The statistical analysis of non-linear longitudinal data in biomedical research using generalized additive models

Beyond repeated measures ANOVA and Linear Mixed Models

Ariel Mundo*   Timothy J. Muldoon†   John R. Tipton‡

Contents

1 Background

A longitudinal study is defined as that which is designed to repeatedly measure a variable of interest in a group (or groups) of subjects. In biomedical research, this type of study is employed when the intention is to observe the evolution of the effect of a certain treatment across time, rather than analyzing it at a single time point (a cross-sectional study). Clinical studies of breast and neck cancer have used a longitudinal approach [1], [2]. In the first case, weekly measurements of skin toxicities in patients with radiation-induced dermatitis were taken for up to 8 weeks; whereas in the latter mouth opening was assessed at 6,12, 18, 24 and 36 months after radiotherapy (RT). Longitudinal studies have beenn also used to analyze tumor response [3]–[6], antibody expression [7], [8], and cell metabolism [9], [10].

Traditionally, a "frequentist" approach is used in biomedical research to derive inferences from a longitudinal study. Such statistical view derives its name from the fact that it regards probability as a limiting frequency[11] and its application is based on a null hypothesis test using the *analysis of variance over repeated measures* (repeated measures ANOVA or rm-ANOVA). This methodology makes two key assumptions regarding longitudinal data: constant correlation exists across same-subject measurements, and observations

---

*Department of Biomedical Engineering, University of Arkansas, Fayetteville

†Department of Biomedical Engineering, University of Arkansas, Fayetteville

‡Department of Mathematical Sciences, University of Arkansas, Fayetteville

from each subject are obtained at all time points through the study (a condition also known as complete observations) [12], [13].

However, constant correlation is frequently unjustified as its value tends to diminish between measures when the time interval between them increases [14]. In the case of the number of observations during the study, different situations can arise that prevent the collection of complete measurements for all subjects: in a clinical trial voluntary withdrawal from one or multiple patients can occur, whereas attrition in animals due to injury or weight loss can occur in preclinical experiments, and it is even possible that unexpected complications with equipment or supplies arise, preventing the researcher from collecting measurements at a certain time point and therefore violating the *complete observations* assumption of rm-ANOVA.

When incomplete observations occur, an rm-ANOVA requires the exclusion of all subjects with missing observations from the analysis. This can result in increased costs if the desired statistical power is not met with the remaining observations, as it would be necessary to enroll more subjects. At the same time, if the excluded observations contain insightful information that is not used, their elimination from the analysis may limit the demonstration of significant differences between groups. Additionally, rm-ANOVA uses a *post hoc* analysis to assess differences between the measured response in different groups. A *post hoc* analysis is based on multiple repeated comparisons to estimate a *p-value*, a metric that is widely used as a measure of significance. Because the *p-value* is highly variable, multiple comparisons can inflate the false positivity rate [15], [16], consequently biasing the conclusions of the study.

During the last decade, the biomedical community has started to recognize the limitations of rm-ANOVA in the analysis of longitudinal information. This is exemplified by the use of linear mixed effects models (LMEMs) in certain groups to analyze longitudinal data [10], [17]. Briefly, these models incorporate *fixed effects*, which correspond to the levels of experimental factors in the study (e.g. the different drug regimens in a clinical trial), and *random effects*, which account for random variation within the population. These models are more flexible than the traditional rm-ANOVA as they can accommodate missing observations for multiple subjects and allow different modeling strategies for the variability within each measure in every subject [18]. On the other hand, they impose restrictions in the distribution of the errors of the random effects [12].

One final assumption that is not initially evident for both rm-ANOVA and LMEMs models pertains the relationship between the observed effect and the progression of the timeline of the study, which is expected to be linear [18]. This common assumption in both rm-ANOVA ane LMEMs consequently restricts the inferences from the model when the data does not follow a linear trend. In biomedical research, a special case of this behavior in longitudinal data arises in measurements of tumor response in preclinical and clinical settings [3], [10], [17]. These studies have shown that the collected signal does not follow a linear trend over time, and presents extreme variability at different time points, making the fit of and the estimations of an LMEM or rm-ANOVA model inconsistent with the observed variation. In other words, the in both cases the model tries to explain highly-variable data using a linear trend, consequently biasing the estimates. Additionally, although it is possible that a *post hoc* analysis is able to find "significant" *p-values*($p<0.05$) by using multiple comparisons between the model terms, the use of this estimator is limited in such scenario because it is intended to work in models that have a reasonable agreement with the data.

As the "frequentist" rm-ANOVA and the more advanced LMEM approach are both limited in the analysis of non-linear longitudinal information, there is a need for biomedical researchers to explore the use of additional statistical tools that allow the information (and not an assumed trend) to determine the fit of the model, while enabling inferences that are both adequate and consistent from a statistical perspective. In this regard, Generalized Additive Models (GAMs) present an alternative approach to analyze longitudinal data. Although not commonly used in the biomedical community, these non-parametric models have been used to analyze temporal variations in geochemical and palaeoecological data [19]–[21], health-environment interactions [22] and political science [23] . There are several advantages of GAMs over LMEMs and rm-ANOVA models: They have an extensively developed theoretical background, enable the data to dictate the trend of the model, can accommodate missing observations and do not require constant correlation between repeated measurements [24].Therefore, GAMs can provide a more suitable statistical method to analyze non-linear biomedical longitudinal data.

The current advances in programming languages designed for statistical analysis (specifically R), have eased the computational implementation of more complex models beyond LMEMs. In particular, R has an extensive collection of documentation and functions to fit GAMs in the package *mgcv* [24]–[26] that not only speed up the initial stages of the analysis but enable the use of advanced modeling structures (e.g. hierarchical models, confidence interval comparisons) without requiring advanced programming skills from the user. This programming language is also able to simulate data, an emerging strategy used to test statistical models [27]. This allows to create and explore different alternatives for analysis without collecting information in the field, and reduces the time window between experiment design and its implementation.

Therefore, the purpose of this study is to provide biomedical researchers with a clear understanding of the theory and the practical implementation of GAMs to analyze longitudinal data using by focusing on four areas. First, the limitations of rm-ANOVA and LMEMs to analyze longitudinal data are contrasted in the cases of missing observations, assumption of linearity and correlation structures. Secondly, the key theoretical elements of GAMS are presented without using extensive derivations or complex mathematical notation that can obscure their understanding. Thirdly, simulated data that follows the trend of previously reported values [17] is used to illustrate the type of non-linear longitudinal data that occurs in biomedical research and the implementation of GAMs in such scenario. Finally, reproducibility is emphasized by providing the code to generate the data and the implementation of GAMs in R, in conjunction with a step-by-step guide to fit models of increasing complexity.

In summary, the exploration of modern statistical techniques to analyze longitudinal data may allow biomedical researchers to consider the use of GAMs instead of rm-ANOVA or LMEMs when the data does not follow a linear trend, and will also help to improve the standards for reproducibility in biomedical research.

2 Challenges presented by longitudinal studies

2.1 The "frequentist" case for longitudinal data

The *repeated measures analysis of variance* (rm-ANOVA) is the standard statistical analysis for longitudinal data in biomedical research, but certain assumptions are necessary to make the model valid. From a practical view, they can be divided in three areas: linear relationship between covariates and response, constant correlation between measurements, and complete observations for all subjects. Each one of these assumptions is discussed below.

2.2 Linear relationship

In a biomedical longitudinal study, two or more groups of subjects (patients, mice, samples) are subject to a different treatments (e.g. a "treatment" group receives a novel drug vs. a "control" group that receives a placebo), and measurements from each subject within each group are collected at specific time points. The collected response has both *fixed* and *random* components. The *fixed* component can be understood as a constant value in the response which the researcher is interested in measuring, i.e, the effect of the novel drug in the "treatment" group.The *random* component can be defined as "noise" caused by some inherent variability within the study. For example, if the blood concentration of the drug is measured in certain subjects in the early hours of the morning while others are measured in the afternoon, it is possible that the difference in the collection time of the measurement introduces some "noise" in the signal. As their name suggests, this "random" variability needs to be modeled as a variable rather than as a constant value.

Mathematically speaking, a rm-ANOVA model can be written as:

$$y_i = \mu + \beta_1 * time_t + \beta_2 * treatment_j + \beta_3 * time_t * treatment_j + \pi_{ij} + \epsilon_{tij} \tag{1}$$

In this model y_i is the response by subject i, which can be decomposed in a mean value μ, *fixed effects* of time ($time_t$), treatment ($time_j$) and their interaction; a *random effect* π_{ij} of each subject within each group, and

3

errors per subject per time per group ϵ_{tij}. Suppose two treatments groups are used in the study. Models per group can be generated if 0 is used for the first treatment group (Group A) and 1 for the second treatment group (Group B). The linear models then become:

$$y = \begin{cases} \mu + \beta_1 * time_t + \epsilon_{ti} & \text{if Group A} \\ \mu + \beta_1 * time_t + \beta_2 * treatment_j + \beta_3 * time_t * treatment_j + \pi_i + \epsilon_{tij} & \text{if Group B} \end{cases} \tag{2}$$

Furthermore, (2) can be re-written under the generalized linear model (GLM) framework [28]. The main difference in this case is that a GLM model uses a *linking function* to connect the observed value to the linear model.

Thus,

$$y = \mu + \boldsymbol{\beta}_1 * f(time) + \beta_2 * treatment + \boldsymbol{\beta}_3 * f(time) * treatment \tag{3}$$

The linear model in (1) is a GLM where $f(time) = time$ with the identity function $f(x) = x$. If the GLM is then written in terms of two treatment groups as in (2), the model becomes:

$$y = \begin{cases} \mu + \boldsymbol{\beta}_1 * f(time) & \text{if Group A} \\ \mu + \boldsymbol{\beta}_1 * f(time) + \beta_2 * treatment + \boldsymbol{\beta}_3 * f(time) * treatment & \text{if Group B} \end{cases} \tag{4}$$

Regardless of its presentation, (1), (2) and (4) show that the model expects a linear relationship between the covariates and the response. When the data does not follow a linear trend, the fit that the model produces does not accurately represent the changes in the data. To exemplify this consider simulated longitudinal data where the a normally distributed response of two groups of 10 subjects each is obtained, and an rm-ANOVA model identical to (1) is fitted to the data using the package *nlme* (the code for the simulated data and models is available in the Appendix).

It is clear from 1 that the fit produced by the rm-ANOVA model is good as the predictions for the mean response are identical to the "truth" of the simulated data.

However, consider the case when the data follows a non-linear trend, such as the one simulated in 2. Here, the simulated data follows a quadratic behavior. It is clear from the figure that changes in each group occur through the timeline, although the final mean value is the same as the initial value. Fitting an rm-ANOVA model like (1) again produces the fit that appears at the bottom of 2.

In this case, when the predictions are compared to the simulated data it is clear that the model is not capturing the changes within each group throughout the timeline. This highlights the limitation of rm-ANOVA and LMEMs with longitudinal non-linear data, where the variations are not captured and modeled properly.

2.3   Covariance in rm-ANOVA and LMEMs

In a longitudinal study there is an expected *variance* between repeated measurements on the same subject, and because repeated measures occur in the subjects within each group, there is a *covariance* between measurements at each time point within each group. The *covariance matrix* (also known as the variance-covariance matrix) is a matrix that captures the variation between and within subjects in a longitudinal study[29] (For an in-depth analysis of the covariance matrix see [30], [31]).

In the case of an rm-ANOVA analysis, it is assumed that the covariance matrix has a specific construction known as *compound symmetry* (also known as "sphericity" or "circularity"). Under this assumption, the between-subject variance and within-subject correlation are constant across time [31]–[33]. However, it has been shown that this condition is frequently unjustified because the correlation between measurements tends to change over time [34]; and it is higher between consecutive measurements [12], [14]. Although corrections
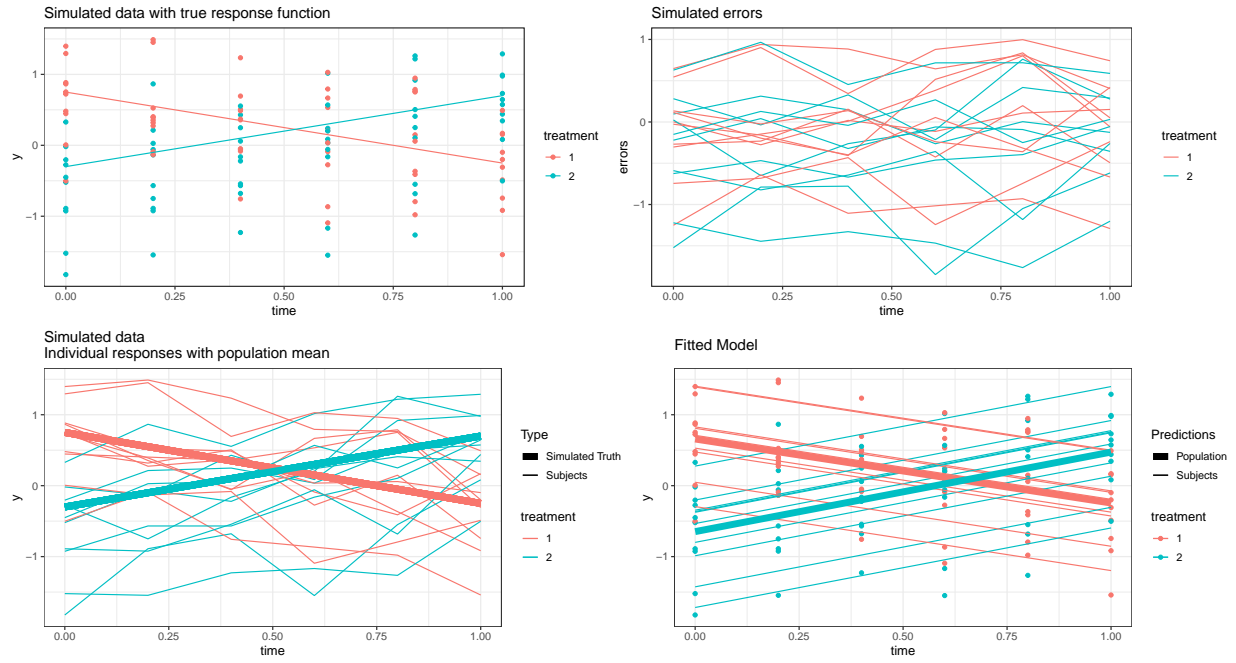
Figure 1: Simulated linear responses from two groups with a rm-ANOVA model fitted. Top row, simulated data, lines represent mean response. Bottom row, fitted model, thick lines represent predicted mean response
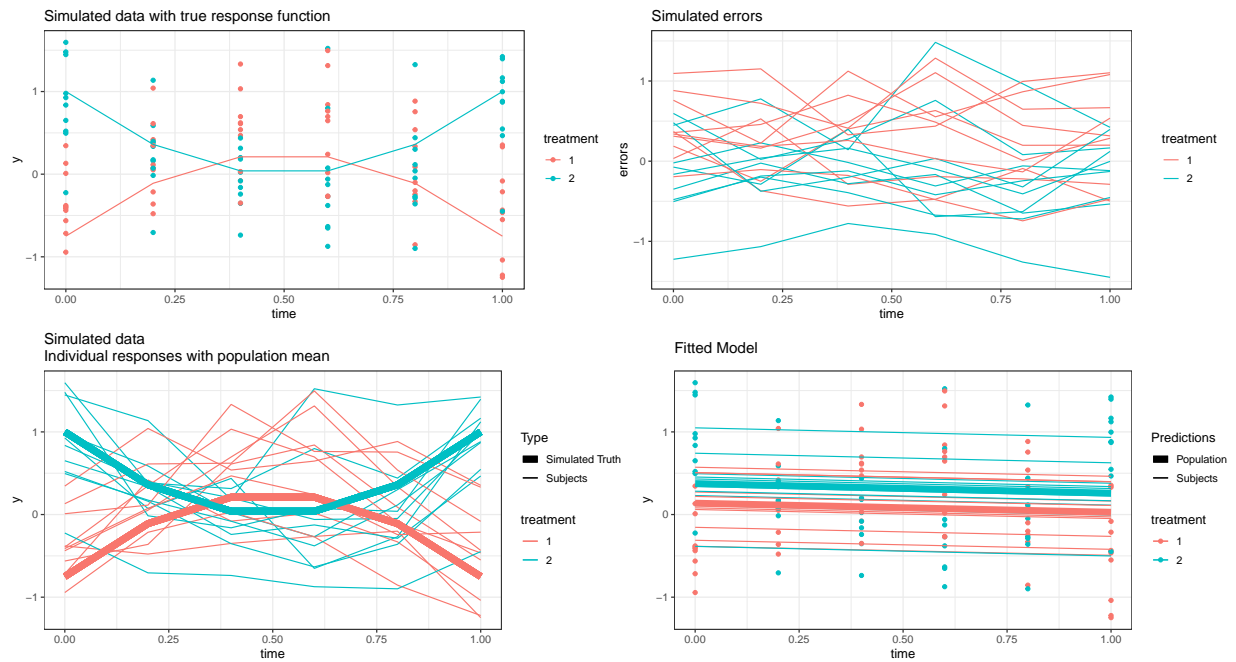

Figure 2: Simulated quadratic responses from two groups with an rm-ANOVA model fitted. Top row, simulated data, lines represent mean response. Bottom row, fitted model, thick lines represent predicted mean response

can be made (such as Huyhn-Feldt or Greenhouse-Geisser) the effectiveness of each correction is limited because it depends on the size of the sample,the number of repeated measurements[27], and they are not robust if the group sizes are unbalanced [35]. In other words, if the data does not present constant correlation between repeated measurements, the assumptions required for an rm-ANOVA model are not met and the use of corrections may still not provide a reasonable adjustment that makes the model valid.

In the case of LMEMs, one key advantage over rm-ANOVA is that they allow different structures for the variance-covariance matrix including exponential, autoregressive of order 1, rational quadratic and others [18]. Nevertheless, the analysis required to determine an appropriate variance-covariance structure for the data can be a long process by itself. Overall, the spherical assumption for rm-ANOVA may not capture the natural variations of the correlation in the data, and can bias the inferences from the analysis.

2.4   Missing observations

Missing observations are an issue that arises frequently in longitudinal studies. In biomedical research, this situation can be caused by reasons beyond the control of the investigator [molenberghs2004].Dropout from patients, or attrition or injury in animals are among the reasons for missing observations. Statistically, missing information can be classified as *missing at random* (MAR), *missing completely at random* (MCAR), and *missing not at random* (MNAR) [31]. In a MAR scenario, the pattern of the missing information is related to some variable in the data, but it is not related to the variable of interest [36]. If the data are MCAR, this means that the missigness is completely unrelated to the collected information [37], and in the case of MNAR the missing values are dependent on their value. An rm-ANOVA model assumes complete observations for all subjects, and therefore subjects with one or more missing observations are excluded from the analysis. This is inconvenient because the remaining subjects might not accurately represent the population, and statistical power is affected by this reduction in sample size [38].

In the case of LMEMs, inferences from the model are valid when missing observations in the data exist that are MAR [30]. The pattern of missing observations can be considered MAR if the missing observations are not related any of the other variables measured in the study [34]. For example, if attrition occurs in all mice that had lower weights at the beginning of a chemotherapy response study, the missing data can be considered MAR because the missigness is unrelated to other variables of interest.

This section has presented the assumptions of rm-ANOVA to analyze longitudinal information and its differences when compared to LMEMs regarding to missing data and the modeling of the covariance matrix. Of notice, LMEMs offer a more robust and flexible approach than rm-ANOVA and if the data follows a linear trend, they provide an excellent choice to derive inferences from a repeated measures study. However, when the data presents high variability, LMEMs fail to capture the non-linear trend of the data. To analyze such type of data, we present generalized additive models (GAMs) as an alternative in the following section.

3   GAMs as a special case of Generalized Linear Models

A GAM is a special case of the Generalized Linear Model (GLM), a framework that allows for response distributions that do not follow a normal distribution [24], [39]. Following the notation by Simpson [21] A GAM model can be represented as:

$$y_t = \beta_0 + f(x_t) + \epsilon_t \tag{5}$$

Where y_t is the response at time t, β_0 is the expected value at time 0, the change of y over time is represented by the function $f(x_t)$ and ϵ_t represents the residuals.

In contrast to rm-ANOVA or LMEMs, GAMs use *smooth functions* to model the relationship between the covariates and the response. This approach is more advantageous as it does not restrict the model to a linear relationship. One possible function for $f(x_t)$ is a polynomial, but a major limitation is that they create a

"global" fit as they assume that the same relationship exists everywhere, which can cause problems with the fit [23].

To overcome this limitation, the smooth functions in GAMs are represented using *basis functions* over evenly spaced ranges of the covariates known as *knots*. The *basis function* used is a cubic spline, which is a smooth curve constructed from cubic polynomials joined together[21], [24]. Cubic splines have a long history in their use to solve non-parametric statistical problems and are the default choice to fit GAMs as they are the simplest option to obtain visual smoothness [40]. Therefore, GAMs overcome the limitation that occurs in LMEMs and rm-ANOVA when the data is non linear, such as 2. Regarding longitudinal data, Pedersen et al [20] demonstrated the capabilities of GAMs in this area using ecological data.

The use of GAMs to analyze biomedical longitudinal data, and the impact of missing observations in the fit of the model will be examined in detail in the following section using simulation.

References

[1] T. T. Sio *et al.*, "Repeated measures analyses of dermatitis symptom evolution in breast cancer patients receiving radiotherapy in a phase 3 randomized trial of mometasone furoate vs placebo (n06c4 [alliance])," *Supportive Care in Cancer*, vol. 24, no. 9, pp. 3847–3855, 2016.

[2] J. Kamstra, P. Dijkstra, M. Van Leeuwen, J. Roodenburg, and J. Langendijk, "Mouth opening in patients irradiated for head and neck cancer: A prospective repeated measures study," *Oral Oncology*, vol. 51, no. 5, pp. 548–555, 2015.

[3] D. Roblyer *et al.*, "Optical imaging of breast cancer oxyhemoglobin flare correlates with neoadjuvant chemotherapy response one day after starting treatment," *Proceedings of the National Academy of Sciences*, vol. 108, no. 35, pp. 14626–14631, 2011.

[4] A. Tank *et al.*, "Diffuse optical spectroscopic imaging reveals distinct early breast tumor hemodynamic responses to metronomic and maximum tolerated dose regimens," *Breast cancer research*, vol. 22, no. 1, pp. 1–10, 2020.

[5] M. V. Pavlov *et al.*, "Multimodal approach in assessment of the response of breast cancer to neoadjuvant chemotherapy," *Journal of biomedical optics*, vol. 23, no. 9, p. 091410, 2018.

[6] V. Demidov *et al.*, "Preclinical longitudinal imaging of tumor microvascular radiobiological response with functional optical coherence tomography," *Scientific reports*, vol. 8, no. 1, pp. 1–12, 2018.

[7] G. Ritter, L. S. Cohen, C. Williams, E. C. Richards, L. J. Old, and S. Welt, "Serological analysis of human anti-human antibody responses in colon cancer patients treated with repeated doses of humanized monoclonal antibody a33," *Cancer Research*, vol. 61, no. 18, pp. 6851–6859, 2001.

[8] E. M. Roth *et al.*, "Antidrug antibodies in patients treated with alirocumab," 2017.

[9] J. D. Jones, H. E. Ramser, A. E. Woessner, and K. P. Quinn, "In vivo multiphoton microscopy detects longitudinal metabolic changes associated with delayed skin wound healing," *Communications biology*, vol. 1, no. 1, pp. 1–8, 2018.

[10] M. C. Skala, A. N. Fontanella, L. Lan, J. A. Izatt, and M. W. Dewhirst, "Longitudinal optical imaging of tumor metabolism and hemodynamics," *Journal of biomedical optics*, vol. 15, no. 1, p. 011112, 2010.

[11] E.-J. Wagenmakers, M. Lee, T. Lodewyckx, and G. J. Iverson, "Bayesian versus frequentist inference," in *Bayesian evaluation of informative hypotheses*, Springer, 2008, pp. 181–207.

[12] R. Gueorguieva and J. H. Krystal, "Move over anova: Progress in analyzing repeated-measures data andits reflection in papers published in the archives of general psychiatry," *Archives of general psychiatry*, vol. 61, no. 3, pp. 310–317, 2004.

[13] P. Schober and T. R. Vetter, "Repeated measures designs and analysis of longitudinal data: If at first you do not succeed—try, try again," *Anesthesia and analgesia*, vol. 127, no. 2, p. 569, 2018.

[14] C. Ugrinowitsch, G. W. Fellingham, and M. D. Ricard, "Limitations of ordinary least squares models in analyzing repeated measures data," *Medicine and science in sports and exercise*, vol. 36, pp. 2144–2148, 2004.

[15] C. Liu, T. P. Cripe, and M.-O. Kim, "Statistical issues in longitudinal data analysis for treatment efficacy studies in the biomedical sciences," *Molecular Therapy*, vol. 18, no. 9, pp. 1724–1730, 2010.

[16] L. G. Halsey, D. Curran-Everett, S. L. Vowler, and G. B. Drummond, "The fickle p value generates irreproducible results," *Nature methods*, vol. 12, no. 3, pp. 179–185, 2015.

[17] K. Vishwanath, H. Yuan, W. T. Barry, M. W. Dewhirst, and N. Ramanujam, "Using optical spectroscopy to longitudinally monitor physiological changes within solid tumors," *Neoplasia*, vol. 11, no. 9, pp. 889–900, 2009.

[18] J. Pinheiro and D. Bates, *Mixed-effects models in s and s-plus*. Springer Science & Business Media, 2006.

[19] N. L. Rose, H. Yang, S. D. Turner, and G. L. Simpson, "An assessment of the mechanisms for the transfer of lead and mercury from atmospherically contaminated organic soils to lake sediments with particular reference to Scotland, UK," *Geochimica et Cosmochimica Acta*, vol. 82, pp. 113–135, 2012.

[20] E. J. Pedersen, D. L. Miller, G. L. Simpson, and N. Ross, "Hierarchical generalized additive models in ecology: An introduction with mgcv," *PeerJ*, vol. 7, p. e6876, 2019.

[21] G. L. Simpson, "Modelling palaeoecological time series using generalised additive models," *Frontiers in Ecology and Evolution*, vol. 6, p. 149, 2018.

[22] L. Yang, G. Qin, N. Zhao, C. Wang, and G. Song, "Using a generalized additive model with autoregressive terms to study the effects of daily temperature on mortality," *BMC Medical Research Methodology*, vol. 12, no. 1, p. 165, 2012.

[23] N. Beck and S. Jackman, "Beyond linearity by default: Generalized additive models," *American Journal of Political Science*, pp. 596–627, 1998.

[24] S. N. Wood, *Generalized additive models: An introduction with r*. CRC press, 2017.

[25] S. N. Wood, N., Pya, and B. Säfken, "Smoothing parameter and model selection for general smooth models (with discussion)," *Journal of the American Statistical Association*, vol. 111, pp. 1548–1575, 2016.

[26] S. N. Wood, N., Pya, and B. Säfken, "Smoothing parameter and model selection for general smooth models (with discussion)," *Journal of the American Statistical Association*, vol. 111, pp. 1548–1575, 2016.

[27] N. Haverkamp and A. Beauducel, "Violation of the sphericity assumption and its effect on type-i error rates in repeated measures anova and multi-level linear models (mlm)," *Frontiers in psychology*, vol. 8, p. 1841, 2017.

[28] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.

[29] R. D. Wolfinger, "Heterogeneous variance: Covariance structures for repeated measures," *Journal of agricultural, biological, and environmental statistics*, pp. 205–230, 1996.

[30] B. T. West, K. B. Welch, and A. T. Galecki, *Linear mixed models: A practical guide using statistical software*. CRC Press, 2014.

[31] R. E. Weiss, *Modeling longitudinal data*. Springer Science & Business Media, 2005.

[32] S. Geisser, S. W. Greenhouse, and others, "An extension of box's results on the use of the F distribution in multivariate analysis," *The Annals of Mathematical Statistics*, vol. 29, no. 3, pp. 885–891, 1958.

[33] H. Huynh and L. S. Feldt, "Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs," *Journal of educational statistics*, vol. 1, no. 1, pp. 69–82, 1976.

[34] S. E. Maxwell, H. D. Delaney, and K. Kelley, *Designing experiments and analyzing data: A model comparison perspective.* Routledge, 2017.

[35] H. Keselman, J. Algina, and R. K. Kowalchuk, "The analysis of repeated measures designs: A review," *British Journal of Mathematical and Statistical Psychology*, vol. 54, no. 1, pp. 1–20, 2001.

[36] J. Scheffer, "Dealing with missing data," 2002.

[37] R. F. Potthoff, G. E. Tudor, K. S. Pieper, and V. Hasselblad, "Can one assess whether missing data are missing at random in medical studies?" *Statistical methods in medical research*, vol. 15, no. 3, pp. 213–234, 2006.

[38] Y. Ma, M. Mazumdar, and S. G. Memtsoudis, "Beyond repeated-measures analysis of variance: Advanced statistical methods for the analysis of longitudinal data in anesthesia research," *Regional Anesthesia & Pain Medicine*, vol. 37, no. 1, pp. 99–105, 2012.

[39] T. Hastie and R. Tibshirani, "Generalized additive models: Some applications," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 371–386, 1987.

[40] E. J. Wegman and I. W. Wright, "Splines in statistics," *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 351–365, 1983.