Generalized additive models to analyze non-linear trends in biomedical longitudinal data using R: Beyond repeated measures ANOVA and Linear Mixed Models

Ariel I. Mundo, Timothy J. Muldoon*

Department of Biomedical Engineering, University of Arkansas, Fayetteville, AR, USA

tmuldoon@uark.edu

John R. Tipton

Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR, USA

1 Abstract

In biomedical research, the outcome of longitudinal studies has been traditionally analyzed using the *repeated measures analysis of variance* (rm-ANOVA) or more recently, *linear mixed models* (LMEMs). Although LMEMs are less restrictive than rm-ANOVA as they can work with unbalanced data and non-constant correlation between observations, both methodologies assume a linear trend in the measured response. It is common in biomedical research that the true trend response is nonlinear and in these cases the linearity assumption of rm-ANOVA and LMEMs can lead to biased estimates and unreliable inference.

In contrast, GAMs relax the linearity assumption of rm-ANOVA and LMEMs and allow the data to determine the fit of the model while also permitting incomplete observations and different correlation structures. Therefore, GAMs present an excellent choice to analyze longitudinal data with non-linear trends in the context of biomedical research. This paper summarizes the limitations of rm-ANOVA and LMEMs and uses simulated data to visually show how both methods produce biased estimates when used on data with non-linear trends. We present the basic theory of GAMs and using reported trends of oxygen saturation in tumors, we simulate example longitudinal data (2 treatment groups,

10 subjects per group, 5 repeated measures for each group) to demonstrate their implementation in R. We also show that GAMs are able to produce estimates with non-linear trends even when incomplete observations exist (with 40% of the simulated observations missing). To make this work reproducible, the code and data used in this paper are available at: https://github.com/aimundo/GAMs-biomedical-research.

Keywords

longitudinal data; biomedical data; generalized additive models; simulation; R

2 Background

Longitudinal studies are designed to repeatedly measure a variable of interest in a group (or groups) of subjects, with the intention of observing the evolution of effect across time rather than analyzing a single time point (e.g., a cross-sectional study). Biomedical research frequently uses longitudinal studies to analyze the evolution of a "treatment" effect across multiple time points, with subjects of analysis ranging from animals (mice, rats, rabbits), to human patients, cells, or blood samples, among many others. Tumor response[1–4], antibody expression[5,6], and cell metabolism[7,8] are examples of different situations where researchers have used longitudinal designs to study some physiological response. Because the frequency of the measurements in a longitudinal study is dependent on the biological phenomena of interest and the experimental design of the study, the frequency of such measurements can range from minute intervals to study a short-term response such as anesthesia effects in animals[9], to weekly measurements to analyze a mid-term response like the evolution of dermatitis symptoms in breast cancer patients[10], to monthly measurements to study a long-term response such as mouth opening following radiotherapy (RT) in neck cancer patients[11].

Traditionally, a "frequentist" or "classical" statistical paradigm is used in biomedical research to derive inferences from a longitudinal study. The frequentist paradigm regards probability as the limit of the expected outcome when an experiment is repeated a large number of times[12], and such view is applied to the analysis of longitudinal data by assuming a null hypothesis under a statistical model that is often an *analysis of variance over repeated measures* (repeated measures ANOVA or rm-ANOVA). The rm-ANOVA model makes three assumptions regarding longitudinal data: 1) linearity of the response across time, 2) constant correlation across same-subject measurements, and 3) observations from each subject are obtained at all time points through the study (a condition also known as *complete observations*)[13,14].

The expected linear behavior of the response through time is a key requisite in rm-ANOVA[15]. This "linearity assump-

tion" in rm-ANOVA implies that the model is misspecified when the data does not follow a linear trend, which results in unreliable inference. In longitudinal biomedical research, non-linear trends are the norm rather than the exception. A particular example of this non-linear behavior in longitudinal data arises in measurements of tumor response to chemo and/or radiotherapy in preclinical and clinical settings[1,8,16]. 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 rm-ANOVA model inconsistent with the observed variation. Therefore, when rm-ANOVA is used to draw inference of such data the estimates are inevitably biased because the model is only able to accommodate linear trends that fail to adequately represent the biological phenomenon of interest.

A *post hoc* analysis is often used in conjunction with rm-ANOVA to perform repeated comparisons to estimate a *p-value*, which in turn is used as a measure of significance. Although it is possible that a *post hoc* analysis of rm-ANOVA is able to find "significant effects (*p-value*<0.05)" from data with non-linear trends, the validity of such a metric is dependent on how adequate the model fits the data. In other words, *p-values* are valid only if the model and the data have good agreement; if that is not the case, a "Type III" error (known as "model misspecification") occurs[17]. For example, model misspecification will occur when a model that is only able to explain linear responses (such as rm-ANOVA) is fitted to data that follows a quadratic trend, thereby causing the resulting *p-values* and parameter estimates to be invalid[18].

Additionally, the *p-value* itself is highly variable, and multiple comparisons can inflate the false positivity rate (Type I error or α)[19,20], consequently biasing the conclusions of the study. Corrections exist to address the Type I error issue of multiple comparisons (such as Bonferroni[21]), but they in turn reduce statistical power (1-β)[22], and lead to increased Type II error (failing to reject the null hypothesis when it is false)[23,24]. Therefore, the tradeoff of *post hoc* comparisons in rm-ANOVA between Type I, II and III errors might be difficult to resolve in a biomedical longitudinal study where a delicate balance exists between statistical power and sample size.

On the other hand, the assumption of constant correlation in rm-ANOVA (often known as the *compound symmetry assumption*) is typically unreasonable because correlation between the measured responses often diminishes as the time interval between the observation increases[25]. Corrections can be made in rm-ANOVA in the absence of compound symmetry[26,27], but the effectiveness of the correction is limited by the size of the sample, the number of measurements[28], and group sizes[29]. In the case of biomedical research, where living subjects are frequently used, sample sizes are often not "large" due to ethical and budgetary reasons,[30] which might cause the corrections for lack of compound symmetry to be ineffective.

Due to a variety of causes, the number of observations during a study can be different between all subjects. For example, in a clinical trial patients may voluntarily withdraw, whereas attrition due to injury or weight loss in preclinical animal studies is possible. It is even plausible that unexpected complications with equipment or supplies arise that

prevent the researcher from collecting measurements at certain time points. In each of these scenarios, the *complete observations* assumption of classical rm-ANOVA is violated. When incomplete observations occur, a rm-ANOVA model is fit by excluding all subjects with incomplete observations from the analysis[13]. This elimination of partially missing data from the analysis can result in increased costs if the desired statistical power is not met with the remaining observations, because 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.

During the last decade, the biomedical community has started to recognize the limitations of rm-ANOVA in the analysis of longitudinal data. The recognition on the shortcomings of rm-ANOVA is exemplified by the use of linear mixed effects models (LMEMs) by certain groups to analyze longitudinal tumor response data[8,16]. Briefly, LMEMs 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 (e.g., the individual-level differences not due to treatment such as weight or age). When compared to the traditional rm-ANOVA, LMEMs are more flexible as they can accommodate incomplete observations for multiple subjects and allow different modeling strategies for the variability within each measure in every subject[15]. However, LMEMs impose restrictions in the distribution of the random effects, which need to be independent[13,31]. And even more importantly, LMEMs also assume by default a linear relationship between the response and time[15] (polynomial effects can be used, but this approach has its own shortcomings as we discuss in Section 4.2.1) .

As the rm-ANOVA and the more flexible LMEM approaches make overly restrictive assumptions regarding the trend of the response, there is a need for biomedical researchers to explore the use of statistical tools that allow the data (and not a model assumption) to determine the trend of the fitted model and to enable appropriate inference. In this regard, generalized additive models (GAMs) present an alternative approach to analyze longitudinal data. Although not frequently used by the biomedical community, these semi-parametric models are customarily used in other fields to analyze longitudinal data. Examples of the use of GAMs include the analysis of temporal variations in geochemical and palaeoecological data[32–34], health-environment interactions[35] and the dynamics of government in political science[36]. There are several advantages of GAMs over LMEMs and rm-ANOVA models: 1) GAMs can fit a more flexible class of smooth responses that enable the data to dictate the trend in the fit of the model, 2) they can model non-constant correlation between repeated measurements[37], and 3) can easily accommodate incomplete observations. Therefore, GAMs provide a more flexible statistical approach to analyze non-linear biomedical longitudinal data than LMEMs and rm-ANOVA.

The current advances in programming languages designed for statistical analysis (specifically R), have eased the computational implementation of traditional models such as rm-ANOVA and more complex approaches such as LMEMs

and GAMs. In particular, R[38] has an extensive collection of documentation and functions to fit GAMs in the package *mgcv*[37,39] that speed up the initial stages of the analysis and enable the use of advanced modeling structures (e.g. hierarchical models, confidence interval comparisons) without requiring advanced programming skills. At the same time, R has many tools that simplify data simulation.

Data simulation methods are an emerging technique that allow the researcher to create and explore different alternatives for analysis without collecting information in the field, reducing the time window between experiment design and its implementation. In addition, simulation can be also used for power calculations and study design questions[28].

This work provides biomedical researchers with a clear understanding of the theory and the practice of using GAMs to analyze longitudinal data using by focusing on four areas. First, the limitations of LMEMs and rm-ANOVA regarding an expected trend of the response, constant correlation structures, and complete observations are explained in detail. Second, the key theoretical elements of GAMs are presented using clear and simple mathematical notation while explaining the context and interpretation of the equations. Third, we illustrate the type of non-linear longitudinal data that often occurs in biomedical research using simulated data that reproduces patterns in previously reported studies[16]. The simulated data experiments highlight the differences in inference between rm-ANOVA, LMEMs and GAMs on data similar to what is commonly observed in biomedical studies. Finally, reproducibility is emphasized by providing the code to generate the simulated data and the implementation of different models in R, in conjunction with a step-by-step guide demonstrating how to fit models of increasing complexity.

In summary, this work will allow biomedical researchers to identify when the use of GAMs instead of rm-ANOVA or LMEMs is appropriate to analyze longitudinal data, and provide guidance on the implementation of these models to improve the standards for reproducibility in biomedical research.

3 Challenges presented by longitudinal studies

3.1 The repeated measures ANOVA and Linear Mixed Model

The *repeated measures analysis of variance* (rm-ANOVA) and the *linear mixed model* (LMEM) are the most commonly used statistical analysis for longitudinal data in biomedical research. These statistical methodologies require certain assumptions for the model to be valid. From a practical view, the assumptions can be divided in three areas: 1) an assumed relationship between covariates and response, 2) a constant correlation between measurements, and, 3) complete observations for all subjects. Each one of these assumptions is discussed below.

3.2 Assumed relationship

3.2.1 The repeated measures ANOVA case

In a longitudinal biomedical study, two or more groups of subjects (e.g., human subject, mice, samples) are subject to different treatments (e.g., a "treatment" group receives a novel drug or intervention 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 is modeled with *fixed* components. The *fixed* component can be understood as a constant value in the response which the researcher is interested in measuring, i.e., the average effect of the novel drug/intervention in the "treatment" group.

Mathematically speaking, a rm-ANOVA model with an interaction can be written as

$$y_{ijt} = \beta_0 + \beta_1 \times treatment_j + \beta_2 \times time_t + \beta_3 \times time_t \times treatment_j + \varepsilon_{ijt}, \tag{1}$$

In this model y_{ijt} is the response for subject i, in treatment group j at time t, which can be decomposed in a mean value β_0, *fixed effects* of treatment ($treatment_j$), time ($time_t$), and their interaction $time_t \times treatment_j$ which have linear slopes given by β_1, β_2 and β_3, respectively. Independent errors ε_{ijt} represent random variation from the sampling process assumed to be $\overset{i.i.d.}{\sim} N(0, \sigma^2)$ (independently and identically normally distributed with mean zero and variance σ^2). In a biomedical research context, suppose two treatments groups are used in a study (e.g., "placebo" vs. "novel drug", or "saline" vs. "chemotherapy"). Then, the group terms in Equation (1) can be written as below with $treatment_j = 0$ representing the first treatment group (Group A) and $treatment_j = 1$ representing the second treatment group (Group B). With this notation, the linear model then can be expressed as

$$y_{ijt} = \begin{cases} \beta_0 + \beta_2 \times time_t + \varepsilon_{ijt} & \text{if Group A,} \\ \beta_0 + \beta_1 + \beta_2 \times time_t + \beta_3 \times time_t + \varepsilon_{ijt} & \text{if Group B.} \end{cases} \tag{2}$$

To further simplify the expression, substitute $\widetilde{\beta_0} = \beta_0 + \beta_1$ and $\widetilde{\beta_1} = \beta_2 + \beta_3$ in the equation for Group B. This substitution allows for a different intercept and slope for Groups A and B. The model is then written as

$$y_{ijt} = \begin{cases} \beta_0 + \beta_2 \times time_t + \varepsilon_{ijt} & \text{if Group A,} \\ \widetilde{\beta_0} + \widetilde{\beta_1} \times time_t + \varepsilon_{ijt} & \text{if Group B.} \end{cases} \tag{3}$$

Presenting the model in this manner makes clear that when treating different groups, an rm-ANOVA model is able

to accommodate non-parallel lines in each case (different intercepts and slopes per group). In other words, the rm-ANOVA model "expects" a linear relationship between the covariates and the response. This means that either presented as Equations (1), (2) or (3), an rm-ANOVA model is only able to accommodate linear patterns in the data. If the data show non-linear trends, the rm-ANOVA model will approximate this behavior with non-parallel lines.

3.2.2 The Linear Mixed Model (LMEM) Case

A LMEM is a class of statistical models that incorporates *fixed effects* to model the relationship between the covariates and the response, and *random effects* to model subject variability that is not the primary focus of the study but that might be important to account for[15,40]. A LMEM with interaction between time and treatment for a longitudinal study can be written as

$$y_{ijt} = \beta_0 + \beta_1 \times treatment_j + \beta_2 \times time_t + \beta_3 \times time_t \times treatment_j + \alpha_{ij} + \varepsilon_{ijt}. \tag{4}$$

When Equations (1) and (4) are compared, it is noticeable that LMEMs and rm-ANOVA have the same construction regarding the *fixed effects* of time and treatment, but that the LMEM incorporates an additional source of variation (the term α_{ij}). This term α_{ij} corresponds to the *random effect*, accounting for variability in each subject (subject$_i$) within each group (group$_j$). The *random* component can also be understood as modeling some "noise" in the response, but that does not arise from the sampling error term ε_{ijt} from Equations (1) through (3).

For example, if the blood concentration of a drug is measured in certain subjects in the early hours of the morning while other subjects are measured in the afternoon, it is possible that the difference in the collection time introduces some "noise" in the data that needs to be accounted for. As the name suggests, this "random" variability needs to be modeled as a variable rather than as a constant value. The random effect α_{ij} in Equation (4) is assumed to be $\alpha_{ij} \sim N(0, \sigma_\alpha^2)$. In essence, the *random effect* in a LMEM enables fitting models with different intercepts at the subject-level[15]. However, the expected linear relationship of the covariates and the response in Equation (1) and in Equation (4) is essentially the same, representing a major limitation of LMEMs to fit a non-linear response.

Of note, LMEMs are capable of fitting non-linear trends using an "empirical" approach (using polynomial fixed effects instead of linear effects such as in Equation (4)), which is described in detail by Pinheiro and Bates[15]. However, polynomial fits have limited predictive power, cause bias on the boundaries of the covariates[36], and more importantly, their lack of biological or mechanistic interpretation limits their use in biomedical studies[15].

3.3 Covariance in rm-ANOVA and LMEMs

In a longitudinal study there is an expected *covariance* 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[41] (For an in-depth analysis of the covariance matrix see West[40] and Weiss[42]).

In the case of an rm-ANOVA analysis, it is typically 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[26,42,43]. However, it has been shown that this condition is frequently not justified because the correlation between measurements tends to change over time[44]; and is higher between consecutive measurements[13,25]. Although corrections can be made (such as Huyhn-Feldt or Greenhouse-Geisser)[26,27] their effectiveness is dependent on sample size and number of repeated measurements[28], and it has been shown that corrections are not robust if the group sizes are unbalanced[29]. Because biomedical longitudinal studies are often limited in sample size and can have an imbalanced design, the corrections required to use an rm-ANOVA model may not be able to 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[15]. Nevertheless, the analysis required to determine an appropriate variance-covariance structure for the data can be a challenging 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.

3.4 Unbalanced data

In a longitudinal study, it is frequently the case that the number of observations is different across subjects. In biomedical research, this imbalance in sample size can be caused by reasons beyond the control of the investigator (such as dropout from patients in clinical studies and attrition or injury of animals in preclinical research) leading to what is known as "missing", "incomplete", or (more generally speaking) unbalanced data[45]. The rm-ANOVA model is very restrictive in these situations as it assumes that observations exist for all subjects at every time point; if that is not the case 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[46].

On the other hand, LMEMs and GAMs can work with missing observations, and inferences from the model are valid when the imbalance in the observations are *missing at random* (MAR) or *completely missing at random* (MCAR)[40,42]. 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[47]. If the data are MCAR, this means that the missingness is completely unrelated to the collected information[48]. Missing observations can also be *missing not at random* (MNAR) and in the case the missing observations are dependent on their value. 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 missingness is unrelated to other variables of interest.

However, it is worth reminding that "all models are wrong"[49] and that the ability of LMEMs and GAMs to work with unbalanced data does not make them immune to problems that can arise due to high rates of incomplete data, such as sampling bias or a drastic reduction in statistical power. Researchers must ensure that the study design is statistically sound and that measures exist to minimize missing observation rates.

3.5 What does the fit of an rm-ANOVA and LMEM look like? A visual representation using simulated data

To visually demonstrate the limitations of rm-ANOVA and LMEMs for longitudinal data with non-linear trends, this section presents a simulation experiment of a normally distributed response of two groups of 10 subjects each. An rm-ANOVA model (Equation (1)), and a LMEM (Equation (4)) are fitted to each group using R[38] and the package *nlme*[50].

Briefly, two cases for the mean response for each group are considered: in the first case, the mean response in each group is a linear function over time with different intercepts and slopes; a negative slope is used for Group 1 and a positive slope is used for Group 2 (Figure 1A). In the second case, a second-degree polynomial (quadratic) function is used for the mean response per group: the quadratic function is concave down for Group 1 and it is concave up for Group 2 (Figure 1D). In both the linear and quadratic simulated data, the groups start with the same mean value in order to simulate the expected temporal evolution of some physiological quantity, starting at a common initial condition.

Specifically, the rationale for the chosen linear and quadratic functions is the expectation that a measured response in two treatment groups is similar in the initial phase of the study, but as therapy progresses a divergence in the trend of the response indicates a treatment effect. In other words, Group 1 can be thought as a "Control" group and Group 2 as a "Treatment" group. From the mean response per group (linear or quadratic), the variability or "error" of individual responses within each group is simulated using a covariance matrix with compound symmetry (constant variance across time). Thus, the response per subject at each timepoint in both the linear and quadratic simulation corresponds

to the mean response per group plus the error (represented by the points in Figure 1 A, D).

A more comprehensive exploration of the fit of rm-ANOVA and LMEMs for linear and non-linear longitudinal data can be obtained from the code that appears in Appendix B, (Figures B.1 and B.2), where a simulation with compound symmetry and independent errors (errors generated from a normal distribution that are not constant over time) is presented. We are aware that the simulated data used in this section present an extreme case that might not occur frequently in biomedical research, but they are used to 1) present the consequences of modeling non-linear trends in data with a linear model such as rm-ANOVA or a LMEM with "default" (linear) effects and, 2) demonstrate that a visual assessment of model fit is an important tool that helps determine the validity of any statistical assumptions. In Section 5 we use simulated data that does follow reported trends in the biomedical literature to implement GAMs.

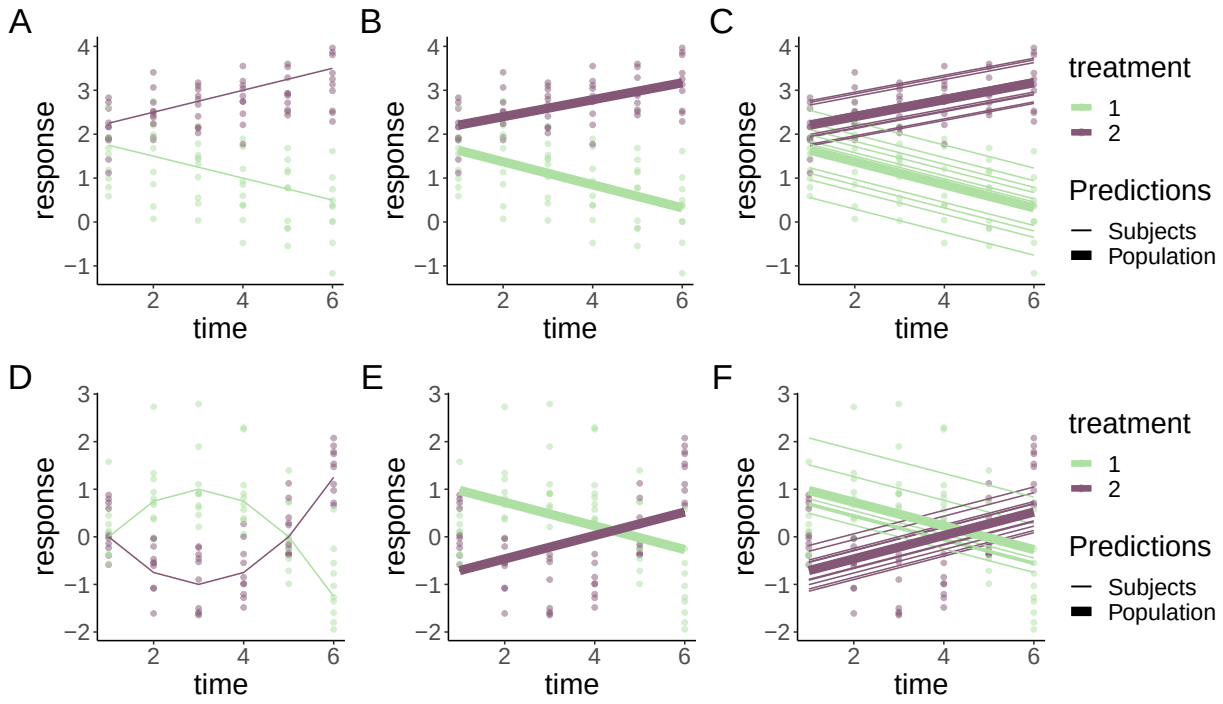

Figure 1: Simulated responses from two groups with correlated errors using a LMEM and a rm-ANOVA model. Top row: linear response, bottom row: quadratic response. **A**: Simulated linear data with known mean response (thick lines) and individual responses (points) showing the dispersion of the data. **D**: Simulated quadratic data with known mean response (thick lines) and individual responses (points) showing the dispersion of the data. **B,E**: Estimates from the rm-ANOVA model for the mean group response (linear of quadratic). Points represent the original raw data. The rm-ANOVA model not only fails to pick the trend of the quadratic data (E) but also assigns a global estimate that does not take into account the between-subject variation. **C, F**: Estimates from the LMEM in the linear and quadratic case (subject: thin lines, population: thick lines) . The LMEM incorporates a random effect for each subject, but this model and the rm-ANOVA model are unable to follow the trend of the data and grossly bias the initial estimates for each group in the quadratic case (bottom row).

The simulation shows that the fits produced by the LMEM and the rm-ANOVA model are good for linear data (1B), as the predictions for the mean response are reasonably close to the "truth" of the simulated data (Figure 1A). Note that because the LMEM incorporates *random effects*, is able to provide estimates for each subject and a "population"

estimate (Figure 1C).

However, consider the case when the data follows a non-linear trend, such as the simulated data in Figure 1D. Here, the mean response per group was simulated using a quadratic function, and errors and individual responses were produced as in Figure 1A. The mean response in the simulated data with quadratic behavior changes in each group through the timeline, and the mean value is the same as the initial value by the fifth time point for each group. Fitting an rm-ANOVA model (Equation (1)) or a LMEM (Equation (4)) to this data produces the fit that appears in Figure 1E, F.

Comparing the fitted responses of the LMEM and the rm-ANOVA models used in the simulated quadratic data (Figure 1E, F) indicates that the models are not capturing the changes within each group. Specifically, note that the fitted mean response of both models shows that the change (increase for Treatment 1 or decrease for Treatment 2) in the response through time points 2 and 4 is not being captured.

The LMEM is only able to account for between-subject variation by providing estimates for each subject (Figure 1F), but both models are unable to capture the fact that the initial values are the same in each group, and instead fit non-parallel lines that have initial values that are markedly different from the "true" initial values in each case (compare Figure 1D with Figure 1E, F). If such a change has important physiological implications, both rm-ANOVA and LMEMs omit it from the fitted mean response. Thus, even though the model correctly detects a divergence between treatment groups, the exact nature of this difference is not correctly identified, limiting valuable inferences from the data. It could be argued that a LMEM with quadratic effects should have been used to fit the data in Figure1F. However, because in reality the true function is not known, choosing a polynomial degree causes more questions (e.g., is it quadratic?, cubic?, or a higher degree?). Additionally, polynomial effects have other limitations, which we cover in Section 4.2.1.

This section has used simulation to better convey and visualize the limitations of linearity and correlation in the response in data with non-linear trends using an rm-ANOVA model and a LMEM, where the main issue is the expected linear trend in the response. Notice that the model misspecification is easily noticeable if the model fit and the response are visualized. In the following section, we provide a brief overview of linear models, general linear models and generalized linear mixed models before presenting the theory of GAMs, a class of semi-parametric models that can fit non-linear trends in data and that overcome the limitations of rm-ANOVA and LMEMs in the analysis of biomedical longitudinal data.

4 Linear Models, and beyond

Linear models (LMs) are those that assume a normal (Gaussian) distribution of the errors, and only incorporate *fixed effects* (such as rm-ANOVA). These are by far the models most commonly used to analyze data within the biomedical research community. On the other hand, Linear Mixed Effect Models (LMEMs) also incorporate *random effects*, as it has been described in Section 3.2.2.

In reality, rm-ANOVA and LMEMs are just *special cases* of a broader class of models (General Linear Models and Generalized Linear Mixed Models, respectively). In order to fully capture the constraints of such models and to understand how GAMs overcome those limitations this section will briefly provide an overview of the different classes of models and indicate how rm-ANOVA, LMEMs, and GAMs fit within this framework.

4.1 Generalized Linear Models (GLMs)

A major limitation of LMs is their assumption of normality in the errors. If the residuals are non-normal, a transformation is necessary in order to properly fit the model. However, transformation can lead to poor model performance[51], and can cause problems with the biological interpretation of the model estimates. McCullagh and Nelder[52] introduced General Linear Models (GLMs) as an extension of LMs, where the errors do not need to be normally distributed. To achieve this, consider the following model

$$y_{ijt} \sim \mathscr{D}(\mu_{ijt}, \phi), \tag{5}$$

where y_{ijt} is the observation i in group j at time t, that is assumed to come from some distribution of the exponential family $\mathscr{D}$, with some mean μ_{ijt}, and potentially, a dispersion parameter ϕ (which in the Gaussian case is the variance σ^2). The mean (μ_{ijt}) is also known as the *expected value* (or *expectation*) $E(y_{ijt})$ of the observed response y_{ijt}.

Then, the *linear predictor* η, which defines the relationship between the mean and the covariates can be defined as

$$\eta_{ijt} = \beta_0 + \beta_1 \times treatment_j + \beta_2 \times time_t + \beta_3 \times time_t \times treatment_j, \tag{6}$$

where η_{ijt} is the linear predictor for each observation i in each group j, at each timepoint t. Following the notation from Equation (1) the model parameters for each group are β_0 (the intercept), β_1, β_2, and β_3; $time_t$ represents the covariates from each subject in each group at each time point, $treatment_j$ represents the different treatment levels, and $time_t \times treatment_j$ represents their interaction.

Finally,

$$E(y_{ijt}) = \mu_{ijt} = g^{-1}(\eta_{ijt}), \tag{7}$$

where $E(y_{ijt})$ is the expectation, and g^{-1} is the inverse of a *link function* (g). The link function transforms the values from the response scale to the scale of the linear predictor η (Equation (6)). Therefore, it can be seen that LMs (such as rm-ANOVA) are a special case of GLMs where the response is normally distributed.

4.2 Generalized linear mixed models (GLMMs)

Although GLMs relax the normality assumption, they only accommodate fixed effects. Generalized Linear Mixed Models (GLMMs) are an extension of GLMs that incorporate *random effects*, which have an associated probability distribution[53]. Therefore, in GLMMs the linear predictor takes the form

$$\eta_{ijt} = \beta_0 + \beta_1 \times treatment_j + \beta_2 \times time_t + \beta_3 \times time_t \times treatment_j + \alpha_{ij}, \tag{8}$$

where α_{ij} corresponds to the random effects that can be estimated within each subject in each group, and all the other symbols correspond to the notation of Equation (6). Therefore, LMEMs are special case of GLMMs where the distribution of the response is normally distributed[52], and GLMs are a special case of GLMMs where there are no random effects. In-depth and excellent discussions about LMs, GLMs and GLMMs can be found in Dobson[54] and Stroup[55].

4.2.1 GAMs as a special case of Generalized Linear Models

4.2.1.1 GAMs and Basis Functions Notice that in the previous sections, the difference between GLMs and GLMMs resides on their linear predictors (Equations (6), (8)). Generalized additive models (GAMs) are an extension of the GLM family that allow the estimation of smoothly varying trends where the relationship between the covariates and the response is modeled using *smooth functions*[34,37,56]. In a GAM, the linear predictor has the form

$$\eta_{ijt} = \beta_0 + \beta_1 \times treatment_j + f(time_t|\beta_j), \tag{9}$$

where β_0 is the intercept, and β_1 is the coefficient for each treatment group. Notice that the construction of the predictor is similar to that of Equation (6), but in this case the parametric terms involving the effect of time, and the interaction

between time and treatment have been replaced by the smooth term $f(time_t|\beta_j)$. The smooth term $f(time_t|\beta_j)$ gives a different smooth response for each treatment. [1] A GAM version of a linear model can be written as

$$y_{ijt} = \beta_0 + \beta_1 \times treatment_j + f(time_t \mid \beta_j) + \varepsilon_{ijt}, \tag{10}$$

where y_{ijt} is the response at time t of subject i in group j, and ε_{ijt} represents the deviation of each observation from the mean.

In contrast to the linear functions used to model the relationship between the covariates and the response in rm-ANOVA or LMEM, the use of smooth functions in GAMs is advantageous as it does not restrict the model to a linear relationship, although a GAM can estimate a linear relationship if the data is consistent with a linear response. One possible set of functions for $f(time_t \mid \beta_j)$ that allow for non-linear responses are polynomials (which can also be used in LMEMs), but a major limitation is that polynomials create a "global" fit as they assume that the same relationship exists everywhere, which can cause problems with inference[36]. In particular, polynomial fits are known to show boundary effects because as t goes to $\pm\infty$, $f(time_t \mid \beta_j)$ goes to $\pm\infty$ which is almost always unrealistic and causes bias at the endpoints of the time period.

The smooth functional relationship between the covariates and the response in GAMs is specified using a semi-parametric relationship that can be fit within the GLM framework, using a *basis function* expansion of the covariates and estimating random coefficients associated with these basis functions. A *basis* is a set of functions that spans the mathematical space within which the true but unknown $f(time_t)$ is thought to exist[34]. For the linear model in Equation (1), the basis coefficients are β_1, β_2 and β_3 and the basis vectors are $treatment_j$, $time_t$, and $time_t \times treatment_j$. The basis function then, is the linear combination of basis coefficients and basis vectors that map the possible relationship between the covariates and the response[57], which in the case of Equation (1) is restricted to a linear family of functions. In the case of Equation (10), the basis functions are contained in the expression $f(time_t \mid \beta_j)$, which means that the model allows for non-linear relationships among the covariates.

Splines (which derive their name from the physical devices used by draughtsmen to draw smooth curves) are commonly used as basis functions that have a long history in solving semi-parametric statistical problems and are often a default choice to fit GAMs as they are a simple, flexible, and powerful option to obtain smoothness[58]. Although different types of splines exist, cubic, thin plate splines, and thin plate regression splines will be briefly discussed next to give a general idea of these type of basis functions, and their use within the GAM framework.

Cubic splines (CS) are smooth curves constructed from cubic polynomials joined together in a manner that enforces

[1] If the smooth term represented a linear relationship, then $f(time_t|\beta_j) = \beta_2 \times time_t + \beta_3 \times time_t \times treatment_j$; however, in general, the smooth term is a more flexible function than a linear relationship, with parameter vectors β_j for each treatment

smoothness. The use of CS as smoothers in GAMs was discussed within the original GAM framework[56], but they are limited by the fact that their implementation requires the selection of some points along the covariates (known as 'knots', the points where the basis functions meet) to obtain the finite basis, which affects model fit[59]. A solution to the knot placement limitation of CS is provided by thin plate splines (TPS), which provide optimal smooth estimation without knot placement, but that are computationally costly to calculate[37,59]. In contrast, thin plate regression splines (TPRS) provide a reasonable "low rank" (truncated) approximation to the optimal TPS estimation, which can be implemented in an computationally efficient[59]. Like TPS, TPRS only requires specifying the number of basis functions to be used to create the smoother (for mathematical details on both TPS and TPRS see Wood[37,59]).

To further clarify the concept of basis functions and smooth functions, consider the simulated response for Group 1 that appears in Figure 1D. The simplest GAM model that can be used to estimate such response is that of a single smooth term for the time effect; i.e., a model that fits a smooth to the trend of the group through time. A computational requisite in *mgcv* is that the number of basis functions to be used to create the smooth cannot be larger than the number of unique values from the independent variable. Because the data has six unique time points, we can specify a maximum of six basis functions (including the intercept) to create the smooth. It is important to note that is not necessary to specify a number of basis equal to the number of unique values in the independent variable; fewer basis functions can be specified to create the smooth as well, as long as they reasonably capture the trend of the data.

Here, the main idea is that the resulting smooth matches the data and approximates the true function without becoming too "wiggly" due to the noise present. A detailed exploration of wiggliness and smooth functions is beyond the scope of this manuscript, but in essence controlling the wiggliness (or "roughness") of the fit is achieved by using a *smoothness parameter* (λ), which is used to penalize the likelihood by multiplying it with the integrated square of the second derivative of the spline smooth. The second derivative of the spline smooth is a measure of curvature, or the rate of change of the slope[34,37], and increasing the penalty by increasing λ results in models with less curvature. As λ increases, the parameter estimates are penalized (shrunk towards 0) where the penalty reduces the wiggliness of the smooth fit to prevent overfitting. In other words, a low penalty estimate will result in wiggly functions whereas a high penalty estimate provides evidence that a linear response is appropriate.

With this in mind, if four basis functions (plus the intercept) are used to fit a GAM for the data of Group 1 (concave down trend) that appears in Figure 1D, the resulting fitting process is shown in Figure 2. In Figure 2A the four basis functions (and the intercept) are shown. Each of the five basis functions is evaluated at six different points (because there are six points on the timeline). The coefficients for each of the basis functions of Figure 2A are estimated using a penalized regression with smoothness parameter λ, that is estimated when fitting the model. The penalized coefficient estimates fitted are shown in Figure 2B.

To get the weighted basis functions, each basis (from Figure 2A) is multiplied by the corresponding coefficients in

Figure 2B, thereby increasing or decreasing the original basis functions. Figure 2C shows the resulting weighted basis functions. Note that the magnitude of the weighting for the first basis function has resulted in a decrease of its overall contribution to the smoother term (because the coefficient for that basis function is negative and its magnitude is less than one). On the other hand, the third basis function has roughly doubled its contribution to the smooth term. Finally, the weighted basis functions are added at each timepoint to produce the smooth term. The resulting smooth term for the effect of *time* is shown in Figure 2D (brown line), along the simulated values which appear as points.

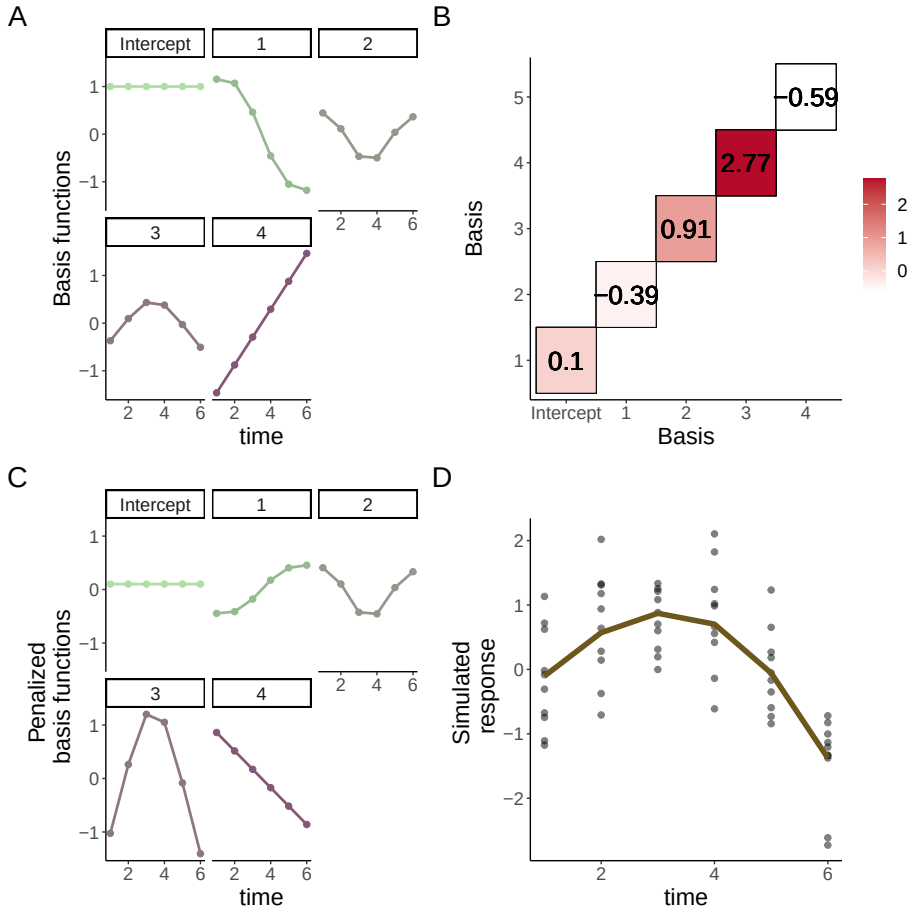

Figure 2: Basis functions for a single smoother for time. **A**: Basis functions for a single smoother for time for the simulated data of Group 1 from Figure 2. **B**: Matrix of basis function weights. Each basis function is multiplied by a coefficient which can be positive or negative. The coefficient determines the overall effect of each basis in the final smoother. **C**: Weighted basis functions. Each of the four basis functions (and the intercept) of panel A has been weighted by the corresponding coefficient shown in Panel B. Note the corresponding increase (or decrease) in magnitude of each weighted basis function. **D**: Smoother for time and original data points. The smoother (line) is the result of the sum of each weighted basis function at each time point, with simulated values for Group 1 shown as points.

4.2.2 A Bayesian interpretation of GAMs

Bayes' theorem states that the probability of an event can be calculated using prior knowledge and observed data[60]. In the case of data that shows non-linear trends, the prior that the *true* trend of the data is likely to be smooth rather than "wiggly" introduces the concept of a prior distribution for wiggliness (and therefore a Bayesian view) of GAMs[37]. Moreover, GAMs are considered "empirical" Bayesian models when fitted using the package *mgcv* because the smoothing parameters are estimated from the data (and not from a posterior distribution as in the "fully Bayesian" case, which can be fitted using JAGS, Stan, or other probabilistic programming language)[61]. Therefore, the confidence intervals (CIs) calculated by default for the smooth terms using *mgcv* are considered empirical Bayesian credible intervals[33], which have good *across the function* ("frequentist") coverage[37].

To understand across the function coverage, recall that a CI provides an estimate of the region where the "true" or "mean" value of a function exists, taking into account the randomness introduced by the sampling process. Because random samples from the population are used to calculate the "true" value of the function, there is inherent variability in the estimation process and the CI provides a region with a nominal value (usually, 95%) where the function is expected to lie. In an *across the function* CI (like those estimated by default for GAMs using *mgcv*), if we average the coverage of the interval over the entire function we get approximately the nominal coverage (95%). In other words, we expect that about 95% of the points that compose the true function will be covered by the across the function CI. As a consequence, some areas of the CI for the function have more than nominal coverage and some areas less than the nominal coverage.

Besides the across the function CI, "simultaneous" or "whole function" CIs can also be computed, which contain the *whole function* with a specified probability[37]. Suppose we chose a nominal value (say, 95%) and compute a simultaneous CI; if we obtain 100 repeated samples and compute a simultaneous CI in each case, we would expect that the true function lies completely within the computed simultaneous CI in 95 of those repeated samples. Briefly, to obtain a simultaneous CI we simulate 10,000 draws from the empirical Bayesian posterior distribution of the fitted smooths. Then, we obtain the maximum absolute standardized deviation of the differences in smooth estimates which is used to correct the coverage of the across the function CI[62] in a similar fashion to how *q-values* correct *p-values* to control false positive discovery rates[63].

In-depth theory of the Bayesian interpretation of GAMs and details on the computation of simultaneous and across the function CIs are beyond the scope of this paper, but can be found in Miller[61], Wood[37], Simpson[34], Marra[64], and Ruppert[62]. What we want to convey is that a Bayesian interpretation of GAMs allows for robust estimation using simultaneous empirical Bayesian CIs, as their estimates can be used to make comparisons between different groups in a similar way that multiple comparisons adjustments make inference from ANOVA models more reliable.

With this in mind, in the next section we consider the use of GAMs to analyze longitudinal biomedical data with non-linear trends, and use simultaneous empirical Bayesian CIs to assess significance between treatment groups.

5 The analyisis of longitudinal biomedical data using GAMs

The previous sections provided the basic understanding of the GAM framework and how these models are more advantageous to analyze non-linear longitudinal data when compared to rm-ANOVA or LMEMs. This section will use simulation to present the practical implementation of GAMs for longitudinal biomedical data using R and the package `mgcv`. A brief guide for model selection and diagnostics appears in Appendix A, and the code for the simulated data and figures can be found in Appendix B.

5.1 Simulated data

The simulated data is based on the reported longitudinal changes in oxygen saturation (StO_2) in subcutaneous tumors (Figure 3C in Vishwanath et. al.[16]), where diffuse reflectance spectroscopy was used to quantify StO_2 changes in both groups at the same time points (days 0, 2, 5, 7 and 10). In the "Treatment" group (chemotherapy) an increase in StO_2 is observed through time, while a decrease is seen in the "Control" (saline) group. Following the reported trend, we simulated 10 normally distributed observations at each time point with a standard deviation (SD) of 10% (matching the SD in the original paper). The simulation based on the real data appears in Figure 3A.

5.2 An interaction GAM for longitudinal data

An interaction effect is typically the main interest in longitudinal biomedical data, as the interaction takes into account treatment, time, and their combination. In a practical sense, when a GAM is implemented for longitudinal data, a smooth can be added to the model for the *time* effect for each treatment to account for the repeated measures over time. Although specific methods of how GAMs model correlation structures is a topic beyond the scope of this paper, it suffices to say that GAMs are flexible and can handle correlation structures beyond compound symmetry. A detailed description on the close relationship between basis functions and correlation functions can be found in Hefley et. al.[57].

For the data in Figure 3A, the main effect of interest is how StO_2 changes over time for each treatment. To estimate this, the model incorporates separate smooths for each *Group* as a function of *Day*. The main thing to consider is that model syntax accounts for the fact that one of the variables is numeric (*Day*) and the other is a factor (*Group*). Because the smooths are centered at 0, the factor variable needs to be specified as a parametric term in order to identify any differences between the group means. Using R and the package `mgcv` the model syntax is:

```
gam_02 <- gam(StO2_sim ~ Group + s(Day, by = Group, k = 5), method ='REML'
    , data = dat_sim)
```

This syntax specifies that `gam_02` (named this way so it matches the model workflow from Appendix A) contains the fitted model, and that the change in the simulated oxygen saturation (`StO2_sim`) is modeled using independent smooths over `Day` for each `Group` (the parenthesis preceded by `s`) using four basis functions (plus intercept). The smooth is constructed by default using TPRS, but other splines can be used if desired, including Gaussian process smooths[34] (a description of all the available smooths can be found by typing `?mgcv::smooth.terms` in the Console). Finally, the parametric term `Group` is added to quantify overall mean differences in the effect of treatment between groups, as we have indicated above.

Although the default `method` used to estimate the smoothing parameters in *mgcv* is generalized cross validation (GCV), Wood[37] showed the restricted maximum likelihood (REML) to be more resistant to overfitting while also easing the quantification of uncertainty in the smooth parameters; therefore in this manuscript REML is always used for smooth parameter estimation. An additional argument (`family`) allows to specify the expected distribution of the response, but it is not used in this model because we expect a normally-distributed response (which is the default `family` in *mgcv*).

When the smooths are plotted over the raw data, it is clear that the model has been able to capture the trend of the change of StO_2 for each group across time (Figure 3B). Model diagnostics can be obtained using the `gam.check` function, and the function `appraise` from the package *gratia*[65] as we show in Appendix A. Additional discussions on model selection can be found in Wood[37] and Harezlak[66].

One question that might arise at this point is "what is the fit that an rm-ANOVA model produces for the simulated data?". The fit of an rm-ANOVA model, which corresponds to Equation (1), is presented in Figure 3C. This is a typical case of model misspecification: The slopes of each group are different, which would lead to a *p-value* indicating significance for the treatment and time effects, but the model is not capturing the changes that occur at days 2 and between days 5 and 7, whereas the GAM model is able to reliably estimate the trend over all timepoints (Figure 3B) .

Because GAMs do not require equally-spaced or complete observations for all subjects (as rm-ANOVA does), they are advantageous to analyze longitudinal data where unbalanced data exists. The rationale behind this is that GAMs are able to pick the trend in the data even when some observations are incomplete. However, this usually causes the resulting smooths to have wider confidence intervals and less ability to discern differences in trends. To exemplify this, consider the random deletion of 40% of the simulated StO_2 values from Figure 3A. If the same interaction GAM (`gam_02`) is fitted to this data with unbalanced observations, the resulting smooths appear in (Figure 3D). Note that

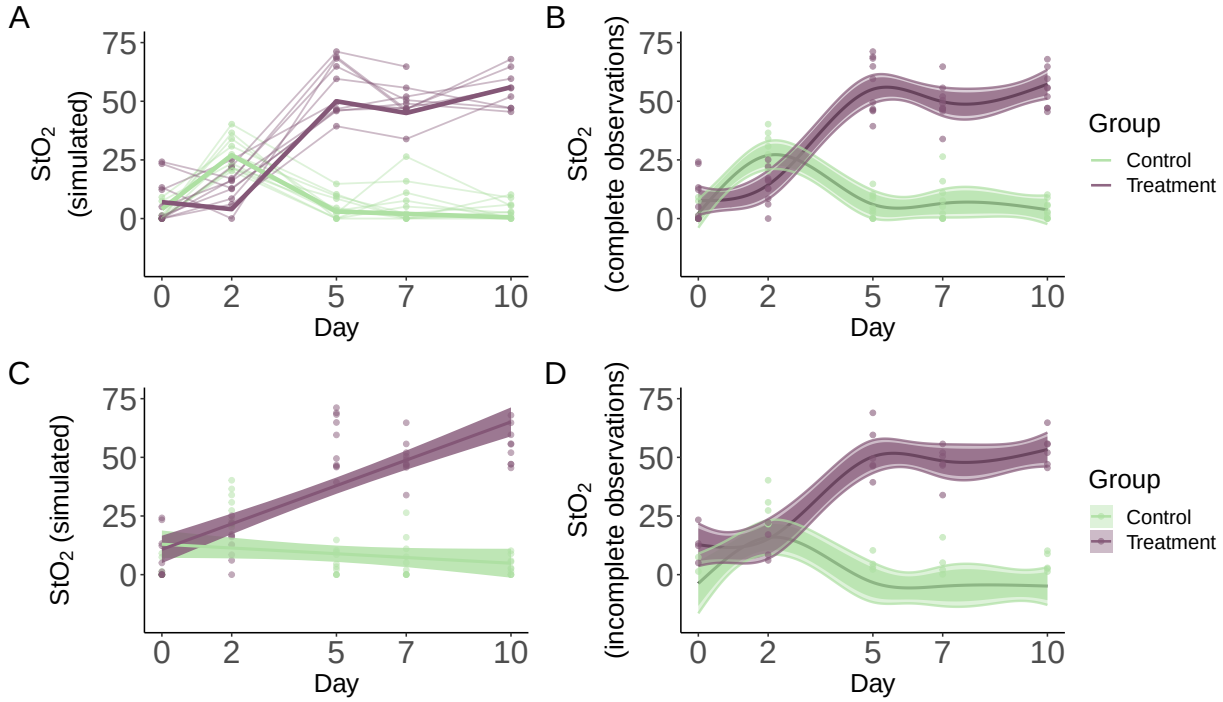

Figure 3: Simulated data and smooths for oxygen saturation in tumors. **A**: Simulated data (thin lines) that follows previously reported trends (thick lines) in tumors under chemotherapy (Treatment) or saline (Control) treatment. Simulated data is from a normal distribution with standard deviation of 10% with 10 observations per time point. **B**: Smooths from the GAM model for the full simulated data with interaction of Group and Treatment. Lines represent trends for each group, shaded regions are 95% across the function (narrow region) and simultaneous (wide region) confidence intervals. **C**: The rm-ANOVA model for the simulated data, which does not capture the changes in each group over time. **D**: Smooths for the GAM model for the simulated data with missing observations (40%). Lines represent trends for each group, shaded regions are 95% across the function (narrow region) and simultaneous (wide region) confidence intervals.

the model is still able to show a different trend for each group, but with a somewhat more linear profile in some areas.

Additionally, note that in Figure 3B,D we show two CIs for each of the fitted smooths (shaded regions). The across the function CIs are represented by the narrow regions and because the simultaneous CIs contain the whole function on a nominal value, they are wider than the across the function CI, resulting in the wide shaded regions. For the dataset with incomplete observations, the CIs for the smooths overlap during the first 3 days because the estimates are less robust with fewer data points, and the trend is less pronounced than in the full dataset. However, the overall trend of the data is picked by the model in both cases, with as few as 4 observations per group at certain time points.

5.3 Determination of significance in GAMs for longitudinal data

At the core of a biomedical longitudinal study lies the question of a significant difference between the effect of two or more treatments in different groups. In linear models (such as rm-ANOVA), if there is a significant *p-value* after a *post-hoc* analysis we then can make inference about the effect size using the slope or the intercept from the model. In GAMs however, there is no single *p-value* to determine the significance of an effect as in linear models. Therefore, the coefficients of GAMs do not provide a simple interpretation as in the linear model case, but the changes in slope at specific timepoints can be used to determine the instantaneous effect size. In essence, the idea behind the estimation of significance in GAMs across different treatment groups is that the difference between the separate smoothers per group (such as in `gam_02`) can be computed pairwise, followed by the estimation of an empirical Bayesian simultaneous CI around this difference.

The pairwise difference in smooths can be conceptualized in the following manner: Different time trends in each group are an indication of an effect by the treatment as in Figure 3A, where the chemotherapy causes StO_2 to increase over time. With this expectation of different trends in each group, computing the difference between the trends will identify if the observed difference is significant. A difference between groups with similar trends is unlikely to be distinguishable from zero, which would indicate that the treatment is not causing a change in the response in one of the groups (assuming the other group is a Control or Reference group). Therefore, the computation of both the difference between the smooths and the corresponding simultaneous empirical Bayesian CI around this difference is able to provide an estimation of when and by how much there is a significant difference between the different groups. Additionally, the correction provided by the simultaneous empirical Bayesian CI makes the estimation robust as we know that on average, the simultaneous CI will contain the *whole function* at a nominal value (say, 95%).

To visualize this, consider the calculation of pairwise differences for the fitted smooths that appear in Figure 3B, D. Figure 4 shows the difference between each treatment group trend for the full and missing datasets with a simultaneous CI computed around the difference. Here, the “Control” group is used as the reference to which “Treatment” group

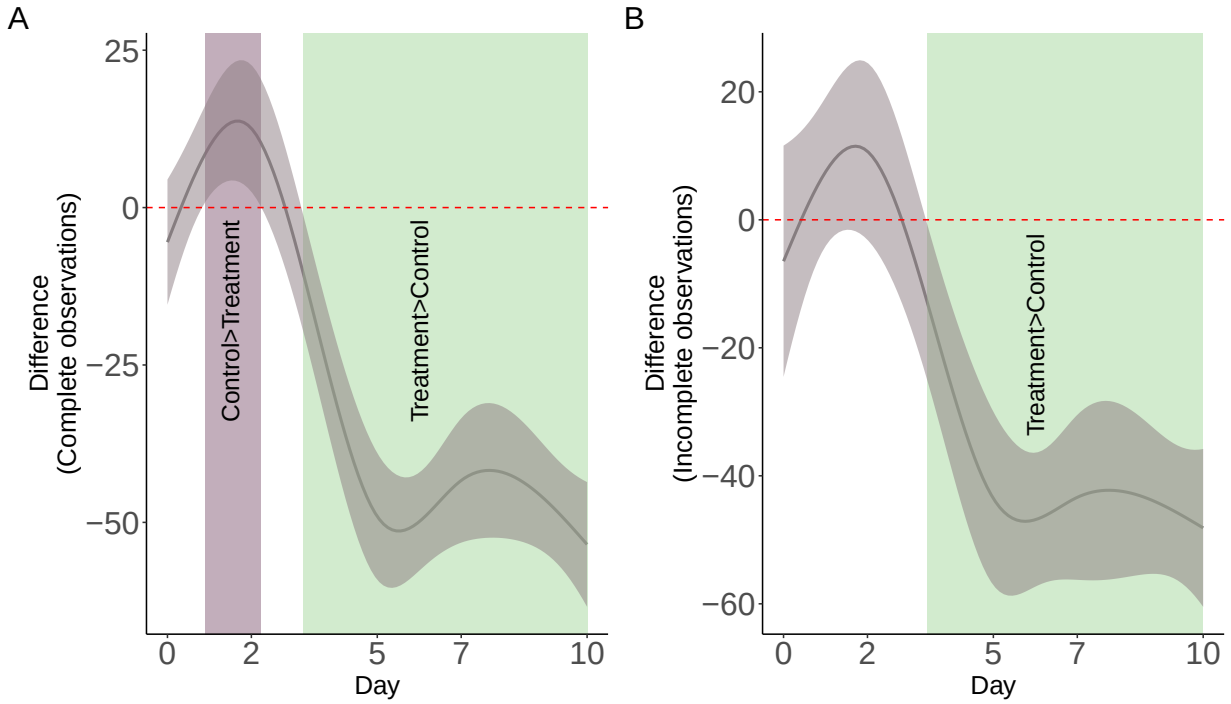

Figure 4: Pairwise comparisons for smooth terms. **A**: Pairwise comparisons for the full dataset. **B**: Pairwise comparisons for the dataset with incomplete observations. Significant differences exist where the 95% empirical Bayesian simultaneous CI does not cover 0. In both cases the effect of treatment is significant after day 3. For the difference, we have included the means so the value of the difference has direct correspondence with the scale of the response (Figure 3A).

is being compared. Notice that because we have included the means of each group, there is correspondence between the scale of the original data and the scale of the pairwise comparisons. This can be seen in Figure 3B, where at day 5 there is essentially a difference of 50% between StO_2 in both groups, which corresponds to the -50% difference in 4A. However, if there are multiple parametric terms in the model (more factors that need to be specified), such inclusion of the means can become problematic. However, we believe that the model we have presented here suffices in a wide range of situations where adding the group means is a relatively easy implementation that can help better visualize the estimation from the model from a biological perspective.

In Figure 4A, the shaded regions over the confidence interval (where the CI does not cover 0) indicate the time interval where each group has a different mean effect than the other. Notice that the shaded region between days 1 and ≈ 2 for the full dataset indicates that through that time, the "Control" group has higher mean StO_2, but as therapy progresses the effect is reversed and by day ≈ 3 it is the "Treatment" group that statistically on average has greater StO_2. This would suggest that the effect of chemotherapy in the "Treatment" group becomes significant after day 3 for the given model. Moreover, notice that although there is no actual measurement at day 3, the model is capable of providing an estimate of when the shift in mean StO_2 occurs.

On the data with missing observations (Figure 3D), the smooth pairwise comparison (Figure 4B) shows that because the confidence intervals overlap zero for the first two days there is no statistically significant difference between the groups. However, because the model is still able to pick the overall trend in StO_2, the pairwise comparison is able to estimate the change on day 3 where the mean difference between groups becomes statistically significant as in the full dataset smooth pairwise comparison.

For biomedical studies, the ability of smooth comparisons to provide an estimate of *when* and by *how much* a biological process becomes significant is advantageous because it can help researchers gain insight on metabolic changes and other biological processes that can be worth examining, and can help refine the experimental design of future studies in order to obtain measurements at time points where a significant change might be expected.

6 Discussion

Biomedical longitudinal data is particularly challenging to analyze due to the frequency of incomplete observations and different correlation structures in the data, which limit the use of rm-ANOVA. Although LMEMs can handle unbalanced observations and different correlation structures, both LMEMs and rm-ANOVA yield biased estimates when they are used to fit data with non-linear trends as we have visually demonstrated in Section 3.5. When these models do not capture the non-linear trend of the data, this results in a "model misspecification error". This "model

misspecification" error, also is known as a "Type III" error[17] is particularly important because although the *p-value* is the common measure of statistical significance, the validity of its interpretation is determined by the agreement of the data and the model.

Guidelines for statistical reporting in biomedical journals exist (the SAMPL guidelines)[67] but they have not been widely adopted and in the case of longitudinal data, we consider that researchers would benefit from reporting a visual assessment of the correspondence between the model fit and the data, rather than merely relying on a R^2 or *p-value* value, whose interpretation is not clear in the case of a Type III error.

In this paper we have presented GAMs as a suitable method to analyze longitudinal data with non-linear trends. It is interesting to note that although GAMs are a well established method to analyze temporal data in different fields (e.g., which are palaeoecology, geochemistry, and ecology)[33,57] they are not routinely used in biomedical research despite an early publication from Hastie and Tibshirani that demonstrated their use in medical research[68]. This is possibly due to the fact that the theory behind GAMs can seem very different from that of rm-ANOVA and LMEMs.

However, in Section 4.2.1 we demonstrated that at its core the principle underlying GAMs is quite simple: Instead of using a linear relationship to model the response (as rm-ANOVA and LMEMs do), GAMs use basis functions to build smooths that are capable of learning non-linear trends in the data. The use of basis functions is a major advantage over models where the user has to know the non-linear relationship *a priori*, such as in the case of polynomial effects in LMEMs where in addition, there is no biological interpretation for such polynomial assumptions. This does not mean, however, that as any other statistical model GAMs do not have certain limitations. In particular, beyond the range of the data GAMs only reflect the assumptions built into the basis functions, be that flat values or linear extrapolation of the slope. Therefore, researchers need to be careful when using GAMs for extrapolating purposes. In addition, both polynomial and GAMs show higher variance in estimates near the boundary of the data, but additive models generally have less variance than polynomials. However, because GAMs let the data speak for themselves, they provide estimates that are consistent with non-linear trends and therefore can be used to obtain an accurate representation of the effect of time in a biological process.

Beyond the theory, from a practical standpoint is equally important to demonstrate how GAMs are computationally implemented. We have provided an example on how GAMs can be fitted using simulated data that follows trends reported in biomedical literature[16] using R and the package *mgcv*[37] in Section 5, while a basic workflow for model selection is in Appendix A.

One of the features of GAMs is that their Bayesian interpretation allows for inference about differences between groups without the need of a *p-value*, thereby providing a time-based estimate of shifts in the response that can be directly tied to biological values as the pairwise smooth comparisons in Figure 4 indicate. The GAM is therefore

able to identify changes between the groups at time points where data was not directly measured and in the case of incomplete observations ($\approx$ day 3 in Figure 4 A, B). This more nuanced inference can be used by researchers as feedback on experiment design and to further evaluate important biological changes in future studies.

We have used R as the software of choice for this paper because it provides a fully developed environment to fit GAMs, enables simulation (which is becoming increasingly used for exploratory statistical analysis and power calculations), and provides powerful and convenient methods of visualization, which are key aspects that biomedical researchers might need to consider to make their work more reproducible. In this regard, reproducibility is still an issue in biomedical research[69,70], but it is becoming apparent that what other disciplines have experienced in this aspect is likely to impact this field sooner rather than later. Researchers need to plan on how they will make their data, code, and any other materials open and accessible as more journals and funding agencies recognize the importance and benefits of open science in biomedical research. We have made all the data and code used in this paper accessible, and we hope that this will encourage other researchers to do the same with future projects.

7 Conclusion

We have presented GAMs as a method to analyze longitudinal biomedical data. Future directions of this work will include simulation-based estimations of statistical power using GAMs, as well as demonstrating the prediction capabilities of these models using large datasets. By making the data and code used in this paper accessible, we hope to address the need of creating and sharing reproducible work in biomedical research.

8 Acknowledgements

This work was supported by the National Science Foundation Career Award (CBET 1751554, TJM) and the Arkansas Biosciences Institute.

9 Declaration of Conflicting Interests

The Authors declare that there is no conflict of interest.

Supplementary Materials

Two Appendices which contain a basic workflow to implement GAMs in R and all the code used to create this manuscript are available as PDFs in the Supplementary Material. A GitHub repository containing all the code used for this paper along with detailed instructions for its use is available at https://github.com/aimundo/GAMs-biomedical-research.

---

10 References

1. Roblyer D, Ueda S, Cerussi A, 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*. 2011;108(35):14626-14631. doi:10.1073/pnas.1013103108
2. Tank A, Peterson HM, Pera V, et al. Diffuse optical spectroscopic imaging reveals distinct early breast tumor hemodynamic responses to metronomic and maximum tolerated dose regimens. *Breast Cancer Research*. 2020;22(1). doi:10.1186/s13058-020-01262-1
3. Pavlov MV, Kalganova TI, Lyubimtseva YS. Multimodal approach in assessment of the response of breast cancer to neoadjuvant chemotherapy. *Journal of Biomedical Optics*. 2018;23(09):1. doi:10.1117/1.jbo.23.9.091410
4. Demidov V, Maeda A, Sugita M, et al. Preclinical longitudinal imaging of tumor microvascular radiobiological response with functional optical coherence tomography. *Scientific Reports*. 2018;8(1). doi:10.1038/s41598-017-18635-w
5. Ritter G, Cohen L, Williams C, Richards E, Old L, Welt S. Serological analysis of human anti-human antibody responses in colon cancer patients treated with repeated doses of humanized monoclonal antibody A33. *Cancer Research*. 2001;61(18):6851-6859.
6. Roth EM, Goldberg AC, Catapano AL, et al. Antidrug antibodies in patients treated with alirocumab. *New England Journal of Medicine*. 2017;376(16):1589-1590. doi:10.1056/nejmc1616623
7. Jones JD, Ramser HE, Woessner AE, Quinn KP. In vivo multiphoton microscopy detects longitudinal metabolic changes associated with delayed skin wound healing. *Communications Biology*. 2018;1(1). doi:10.1038/s42003-018-0206-4
8. Skala MC, Fontanella A, Lan L, Izatt JA, Dewhirst MW. Longitudinal optical imaging of tumor metabolism and hemodynamics. *Journal of Biomedical Optics*. 2010;15(1):011112. doi:10.1117/1.3285584

9. Greening GJ, Miller KP, Spainhour CR, Cato MD, Muldoon TJ. Effects of isoflurane anesthesia on physiological parameters in murine subcutaneous tumor allografts measured via diffuse reflectance spectroscopy. *Biomedical Optics Express*. 2018;9(6):2871. doi:10.1364/boe.9.002871

10. Sio TT, Atherton PJ, Birckhead BJ, 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*. 2016;24(9):3847-3855. doi:10.1007/s00520-016-3213-3

11. Kamstra JI, Dijkstra PU, Leeuwen M van, Roodenburg JLN, Langendijk JA. Mouth opening in patients irradiated for head and neck cancer: A prospective repeated measures study. *Oral Oncology*. 2015;51(5):548-555. doi:10.1016/j.oraloncology.2015.01.016

12. Wagenmakers E-J, Lee M, Lodewyckx T, Iverson GJ. Bayesian versus frequentist inference. In: *Bayesian Evaluation of Informative Hypotheses*. Springer New York; 2008:181-207. doi:10.1007/978-0-387-09612-4_9

13. Gueorguieva R, Krystal JH. Move over ANOVA. *Archives of General Psychiatry*. 2004;61(3):310. doi:10.1001/archpsyc.61.3.310

14. Schober P, Vetter TR. Repeated measures designs and analysis of longitudinal data. *Anesthesia & Analgesia*. 2018;127(2):569-575. doi:10.1213/ane.0000000000003511

15. Pinheiro J, Bates D. *Mixed-effects models in S and S-PLUS*. Springer Science; Business Media; 2006. doi:https://doi.org/10.1007/b98882

16. Vishwanath K, Yuan H, Barry WT, Dewhirst MW, Ramanujam N. Using optical spectroscopy to longitudinally monitor physiological changes within solid tumors. *Neoplasia*. 2009;11(9):889-900. doi:10.1593/neo.09580

17. Dennis B, Ponciano JM, Taper ML, Lele SR. Errors in statistical inference under model misspecification: Evidence, hypothesis testing, and AIC. *Frontiers in Ecology and Evolution*. 2019;7. doi:10.3389/fevo.2019.00372

18. Wang B, Zhou Z, Wang H, Tu XM, Feng C. The p-value and model specification in statistics. *General Psychiatry*. 2019;32(3):e100081. doi:10.1136/gpsych-2019-100081

19. Liu C, Cripe TP, Kim M-O. Statistical issues in longitudinal data analysis for treatment efficacy studies in the biomedical sciences. *Molecular Therapy*. 2010;18(9):1724-1730. doi:10.1038/mt.2010.127

20. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle p value generates irreproducible results. *Nature Methods*. 2015;12(3):179-185. doi:10.1038/nmeth.3288

21. Abdi H. Holm's sequential Bonferroni procedure. *Encyclopedia of Research Design*. 2010;1(8):1-8. doi:10.4135/9781412961288.n178

22. Nakagawa S. A farewell to bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology*. 2004;15(6):1044-1045. doi:10.1093/beheco/arh107

23. Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*. 2012;5(2):189-211. doi:10.1080/19345747.2011.618213

24. Albers C. The problem with unadjusted multiple and sequential statistical testing. *Nature Communications*. 2019;10(1). doi:10.1038/s41467-019-09941-0

25. Ugrinowitsch C, Fellingham GW, Ricard MD. Limitations of ordinary least squares models in analyzing repeated measures data. *Medicine & Science in Sports & Exercise*. Published online December 2004:2144-2148. doi:10.1249/01.mss.0000147580.40591.75

26. Huynh H, Feldt LS. Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*. 1976;1(1):69-82. doi:10.3102/10769986001001069

27. Greenhouse SW, Geisser S. On methods in the analysis of profile data. *Psychometrika*. 1959;24(2):95-112. doi:10.1007/bf02289823

28. Haverkamp N, Beauducel A. 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*. 2017;8. doi:10.3389/fpsyg.2017.01841

29. Keselman HJ, Algina J, Kowalchuk RK. The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology*. 2001;54(1):1-20. doi:10.1348/000711001159357

30. Charan J, Kantharia N. How to calculate sample size in animal studies? *Journal of Pharmacology and Pharmacotherapeutics*. 2013;4(4):303. doi:10.4103/0976-500x.119726

31. Barr DJ, Levy R, Scheepers C, Tily HJ. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*. 2013;68(3):255-278. doi:10.1016/j.jml.2012.11.001

32. Rose NL, Yang H, Turner SD, Simpson GL. 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*. 2012;82:113-135. doi:10.1016/j.gca.2010.12.026

33. Pedersen EJ, Miller DL, Simpson GL, Ross N. Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ*. 2019;7:e6876. doi:10.7717/peerj.6876

34. Simpson GL. Modelling palaeoecological time series using generalised additive models. *Frontiers in Ecology and Evolution*. 2018;6. doi:10.3389/fevo.2018.00149

35. Yang L, Qin G, Zhao N, Wang C, Song G. Using a generalized additive model with autoregressive terms to study the effects of daily temperature on mortality. *BMC Medical Research Methodology*. 2012;12(1). doi:10.1186/1471-2288-12-165

36. Beck N, Jackman S. Beyond linearity by default: Generalized additive models. *American Journal of Political Science*. 1998;42(2):596. doi:10.2307/2991772

37. Wood SN. *Generalized Additive Models*. Chapman; Hall/CRC; 2017. doi:10.1201/9781315370279

38. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2020. https://www.R-project.org/

39. Wood SN, Pya N, Säfken B. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*. 2016;111(516):1548-1563. doi:10.1080/01621459.2016.1180986

40. West BT, Welch KB, Galecki AT. *Linear Mixed Models: A Practical Guide Using Statistical Software, Second Edition*. Taylor & Francis; 2014. https://books.google.com/books?id=hjT6AwAAQBAJ

41. Wolfinger RD. Heterogeneous variance: Covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*. 1996;1(2):205. doi:10.2307/1400366

42. Weiss RE. *Modeling Longitudinal Data*. Springer New York; 2005. doi:10.1007/0-387-28314-5

43. Geisser S, Greenhouse SW. An extension of Box's results on the use of the F distribution in multivariate analysis. *The Annals of Mathematical Statistics*. 1958;29(3):885-891. doi:10.1214/aoms/1177706545

44. Maxwell SE, Delaney HD, Kelley K. *Designing Experiments and Analyzing Data*. Routledge; 2017. doi:10.4324/9781315642956

45. Molenberghs G. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*. 2004;5(3):445-464. doi:10.1093/biostatistics/kxh001

46. Ma Y, Mazumdar M, Memtsoudis SG. Beyond repeated-measures analysis of variance. *Regional Anesthesia and Pain Medicine*. 2012;37(1):99-105. doi:10.1097/aap.0b013e31823ebc74

47. Scheffer J. Dealing with missing data. *Research Letters in the Information and Mathematical Sciences*. 2002;3:153-160.

48. Potthoff RF, Tudor GE, Pieper KS, Hasselblad V. Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research*. 2006;15(3):213-234. doi:10.1191/0962280206sm448oa

49. Box GEP. Science and statistics. 1976;71(356):791-799. doi:10.1080/01621459.1976.10480949

50. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*.; 2020. https://CRAN.R-project.org/package=nlme

51. OHara R, Kotze J. Do not log-transform count data. Published online January 2010. doi:10.1038/npre.2010.4136.1

52. Nelder JA, Wedderburn RWM. Generalized linear models. *Journal of the Royal Statistical Society Series A (General)*. 1972;135(3):370. doi:10.2307/2344614

53. McCulloch C. *Generalized, Linear, and Mixed Models*. John Wiley & Sons; 2001.

54. Dobson A. *An Introduction to Generalized Linear Models*. CRC Press; 2008.

55. Stroup W. *Generalized Linear Mixed Models : Modern Concepts, Methods and Applications*. CRC Press, Taylor & Francis Group; 2013.

56. Hastie T, Tibshirani R. Generalized additive models: Some applications. *Journal of the American Statistical Association*. 1987;82(398):371-386. doi:10.1080/01621459.1987.10478440

57. Hefley TJ, Broms KM, Brost BM, et al. The basis function approach for modeling autocorrelation in ecological data. *Ecology*. 2017;98(3):632-646. doi:10.1002/ecy.1674

58. Wegman EJ, Wright IW. Splines in statistics. *Journal of the American Statistical Association*. 1983;78(382):351-365. doi:10.1080/01621459.1983.10477977

59. Wood SN. Thin plate regression splines. 2003;65(1):95-114. doi:10.1111/1467-9868.00374

60. McElreath R. *Statistical Rethinking*. Chapman; Hall/CRC; 2018. doi:10.1201/9781315372495

61. Miller DL. Bayesian views of generalized additive modelling. *arXiv preprint arXiv:190201330*. Published online 2019.

62. Ruppert D. *Semiparametric Regression*. Cambridge University Press; 2003.

63. Storey JD. The positive false discovery rate: A bayesian interpretation and the q-value. *The Annals of Statistics*. 2003;31(6). doi:10.1214/aos/1074290335

64. Marra G, Wood SN. Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*. 2012;39(1):53-74. doi:10.1111/j.1467-9469.2011.00760.x

65. Simpson GL. *Gratia: Graceful 'Ggplot'-Based Graphics and Other Functions for GAMs Fitted Using 'Mgcv'.*; 2020. https://CRAN.R-project.org/package=gratia

66. Harezlak J, Ruppert D, Wand MP. *Semiparametric Regression with R*. Springer New York; 2018. doi:10.1007/978-1-4939-8853-2

67. Lang TA, Altman DG. Basic statistical reporting for articles published in biomedical journals: The "statistical analyses and methods in the published literature" or the SAMPL guidelines. *International Journal of Nursing Studies*. 2015;52(1):5-9. doi:10.1016/j.ijnurstu.2014.09.006

68. Hastie T, Tibshirani R. Generalized additive models for medical research. *Statistical Methods in Medical Research*. 1995;4(3):187-196. doi:10.1177/096228029500400302

69. Begley CG, Ioannidis JPA. Reproducibility in science. *Circulation Research*. 2015;116(1):116-126. doi:10.1161/circresaha.114.303819

70. Weissgerber TL, Garcia-Valencia O, Garovic VD, Milic NM, Winham SJ. Why we need to report more than 'data were analyzed by t-tests or ANOVA'. *eLife*. 2018;7. doi:10.7554/elife.36163