr g gcorr;
repeated TC / type=ar(1) subject=subject r rcorr;
*.......The differerence in the slope..............;
estimate "diff T-12"  T*group 1 -1 ;
*...The differerence in mean response at time 12, 24 and 36..;
estimate "diff T-12"  group 1 -1 T*group 12 -12 /alpha=0.0167;
estimate "diff T-24"  group 1 -1 T*group 24 -24/alpha=0.0167;
estimate "diff T-36"  group 1 -1 T*group 36 -36/alpha=0.0167;
run;quit;

*...simulation: sample size (5 and 10 per group) for rho (0.1, 0.3, 0.5, 0.7, 0.9)..;
*... for different number of subjects per group, one need to change ns column...;
*.... by filling in 2N................................;
data jo.set0;
input set g1mu g2mu vg_1 vg_2 d11 d22 drho nrun ns nr arrho v_res;
cards;
 1 96.1018 98.4641 -0.022 -0.3660 38.2287 0.0121 -0.5700 1000 10 13 0.1 11.4549
 2 96.1018 98.4641 -0.022 -0.3660 38.2287 0.0121 -0.5700 1000 10 13 0.3 11.4549
 3 96.1018 98.4641 -0.022 -0.3660 38.2287 0.0121 -0.5700 1000 10 13 0.5 11.4549
 4 96.1018 98.4641 -0.022 -0.3660 38.2287 0.0121 -0.5700 1000 10 13 0.7 11.4549
 5 96.1018 98.4641 -0.022 -0.3660 38.2287 0.0121 -0.5700 1000 10 13 0.9 11.4549
 6 96.1018 98.4641 -0.022 -0.3660 38.2287 0.0121 -0.5700 1000 20 13 0.1 11.4549
 7 96.1018 98.4641 -0.022 -0.3660 38.2287 0.0121 -0.5700 1000 20 13 0.3 11.4549
 8 96.1018 98.4641 -0.022 -0.3660 38.2287 0.0121 -0.5700 1000 20 13 0.5 11.4549
 9 96.1018 98.4641 -0.022 -0.3660 38.2287 0.0121 -0.5700 1000 20 13 0.7 11.4549
 10 96.1018 98.4641 -0.022 -0.3660 38.2287 0.0121 -0.5700 1000 20 13 0.9 11.4549
  ;
run;quit;

*..generating random intercept-b0 random slope--b1.. ;
data jo.set1; set jo.set0;
do run = 1 to nrun;
seed0 = floor(ranuni(345)* 1000);
do subject = 1 to ns;
a = rannor(0);
b=sqrt(1-drho**2);
c = drho*a+b*rannor(0);
b0 = 0 + sqrt(d11)*a;
b1 = 0 + sqrt(d22)*c;
do T = 12 to 36 by 2;
output; end; end; end; run;
```

```
*.generating error term.........;
data jo.set2; set jo.set0;
do run = 1 to nrun;
seed0 = floor(ranuni(345)* 1000);
do subject = 1 to ns;
do j= 1 to nr;
if j = 1 then do;
e = rannor(0)*sqrt(v_res);
s = 0;
output; end; else do;
s = rannor(0)*sqrt((1-arrho*arrho)*v_res);
e = arrho*e+s;
output;
end; end; end; end; run; quit;

*..group indicator..........;
data jo.set3; set jo.set1;
if subject<=(ns/2) then group=1;
else group=2;
run;quit;

proc sort data=jo.set2; by set run subject; run;quit;
proc sort data=jo.set3; by set run subject; run;quit;
data jo.set4; merge jo.set3 jo.set2; run;quit;

data jo.set5; set jo.set4;
if group=1 then y=g1mu + b0 +(T*(vg_1+b1))+e;
if group=2 then y=g2mu + b0 +(T*(vg_2+b1))+e;
run;quit;

data jo.simulation; set jo.set5; TC=T; run;quit;

*....fitting LMM...............;
ods _all_ close; ods exclude all;
ods output ConvergenceStatus =jo.ConvergenceStatus Estimates=jo.Estimates
SolutionF=jo.SolutionF CovParms=jo.CovParms V=jo.v VCorr=jo.vcorr;
proc mixed data=jo.simulation method=reml empirical;
class subject group TC;
model Y = group T*group  / solution;
random intercept T/ type=un subject=subject v vcorr g gcorr ;
repeated TC / type=ar(1) subject=subject r rcorr;
by set run;
*.......The differerence in mean response at time 12, 24 and 36....;
estimate "diff T-12"  group 1 -1 T*group 12 -12 /alpha=0.0167;
estimate "diff T-24"  group 1 -1 T*group 24 -24/alpha=0.0167;
estimate "diff T-36"  group 1 -1 T*group 36 -36/alpha=0.0167;
run;quit;ods;
```

```
*.........T TEST at time 12, 24 and 36 using simulated data.....................;
ods _all_ close; ods exclude all;
ods output ttests=jo.ttest12;
proc ttest data=jo.simulation alpha=0.0167;
class group;
var Y; where T=12;
by set run;run;ods;

ods _all_ close; ods exclude all;
ods output  ttests=jo.ttest24;
proc ttest data=jo.simulation alpha=0.0167;
class group;
var Y;
where T=24;
by set run;run;ods;

ods _all_ close; ods exclude all;
ods output  ttests=jo.ttest36 ;
proc ttest data=jo.simulation alpha=0.0167;
class group;
var Y;
where T=36;
by set run;run;ods;


*.....number measurements: 3,5,7, and 13.....;
*....simulation code LMM with random slope and intercept......;
data set0;
input g1mu g2mu vg_1 vg_2 d11 d22 drho nrun ns nr arrho v_res;
cards;
  96.1018 98.4641 -0.022 -0.3660 38.2287 0.0121 -0.5700 1000 30 13 0.7844 11.4549
 ;
run;quit;

*..generating random intercept-b0 random slope--b1  and error term....;
data set1; set set0;
do run = 1 to nrun;
seed0 = floor(ranuni(345)* 1000);
do subject = 1 to ns;
a = rannor(0);
b=sqrt(1-drho**2);
c = drho*a+b*rannor(0);
b0 = 0 + sqrt(d11)*a;
b1 = 0 + sqrt(d22)*c;
do T = 12 to 36 by 2;
output;
```

```
end;end;end;run;

data set2; set set0;
do run = 1 to nrun;
seed0 = floor(ranuni(345)* 1000);
do subject = 1 to ns;
do j= 1 to nr;
if j = 1 then do;
e = rannor(0)*sqrt(v_res);
s = 0;
output; end;
else do;
s = rannor(0)*sqrt((1-arrho*arrho)*v_res);
e = arrho*e+s;
output;
end;end;end;end;
run;quit;

*..group indicator...........;
data set3; set set1;
if subject<=(ns/2) then group=1;
else group=2;
run;quit;

proc sort data=set2; by run subject; run;quit;
proc sort data=set3; by run subject; run;quit;
data set4; merge set3 set2; run;quit;

data set5; set set4;
if group=1 then y=g1mu + b0 +(T*(vg_1+b1))+e;
if group=2 then y=g2mu + b0 +(T*(vg_2+b1))+e;
run;quit;

data jo.simulation;
set set5; TC=T;
run;quit;

*..... 3 measurements per subject ..............;
data jo.simulation3; set jo.simulation;
if T ^=12 and T ^=24 and T ^=36 then delete;
run;

*....fitting LMM................;
ods _all_ close; ods exclude ALL;
ODS OUTPUT ConvergenceStatus =jo.ConvergenceStatus Estimates=jo.Estimates
SolutionF=jo.SolutionF CovParms=jo.CovParms V=jo.v VCorr=jo.vcorr;
proc mixed data=jo.simulation3 method=reml empirical;
```

```
class subject group TC;
model Y = group T*group  / solution;
random intercept T/ type=un subject=subject v vcorr g gcorr ;
repeated TC / type=ar(1) subject=subject r rcorr;
by run;
*........The differerence in mean response at time 12, 24 and 36....;
estimate "diff T-12"  group 1 -1 T*group 12 -12 /alpha=0.0167;
estimate "diff T-24"  group 1 -1 T*group 24 -24/alpha=0.0167;
estimate "diff T-36"  group 1 -1 T*group 36 -36/alpha=0.0167;
run;quit;ods;

*.......T TEST at time 12, 24 and 36 using simulated data........;
ods _all_ close; ods exclude all;
ods output statistics =jo.stat12 ttests=jo.ttest12;
proc ttest data=jo.simulation3 alpha=0.0167;
class group;
var Y;
where T=12;
by  run;run;

ods _all_ close; ods exclude all;
ods output statistics =jo.stat24 ttests=jo.ttest24;
proc ttest data=jo.simulation3 alpha=0.0167;
class group;
var Y;
where T=24;
by  run;run;

ods _all_ close; ods exclude all;
ods output statistics =jo.stat36 ttests=jo.ttest36 ;
proc ttest data=jo.simulation3 alpha=0.0167;
class group;
var Y;
where T=36;
by run;run;

*........... 5 measurements per subject  ...;
data jo.simulation5; set jo.simulation;
if T ^=12 and T ^=18 and T ^=24 and T ^=30 and T ^=36 then delete;
run;

*....fitting LMM..........;
ods _all_ close; ods exclude ALL;
ODS OUTPUT ConvergenceStatus =jo.ConvergenceStatus Estimates=jo.Estimates
SolutionF=jo.SolutionF CovParms=jo.CovParms V=jo.v VCorr=jo.vcorr;
proc mixed data=jo.simulation5 method=reml empirical;
class subject group TC;
```

```
model Y = group T*group  / solution;
random intercept T/ type=un subject=subject v vcorr g gcorr ;
repeated TC / type=ar(1) subject=subject r rcorr;
by run;
*........The differerence in mean response at time 12, 24 and 36....;
estimate "diff T-12"  group 1 -1 T*group 12 -12 /alpha=0.0167;
estimate "diff T-24"  group 1 -1 T*group 24 -24/alpha=0.0167;
estimate "diff T-36"  group 1 -1 T*group 36 -36/alpha=0.0167;
run;quit;ods;

*.....T TEST at time 12, 24 and 36 using simulated data...........;
ods _all_ close;
ods exclude all;
ods output statistics =jo.stat12 ttests=jo.ttest12;
proc ttest data=jo.simulation5 alpha=0.0167;
class group;
var Y;
where T=12;
by  run;run;

ods _all_ close; ods exclude all;
ods output statistics =jo.stat24 ttests=jo.ttest24;
proc ttest data=jo.simulation5 alpha=0.0167;
class group;
var Y;
where T=24;
by  run;run;

ods _all_ close; ods exclude all;
ods output statistics =jo.stat36 ttests=jo.ttest36 ;
proc ttest data=jo.simulation5 alpha=0.0167;
class group;
var Y;
where T=36;
by run;run;

*............  7 measurements per subject .....;
data jo.simulation7; set jo.simulation;
if T ^=12 and T ^=16 and T ^=20 and T ^=24 and T ^=28
and T ^=32 and T ^=36 then delete;
run;

*....fitting LMM...........;
ods _all_ close; ods exclude ALL;
ODS OUTPUT ConvergenceStatus =jo.ConvergenceStatus Estimates=jo.Estimates
 SolutionF=jo.SolutionF CovParms=jo.CovParms V=jo.v VCorr=jo.vcorr;
proc mixed data=jo.simulation7 method=reml empirical;
```

```
class subject group TC;
model Y = group T*group  / solution;
random intercept T/ type=un subject=subject v vcorr g gcorr ;
repeated TC / type=ar(1) subject=subject r rcorr;
by run;
*........The differerence in mean response at time 12, 24 and 36....;
estimate "diff T-12"  group 1 -1 T*group 12 -12 /alpha=0.0167;
estimate "diff T-24"  group 1 -1 T*group 24 -24/alpha=0.0167;
estimate "diff T-36"  group 1 -1 T*group 36 -36/alpha=0.0167;
run;quit;ods;

*......T TEST at time 12, 24 and 36 using simulated data.......;
ods _all_ close; ods exclude all;
ods output statistics =jo.stat12 ttests=jo.ttest12;
proc ttest data=jo.simulation7 alpha=0.0167;
class group;
var Y;
where T=12;
by  run;run;

ods _all_ close; ods exclude all;
ods output statistics =jo.stat24 ttests=jo.ttest24;
proc ttest data=jo.simulation7 alpha=0.0167;
class group;
var Y;
where T=24;
by  run;run;

ods _all_ close; ods exclude all;
ods output statistics =jo.stat36 ttests=jo.ttest36 ;
proc ttest data=jo.simulation7 alpha=0.0167;
class group;
var Y;
where T=36;
by run;run;

*...........  13 measurements per subject .........;
data jo.simulation13;
set jo.simulation;run;

*....fitting LMM...........;
ods _all_ close; ods exclude ALL;
ODS OUTPUT ConvergenceStatus =jo.ConvergenceStatus Estimates=jo.Estimates
SolutionF=jo.SolutionF CovParms=jo.CovParms V=jo.v VCorr=jo.vcorr;
proc mixed data=jo.simulation13 method=reml empirical;
class subject group TC;
model Y = group T*group  / solution;
```

```
random intercept T/ type=un subject=subject v vcorr g gcorr ;
repeated TC / type=ar(1) subject=subject r rcorr;
by run;
*........The differerence in mean response at time 12, 24 and 36....;
estimate "diff T-12"  group 1 -1 T*group 12 -12 /alpha=0.0167;
estimate "diff T-24"  group 1 -1 T*group 24 -24/alpha=0.0167;
estimate "diff T-36"  group 1 -1 T*group 36 -36/alpha=0.0167;
run;quit;ods;

*..........T TEST at time 12, 24 and 36 using simulated data.....;
ods _all_ close; ods exclude all;
ods output statistics =jo.stat12 ttests=jo.ttest12;
proc ttest data=jo.simulation13 alpha=0.0167;
class group;
var Y;
where T=12;
by  run;run;

ods _all_ close; ods exclude all;
ods output statistics =jo.stat24 ttests=jo.ttest24;
proc ttest data=jo.simulation13 alpha=0.0167;
class group;
var Y;
where T=24;
by  run;run;

ods _all_ close; ods exclude all;
ods output statistics =jo.stat36 ttests=jo.ttest36 ;
proc ttest data=jo.simulation13 alpha=0.0167;
class group;
var Y;
where T=36;
by run;run;
```