Relative efficiency of longitudinal, endpoint, and change score analyses in randomized clinical trials

Yuanjia Wang*

Department of Biostatistics, Mailman School of Public Health, Columbia University

Email: yuanjia.wang@columbia.edu

and Naihua Duan

Department of Psychiatry and Department of Biostatistics

Columbia University, New York, NY 10032

Abstract

In the last two decades, the design of longitudinal studies, especially the sample size determination, has received extensive attention (Overall and Doyle 1994; Hedeker et al. 1999; Roy et al. 2007). However, there is little discussion on the relative efficiency of three strategies widely used to analyze randomized clinical trial data: a full longitudinal analysis using all data measured over time, an endpoint analysis using data measured at the endpoint (the primary time point for outcome evaluation), and a change score analysis using data measured at the baseline and the endpoint. When designing randomized clinical trials, investigators usually need to decide whether they would collect the interim data and if so, which type of analysis among the three would be the primary analysis. In this work, we compare the relative efficiency of detecting an intervention effect in randomized clinical trials using longitudinal, endpoint, and change score analysis, assuming linearity of the outcome trajectory and several commonly used within-individual correlation structures. Our analysis reveals an important and

*Corresponding author

somewhat surprising finding that a full longitudinal analysis using all available data is often less efficient than an endpoint analysis and the change score analysis if the correlation among the repeated measurements is not particularly strong and the drop out rate is not high. We illustrate our findings through the design of two randomized clinical trials.

Key words: Experimental design; Clinical trials; Sample size calculation; Power analysis

1 Introduction

In many randomized clinical trials, repeated measures of an outcome are recorded over the course of study. When designing or analyzing these trials, investigators need to decide whether to choose among a longitudinal analysis that incorporates all data measured over time, an endpoint analysis based on outcomes measured at the endpoint (the study's primary time point for outcome evaluation), or a change score analysis based on change from the baseline to the endpoint. The endpoint analysis is usually a two-sample test comparing group mean outcomes at the endpoint. The change score analysis is usually a two-sample test on the difference scores. The longitudinal analysis is usually based on a linear mixed effects model (Laird and Ware 1982) or a marginal model using general estimating equations (Liang and Zeger 1986) with all available data, including baseline, interim and primary time point data.

Intuitively, it might appear reasonable to assume that the longitudinal analysis would be more efficient than the endpoint or change score analysis because more data is used. However, whether this intuition indeed holds depends on how informative the extra data are. As an example, consider a randomized parallel group clinical trial with two waves of data collection at the baseline and the end of study. In this case, longitudinal analysis and change score analysis coincide. Here it is well-known that longitudinal analysis based on change scores (difference between the two waves) is less efficient than the endpoint analysis if the autocorrelation ρ is less than 0.5. The longitudinal/change scores analysis is inferior

in this case because adding the baseline measures introduces more noise than information when the autocorrelation ρ is low. It appears reasonable that the same trade-off might also hold for studies with more than two waves of data collection. However, explicit relationship between relative efficiency of these analyses and the autocorrelation has not been studied systematically in the literature for more general cases.

Our experience with practical studies reveals that the autocorrelation between the baseline and primary endpoint outcomes is often less than 0.5. For example, in a landmark randomized clinical trial of complicated grief (Shear et al. 2005), the correlation between the baseline outcome measure inventory of complicated grief (ICG) and the 16-week endpoint ICG was only 0.38. Therefore it is important to take autocorrelation into consideration when choosing analysis strategies.

The inferiority of the two-wave longitudinal/change score analysis when the autocorrelation ρ is less than 0.5 did not stop the wide use of of longitudinal studies. Part of the reason for this popularity of longitudinal studies might be because investigators and statisticians anticipated that this inferiority will diminish with more waves of data. However, our findings presented below indicate that this intuition is not always true. For example, with the autoregressive correlation structure (AR1), the breakpoint for the longitudinal analysis to be more efficient than the endpoint analysis remains very close to $\rho = 0.5$ with more waves of data (Figure 1, upper panel); furthermore, longitudinal analysis is less efficient than change score analysis for almost the entire range of autocorrelation (ρ) between the baseline and the primary time point, except when ρ is very close to zero: $\rho <= 0.013$ for four waves, and 0.027 for five waves (Figure 1, bottom panel). Therefore, the intuition that longitudinal analysis is more efficient by virtue of its use of more data is not necessarily valid.

Another part of the reason for the popularity of longitudinal studies might be due to the desire to guard against participant dropout: with longitudinal studies, interim data can be used to extrapolate to the endpoint, say, using a linear trajectory. For the first part of this paper, we focus on studies with little dropout and missing data, to illustrate the relative

efficiency among the three analysis strategies in the most transparent way. In the second part of this paper the impact of missing data on the relative efficiency among the three analysis strategies.

There is a wealth of literature on statistical methods for the analysis of longitudinal data (e.g., Diggle et al. 2002; Fitzmaurice et al. 2004; Hardin and Hilbe 2002). In addition, the design issues of longitudinal studies, especially the sample size determination, has drawn extensive attention over the years (Vonesh and Schork 1986; Muller et al. 1992; Overall and Doyle 1994; Rochon J 1998; Hedeker et al. 1999; Yan and Su 2006; Roy et al. 2007; Bhaumik et al. 2008; Moerbeek 2008; Gibbons et al. 2010). For example, Muller et al. (1992) considered power calculation for multivariate linear models with repeated measures. Hedeker et al. (1999) focused on one-degree-of-freedom contrasts such as time by group interaction, and examined power and sample size for longitudinal studies, allowing for attrition. Roy et al. (2007) extended the work in Hedeker et al. (1999) to multi-level designs where repeated measurements are nested within participants and participants are nested within clusters. However, how the power of longitudinal analysis compares with simple endpoint or change score analysis has not received much attention.

In recent years, the CONSORT statement (http://www.consort-statement.org) has been widely accepted as a guideline for designing and reporting clinical trials, endorsed by major clinical journals such as the Journal of American Medical Association, New England Journal of Medicine, and British Medical Journal. Item 6a of the CONSORT statement recommends "Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed". In addition, it states, "authors should also indicate the pre-specified time point of primary interest." Clearly, the endpoint analysis conducted at the endpoint directly follows the CONSORT guideline, under the implicit assumption that the endpoint is the "time point of primary interest."

In this paper, we compare the relative efficiency of longitudinal, endpoint, and change score analyses for randomized clinical trials. In an endpoint analysis, we focus on testing the

group mean difference between treatment arms at the endpoint. In a change score analysis, we compare the difference between the endpoint and the baseline across treatment groups. In a longitudinal analysis, we focus on testing time by group interaction, which is equivalent to comparing the group mean difference at the endpoint under several common assumptions: the expected outcome trajectory is linear in both groups; participants are randomized to one of the treatment arms (therefore their expected baseline outcomes are the same), and participants in different arms are followed by the same period of time. In this case, testing for time by group interaction is also equivalent to testing the difference in the rate of change of the outcome across groups.

We report a somewhat surprising result that the longitudinal analysis is often less efficient than the endpoint analysis if the autocorrelation among repeated measurements is not particularly strong. Furthermore, in some cases using more than two waves ofdata in the longitudinal analysis does not appear to be more advantageous than a change-score analysis. We first compare the relative efficiency without considering attrition for several widely used within-individual correlation structures and then discuss how drop outs would affect the efficiency.

This work was motivated by the design of the Recovery After an Initial Schizophrenia Episode (RAISE) study and the Healing Emotion After Loss (HEAL) study. In the RAISE study, the investigators are interested in identifying patients with early psychosis to change their prognoses of schizophrenia through coordinated and aggressive treatment in the earliest stages of illness. In the HEAL study, the primary research question is to identify the optimal treatment for individuals suffering from the debilitating condition of complicated grief (Shear et al. 2005). Both studies are multi-site randomized clinical trials with multiple assessments scheduled across the length of study. The investigators are interested in comparing improvement of a range of the functional and disease severity outcomes between two treatment arms. Both the interim outcomes and the end of study outcomes will be available and the investigators would like to choose a powerful analysis as the primary analysis. We

illustrate our methods through the design of the RAISE and HEAL studies.

2 Efficiency comparison without attrition

2.1 General model set up

For simplicity, we consider a longitudinal study with a balanced design (the same assessment schedule is used for all participants) and equally spaced assessment time points. Let i index subjects, j index the assessment time points, and k index treatment groups. We assume equal sample size allocation across two treatment groups. Let J denote the fixed total number of assessments. We assume the time point of primary interest is the last wave, J, the end point. Let μ_{jk} denote the expected outcome for group k at time point j, and let $\mu_k = (\mu_{1k}, \cdots, \mu_{Jk})^T$. The outcome model is

$$y_{ijk} = \mu_{jk} + \varepsilon_{ikj}, \tag{1}$$

where ε_{ikj} are measurement errors with mean zero. We discuss several covariance structures for ε_{ikj} in section 2.3. The expected outcome trajectory is assumed to be linear in each treatment group. We assume the same expected group mean difference at the primary time point when comparing different analyses.

In many clinical studies, the research question can be summarized as examining the significance of a pre-specified contrast

$$l = \sum_{j=1}^{J} c_j(\mu_{j1} - \mu_{j2}) = c^T(\mu_1 - \mu_2), \tag{2}$$

where c_j denote the weight placed on the contrast at time point j, and $c = (c_1, \cdots, c_J)^T$. We give three examples of commonly used contrast in the next section. For a given contrast, the corresponding linear combination of the sample mean,

$$c^T(\bar{y}_1 - \bar{y}_2),$$

where $\bar{y}_k = \frac{1}{nJ}\sum_{ij} y_{ijk}$, is used to test its significance. Denote the standardized effect size as $d_J = (\mu_{1J} - \mu_{2J})/\sigma_{JJ}$ (Cohen 1988). Under the assumption of a linear expected outcome trajectory, we obtain the expected group difference to be

$$d_j = (j-1)/(J-1)d_J. \tag{3}$$

Note that $d_1 = 0$ due to the virtue of randomization.

Overall and Doyle (1994) and Hedeker (1999) showed that the sample size needed in each group at the beginning of the study for testing contrast l is:

$$N_L = 2(z_{\alpha/2} + z_\beta)^2/d_L^2, \quad d_L^2 = l^2/c^T\Sigma c, \tag{4}$$

where $z_{\alpha/2}$ is the upper $(\alpha/2)$th percentile of a standard normal distribution, z_β is the upper βth percentile, and $\Sigma = \{\sigma^2_{jj'}\}_{j,j'=1,\cdots,J}$ denotes the $J \times J$ within-individual covariance matrix of $(\varepsilon_{i1k}, \cdots, \varepsilon_{iJk})^T$ assumed to be homogeneous across treatment groups. In addition, in the design stage of a clinical trial, investigators usually assume a homogenous variance across time points, that is, $\sigma^2_{jj} = \sigma^2, j = 1, \cdots, J$.

It can be seen from (2) that the power for the longitudinal study depends on the linear contrast to be tested (both the contrast weights and the effect sizes at each time point) and the covariance structure among the repeated measurements. We will discuss the specification for these parameters in the next two subsections.

2.2 Three common analysis strategies

There are three strategies commonly used in the analysis of clinical trial data, which we now present in the notation of a pre-specified contrast. In a longitudinal analysis, Hedeker et al. (1999) suggested to specify c_j as orthonormal coefficients, which yields a test of time by group interaction. To be specific, with two time points, the orthonormal coefficients are $c_1 = -1/\sqrt{2}$, and $c_2 = 1/\sqrt{2}$. Therefore the contrast to be tested has a norm of one, and is the difference at two time points of the expected difference of the outcome in two groups (difference of the difference, i.e., interaction effect). For three time points, the time by group

interaction is the difference between time three and time two (in the mean group difference $\mu_{j1} - \mu_{j2}$) minus the difference between time two and time one, and the coefficients are $c_1 = -1/\sqrt{2}$, $c_2 = 0$, and $c_3 = 1/\sqrt{2}$. These coefficients reflect the conventional wisdom that for odd number of time points, the group mean difference at the middle time point does not contribute to testing interaction. In general, the orthonormal contrast for testing interaction has the form

$$c_j = \begin{cases} [j - (J+1)/2]/w_J, & J \text{ odd number} \\ [2j - (J+1)]/w_J, & J \text{ even number,} \end{cases}$$

where w_J is a normalization constant.

In the notation of (2), the change score analysis analysis tests a contrast with $c_1 = -1, c_J = 1$ and $c_j = 0$, for $j \neq 1$ or J. The test statistic is $\frac{1}{n}\{(\sum_i y_{iJ2} - \sum_{i=1}^n y_{i12}) - \sum_i (y_{iJ1} - \sum_{i=1}^n y_{i11})\}$. The endpoint analysis corresponds to $c_j = 0$ for $j = 1, \cdots, J-1$ and $c_J = 1$. The test statistic is $\frac{1}{n}(\sum_i y_{iJ2} - \sum_i y_{iJ1})$. It is easy to see from (4) that the sample size needed in each group for an endpoint analysis is

$$N_E = \frac{2(z_{\alpha/2} + z_\beta)^2 \sigma_{JJ}^2}{(\mu_{1J} - \mu_{2J})^2}. \tag{5}$$

Using the standardized effect size, (5) is also written as $N_E = 2(z_{\alpha/2} + z_\beta)^2 / d_J^2$, which is a well-known result in Fleiss (1986).

2.3 Specification of the covariance structure

There is a variety of correlation structures for repeated measures that are commonly implemented in longitudinal models. One useful correlation structure is the first-order autoregressive structure (AR1), which assumes the correlation decreases exponentially with increasing time interval between two measurements. Therefore, for a balanced design

$$\sigma_{jj'}^2 = \sigma^2 \rho^{\frac{|j-j'|}{(J-1)}}, \quad \rho = \text{corr}(\varepsilon_{i1k}, \varepsilon_{iJk}).$$

Note that ρ is defined as the correlation between the measurement at the baseline and at the primary time point, to give it an interpretation that is independent of the number of intermediate assessments.

Another popular correlation structure is the compound symmetry structure, which assumes correlation between two distinct measures to be a constant regardless of how further apart they are, that is,

$$\sigma^2_{jj'} = \sigma^2\rho, \quad \rho = \text{corr}(\varepsilon_{ijk}, \varepsilon_{ij'k}), \quad \forall j \neq j'.$$

The last correlation structure we investigate is the random effects structure (Laird and Ware 1982). Here the random errors in model (1) decompose as

$$\varepsilon_{ijk} = Z\alpha_i + e_{ijk}$$

where Z is the design matrix for the random effects α_i, and e_{ijk} are independent residual measurement errors. The within-individual variance-covariance matrix can be expressed as

$$\Sigma = Z\Sigma_\alpha Z' + \Omega,$$

where $\Sigma_\alpha = \text{cov}(\alpha_i)$, $e_{ik} = (e_{i1k}, \cdots, e_{iJk})^T$, and $\Omega = \text{cov}(e_{ik})$. A random intercept model with independent residuals yields a compound symmetry covariance structure. To study a mixture of compound symmetry and AR1 correlation, we assume a random intercept model with AR1 residuals. For this case, Z is a vector of J constants of one, Σ_α is a scalar, $\{\sigma^2_\alpha\}$, and Ω has the (j, j')th element $\sigma^2_\epsilon\rho^{|j-j'|/(J-1)}$, where $\rho = \text{corr}(e_{i1k}, e_{iJk})$. We define the correlation due to the random intercept as $\rho_{cs} = \sigma^2_\alpha/(\sigma^2_\alpha + \sigma^2_\epsilon)$.

2.4 Optimal linear contrast

Although here we are interested in testing an intervention effect in a randomized clinical trial by testing time by group interaction, we briefly describe the optimal linear contrast with greatest efficiency for a given covariance structure. We define relative efficiency of longitudinal analysis comparing with the endpoint analysis as the usual Pitman asymptotic relative efficiency (ARE), is the ratio of sample size required for each analysis to reach the same power and effect size. Under the linearity assumption (3), the ARE is

$$\frac{N_L}{N_E} = \frac{c^T\Sigma c}{c^T\Omega c},$$

where

$$\Omega = uu^T, \quad \text{and} \quad u = \left(0, \cdots, \frac{j-1}{J-1}, \cdots, 1\right)^T.$$

Note that the ARE takes the form of a Raleigh quotient (Mardias et al. 1988?). By maximizing reciprocal of the ARE, we obtain the optimal linear contrast for longitudinal analysis to require least number of subjects comparing to the endpoint analysis, therefore is most efficient. The maxima of the ARE is obtained from an eigen-analysis of $\Sigma^{-1}\Omega$. To be specific, solution to the optimization problem

$$\max_c \frac{c^T \Omega c}{c^T \Sigma c}$$

is $c^* = v_1$, where v_1 is the first eigenvector of $\Sigma^{-1}\Omega$ corresponding to the largest eigenvalue, which is the maximized value of ARE.

It may be worth pointing out that the optimal linear contrast may not be the orthonormal weights used in testing interaction. For example, with two time points, the orthonormal weights on expected differences in outcome are placed equally at the baseline and the end of the study. The linear contrast for testing interaction is $c = (-1/\sqrt{2}, 1/\sqrt{2})^T$, while the optimal linear contrast is $c^* = (-1/\sqrt{5}, 4/\sqrt{5})^T$ for $\rho = 0.5$. At this correlation value, the longitudinal analysis is 25% more efficient (ARE*=0.75). In other words, longitudinal analysis would require 25% less sample size than the endpoint analysis to reach the same power. In addition, 0.5 is no longer the break point of the correlation for longitudinal analysis to be more efficient. In fact, at any positive ρ, the longitudinal analysis requires less number of subjects at the beginning of study. This suggests that under certain conditions and for a given covariance structure, one may be able to find an optimal linear contrast so that the longitudinal analysis always beats the endpoint analysis. For three time points and compound symmetry correlation with $\rho = 0.5$, $c^* = (-3/\sqrt{35}, 1/\sqrt{35}, 5/\sqrt{35})^T$, and ARE*=0.727. We will investigate the efficiency of the two analyses with optimal linear contrast in a future work.

2.5 Efficiency comparisons under three correlation structures

2.5.1 AR1 correlation structure

From (4), the necessary sample size to detect a contrast l with power $1-\beta$ is

$$N_L = \frac{2(z_{\alpha/2}+z_{\beta})^2(\sum_{j=1}^{n} c_j^2 + 2\sum_{j<j'} \rho^{\frac{|j-j'|}{(J-1)}} c_j c_{j'})}{(\sum_{j=1}^{n} c_j d_j)^2},$$

where d_j is the standardized effect size measured at time j. Recall the sample size for an endpoint analysis is N_E defined in (5), it follows that the relative efficiency of an endpoint analysis versus a longitudinal analysis is

$$r_{\text{AR1}} = \frac{N_L}{N_E} = \frac{(\sum_{j=1}^{J} c_j^2 + 2\sum_{j<j'} \rho^{\frac{|j-j'|}{J-1}} c_j c_{j'})d_J^2}{(\sum_{j=1}^{n} c_j d_j)^2}, \tag{6}$$

where $r_{\text{AR1}} > 1$ indicates the endpoint analysis to be more efficient.

In Table 1, we summarize the relative efficiency for up to five time points. We can see that the relative efficiency is a power function of the autocorrelation ρ. In Figure 1, we plot the relative efficiency at various values of ρ. For two time points, the longitudinal and change score analysis coincides. The results replicate the well-known fact that the change score (or longitudinal) analysis is more efficient than the endpoint analysis if and only if the autocorrelation between the baseline outcome and endpoint outcome is greater than 0.5. More specifically, the relative efficiency r_{AR1} is a linear function of the autocorrelation ρ: $r_{\text{AR1}} < 1$ for $\rho > 0.5$, indicating the change score (or longitudinal) analysis is more efficient than the endpoint analysis; $r_{\text{AR1}} > 1$ for $\rho < 0.5$, indicating the endpoint analysis is more efficient.

When $J = 3$, since the weight placed on the middle point is zero, r_{AR1} is the same as when $J = 2$. For four or five time points, r_{AR1} first increases and then decreases. It reaches the peak at a correlation of 0.037 and 0.029, respectively, for which values the longitudinal analysis is the least efficient. When there are four time points, at a correlation of 0.037, the longitudinal analysis requires twice of sample size as of an endpoint analysis. For five time points, at the correlation of 0.029, longitudinal analysis requires 1.95 times of sample

size as the endpoint analysis. For $J = 4$ or $J = 5$, when ρ is greater than 0.573 or 0.523, respectively, the longitudinal design is more efficient. Note that the break point correlation for the longitudinal analysis to be more efficient does not always decrease with more waves of data. With $J = 2$, the break point correlation is 0.5, while with $J = 4$, the break point correlation is 0.573.

Next we compare the efficiency of a change score analysis versus a full longitudinal analysis using interim data with $J > 2$. This is equivalent to comparing the rows with $J = 3, 4$, or 5 in Table 1 with the rows with $J = 2$. Again, using an additional wave of data collected at the midpoint of study period does not increase the efficiency, therefore using three waves of data has the same power as a change score analysis. The lower panel of Figure 1 depicts the relative efficiency using four or five waves of data. Note that the advantage of using more waves of data quickly diminishes with increasing correlation among repeated measures, indicating that the additional information in interim data is limited compared to noise. In general, the change score analysis is more efficient than the longitudinal analysis with $J = 4$ or $J = 5$, unless the correlation is very small. Specifically, with AR1 correlation when $\rho > 0.013$ and $\rho > 0.027$, respectively for $J = 4$ and $J = 5$, a change score analysis using just two waves of data is more efficient.

2.5.2 Compound symmetry correlation structure

For compound symmetry structure, the relative efficiency for an endpoint analysis versus a longitudinal analysis is

$$r_{\text{CS}} = \frac{N_L}{N_E} = \frac{(\sum_{j=1}^{J} c_j^2 + 2\rho \sum_{j<j'} c_j c_{j'}) d_J^2}{(\sum_{j=1}^{n} c_j d_j)^2}.$$

It is shown in the appendix that the above formula simplifies to

$$r_{\text{CS}} = \frac{(1-\rho) d_J^2}{(\sum_{j=1}^{J} c_j d_j)^2}. \tag{7}$$

Therefore the relative efficiency is a decreasing linear function of the correlation ρ. It is easy to see that in general, when $\rho \geq 1 - (\sum_j c_j d_j)^2 / d_J^2$, the longitudinal analysis is more

efficient than the endpoint analysis. We summarize the relative efficiency for 2, 3, 4, and 5 time points in the middle panel of Table 1. When ρ is greater than 0.5, 0.5, 0.444 and 0.375, respectively for these time points, the longitudinal design is more efficient. We show in the online appendix that when the number of time points increases, the correlation required for the longitudinal analysis to be more efficient decreases. Specifically, the break point correlation h_J with J time points has the general form of

$$h_J = 1 - \left(\sum_{j=1}^{J} c_j \frac{j-1}{J-1} \right)^2 .$$

Figure A1 in the online appendix shows this relationship for 2 to 10 time points. Note that with ten or more time points, longitudinal analysis is always more efficient.

The upper panel in Figure 2 presents the relationship between the relative efficiency and the correlation. Note that for two or three time points, the relationship is the same as the AR1 correlation. For more than two time points, at the same value of the correlation, the compound symmetry structure requires less sample size for a longitudinal longitudinal analysis than AR1 correlation. Intuitively, this is because for the former, the correlation between any two repeated measurements is a constant, while for the latter, it decreases as the time interval between the observations increases. The higher the correlation between the observations, the more efficient a longitudinal analysis is for examining the time by group interaction.

The lower panel in Figure 2 compares longitudinal analysis with a change score analysis for compound symmetry correlation. We can see that the relative efficiency is a constant. Using four or five waves of data requires 90% or 80% of sample size as a change score analysis using two waves of data, respectively. The improvement in efficiency with additional interim data is only moderate.

2.5.3 Random effects correlation structure

The random effects correlation here can be interpreted as a mixture of AR1 and compound symmetry. In this case, the sample size needed for a longitudinal analysis is

$$N_L = \frac{2(z_\alpha + z_\beta)^2[\sum_{j=1}^{J} c_j^2 + 2\sum_{j<j'}\{\rho_{cs} + (1-\rho_{cs})\rho^{\frac{|j-j'|}{J-1}}\}c_j c_{j'}]}{(\sum_{j=1}^{n} c_j d_j)^2},$$

and the relative efficiency is

$$r_{\text{mix}} = \frac{[\sum_{j=1}^{n} c_j^2 + 2\sum_{j<j'}\{\rho_{cs} + (1-\rho_{cs})\rho^{\frac{|j-j'|}{J-1}}\}c_j c_{j'}]d_n^2}{(\sum_{j=1}^{n} c_j d_j)^2}.$$

The third panel in Table 1 shows the relative efficiency for 2, 3, 4 and 5 time points. The relative sample size is a power function in the AR1 autocorrelation ρ. When $\rho_{cs} = \rho$, the correlation between the baseline and the endpoint outcome is $\rho^* = 2\rho - \rho^2$, which is also the maximum correlation between these two outcomes. With two sources of random errors in this model, we examine the relative efficiency when the two sources of correlation are equal ($\rho_{cs} = \rho$), and summarize results as a function of ρ^*. The upper panel in Figure 3 shows the trend of relative efficiency. It can be seen that the relationship between the relative efficiency and the correlation is similar regardless of the number of waves. When ρ^* is greater than 0.5, 0.5, 0.51, and 0.522 for 2, 3, 4 and 5 time points, the longitudinal analysis is more efficient. The break point for longitudinal analysis to be more beneficial increases slightly with more number of waves. At the same value of correlation between baseline and endpoint, the relative efficiency of a mixture type correlation is in between the compound symmetry and AR1.

Again the lower panel of Figure 3 compares a change score analysis with a longitudinal analysis. Similar to the AR1 correlation, under this random effects correlation, the advantage of using more waves of data also quickly disappears with increasing correlation, which implies limited information contained in the interim data. Specifically, when $\rho > 0.088$ or $\rho > 0.053$, a change score analysis is more efficient for $J = 4$ and $J = 5$, respectively.

3 Efficiency comparison with attrition

In many clinical trials, subjects drop out of the assessments regardless of investigators' best efforts to follow them in the study. In this section we investigate the impact of drop out on the relative efficiency. For simplicity, first assume that the retention rate (one minus drop out rate) is the same for both intervention groups. It is easy to generalize to differential retention rates in the two groups. Let b_j denote the retention rate at the assessment j, then the sample size observed at that point is $b_j N$. When there are drop outs, the sample size formula corresponding to (5) and (4) for an endpoint analysis and a longitudinal analysis are, respectively,

$$N_E = \frac{2(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_1 - \mu_2)^2 b_J},$$

and

$$N_L = \frac{2(z_\alpha + z_\beta)^2 [\sum_{j=1}^J c_j^2/b_j^2 + 2\sum_{j<j'} \sigma_{jj'} c_j c_{j'}/\sqrt{b_j b_{j'}}]}{(\sum_{j=1}^J c_j d_j)^2}.$$

The relative sample sizes for AR1, compound symmetry and a mixture of the two taking into account of drop outs are, respectively,

$$r_{\text{AR1}} = \frac{(\sum_{j=1}^J c_j^2/b_j^2 + 2\sum_{j<j'} \rho^{|j-j'|} c_j c_{j'}/\sqrt{b_j b_{j'}}) d_J^2 b_J}{(\sum_{j=1}^J c_j d_j)^2},$$

$$r_{\text{cs}} = \frac{(\sum_{j=1}^J c_j^2/b_j^2 + 2\rho\sum_{j<j'} c_j c_{j'}/\sqrt{b_j b_{j'}}) d_J^2 b_J}{(\sum_{j=1}^J c_j d_j)^2},$$

and

$$r_{\text{mix}} = \frac{\{\sum_{j=1}^J c_j^2/b_j^2 + 2\sum_{j<j'} (\rho_{cs} + (1-\rho_{cs}))\rho^{|j-j'|} c_j c_{j'}/\sqrt{b_j b_{j'}}\} d_J^2 b_J}{(\sum_{j=1}^J c_j d_j)^2}.$$

Assume the drop out rate increases linearly across time, that is, $b_j = 1 - \dfrac{j-1}{J-1}(1 - b_J)$. Tables 2 shows the numerical values of the break point for longitudinal analysis to be more efficient for the AR1, compound symmetry and random effect correlation. Note the decreasing trend of the break point as the retention rate decreases in all three covariance structures.

Figure A2 in the online appendix shows the relative efficiency as a function of the autocorrelation with the AR1 structure for several retention rates at the end of study. As the final retention rate decreases, smaller correlation is required for the longitudinal analysis to be more efficient than the endpoint analysis. This reflects the benefit of longitudinal analysis utilizing all interim data when there are drop outs. For example, with a 50% retention rate and two time points, when $\rho > 0.25$ the longitudinal design is more efficient. In contrast, when there are no drop outs, the break point is $\rho > 0.5$.

Figures A3 and A4 in the online appendix present the relative efficiency for the compound symmetry and random effect correlation. We see a similar trend between the retention rate and the break point for the longitudinal analysis to be more efficient. Note that for five time points, 25% retention rate and with compound symmetry correlation, under all correlation values the longitudinal analysis is more efficient.

4 Application to the RAISE and HEAL studies

In the RAISE study, a longitudinal design with five equally spaced assessments was used. The investigators are interested in whether an endpoint analysis or a longitudinal analysis should be proposed as the primary analysis. We compared the sample size needed for each of the analysis to have 80% power under various values of attrition rates. For an endpoint analysis with 20% attrition rate and an effect size of 0.4 at the end of the study, the required sample size for comparing the outcome in two groups with a 1:1 sample size ratio is 124 per group. Assume the same effect size at the end of study and a linear drop out rate, for an AR1 structure with autocorrelation of 0.5, the sample size required to compare the rate of change between two groups for a longitudinal analysis is 143 per group. The endpoint analysis was chosen as primary analysis due to its robustness to model misspecification, simplicity in interpretation and good power.

The HEAL study is a randomized treatment study of complicated grief (CG). The primary research question is to compare rate of change in severity scores among individuals with

CG who were assigned to receive escitalopram (ESC) plus complicated grief psychotherapy (CGT) to the rate of change among those who were assigned to receive placebo (PBO) plus CGT. The sample size was 110 each arm. We assume a target power of 80%, a baseline assessment and eleven post-randomization weekly measurements, an AR1 structure with correlation between two adjacent measurements to be 0.93 (estimated from the symptom severity data from a prior study) and between the baseline and the end of study is $0.93^{11} = 0.45$. We also assume that the assessment dropout is distributed uniformly across waves of follow-up assessments. We derive the minimal detection limit (MDL), the smallest effect size that can be detected with 80% power (any effect size larger than MDL has more than 80% power; any effect size smaller than MDL has less than 80% power). With 0% and 10% drop out rate, the MDL in the rate of change was 0.410 and 0.423 points per week, respectively. The expected MDL at the end of the study is 4.51 and 4.65 points, respectively. For an endpoint analysis, the MDL was 0.758 and 0.8 points, respectively. For this example, clearly an endpoint analysis is more powerful, due to a low correlation between the baseline and the end of study outcomes.

5 Discussion

In this work we are interested in the relationship between relative efficiency of endpoint or change score analysis versus longitudinal analysis and correlation among repeated measures for testing a treatment effect in randomized clinical trials. We show that in some cases, it is reasonable to focus on the primary time point outcome or change from the baseline, and not to use the interim measurements. When there is no attrition and with an AR1 correlation structure, the correlation between adjacent outcomes needs to be fairly large for the longitudinal analysis to be more beneficial. For four or five time points, the relative efficiency is a power function with a unique peak between zero and one. When designing a study with little knowledge of the autocorrelation, one can use the "worse-case autocorrelation" which leads to the largest sample size as a benchmark. For example, for four time points, the required

sample size reaches its peak at $\rho = 0.018$, and one can use this as assumed correlation for the sample size calculation.

For compound symmetry structure, the correlation between repeated measurements is a constant across time. With the same number of time points and same correlation, the break point is smaller than the AR1, indicating that compound symmetry structure is more favorable for the longitudinal analysis. The break point decreases with increasing number of time points, implying that the efficiency gain of the longitudinal analysis is greater when there are more measurements per subject. Longitudinal analysis is always more efficient with $J > 10$. However, the compound symmetry structure assumes a constant correlation across all time points, which may not be realistic. For a mixture of compound symmetry and AR1, the break point is between the previous two types of covariance.

When $J > 3$, we also compared efficiency of using two waves of data in a change score analysis (baseline and endpoint outcome) with using additional interim data in a longitudinal analysis. For any correlation structure, using an additional wave at the middle point of study does not improve power for testing time by group interaction. For compound symmetry structure, using four and five waves of data is 10% and 20% more efficient, respectively. For the other two types, using additional waves of data is only beneficial when the correlation between repeated measures is very low. In particular, for AR1 using four waves of data can be much less efficient than just using two waves when correlation is high. Note that for a mixture of AR1 and compound symmetry correlation, using five waves of data can be less informative than four waves when assessments are equally spaced. This might be due to an interplay of compound symmetry and AR1 correlation. A future work on optimal design of assessment time points is in demand.

The endpoint analysis is simple and transparent, and the results are clearly understood without any ambiguity. This analysis does not require any modeling on the trajectory of the outcome, therefore the interpretation of the results is free from such model assumptions. In contrast, longitudinal analysis can be used to evaluate interim treatment effect and cu-

mulative treatment effect, but a parametric longitudinal analysis requires assumptions on the functional form of the trajectories and the covariance structure of the repeated assessments. Validity of the results might be questionable if these assumptions are not satisfied. Of course nonparametric longitudinal analysis (see for example, Wu and Zhang 2006) or marginal approaches without requiring correct specification of correlation structure can be performed when some of these assumptions are in doubt. However, it is well known that nonparametric analysis requires larger sample size to reach the same power comparing to a correctly specified parametric analysis.

The pros and cons of endpoint versus longitudinal analysis depend in part on the anticipated assessment drop-out rate. When there is substantial attrition, the benefit of longitudinal analysis using all interim data becomes more crucial. When attrition rate increases, the break points in correlation in all three types of covariances decreases. In fact, for five time points, with high proportion of missing data (retention rate at the end of study being 25% or less) and compound symmetry covariance, the longitudinal analysis is always more efficient at any correlation. Therefore for studies anticipating large proportion of drop outs, the ability for the longitudinal analysis to capture available information for drop-out cases would be attractive. However, the cost of interim data collection should be evaluated in light of the increased efficiency. For studies making intensive efforts to assess participants irrespective of whether they drop out of the protocol, therefore anticipating a low assessment drop-out rate, the longitudinal analysis has less benefit. In addition, the comparison of longitudinal versus change score analysis reveals limited information in interim data and most of information is retained in the baseline and primary time point data.

One potential strategy combining endpoint and longitudinal analysis is to use endpoint for complete cases, and use longitudinal model to extrapolate incomplete cases, then analyze endpoint data combining complete cases and extrapolated cases. This method may be more favorable than last-observation-carry-forward under very general conditions, and could also be better than the longitudinal analysis per se when the correlation is not very high. Works

along this line is under further investigation.

Some studies repeatedly collect the assessments over time as a retention tool to stay in contact with participants so that they don't dropout (and also to collect updated contact info). However, it is not necessary to administer the full assessment battery which can be burdensome to participants and costly to studies for the purpose of retention, especially if the follow-up requires in-person encounter. A brief follow-up by phone or postcard or electronic means will probably suffice.

Here we assume balanced and equally spaced data, which may be relaxed by incorporating a suitable design matrix. Here we also do not consider designs where subjects are nested in clusters. With such designs, the correlation between clusters may play an important role as well as the correlation between repeated measures on the same subjects (Bhaumik 2008).

In summary, the relative efficiency of longitudinal versus cross-sectional design hinges on the correlation between repeated measurements and attrition rates. When the correlation is weak, a longitudinal analysis is not necessarily more efficient than an endpoint or change score analysis. When there is significant assessment drop out, however, the longitudinal analysis is more favorable. In many practical situations, instead of jumping into the conclusion of choosing a longitudinal analysis, care should be taken when designing a clinical trial.

Acknowledgements

The authors would like to thank Drs. Jeffrey Lieberman, Lisa Dixon and Susan Essock for discussions on the design of the RAISE study (08DS00006, HHSN-271-2009-00020C), and Drs. Kathy Shear, Charles Reynolds, Naomi Simons, Zisook Sidney, and Barry Lebowitz on the design of HEAL study (MH085308-02). Wang's research is supported by NIH grants AG031113-01A2 and NS073670-01.

References

Bhaumik DK, Roy A, Aryal S, Hur K, Duan N, Normand SL and Gibbons RD. (2008). Sample size determination for studies with repeated continuous outcomes. Psychiatric Annals. 38(12): 765771.

Cohen J. (1988). *Statistical power for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Fleiss JL. (1986). The design and analysis of clinical experiments. New York: Wiley.

Gibbons RD, Hedeker D, DuToit S. (2010). Advances in Analysis of Longitudinal Data. Annu Rev Clin Psychol. 6:79-107.

Hedeker D, Gibbons RD, and Waternaux C. (1999). Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. Journal of Educational and Behavioral Statistics. 24:70-93.

Hardin J, Hilbe J. (2002). *Generalized Estimating Equations.* London: Chapman and Hall.

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

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

Muller KE, LaVange LM, Ramey SL, and Ramey CT. (1992). Power calculations for general linear multivariate models including repeated measures applications. Journal of the American Statistical Association. 87:1209-1226.

Overall JE, and Doyle SR. (2008). Estimating sample sizes for repeated measurement designs. Controlled Clinical Trials. 15:100-123.

Moerbeek M. (2008). Powerful and cost-efficient designs for longitudinal intervention studies with two treatment groups. Journal of Educational and Behavioral Statistics. 33:41–61.

Muller KE, LaVange LM, Ramey SL, Ramey CT. (1992). Power calculations for general linear multivariate models including repeated measures applications. Journal of the American Statistical Association. 87:12091226.

Rochon J. (1998). Application of GEE procedures for sample size calculations in repeated measures experiments. Statistics in medicine. 17(14), 1643-1658.

Roy A, Bhaumik DK, Aryal S, Gibbons RD. (2007). Sample size determination for hierarchical longitudinal designs with differential attrition rates. Biometrics. 63: 699-707.

Shear K, Frank E, Houch P, and Reynolds C. (2005). Treatment of Complicated Grief: A Randomized Controlled Trial. Journal of the American Medical Association. 293(21), 2601-2608.

Yan X, Su X. (2006). Sample size determination for clinical trials in patients with nonlinear disease progression. Journal of Biopharmaceutical Statistics. 16:91-105

Vonesh EF, Schork MA. (1986). Sample sizes in multivariate analysis of repeated measurements. Biometrics. 42:601610.

Wu H and Zhang, J. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis Mixed-Effects Modeling Approaches.* Wiley.

Table 1: Relative efficiency of an endpoint analysis versus a longitudinal analysis for various within-subject covariance structures

J	AR1: Relative efficiency ${r_{\text{AR1}}}^*$
2	$2(1-\rho)$
3	$2(1-\rho)$
4	$\frac{9}{5}+\frac{9}{10}\rho^{1/3}-\frac{27}{25}\rho^{2/3}-\frac{81}{50}\rho$
5	$\frac{8}{5}+\frac{32}{25}\rho^{1/4}-\frac{8}{25}\rho^{1/2}-\frac{32}{25}\rho^{3/4}-\frac{32}{25}\rho$
J	Compound Symmetry: Relative efficiency ${r_{\text{cs}}}^*$
2	$2(1-\rho)$
3	$2(1-\rho)$
4	$9(1-\rho)/5$
5	$8(1-\rho)/5$
J	Random effects: Relative efficiency ${r_{\text{mix}}}^\dagger$
2	$2(1-\rho_{cs}-\bar{\rho}_{cs}\rho)$
3	$2(1-\rho_{cs}-\bar{\rho}_{cs}\rho)$
4	$\frac{9}{5}+\frac{9}{10}(\rho_{cs}+\bar{\rho}_{cs}\rho^{1/3})-\frac{27}{25}(\rho_{cs}+\bar{\rho}_{cs}\rho^{2/3})-\frac{81}{50}(\rho_{cs}+\bar{\rho}_{cs}\rho)$
5	$\frac{8}{5}+\frac{32}{25}(\rho_{cs}+\bar{\rho}_{cs}\rho^{1/4})-\frac{8}{25}(\rho_{cs}+\bar{\rho}_{cs}\rho^{1/2})-\frac{32}{25}(\rho_{cs}+\bar{\rho}_{cs}\rho^{3/4})-\frac{32}{25}(\rho_{cs}+\bar{\rho}_{cs}\rho)$

*: For AR1 and compound symmetry covariance, $\rho = \text{corr}(\varepsilon_{i1k}, \varepsilon_{iJk})$

† : For random effects covariance, $\rho = \text{corr}(e_{i1k}, e_{iJk})$, $\rho_{cs} = \sigma^2_\alpha/(\sigma^2_\alpha + \sigma^2_\epsilon)$, and $\bar{\rho}_{cs} = 1 - \rho_{cs}$

Table 2: Correlation required for longitudinal design to be more efficient with various within-subject covariance structures

	AR1 covariance*			
J	Retention 100%	Retention 75%	Retention 50%	Retention 25%
2, 3	$\rho > 0.5$	$\rho > 0.433$	$\rho > 0.353$	$\rho > 0.25$
4	$\rho > 0.573$	$\rho > 0.445$	$\rho > 0.361$	$\rho > 0.226$
5	$\rho > 0.523$	$\rho > 0.458$	$\rho > 0.373$	$\rho > 0.218$
	Compound Symmetry*			
J	Retention 100%	Retention 75%	Retention 50%	Retention 25%
2, 3	$\rho > 0.5$	$\rho > 0.433$	$\rho > 0.353$	$\rho > 0.25$
4	$\rho > 0.444$	$\rho > 0.368$	$\rho > 0.268$	$\rho > 0.103$
5	$\rho > 0.375$	$\rho > 0.287$	$\rho > 0.163$	$\rho > 0$
	Random Effects†			
J	Retention 100%	Retention 75%	Retention 50%	Retention 25%
2, 3	$\rho > 0.5$	$\rho > 0.433$	$\rho > 0.352$	$\rho > 0.248$
4	$\rho > 0.510$	$\rho > 0.443$	$\rho > 0.358$	$\rho > 0.219$
5	$\rho > 0.522$	$\rho > 0.457$	$\rho > 0.370$	$\rho > 0.206$

*: For AR1 and compound symmetry covariance, $\rho = \text{corr}(\varepsilon_{i1k}, \varepsilon_{iJk})$

† : For random effects covariance, $\rho = \text{corr}(e_{i1k}, e_{iJk})$, $\rho_{cs} = \sigma^2_\alpha/(\sigma^2_\alpha + \sigma^2_\epsilon)$, and here we examine scenarios when $\rho = \rho_{cs}$.

Figure 1: Relative efficiency of an endpoint analysis versus a longitudinal analysis (upper panel) and of a change score analysis versus a longitudinal analysis (lower panel) with the AR1 correlation structure

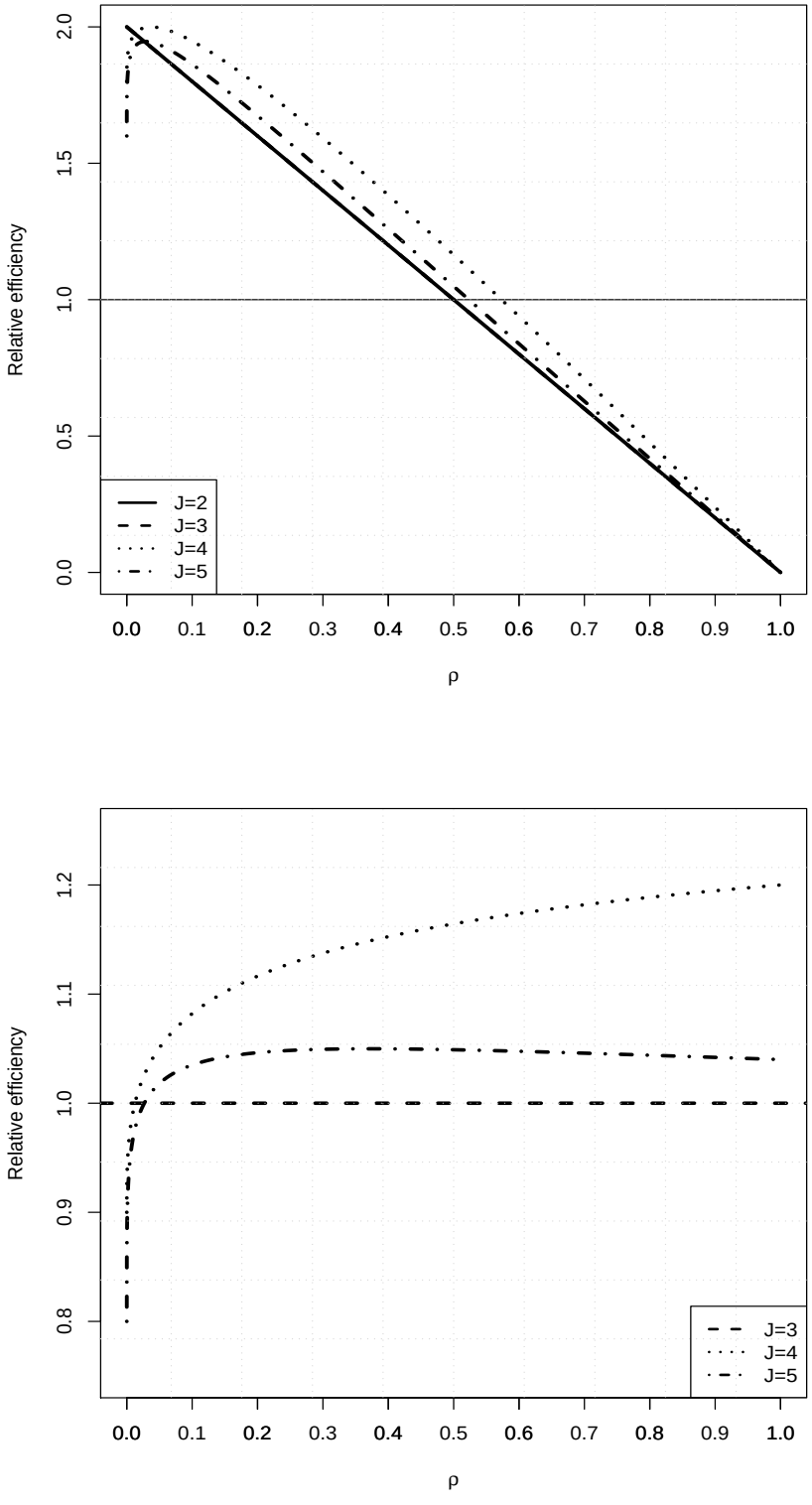

Figure 2: Relative efficiency of an endpoint analysis versus a longitudinal analysis (upper panel) and of a change score analysis versus a longitudinal analysis (lower panel) with the compound symmetry correlation structure

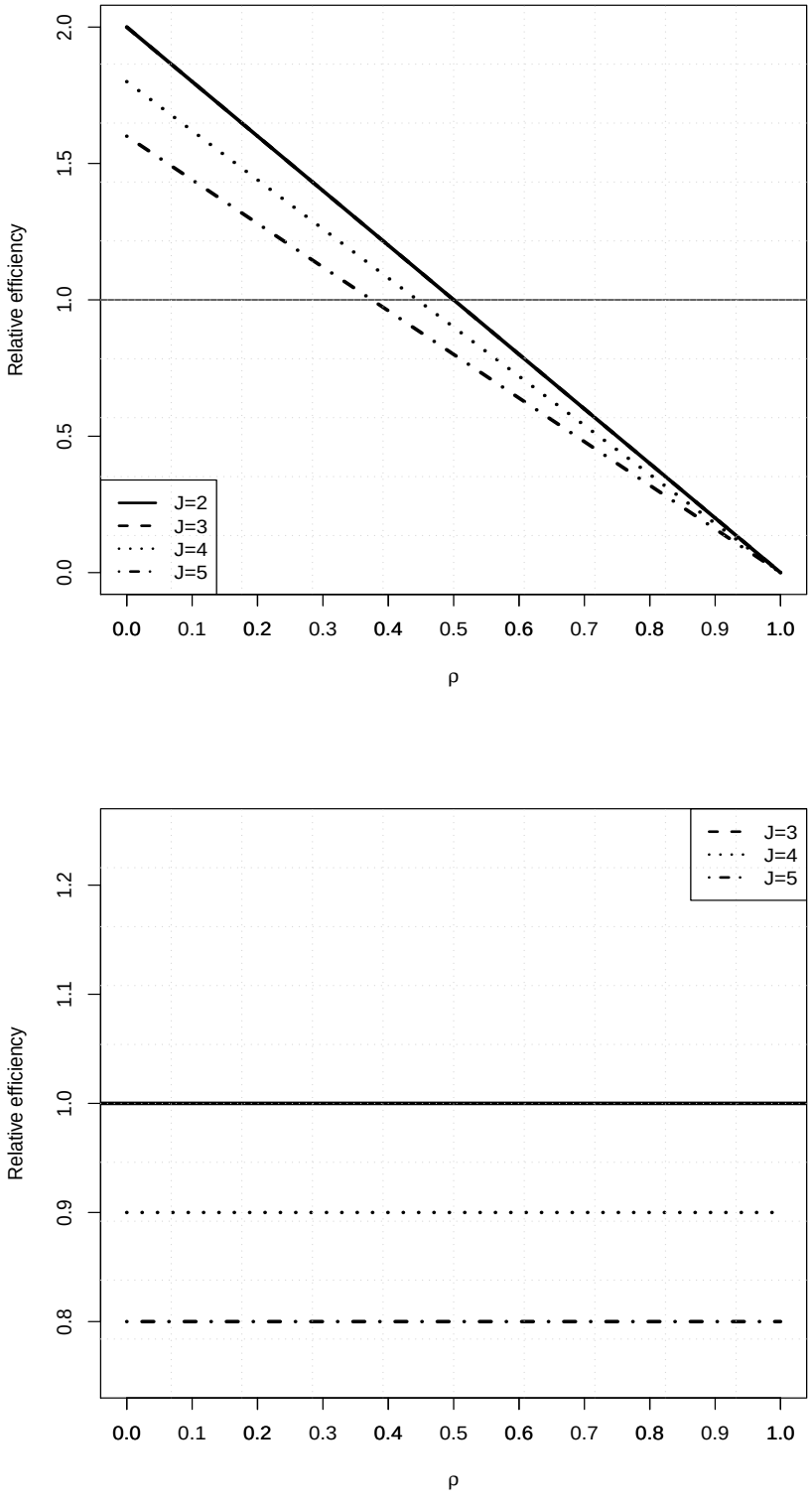

Figure 3: Relative efficiency of an endpoint analysis versus a longitudinal analysis (upper panel) and of a change score analysis versus a longitudinal analysis (lower panel) with the random effects correlation structure