Statistical Primer: Performing repeated measures analysis

Short title: Performing repeated measures analysis

Graeme L Hickey[1]*
Mostafa M Mokhles[2]
David J Chambers[3]
Ruwanthi Kolamunnage-Dona[1]

[1] University of Liverpool, Department of Biostatistics, Institute of Translational Medicine, Waterhouse Building, 1-5 Brownlow Street, Liverpool, L69 3GL, UK

[2] Erasmus University Medical Centre, Department of Cardiothoracic Surgery, Rotterdam, The Netherlands

[3] King's College London, Cardiac Surgical Research, The Rayne Institute, Lambeth Wing, Guy's and St Thomas' NHS Foundation Trust, St Thomas' Hospital, London SE1 7EH, UK

Meeting presentation: this was presented at the European Association for Cardio-Thoracic Surgery Annual Meeting, Vienna, Austria, 7-11 October 2017

*** Address for Correspondence**
Department of Biostatistics, Institute of Translational Medicine
University of Liverpool, Waterhouse Building
1-5 Brownlow Street
Liverpool, L69 3GL
United Kingdom
Tel: +44 (0)151 794 9737
Email: graeme.hickey@liverpool.ac.uk

Word count (article text): 2034
Word count (abstract): 143
Number of figures/tables: 4

SUMMARY

Purpose: Longitudinal data arises when repeated measurements are taken on the same individuals over time. Inference about between group differences of within-subject change is usually of interest. This statistical primer for cardiothoracic and vascular surgeons aims to provide a short and practical introduction of biostatistical methods on how to analyse repeated measures data.

Methods: Several methodological approaches for analysing repeated measures will be introduced, ranging from simple approaches to advanced regression modelling. Design considerations of studies involving repeated measures are discussed and the methods illustrated with a dataset measuring coronary sinus potassium in dogs after occlusion.

Conclusion: Cardiothoracic and vascular surgeons should be aware of the myriad of approaches available to them for analysing repeated measures data, including the relative merits and disadvantages of each. It is important to present effective graphical displays of the data, and to avoid arbitrary cross-sectional statistical comparisons.

Key words: statistics; repeated measurements; serial measurements; longitudinal data

INTRODUCTION

Repeated measures data—also known as longitudinal data and serial measures data—are routinely analysed in many studies [1]. The data can be collected both prospectively and retrospectively, allowing for changes over time and its variability within individuals to be distinguished; for example, echocardiographic measurements recorded at different follow-up times after allograft implantation, or Interleukin-6 measured in rats at pre-specified times following cardiopulmonary bypass. The guidelines for reporting mortality and morbidity after cardiac valve interventions also propose the use of longitudinal data analysis for repeated measurement data in patient undergoing cardiovascular surgery [2].

The focus of this Statistical Primer will be on measurements repeatedly recorded over time, although repeated measures can occur in other circumstances, for example when the conditions are changed (e.g. treatment) and the same patients are measured under each experimental condition. Unlike measurements taken on different patients, repeated measures data, however, are not independent. In other words, repeated observations on the same individual will be more similar to each other than to observations on other individuals. This necessitates statistical methodology that can account for this dependency.

DESIGN CONSIDERATIONS

Balanced versus unbalanced data

When subjects are measured at a fixed number of time points that are common to all subjects, then the data are said to be balanced. For example, rats might be tested at times 0, 2-hours, 6-hours, 12-hours, and 24-hours. In some designed studies, these measurements may be *mistimed*, e.g. in human studies where patients are delayed returning to clinic for scheduled follow-up appointments. In some observational studies, i.e. naturalistic cohort studies, measurement times will often vary between subjects and can vary substantially in the number of measurements recorded. Moreover, the patients may have different durations of follow-up observation for various reasons, and may be censored due to terminal events. This would be classed as unbalanced data, and precludes the use of certain statistical methodologies. For balanced and unbalanced measurements, the datasets are often stored in so-called ‘wide format’ (**Table S1a**) and ‘long format’ (**Table S1b**), respectively.

Missing data

Missing data are not uncommon in longitudinal outcome studies. For example, if a patient fails to attend a scheduled appointment, then measurements cannot be taken, and the observation is deemed missing or incomplete. Approaches to handling missing data include complete-case analysis, i.e. deleting patients with one or more missing measurement values; last observation carried forward (LOCF) or interpolation methods; and other imputation techniques. Assumptions about the mechanism leading to missing data dictates the appropriateness of different techniques; however, in general it is widely accepted that simple techniques such as complete-case analysis and LCOF lead to serious bias, and therefore should be avoided. Alternative methods are discussed elsewhere [3].

METHODOLOGY

Two-stage methods

For balanced data, the comparison of treatments might be done by performing separate statistical tests at each time point (**Figure 2A**). However, this approach is inappropriate as it often fails to address relevant research questions and is subject to statistical deficiencies such as ignoring that observations on a given subject are likely to be correlated, and multiple testing [4]. Additionally, the accompanying presentation is frequently inadequate [5], as illustrated in the example shown in **Figure 2A**. One alternative approach is to *reduce* the data for each subject to a *single* meaningful statistic, which are then analysed using standard methods for independent groups, e.g. the independent samples *t*-test [4]. The choice of statistic will depend on the data and the study question, in particular whether the data display a growth-like pattern or a peaked-like pattern; see **Table S2** for examples. Even when not used for the primary analysis, such reduced data summary statistics can be useful, yet it must still be recognised that there might be some information loss with this approach.

Repeated measures analysis of variance (RM-ANOVA)

RM-ANOVA can only be applied for balanced data [6]. When there is also a between group variable (e.g. treatment) the standard RM-ANOVA decomposes the total variation into (i) between subject variation due to treatment effect; (ii) time effect; (iii) time-and-treatment effect; and (iv) the residual error variation. This can be leveraged to test different hypotheses, respectively: (a) an overall treatment effect; (b) differences in outcomes over

time; (c) a different effect of treatment over time. The latter derives from the interaction between time and treatment, which if zero would imply effects are parallel through all time points. In addition to the usual assumption imposed on ANOVA, RM-ANOVA depends on the assumption of sphericity. Effectively, this can be considered as being equivalent to equal variability of measurements at each time (i.e. homogeneity) and equal correlations between any pair of time points (e.g. $\mathrm{corr}\left(y_{\mathrm{time}_1}, y_{\mathrm{time}_2}\right) \approx \cdots \approx \mathrm{corr}\left(y_{\mathrm{time}_1}, y_{\mathrm{time}_3}\right)$ for measurements y recorded at times 1, 2, 3, …). This assumption is restrictive for longitudinal data, since measurements taken closely together are often more correlated than those taken at larger time intervals [7]. Violation of this assumption typically results in an inflated type I error rate and can bias the interaction effect [7]. If used, it is essential that this assumption is checked and reported. Typically, this is achieved through Mauchly's epsilon test; however, this test is known to have low power. When sphericity is violated, there are several corrections to the degrees of freedom of the *F*-test that can be used [8], including Greenhouse-Geisser and Huynh-Feldt methods.

Linear mixed models (LMMs)

Linear mixed models are extensions of more conventional linear models. Let Y_{ij} denote the observed outcome measured on subject i ($i = 1, \ldots, n$) at time t_{ij} ($j = 1, \ldots, n_i$), where n_i is the number of measurements for subject i. By pooling the data, one can fit a linear regression model

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij},$$

where ε_{ij} is a measurement error term (or residual), which allows for the outcome to randomly vary above or below the mean value for each time point. Here, β_1 represents the population slope (**Figure 1A**, black line): the constant effect on the outcome corresponding to a one-unit increase in time. LMMs can also be fitted to unbalanced datasets with irregularly spaced time points (**Figure 1B**), hence each measurement time (t_{ij}) being allowed to be different between subjects in model above. Linear mixed models are predicated on the idea that each subject has their own mean response profile which deviates randomly from the average (overall) trajectory [9]. That is, for each subject i, we extend the model above by including a random intercept b_{0i} and a random slope b_{1i}:

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) t_{ij} + \varepsilon_{ij},$$

where (b_{0i}, b_{1i}) are called subject-specific random effects, and assumed to follow a zero-mean multivariate normal distribution and be correlated. An intuitive graphical representation of this is shown in **Figure 1A**. Here, β_0 and β_1, averaged across all subjects, have the same interpretation, i.e. fixed population-level intercept and slope effects, as for the simple linear regression model. The combination of fixed and random effects is why we refer to this model as a *mixed effects* model, which are also sometimes referred to as multi-level models, random-effects models, random growth-curve models, etc. As well as allowing for subject-specific trajectories, the random effects also ensures that observations within-subjects are more correlated than observations between-subjects, with the case presented here allowing for heterogeneity over time. In the above we assumed time was measured continuously and linearly; however, we might relax this assumption by treating time as measured categorically (providing the data are balanced) or through spline functions, which allow for smooth regression curves that capture nonlinearity [10]. In such cases, we can include additional higher-order random effects; the linear model was presented here for purposes of demonstration. LMMs can also include other adjustment covariates, including time-varying covariates. In particular, one might want to adjust for the baseline measurement of Y rather than treat it as an outcome at the baseline time point, i.e. before treatment intervention [11].

EXAMPLE

As an example, we consider data from Grizzle and Allen [12], who describe a laboratory experiment that collected serial measurements of coronary sinus potassium (CSP) (mEq/L) from four groups of dogs. The groups were:

- Control group: *N*=9 untreated dogs with coronary occlusion.
- ECD (3-weeks) group: *N*=10 dogs given extrinsic cardiac denervation (ECD) 3-weeks prior to coronary occlusion
- ECD (0-weeks) group: *N*=8 dogs treated similarly to above, but given ECD immediately prior to coronary occlusion.
- Sympathectomy group: *N*=9 dogs treated with bilateral thoracic sympathectomy and stellectomy three weeks prior to coronary occlusion.

The response variable was recorded at times 1, 3, 5, 7, 9, 11, and 13 minutes. Before we analyse the data, we inspect the data graphically (**Figure 2B**), where we observe a growth-like trend and substantial between-subject heterogeneity.

If the primary scientific objective was to describe changes in CSP over the 12-minute follow-up period and determine whether the pattern of change differed between groups, then we could fit a linear mixed model including treatment effect and time as a continuous covariate with an interaction term to capture non-parallel growth trends. Despite **Figure 2B** indicating some non-linearity towards the end of the study follow-up, we note that we've made a strong assumption of linearity in this example. Fitting this model (**Table 2**) indicates that there is a significant increase in CSP during follow-up in the control group (i.e. a significant effect for time; 0.08 [95% CI: 0.05 to 0.12]), and no discernible difference from this trend in group ECD (0-weeks) (i.e. non-significant interaction term with time; -0.02 [95% CI: -0.08 to 0.03]). The ECD (3-weeks) group interaction term is significant (P<0.001), and despite not reaching significance, there was a tendency for CSP to be reduced over time in sympathectomy group (-0.05; 95% CI: -0.10 to 0.00). Moreover, both terms are negative, which is consistent with **Figure 2B** where the time course for these two groups are relatively flat. We could formally test this using appropriate contrasts. One could also perform *post hoc* tests to establish treatment effect differences at each measurement time (**Figure 2A**), but one would need to correct for multiple comparisons (not implemented here). Neither group admitted a significant main treatment effect relative to the control group. Code to fit this model using the R statistical software package are shown in the **Appendix**.

Since the data are consistent with a linear growth-like pattern, one might consider comparing a summary statistic approach. For example, a comparison of the slopes (see **Table S2**) would reveal whether there was a significant difference in the rate of change in CSP between groups. A Kruskal-Wallis test applied to the 4-groups of slopes suggests a significant difference (**Table 2**, **Figure 2C**), with the median slopes (first, third quartiles) being 0.098 (0.086, 0.104), -0.003 (-0.012, -0.002), 0.054 (0.024, 0.125), and -0.009 (-0.021 to 0.089) in the control, ECD (3-weeks), ECD (0-weeks), and sympathectomy groups, respectively.

DISCUSSION

Despite RM-ANOVA being a common choice for analysing repeated measures in the EJCTS and ICVTS, there are many alternative approaches. Linear mixed models represent the most sophisticated of the models discussed, and are more amenable to real-world clinical data as opposed to highly controlled experimental study designs. Hence, there have been calls for some time to abandon less versatile methods [7]. The integration of these model fitting methods into routine statistical software therefore removes a major barrier to applied researchers. Moreover, one can extend mixed models to incorporate more flexible correlation structures [13], non-continuous outcomes (e.g. binary), and non-linear outcomes [14], In some cases, there might be multivariate longitudinal data (multiple repeated measures outcomes), which may even be correlated with a time-to-event outcome, giving rise to so-called *joint models* [9,15]. On the other hand, two-stage approaches offer a simpler—both mathematically and intuitively—approach that can provide insight into data profiles and complement more rigorous modelling approaches. We only addressed a subset of the methodological tools available. Other such methods have not been discussed here, including generalised estimating equations, MANOVA [7], generalised least squares [10], and empirical Bayes [8].

Despite repeated measures data being routinely collected at follow-up, particularly in long-term observational studies, the situation of only analysing baseline (preoperative) and a single postoperative value—typically the last follow-up measurement—remains commonplace in the EJCTS and ICVTS, even though this may not be the most appropriate method. Whatever the choice of methodology employed, it is essential that the data, study design, methods, supporting assumptions, and any post hoc analyses are well described and justified to facilitate reproducibility, to provide opportunity for readers to critique the analysis [16], and to avoid misinterpretation due to overlapping terminology [8]. Graphs are a highly effective way of summarising and presenting repeated measures data; however, it is essential that they are presented on common axes scales, appropriately summarised and described (e.g. defining any error bars) [4]. Nonetheless, figures such as those shown in **Figure 2A** should be avoided. It is important to consider distributional assumptions (e.g. normality in the RM-ANOVA) or that the growth-curve is approximately linear if calculating it as a summary measure. When these assumptions are violated, transformations or alternative models might be considered. In addition, we recommend more thought is given to sample size determination during study design [17].

DECLARATIONS

Conflicts of interest: none to declare.

Data availability: the laboratory experiment data is provided in Grizzle and Allen [12], and downloaded from supplementary data files of Davis [18] at http://www.springer.com/gb/book/9780387953700 [accessed 5th August 2017].

Acknowledgements: GLH is funded by the Medical Research Council (MRC), grant number MR/M013227/1.

REFERENCES

1. Fitzmaurice GM, Ravichandran C. A primer in longitudinal data analysis. Circulation. 2008;118: 2005–2010.
2. Akins CW, Miller DC, Turina MI, Kouchoukos NT, Blackstone EH, Grunkemeier GL, et al. Guidelines for reporting mortality and morbidity after cardiac valve interventions. Eur J Cardio-Thoracic Surg. 2008;33: 523–528.
3. Spratt M, Carpenter JR, Sterne JAC, Carlin JB, Heron J, Henderson J, et al. Strategies for multiple imputation in longitudinal studies. Am J Epidemiol. 2010;172: 478–487.
4. Matthews JNS, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. BMJ. 1990;300: 230–5.
5. Drummond GB, Vowler SL. Show the data, don't conceal them. Br J Pharmacol. 2011;163: 208–210.
6. Sullivan LM. Repeated measures. Circulation. 2008;117: 1238–1243.
7. Gueorguieva R, Krystal JH. Move over ANOVA: progress in analyzing repeated-measures data and its reflection in papers published in the Archives of General Psychiatry. Arch Gen Psychiatry. 2004;61: 310–317.
8. Keselman HJ, Algina J, Kowalchuk RK. The analysis of repeated measures designs: A review. Br J Math Stat Psychol. 2001;54: 1–20.
9. Andrinopoulou E-R, Rizopoulos D, Jin R, Bogers AJJC, Lesaffre E, Takkenberg JJM. An introduction to mixed models and joint modeling: analysis of valve function over time. Ann Thorac Surg. 2012;93: 1765–1772.
10. Harrell Jr FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Second Ed. New York: Springer;

2001.

11. Liu GF, Lu K, Mogg R, Mallick M, Mehrotra D V. Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? Stat Med. 2009;28: 2509–2530. doi:10.1002/sim

12. Grizzle JE, Allen DM. Analysis of growth and dose response curves. Biometrics. 1969;25: 357–381.

13. Littell RC, Pendergast J, Natarajan R. Tutorial in biostatistics: modelling covariance structure in the analysis of repeated measures data. Stat Med. 2000;19: 1793–1819.

14. Mokhles MM, Rajeswaran J, Bekkers JA, Borsboom GJJM, Roos-Hesselink JW, Steyerberg EW, et al. Capturing echocardiographic allograft valve function over time after allograft aortic valve or root replacement. J Thorac Cardiovasc Surg. 2014;148: 1921–1928.e3.

15. Hickey GL, Philipson P, Jorgensen A, Kolamunnage-Dona R. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. BMC Med Res Methodol. BMC Medical Research Methodology; 2016;16: 1–15.

16. Maurissen JP, Vidmar TJ. Repeated-measure analyses: which one? A survey of statistical models and recommendations for reporting. Neurotoxicol Teratol. Elsevier Inc.; 2016;59: 78–84.

17. Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. Second Ed. 2004.

18. Davis CS. Statistical Methods for the Analysis of Repeated Measurements. New York, NY: Springer; 2002.

19. Lim E, Ali A, Theodorou P, Sousa I, Ashrafian H, Chamageorgakis T, et al. Longitudinal study of the profile and predictors of left ventricular mass regression after stentless aortic valve replacement. Ann Thorac Surg. 2008;85: 2026–2029.

FIGURE LEGENDS

Figure 1. Panel A: a graphical representation of a linear mixed effects model. The mean trajectories of two hypothetical patients (A and B; coloured lines) and the mean trajectory averaged over the complete sample of patients (black line) are shown. **Panel B**: longitudinal study dataset exploring the long-term profile of rate of left ventricular mass regression with time after aortic valve replacement with a stentless or a homograft valve. Smoothed lines represent average profiles stratified by valve type, estimated using the LOESS method. Data originally analysed in Lim et al. [19].

Figure 2. Panel A: a so-called 'dynamite plot' showing the mean (height of bars) longitudinal measurement values for different treatment groups at each measurement time, together with the standard deviation (SD; error bar: ± 1 SD). Kruskal Wallis rank-sum tests comparing the outcome between the four treatment groups: # = $P<0.1$, * = $P<0.05$, ** = $P<0.01$, *** = $P<0.001$. **Panel B**: serial measurements of coronary sinus potassium (CSP) (mEq/L) from four groups of dogs. Each translucent line represents a single dog, whilst line colours denote treatment group. Mean profiles (bold lines) are overlaid to summarise the average group trajectories. **Panel C**: a graphical display of the summary statistic slopes method, estimated by fitting separate linear regression lines to each dog (cf. Panel A) and extracting the estimated slopes. The slopes for each treatment group are summarised here as boxplots.

318 **Table 1**. Methodologies for analysing repeated measures data, their advantages and disadvantages, and some software options.

Method	Advantages	Disadvantages	Software
Two-stage methods	• Analysis is based on familiar univariate analysis methods • Data summary methods may facilitate interpretation, e.g. AUC and rate of change are well-understood concepts in biomedicine research • Multiple summary methods can be used	• Can be difficult to specify the correct summary statistic in advance • Reduced data summary statistics are relatively less efficient • Reduced data summary statistics can lose information or fail to capture features of the time course • Summary methods not readily implemented in statistical software, but the summary measures are generally rudimentary to calculate • Missing data can result in sample bias	• Standard tests for independent groups (e.g. *t*-test, ANOVA, Mann-Whitney *U*-test, Kruskal-Wallis test) are standard in all statistics software packages • Summary statistics can be calculated 'by hand' or using a simple programme written in a spreadsheet or statistics package
RM-ANOVA	• Includes the data at all time points • Simple to implement, and conceptually an extension of the ubiquitous ANOVA	• Requires complete data on each subject • Depends on restrictive sphericity assumption, which is highly questionable for longitudinal data • Cannot handle mistimed / unbalanced measurements • Results provide limited information on how the groups differ, often requiring *post hoc* analyses	• SPSS: 'General Linear Model: Repeated Measures' • SAS: PROC GLM • R: aov, Anova (in the car[1] package), ezANOVA (in the ez[2] package) • Stata: anova
LMMs	• Includes the data at all time points • Missing data can be straightforwardly handled if missing (completely) at random • Allows flexible modelling of the time effect	• Implementation and complexity of fitting is relatively more difficult • Assumptions can be harder to assess	• SPSS: 'Mixed Models' • SAS: PROC MIXED • R: lme (nlme[3] package) or lmer (lme4[4] package)

[1] Fox J, Weisberg S (2011). *An R Companion to Applied Regression*, Second Edition. Thousand Oaks CA: *Sage*.

[2] Lawrence MA (2016). ez: Easy Analysis and Visualization of Factorial Experiments. R package version 4.4-0. https://CRAN.R-project.org/package=ez

[3] Pinheiro JC, Bates DM (2000). *Mixed-Effects Models in S and S-PLUS*. New York: *Springer Verlag*.

[4] Bates D, Maechler M, Bolker B, Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

	• Permits unbalanced data with greatly different numbers of measurements per subject • Allows for time-varying covariates • Permits estimation of individual trends • Can be augmented with more complex covariance structures that captures more features of the correlation patterns, and hierarchically	• Stata: `xtmixed`

320 **Table 2.** Results from analysis of laboratory experiment longitudinal data.

Linear mixed effects model[a]				
	Estimate	**SE**	**95% CI**	***P***
Intercept	4.05	0.17	(3.72 to 4.37)	<0.001
Group				
ECD (3-weeks)	-0.44	0.23	(-0.90 to 0.03)	0.064
ECD (0-weeks)	-0.33	0.24	(-0.82 to 0.17)	0.19
Sympathectomy	-0.32	0.23	(-0.80 to 0.15)	0.18
Time (mins)	0.08	0.02	(0.05 to 0.12)	<0.001
Time * ECD (3-weeks)	-0.09	0.03	(-0.14 to -0.04)	<0.001
Time * ECD (0-weeks)	-0.02	0.03	(-0.08 to 0.03)	0.43
Time * Sympathectomy	-0.05	0.03	(-0.10 to 0.00)	0.054
Summary statistic (Kruskal-Wallis rank-sum tests)				
	df		χ^2	***P***
Slope	3		8.53	0.036
Final value	3		11.14	0.011

Notation: CSP–coronary sinus potassium; SE–standard error; CI–confidence interval; ECD–

322 extrinsic cardiac denervation; df–degrees of freedom; χ^2–chi-square statistic.

[a] Fitted by restricted maximum likelihood.