



Published in final edited form as:

Reg Anesth Pain Med. 2012 January ; 37(1): 99–105. doi:10.1097/AAP.0b013e31823ebc74.

Beyond Repeated measures ANOVA: advanced statistical methods for the analysis of longitudinal data in anesthesia research

Yan Ma, PhD^{* **}, Madhu Mazumdar, PhD^{**}, and Stavros G. Memtsoudis, MD, PhD^{***}

^{*}Research Division, Hospital for Special Surgery, Weill Medical College of Cornell University

^{**}Department of Public Health, Weill Medical College of Cornell University

^{***}Department of Anesthesiology, Hospital for Special Surgery, Weill Medical College of Cornell University

Abstract

Background and Objectives—Research in the field of anesthesiology relies heavily on longitudinal designs for answering questions about long-term efficacy and safety of various anesthetic and pain regimens. Yet, anesthesiology research is lagging in the use of advanced statistical methods for analyzing longitudinal data. The goal of this paper is to increase awareness of the advantages of modern statistical methods and promote their use in anesthesia research.

Methods—Here we introduce 2 modern and advanced statistical methods for analyzing longitudinal data: the generalized estimating equations (GEE) and mixed effects models (MEM). These methods are compared to the conventional repeated measures ANOVA (RM-ANOVA) through a clinical example with 2 types of endpoints (continuous and binary). In addition, we compare GEE and MEM to RM-ANOVA through a simulation study with varying sample sizes, varying number of repeated measures, and scenarios with and without missing data.

Results—In the clinical study, the 3 methods are found to be similar in terms of statistical estimation, while the parameter interpretations are somewhat different. The simulation study shows that the methods of GEE and MEM are more efficient in that they are able to achieve higher power with smaller sample size or lower number of repeated measurements in both complete and missing data scenarios.

Conclusions—Based on their advantages over RM-ANOVA, GEE and MEM should be strongly considered for the analysis of longitudinal data. In particular, GEE should be utilized to explore overall average effects, and MEM should be employed when subject-specific effects (in addition to overall average effects) are of primary interest.

Introduction

Longitudinal study (LS) design, involving consecutive measurements on the same individual, has become increasingly popular for examining trends in outcomes over time.

Address correspondence to Dr. Ma: Research Division, Hospital for Special Surgery, 535 East 70th St., New York 10021, Phone: 646 797 8323, Fax: 212 774 7895, may@hss.edu.

Attribute to: Research Division, Hospital for Special Surgery, Department of Public Health, Weill Medical College of Cornell University, and Department of Anesthesiology, Hospital for Special Surgery, Weill Medical College of Cornell University.

Financial disclosure: This study was performed with funds from Clinical TRANSLATIONAL Science Center (CTSC) (NIH UL1-RR024996) (Yan Ma and Madhu Mazumdar) and Center FOR Education and Research in Therapeutics (CERTs) (AHRQ RFA-HS-05-14) (Madhu Mazumdar). No conflicts of interest arise from any part of this study for any of the authors.

Compared to a cross-sectional study, which assesses the outcome at a single point in time, a LS can provide information about changes in both individual and average group outcomes over time. Distinctive features of LS include correlated observations (due to outcome variable measurements at multiple time points), high possibility of missing data (due to the rigorous follow-up needed for each subject), and existence of multiple covariates. If these complexities are properly addressed, a LS has the advantage of being able to answer important and clinically relevant questions with higher precision than a study with simpler design.

LS design has been utilized in many disciplines, including highly influential trials with multi-year follow-up like the Multicenter AIDS Cohort Study (MACS)¹ and the Framingham Heart Study (FHS)². These studies have provided the knowledge base for the natural history and development of treatment for AIDS and cardiovascular diseases. In addition, national funding agencies are particularly interested in studies proposing longitudinal designs. All of these factors have resulted in the growing popularity of longitudinal design and have become a driving force for the development of new statistical methods during the past few decades. The relatively new and more advanced methods of Generalized Estimating Equations (GEE)³ and Mixed Effects Models (MEM)⁴ have started to replace the traditional methods of Repeated Measures ANalysis Of VAriance (RM-ANOVA)⁵ and t-test, as the older methods are not flexible enough to accommodate all of the special features of longitudinal designs.

Longitudinal studies are particularly important in the field of anesthesiology as researchers are often interested in the assessment of efficacy and safety of various anesthetic and pain regimens. In a literature review of the last 2 issues of *Regional Anesthesia and Pain Medicine* in 2010, we found a total of 14 longitudinal studies⁶⁻¹⁹ (74%) out of 19 published studies (16 original articles and 3 ultrasound articles only). Most of these studies were characterized by a small to moderate number of repeated measures and sample sizes (Table 1). Although longitudinal designs were widely adopted, advanced statistical methods were commonly underutilized. The majority of these studies were analyzed using traditional methods (11/14), such as RM-ANOVA (5/11), paired t-test, and Mann-Whitney U test. Only 2 of the 19 studies implemented GEE, and only 1 utilized MEM. Further, missing data (although basically inevitable in LS) was discussed in only 1 of these studies.

Therefore, in this manuscript, we 1) clarify the concept of hypothesis testing for a two-group comparative trial performed in a longitudinal manner, 2) introduce RM-ANOVA, GEE, and MEM in simple terms for analyzing data from a clinical study with 2 types of endpoints (continuous and binary) and compare the results from each method, and 3) contrast the operating characteristics of these 3 methods for varying sample sizes and varying number of repeated measurements with and without missing data. We accomplish the third task by using simulated data to cover a variety of scenarios with known results. Data simulation is a common technique because it is impossible to find real data for each scenario under discussion. However, we guided our simulation parameters with findings from the literature to help mimic real data.

Methods and Materials

In this paper, we consider a two-group comparative (ie, treatment vs. control) LS in both clinical and simulated examples. Although here we use a two-group design, statistical inferences can be readily generalized for LS involving a different number of groups.

Hypothesis Testing for a two-group Comparative LS

The most common research question in two-group comparative LS is whether the treatment group follows the same trend in the mean outcome over time as the control group. This is equivalent to testing the significance of the interaction effect between treatment and time. To prepare for hypothesis testing at the analysis stage, a study needs to set a null hypothesis (H_0) and alternative hypothesis (H_1) prior to start. The null hypothesis, as the name suggests, specifies that there is no difference in outcome between the treatment and control groups at each time point (ie, no interaction effect). Ideally, the alternative hypothesis will specify the smallest clinical effect of the treatment that, if proven, will prompt the treatment to be adopted in clinical care. For simplicity, we assume the mean treatment and control group effects follow linear trends over time (Fig. 1). Therefore, the hypothesis test for an interaction effect can be converted to a comparison of the slopes of the 2 straight lines (S_t : slope of treatment, S_c : slope of control). As illustrated in Figure 1, when there is no interaction effect (H_0), the 2 lines are parallel (ie, $S_t=S_c$); otherwise $S_t \neq S_c$. The difference between S_t and S_c represents the magnitude of an interaction effect.

As shown in Figure 2, there are 2 correct and 2 incorrect actions (types of error) associated with hypothesis testing. Type I error, or false positive, is the probability of rejecting the null hypothesis of equal slopes when, in fact, $S_t=S_c$. Type II error, or false negative, is the probability of accepting the null hypothesis of equal slopes when, in actuality, $S_t \neq S_c$. Power, the complement of the probability of Type II error (= 1-Type II error or false negative), is the probability of detecting a difference between the 2 slopes when, in fact, $S_t \neq S_c$. By convention, to help prevent the investigator from making false claims, the rates of Type I error and Type II error are set at low levels of 5% and 20% (ie, 80% power), respectively. We can ultimately accept or reject the null hypothesis based upon statistical analysis of the collected data.

Repeated Measures ANOVA

RM-ANOVA relates the study outcome variable to a set of covariates (eg, treatment group, time) and compares the mean outcome at multiple time points or between groups. Although RM-ANOVA (one of the earliest proposed methods for analyzing correlated responses) has gained widespread popularity, it has several unattractive features. First, RM-ANOVA requires the outcome variable to be quantitative (ie, a continuous variable) and normally distributed. It also requires the covariates to be discrete (ie, categorical variables). Second, RM-ANOVA requires that the outcome have constant variance across time points as well as constant correlation between any 2 time points (ie, assumption of sphericity). The assumption of constant correlation of repeated measures is often unrealistic in medical research as repeated measures often become less correlated with increasing time from treatment. This kind of violation of the sphericity assumption may cause inflated Type I error²⁰. Third, RM-ANOVA can only handle longitudinal studies in which all subjects have the same numbers of repeated measurements. Specifically, RM-ANOVA excludes those subjects who have missing observations at 1 or more time points (a common occurrence in a LS). Inclusion of only those subjects who have “complete” data for all variables has unfavorable consequences. The group of subjects with “complete” data may not represent a random sample from the target population, thus producing biased results. Further, statistical power is reduced by this artificial attrition in sample size.

Generalized estimating equations (GEE)

The GEE method focuses on average changes in response over time and the impact of covariates on these changes. The method models the mean response as a linear function of covariates of interest via a transformation or link function. To accommodate various types of outcomes that are not necessarily normally distributed, different link functions are employed

for modeling the relationship between outcome and covariates. For example, an identity link function is used for a continuous outcome, a logit link function for a binary outcome, and a log link function for count data.²¹ These transformations can be considered repeated measures analogs of linear regression, logistic regression, and Poisson regression, respectively. In addition, to account for variation in correlation between repeated measures, GEE allows specification of the correlation structure from a wide variety of choices. Popular choices, among others, include the compound symmetry (CS) correlation structure and the autoregressive (AR(1)) correlation structure. The CS correlation structure assumes a common correlation for any pair of responses at different time points, while the AR correlation structure assumes that measurements closer in time have a higher correlation than those that are further apart. GEE also has appealing and robust properties in parameter estimation. Unlike RM-ANOVA, GEE does not require the outcome variable to have a particular distribution. This feature can greatly benefit studies in which data are skewed or the distribution of data is difficult to verify due to a small sample size.

Mixed effects models (MEM)

MEM describes how the response of the individual participant changes over time. It takes into account between-individual heterogeneity by adding random effects to a subset of covariates of interest. These added random effects allow covariate coefficients to vary randomly from 1 individual to another, thereby providing an individual response trajectory over time. The most common MEM in longitudinal studies are those with random effects attached to baseline values or time dependent variables (eg, postoperative day), reflecting heterogeneity among individual responses at baseline (eg, heterogeneous pain scores at baseline), or variation between individual trajectories over time (eg, heterogeneous rates of change in pain). In addition, like GEE, MEM allows specification of the correlation structure between repeated measurements from similar choices, such as the CS and AR(1).

Pattern of Missing data

In the statistical literature, missing completely at random (MCAR) and missing at random (MAR)²² are 2 popular data mechanisms in the context of GEE and MEM. Data are MCAR if the occurrence of missing data is independent of both observed and unobserved outcomes. For example, data missing from a patient who has dropped out of a longitudinal trial because he/she has relocated is considered MCAR. This ‘missingness’ has nothing to do with the treatment effect and its outcome. Alternatively, when missing data is dependent solely upon the observed outcomes, data are considered MAR. For example, when a patient drops out of a trial due to treatment-related adverse effects, any data missing for this patient is classified as MAR. The latter is considered a more serious kind of ‘missingness,’ so special methodological adjustments must be made for data with this issue.

Clinical data example and analysis

A study⁷ conducted in the Department of Anesthesiology at the Hospital for Special Surgery and published in a 2010 issue of RAPM is used for illustration. Thirty-four patients undergoing unilateral total knee arthroplasty (TKA) under tourniquet ischemia were enrolled with 50:50 randomization to either an episode of limb preconditioning before induction of ischemia for surgery or to a control group with no preconditioning. C-reactive protein (CRP) level and postoperative pain scores were 2 outcomes of interest. CRP, a marker of inflammation, was measured at baseline, 6 hours, 12 hours and 24 hours postoperatively. CRP will be used as the continuous outcome in this example. A median pain score for each patient was also obtained for every 6-hour interval postoperatively during the first 48 hours. We convert the pain score to a binary variable for this example by considering a pain indicator to be ‘1’ if a patient’s median pain score at any time point is greater than 0, and ‘0’ in all other cases (implying no pain).

This example provides an illustration of longitudinal data analysis for a continuous outcome (ie, CRP) and a binary outcome (ie, pain). There were also missing CRP and pain scores at various time points for 7 and 26 patients, respectively. For illustrative purposes, only time, treatment group, and the interaction between time and treatment were included in the models. All 3 methods were used to model the continuous outcome (CRP), but only GEE and MEM were used to model the binary outcome (pain) because RM-ANOVA cannot handle non-continuous outcomes. All statistical analyses were performed in SAS version 9.2 (SAS Institute, Cary, NC).

Simulated Data Generation, Analysis, and Reporting

Guided by the literature review, we generated data using 2 sample size settings [low (~8 per group) and moderate (~20 per group)] with 4 repeated measurements. We induced scenarios with complete data and with incomplete data for 20% of the subjects at different time points. We contrasted the operating characteristics of the 3 methods in terms of empirical Type I error and power to detect a significant interaction effect between treatment and time. Mean outcomes for the treatment and control group over time are represented by linear trends (Fig. 1). Data were generated for Type I error analysis by assuming the slopes of the treatment and control were equal (ie, no interaction effect, Fig. 1a), and for power analysis by assuming the 2 slopes were different (ie, interaction effect exists, Fig. 1b), respectively. For illustrative purposes we chose $St=1.55$ and $Sc=1$ in order to have a power of 80% with complete data for a moderate sample size ($n=20$). When $St=1.55$ and $Sc=1$, we mimicked a clinical study, where the outcome of 1 group increased faster than the other by 55% per time unit. To further evaluate the impact of the number of repeated measures (r) and sample size (n) on power, we simulated additional data over a wide range of n (8, 20, 30, 40, 50) and r (4, 6, 8, 10, 12). Data generation and statistical analyses were performed in R (R Foundation for Statistical Computing, Vienna, Austria) and SAS version 9.2 (SAS Institute, Cary, NC), respectively. Statistical significance was set at 0.05.

Results

Clinical Data-Continuous outcome

Shown in Figure 3 are sample mean curves for treatment and control groups as well as subject-specific curves for 3 randomly-selected patients from each group. The subject-specific curves surrounding the sample mean curves display the between-patient variations in baseline CRP and trajectory of CRP level over time. Table 2 reports the findings from all 3 methods in treatment effect, time effect, and interaction effect of treatment and time. While GEE and MEM estimate the magnitude of these effects and test their significances, RM-ANOVA only assesses significance. The reported MEM has random effects attached to baseline CRP and time in order to take into account the between-patient variations, as illustrated in Figure 3. There was no significant treatment effect or treatment by time interaction effect, meaning the mean values of the 2 groups were close and followed similar trends over time. This is akin to the situation illustrated in Figure 1(a), where 2 lines have the same slopes, meaning no interaction effect. The assumption of sphericity in RM-ANOVA is violated ($p<0.001$), implying non-constant variances and correlations over time. The AR(1) correlation structure is specified in GEE and MEM to provide a more realistic modeling strategy.

Clinical Data-Binary outcome

Treatment, time, and the interaction of treatment and time are included as covariates with a binary outcome variable (pain versus no pain) in analysis using GEE and MEM. Random effects are attached to time in MEM to account for between-patient variations in odds of having postoperative pain over time. The AR(1) correlation structure is specified in both

GEE and MEM because probabilities of receiving postoperative pain tend to be more correlated when assessments are closer in time. Table 3 reports the estimated covariate coefficients. The treatment and time variables are significant in both models, but the interaction effect is not significant. Although the estimates of parameters from GEE and MEM are similar, they should be interpreted differently. GEE focuses on group average while MEM targets the individual when outcome is non-continuous. This important distinction will be further illustrated in the Discussion.

Simulation Results

Shown in Table 4 are levels of empirical Type I error and power to detect the significance of an interaction effect between time and treatment group for complete and incomplete simulation data. For a small sample size of $n=8$ per group, with both missing and non-missing data, GEE has the lowest Type I error rates (close to the nominal significance level of 5%). In contrast, RM-ANOVA has a somewhat lower Type I error rate when data are MAR, and MEM has higher Type I error rates in all data situations. When the sample size is increased to $n=20$ per group, the Type I error rates of GEE remain the lowest, while those of MEM are significantly reduced, and those of RM-ANOVA are slightly inflated (especially when data are MAR). In terms of the power analysis, MEM has the highest power when data from a small sample size ($n=8$) are either complete or incomplete. RM-ANOVA has approximately 30% and 50% less power than GEE and MEM, respectively, even when data are complete. When the sample size increases, power increases substantially with all 3 methods. GEE and MEM tend to have similarly high levels of power. RM-ANOVA has lower power in all cases, a disadvantage which is more prominent when data are MCAR and MAR.

Figure 4 depicts the trend of power over a wide range of sample sizes (n) and number of repeated measures (r). When the sample size is small ($n=8$), power analysis is conducted by increasing r ($r=4, 6, 8, 10, 12$). Power is improved in all 3 methods when r increases (Fig 4(a)). MEM always achieves the highest power, while RM-ANOVA receives the lowest. On the other hand, when the number of repeated measurements is small ($r=4$), power is improved by increasing the sample size (Fig. 4(b)). The power analysis is based on complete data, but results should be similar when there are missing data (as illustrated in Table 4).

Discussion and Recommendations

Through the introduction and application of the 3 methods, we were able to show that GEE and MEM are more flexible than RM-ANOVA for handling different types of outcomes (ie, continuous and non-continuous) and modeling a wide variety of correlation patterns between repeated measures. There are other important distinctions among the 3 approaches.

Different ways of handling missing data

Missing data are practically inevitable in LS, thus leading to unbalanced designs. If a subject has any missing values, RM-ANOVA will exclude the individual from the analysis entirely. In contrast, GEE and MEM take all available data into account in an unbalanced design, leading to more efficient effect estimates (eg, treatment effect). In particular, when data are MCAR, individuals with missing data are considered a random subset of the sample. Thus, when data are MCAR, statistical inferences based on either GEE or MEM are valid. But weighted GEE is recommended when missing data are MAR, as non-weighted GEE may provide biased parameter estimates.²³ In comparison, the likelihood-based MEM can generate valid inferences even when data are MAR.²⁴ In the simulation study, we found RM-ANOVA to be associated with the lowest power for both complete and incomplete data. For example, when the sample size was moderate (ie, $n=20$) and missing data were MCAR,

RM-ANOVA had approximately 19% and 23% less power than GEE and MEM, respectively. When the sample size was small (i.e., $n=8$), the problem was exacerbated, and RM-ANOVA might have up to 50% less power than MEM.

The clinical study data contained missing pain scores due to sleep or other activity at the time of data collection; missing CRP occurred due to inadequate blood sample quality (ie, insufficient quantity, clotting of sample) and consequent inability to process samples appropriately. Therefore, missing outcomes occurred randomly and can be considered MCAR. Thus, GEE and MEM are appropriate methods for handling the missing data in this study. In particular, there are 6 patients with 7 missing CRPs in total across time points. RM-ANOVA automatically deleted these 6 patients, resulting in a loss of 24 observations (ie, 6 patients \times 4 repeated measures per patient), among which 17 observations are not missing. In contrast, GEE and MEM utilized all available data. With more data being incorporated into the analysis, GEE and MEM are expected to provide more accurate statistical inferences than RM-ANOVA, as illustrated in the simulation study.

Different interpretations

Besides the noted distinctions, the 3 methods were used to answer different research questions. With data collected from every individual at every time point, we wanted to make statistical inferences regarding the change in mean response over time (a population-averaged inference) or the individual trajectory over time (a subject-specific inference). Both RM-ANOVA and GEE measured population-averaged effects of covariates of interest, such as the average effect of an analgesic regimen on a population's mean pain score over time. MEM, by contrast, could identify subject-specific effects of covariates on the changes in the response over time. Therefore, MEM would be useful in a setting where an intervention is likely to affect some individuals differently than others as compared to RM-ANOVA and GEE, which do not take individual response into account in their interpretations. An example would be a study that evaluates complications of a particular regional anesthetic technique in sub populations that may differ in results from those of the average patient, ie, outliers. In this scenario, for instance, patients with diabetes, who have higher rates of preexisting neuropathy, may have very different results than the average non-diabetic patient. MEM, in this setting, could allow for a more nuanced analysis of individuals of this sub-population, such as predicting individual risk of complications.

In our clinical data example for the continuous outcome, CRP, GEE and MEM produced similar parameter estimates, and both had population-averaged interpretations. For example, average increases of 7.19 (15.72–8.63) pg/mL and 7.89 (15.21–7.83) pg/mL CRP at 6 hours postoperatively in the treatment group compared to the control group, respectively, were identified using GEE and MEM. By adding the predicted random intercept and random slope from MEM to the population mean coefficient estimates in Table 2 for a specific individual, the predicted trajectory of that individual could be obtained. For the binary outcome of pain, MEM had only a subject-specific interpretation. Therefore, in this study, distinct interpretations were derived from the GEE and MEM. For example, the significant treatment effect in GEE implied that patients with preconditioning were 74% (odds ratio $(OR)=e^{(-1.4+0.07)}=0.26$) less likely on average to experience postoperative pain at 6 hours compared to those in the control group. In contrast, the treatment effect in MEM does not have population-averaged interpretation. The significant treatment effect in MEM implies that a patient's odds of experiencing postoperative pain at 6 hours decreased by 79% when treated with preconditioning compared to without preconditioning ($OR=e^{(-1.62+0.07)}=0.21$). Thus, the answer to the question "How beneficial is preconditioning?" will depend on whether the research interest is the impact on the study population, or on a randomly-selected individual from the population. Furthermore, the coefficient estimates from MEM have greater magnitudes than those from GEE. This confirms the early findings in the

literature that the greater the underlying variation among individuals, the greater the discrepancy between the coefficient estimates from the 2 approaches.

Statistical power in LS---some recommendations through simulation study

In addition, through the simulation study we found that sample size and the number of repeated measures are the key parameters in determining power. Power can be improved by increasing either sample size or the number of repeated measures. MEM and GEE are more efficient than RM-ANOVA as smaller sample sizes or numbers of repeated measures are required to achieve 80% power (Fig. 4). In particular, when sample size is small ($n=8$) and the number of repeated measures is 6 or more, the power of MEM or GEE is around or above 80%. However, RM-ANOVA can achieve 80% or higher power only when there are 8 or more repeated measures. When the number of repeated measures is 4 and sample size is at least 20, the power of MEM or GEE is around or above 80%. In contrast, RM-ANOVA would need an approximate sample size of 30 to reach 80% power. Hence, when statistical power is an issue at the design stage of a LS, researchers may choose to either increase the number of repeated measures or the sample size, whichever is feasible for the study. Furthermore, caution should be applied for interpretation of negative results from all 3 methods when power is low due to small sample size and number of repeated measures. Non-significant parameter estimates may become significant if power is increased.

In conclusion, GEE helps estimate the average change per group while MEM highlights subject-specific inference. These advanced statistical methods should be highly recommended since they are readily available in all major statistical software. It is essential to report the frequency and pattern of missing data from a study utilizing longitudinal design.

Acknowledgments

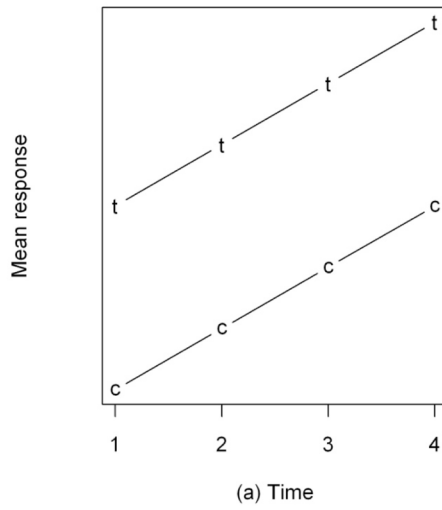
We sincerely thank Kara Fields at Columbia University for her valuable editorial assistance.

References

1. Kaslow RA, Ostrow DG, Detels R, Phair JP, Polk BF, Rinaldo CR Jr. The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. *Am J Epidemiol.* 1987; 126:310–318. [PubMed: 3300281]
2. Dawber TR, Kannel WB, Lyell LP. An approach to longitudinal studies in a community: the Framingham Study. *Ann N Y Acad Sci.* 1963; 107:539–556. [PubMed: 14025561]
3. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics.* 1986; 42:121–130. [PubMed: 3719049]
4. Demidenko, E. *Mixed models: theory and applications.* Wiley; Hoboken, New Jersey: 2004.
5. Girder, ER. *Sage university paper series on quantitative applications in the social sciences, 07–084.* Newbury Park, CA: Sage; 1992. ANOVA: repeated measures.
6. Hashimoto T, Iwasaki T, Takahashi Y, Kimura Y, Nagao M, Yamaguchi S, Hamaguchi S, Kitajima T. Comparison of bupivacaine, ropivacaine, and levobupivacaine in an equal dose and concentration for sympathetic block in dogs. *Reg Anesth Pain Med.* 2010; 35:409–411. [PubMed: 20814280]
7. Memtsoudis SG, Della Valle AG, Jules-Elyse K, Poultsides L, Reid S, Starcher B, Ma Y, Sculco TP. Perioperative inflammatory response in total knee arthroplasty patients : impact of limb preconditioning. *Reg Anesth Pain Med.* 2010; 35:412–416. [PubMed: 20830870]
8. Memtsoudis SG, Salvati EA, Go G, Ma Y, Sharrock NE. Perioperative pulmonary circulatory changes during bilateral total hip arthroplasty under regional anesthesia. *Reg Anesth Pain Med.* 2010; 35:417–421. [PubMed: 20814281]

9. Parrington SJ, O'Donnell D, Chan VWS, Brown-Shreves D, Subramanyam R, Qu M, Brull R. Dexamethasone added to mepivacaine prolongs the duration of analgesia after supraclavicular brachial plexus blockade. *Reg Anesth Pain Med.* 2010; 35:422–426. [PubMed: 20814282]
10. Brummett CM, Amodeo FS, Janda AM, Padda AK, Lydic R. Perineural dexmedetomidine provides an increased duration of analgesia to a thermal stimulus when compared with a systemic control in a rat sciatic nerve block. *Reg Anesth Pain Med.* 2010; 35:427–431. [PubMed: 20814283]
11. Chiono J, Bernard N, Bringuier S, Biboulet P, Choquet O, Morau D, Capdevila X. The ultrasound-guided transverses abdominis plane block for anterior iliac crest bone graft postoperative pain relief: a prospective descriptive study. *Reg Anesth Pain Med.* 2010; 35:520–524. [PubMed: 20975466]
12. Otani T, Shinozaki M, Takahashi Y, Ikeda T, Kimura Y, Nagao M, Yamaguchi S, Kitajima T. Comparison of 2% mepivacaine, clipping, and radiofrequency thermocoagulation for duration and magnitude of action in peripheral arterial blood flow induced by sympathetic block in anesthetized dogs. *Reg Anesth Pain Med.* 2010; 35:525–528. [PubMed: 20975467]
13. Renes SH, van Geffen GJ, Rettig HC, Gielen MJ, Scheffer GJ. Minimum effective volume of local anesthetic for shoulder analgesia by ultrasound-guided block at root c7 with assessment of pulmonary function. *Reg Anesth Pain Med.* 2010; 35:529–534. [PubMed: 20975468]
14. Buvanendran A, Kroin JS, Kari M, Tuman KJ. Can a single dose of 300 mg of pregabalin reach acute antihyperalgesic levels in the central nervous system? *Reg Anesth Pain Med.* 2010; 35:535–538. [PubMed: 20975469]
15. Tran De QH, Dugani S, Finlayson RJ. A randomized comparison between ultrasound-guided and landmark-based superficial cervical plexus block. *Reg Anesth Pain Med.* 2010; 35:539–543. [PubMed: 20975470]
16. Morau D, Levy F, Bringuier S, Biboulet P, Choquet O, Kassim M, Bernard N, Capdevila X. Ultrasound-guided evaluation of the local anesthetic spread parameters required for a rapid surgical popliteal sciatic nerve block. *Reg Anesth Pain Med.* 2010; 35:559–564. [PubMed: 20975475]
17. Deegan CA, Murray D, Doran P, Moriarty DC, Sessler DI, Mascha E, Kavanagh BP, Buggy DJ. Anesthetic technique and the cytokine and matrix metalloproteinase response to primary breast cancer surgery. *Reg Anesth Pain Med.* 2010; 35:490–495. [PubMed: 20975461]
18. Plancarte R, Guajardo-Rosas J, Reyes-Chiquete D, Chejne-Gomez F, Plancarte A, Gonzalez-Buendia NI, Cerezo-Camacho O, Lee A, Medina-Santillan R. Management of chronic upper abdominal pain in cancer transdiscal blockade of the splanchnic nerves. *Reg Anesth Pain Med.* 2010; 35:500–506. [PubMed: 20975463]
19. Wallace MS, Kopecky EA, Ma T, Brothy F, Campbell JC. Evaluation of the depth and duration of anesthesia from heated lidocaine/tetracaine (synera) patches compared with placebo patches applied to health adult volunteers. *Reg Anesth Pain Med.* 2010; 35:507–513. [PubMed: 20975464]
20. Huynh H, Feldt LS. Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *J Educ Statistics.* 1976; 1:69–82.
21. McCullagh, P.; Nelder, JA. *Generalized Linear Models.* 2. Chapman & Hall/CRC; London: 1989.
22. Little, RJA.; Rubin, DB. *Statistical analysis with missing data.* Wiley; Hoboken, New Jersey: 2002.
23. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc.* 1995; 90:106–121.
24. DeSouza, Cm; Legedza, AT.; Sankoh, AJ. An overview of practical approaches for handling missing data in clinical trials. *J Biopharm Stat.* 2009; 19:1055–1073. [PubMed: 20183464]

No interaction effect($St=Sc=1$)



Interaction effect($St=1.55, Sc=1$)

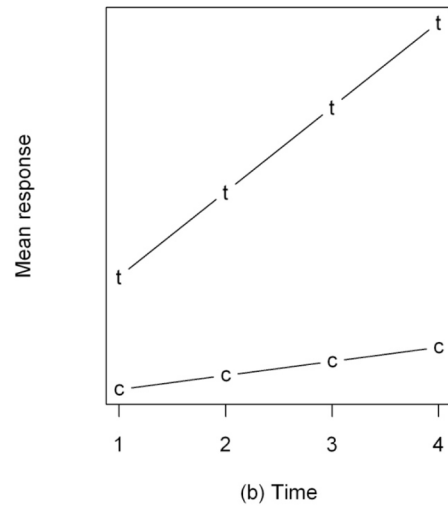


Figure 1. Underlying Setting for Generation of Simulated Data (St : slope of the treatment line; Sc : slope of the control line).

Decisions based on data analysis	The truth in the population	
	H0 True	H0 False
Reject H0 (accept H1)	Type I error (false positive)	Correct action
Fail to reject H0 (accept H0)	Correct action	Type II error (false negative)

Figure 2.

Relations between truth/falseness of the null hypothesis and outcomes of the test (H0:null hypothesis; H1:alternative hypothesis).

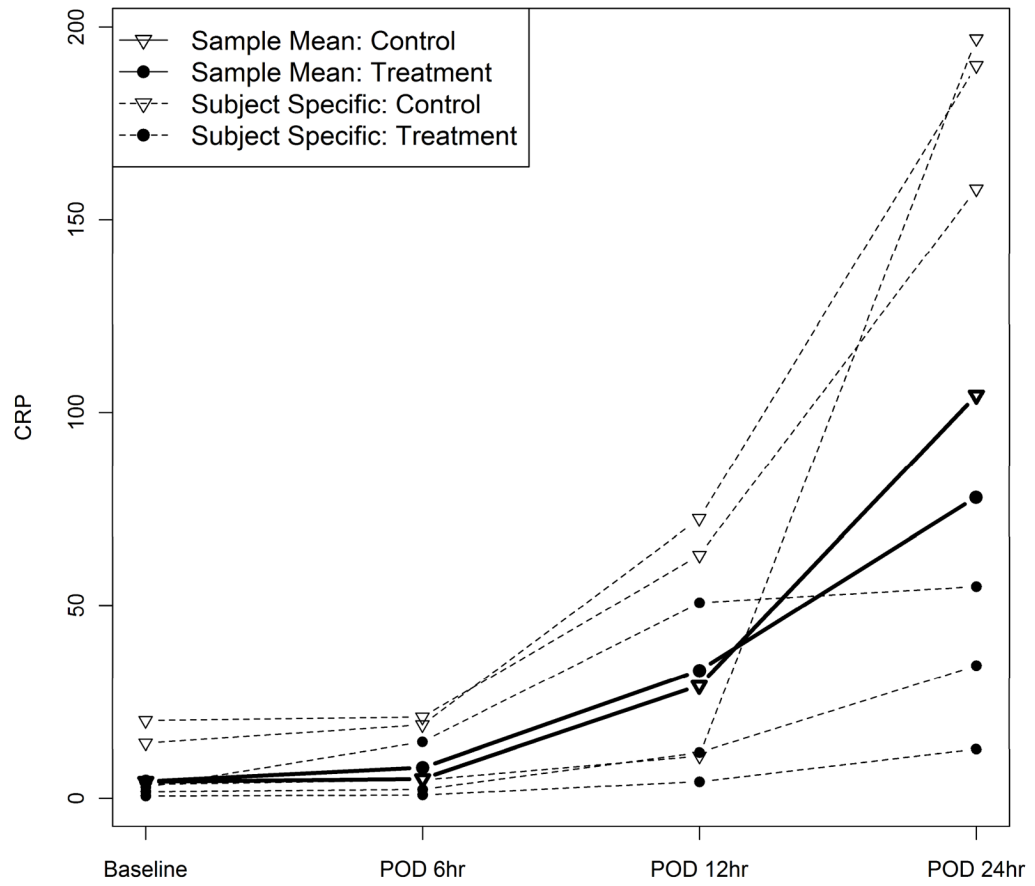


Figure 3. Sample mean curves of CRP with randomly selected individual profiles.

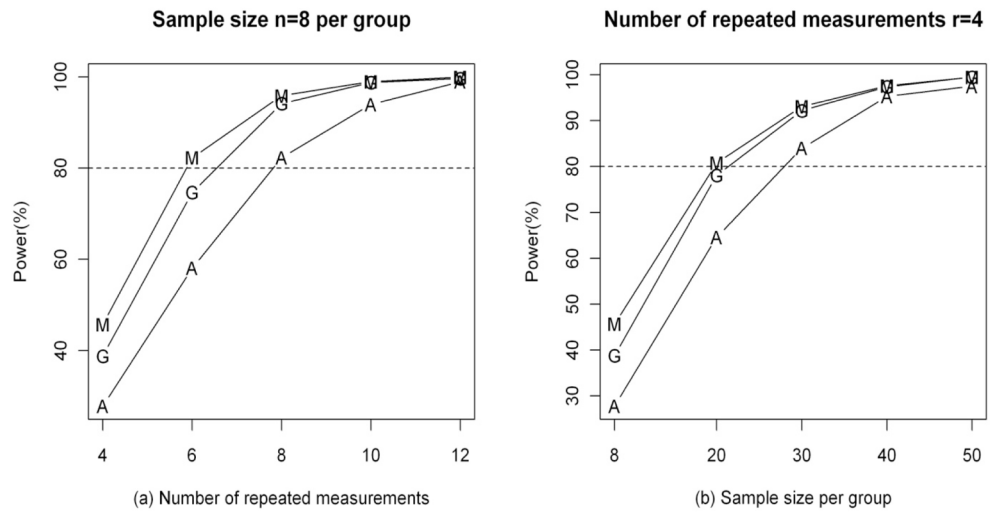


Figure 4. Simulated power analysis under complete data (A=RM-ANOVA, G=GEE, M=MEM).

Table 1Literature review of two issues of *Regional Anesthesia and Pain Medicine*(2010)

First Author	Number of Repeated measures	Number of Groups	Sample Size per Group	Statistical Method Utilized
Hashimoto ⁶	14	3	8/8/8	RM-ANOVA *
Memsoudis ⁷	4	2	17/17	GEE *
Memsoudis ⁸	6	1	24	GEE
Parrington ⁹	5	2	21/24	other **
Brummett ¹⁰	11	5	54 (in total)	RM-ANOVA
Chiono ¹¹	6	1	33	other
Otani ¹²	14	3	8/8/8	RM-ANOVA
Renes ¹³	9	1	20	RM-ANOVA
Buvanendran ¹⁴	4	1	9	other
Tran ¹⁵	3	2	20/20	other
Morau ¹⁶	4	1	100	other
Deegan ¹⁷	4	2	15/17	other
Plancarte ¹⁸	5	1	109	RM-ANOVA
Wallace ¹⁹	5	2	12/12	MEM *

* RM-ANOVA: repeated measures ANOVA; GEE: generalized estimating equations; MEM: mixed effects model

** Other statistical methods include paired t-test, Mann-Whitney U test, etc.

Table 2Estimated effects of treatment, time, and treatment \times time on the continuous outcome CRP

RM-ANOVA			
Effect	Treatment	Time	Treatment \times Time
P-value	0.23	<0.001	0.16
GEE			
Effect	Treatment	Time	Treatment \times Time
Parameter Estimate	15.72	33.1	-8.53
Standard error	11.46	4.92	7.51
P-value	0.17	<0.001	0.25
MEM			
Effect	Treatment	Time	Treatment \times Time
Parameter Estimate	15.21	33.1	-7.83
Standard error	13.42	4.19	5.94
P-value	0.26	<0.001	0.19

Table 3

Estimated effects of treatment, time, and treatment \times time on the binary outcome Pain

Effect\Method	GEE			MEM		
	Parameter Estimate	Standard error	P-value	Estimate	Standard error	P-value
Treatment	-1.4	0.7	0.04	-1.62	0.74	0.03
Time	0.26	0.1	0.01	0.32	0.1	0.003
Treatment \times Time	0.07	0.14	0.63	0.07	0.15	0.65

Table 4

Simulation Study Results: Empirical Type I error and power for detecting significance of the Treatment and Time interaction effect

Sample size	Missing data mechanism	RM-ANOVA	GEE	MEM
Type I error (%)				
N=8	Complete	4.7	4.9	8.1
	MCAR	5	5.2	7.7
	MAR	4.4	4.9	8.2
N=20	Complete	5.5	4.9	5.1
	MCAR	5.1	4.9	5.8
	MAR	6.1	4.8	5.4
Power (%)				
N=8	Complete	27.8	38.8	45.7
	MCAR	23.7	31.7	41.1
	MAR	23.7	33.5	41.7
N=20	Complete	64.6	78.1	80.8
	MCAR	56.5	69.4	73.5
	MAR	58.3	71.5	74.4

* Type I error of 5% and Power of 80% is optimal