Generalized additive models to analyze biomedical non-linear longitudinal data in 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 in terms of correlation and missing observations, both methodologies share an assumption of linearity in the measured response, which results in biased estimates and unreliable inference when they are used to analyze data where the trends are non-linear, which is a common occurrence in biomedical research.

In contrast, generalized additive models (GAMs) relax the linearity assumption, and allow the data to determine the fit of the model while permitting missing observations and different correlation structures. Therefore, GAMs present an excellent choice to analyze non-linear longitudinal data 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 non-linear data. 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 that are consistent with the trends of non-linear data even in the case when missing 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; and in such studies the subjects of analysis range 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 the 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 key 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 assumption" in rm-ANOVA implies that the model is misspecified when the data does not follow a linear trend, which results in unreliable inference. In biomedical research, non-linear trends are the norm rather than the exception in longitudinal studies. 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" *p-values*(p<0.05) from data that shows non-linear trends, the validity of such 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 the null hypothesis 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 vary 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 missing data 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 missing 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 missing 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 errors of the random effects, which need to be normally distributed and independent.[13,31] And even more importantly, LMEMs also assume a linear relationship between the response and time,[15] making them unsuitable to analyze non-linear data.

As the rm-ANOVA and the more flexible LMEM approaches make overly restrictive assumptions regarding the linearity of the response, there is a need for biomedical researchers to explore the use of additional statistical tools that allow the data (and not an assumption in trend) to determine the trend of the fitted model, 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 missing observations. Therefore, GAMs can 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 not only speed up the initial stages of the analysis but also enable the use of advanced modeling structures (e.g. hierarchical models, confidence interval comparisons) without requiring advanced programming skills from the user. At the same time, R has many tools that simplify data simulation, an emerging strategy used to test statistical models.[28] Data simulation methods 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, and simulation can be also used for power calculations and study design questions.

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 linearity of response, constant correlation structures and missing 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) linear 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 Linear 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:

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

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 time ($time_t$), treatment ($treatment_j$) and their interaction $time_t * treatment_j$ which have linear slopes given by β_1, β_2 and β_3, respectively. Independent errors ε_{ijt} represent random variation not explained by the *fixed* effects, and are assumed to be $\sim N(0, \sigma^2)$ (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

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

To further simplify the expression, substitute $\widetilde{\beta}_0 = \beta_0 + \beta_2$ and $\widetilde{\beta}_1 = \beta_1 + \beta_3$ 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

$$y_{ijt} = \begin{cases} \beta_0 + \beta_1 \times time_t + \varepsilon_{ijt} & \text{if Group A} \\ \widetilde{\beta}_0 + \widetilde{\beta}_1 \times time_t + \varepsilon_{ijt} & \text{if Group B} \end{cases} \tag{3}$$

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 Equation (1), Equation (2) or Equation (3), an rm-ANOVA model is only able to accommodate linear patterns in the data. If the data show non-linear behavior, the rm-ANOVA model will approximate this behavior with non-parallel lines.

3.2.2 The Linear Mixed Model Case (LMEM)

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 distinguish.[15,40] A LMEM with interaction between time and treatment for a longitudinal study can be written as:

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

When Equation (1) and Equation (4) are compared, it is easily noticeable that LMEM 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 μ_{ij}). This term μ_{ij} is the one that 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 used to model some "noise" in the response, but that is intended to be analyzed and disentangled from the "global noise" term ε_{ijt} from Equation (1).

For example, if the blood concentration of the 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. As the name suggests, this "random" variability needs to be modeled as a variable rather than as a constant value. The *random effect* μ_{ij} in Equation (4) is assumed to be $\mu_{ij} \sim N(0, \sigma_\mu^2)$. In essence, the *random effect* in a LMEM enables to fit models with different slopes 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.

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).[40,42]

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 it is higher between consecutive measurements.[13,25] Although corrections can be made (such as Huyhn-Feldt or Greenhouse-Geisser)[26,27] the effectiveness of each correction is limited because it depends on the size of the sample,the number of repeated measurements[28], and they 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 missigness 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 missing 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 do an rm-ANOVA and LMEM fit look like? A visual representation using simulated data

To visually demonstrate the limitations of rm-ANOVA an LMEMs for non-linear longitudinal data, 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 responses 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 1C). In both the linear and quadratic simulated data, the groups start with the same mean value at the first time point. This is intentional in order to simulate the expected temporal evolution of some physiological quantity, which is typical in biomedical experiments where a strong non-linear trend is present.

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 in both the linear and quadratic simulation corresponds to the mean response per group plus the error (Figure 1 B,D).

A more comprehensive exploration of the fit of rm-ANOVA and LMEMs for linear and non-linear longitudinal data appears in the Appendix (Figure A.1 and Figure A.2), where simulation with compound symmetry and independent errors (errors generated from a normal distribution that are not constant over time) and the plot of simulated errors, and fitted parameters in 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 as a representation of the consequences of modeling non-linear data with a linear model such as rm-ANOVA or LMEMs. Of notice, in Section 6 we use simulated data that does follow reported trends in the biomedical literature to implement GAMs.

The simulation shows that the fits produced by the LMEM and the rm-ANOVA model are good for linear data, as the predictions for the mean response are reasonably close to the "truth" of the simulated data (Figure 1A). When the

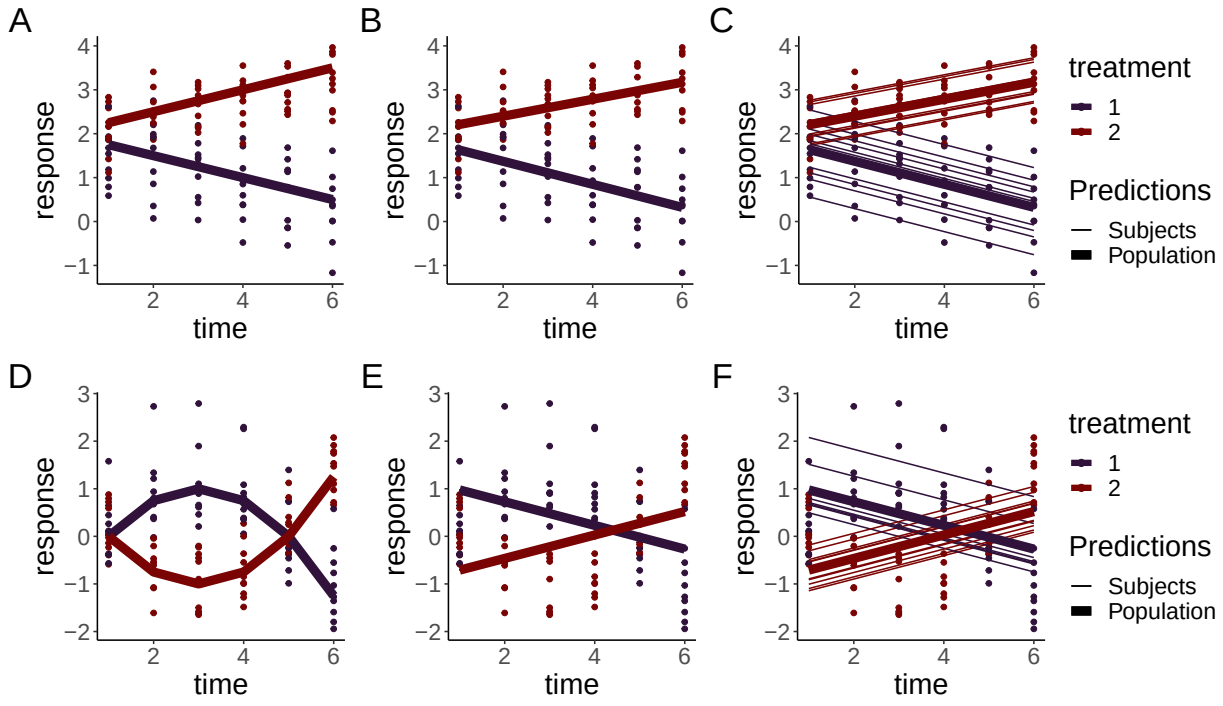

Figure 1: Simulated responses from two groups with correlated errors using a LMEM and a rm-ANOVA model. Top row: linear response, bottom row: quadratic response. A: Simulated linear data with known mean response (thin lines) and individual responses (points) showing the dispersion of the data. D: Simulated quadratic data with known mean response (thin 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 (D) but also assigns a global estimate that does not take between-subject variation. C, F: Estimates from the LMEM in the linear and quadratic case. 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).

linearity and compound symmetry assumptions are met, the rm-ANOVA model approximates well the global trend by group (Figure 1B). 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.

This section has used simulation to better convey the limitations of linearity and correlation in the response in non-linear data. The models fitted to the simulated data were an rm-ANOVA model and a LMEM, where the main issue is the expected linear trend in the response. In the following section, we present generalized additive models (GAMs) as a data-driven alternative method to analyze longitudinal non-linear data that overcomes the linearity assumption.

4 GAMs as a special case of Generalized Linear Models

4.1 GAMs and Basis Functions

Generalized linear models (GLMs) are a family of models (which include rm-ANOVA and LMEMs) that fit a linear response function to data that may not have normally distributed errors.[51] In contrast, GAMs are a family of regression-based methods for estimating smoothly varying trends and are a broader class of models that contain the GLM family as a special case[34,37,52]. A GAM model can be written as:

$$y_{ijt} = \beta_0 + f(x_t \mid \beta_j) + \varepsilon_{ijt} \tag{5}$$

Where y_{ijt} is the response at time t of subject i in group j, β_0 is the expected value at time 0, the change of y_{ijt} over time is represented by the *smooth function* $f(x_t \mid \beta_j)$ with inputs as the covariates x_t and parameters β_j, and ε_{ijt} represents the residual error.

In contrast to the linear functions used to model the relationship between the covariates and the response in rm-ANOVA or LMEM, GAMs use more flexible *smooth functions*. This approach 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 $f(x_t \mid \beta_j)$ that allow for non-linear responses are polynomials, 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 $\pm\infty$, $f(x_t \mid \beta_j)$ 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, by using *basis function* expansions of the covariates and by 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 $f(x_t \mid \beta_j)$ 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 $time_t$, $treatment_j$ and $time_t \times treatment_j$. The basis function then, is the combination of basis coefficients and basis vectors that map the possible relationship between the covariates and the response,[53] which in the case of Equation (1) is restricted to a linear family of functions. In the case of Equation (5), the basis functions are contained in the expression $f(x_t \mid \beta_j)$, 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*, as they 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.[54] 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 smoothness. The use of CS as smoothers in GAMs was discussed within the original GAM framework,[52] 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 bending of the smooth will occur) to obtain the reduced basis, which could affect the model fit.[55] A solution to the "knot" placement of CS is provided by thin plate splines (TPS), which provide optimal smooth

estimation without knot placement, but that are computationally costly to calculate.[37,55]

In contrast, thin regression splines (TPRS) provide a reasonable "low rank" (truncated) approximation to the optimal TPS estimation, which can be implemented in an efficient computational manner.[55] Like TPS, TPRS only require the number of basis to be used to create the smoother (for mathematical details on both TPS and TPRS see refs.[55] and).[37]

To further clarify the concept of basis functions and smooth functions, consider the simulated response for Group 1 in Figure 1C. 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 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 an equal number of basis to the number of unique values in the independent variable; less basis functions can be specified to create the smooth as well, as long as they reasonably capture the trend of the data).

If five basis functions are used to fit a GAM for the data that appears in Figure 1C, the resulting fitting process is shown in Figure 2A. The four basis functions (and the intercept) are shown. Each of the basis functions is composed of six different points (because there are six points on the timeline). To control the "wiggliness" of the fit, each of the basis functions of Figure 2A is weighted by multiplying it by a coefficient according to the matrix of Figure 2B. The parameter estimates are penalized (shrunk towards 0) where the penalty reduces the "wiggliness" of the smooth fit to prevent overfitting. A weak penalty estimate will result in wiggly functions whereas a strong penalty estimate provides evidence that a linear response is appropriate.

To get the weighted basis functions, each basis (from Figure 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 less than 1). 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 (orange line), along the simulated values per group, which appear as points.

5 A Bayesian interpretation of GAMs

Bayes' theorem states that the probability of an event can be calculated using prior knowledge or belief.[56] In the case of data that shows non-linear trends, the belief that the *true* trend of the data is likely to be smooth rather than extremely "wiggly" introduces the concept of a prior distribution for wiggliness (and therefore a Bayesian view) of

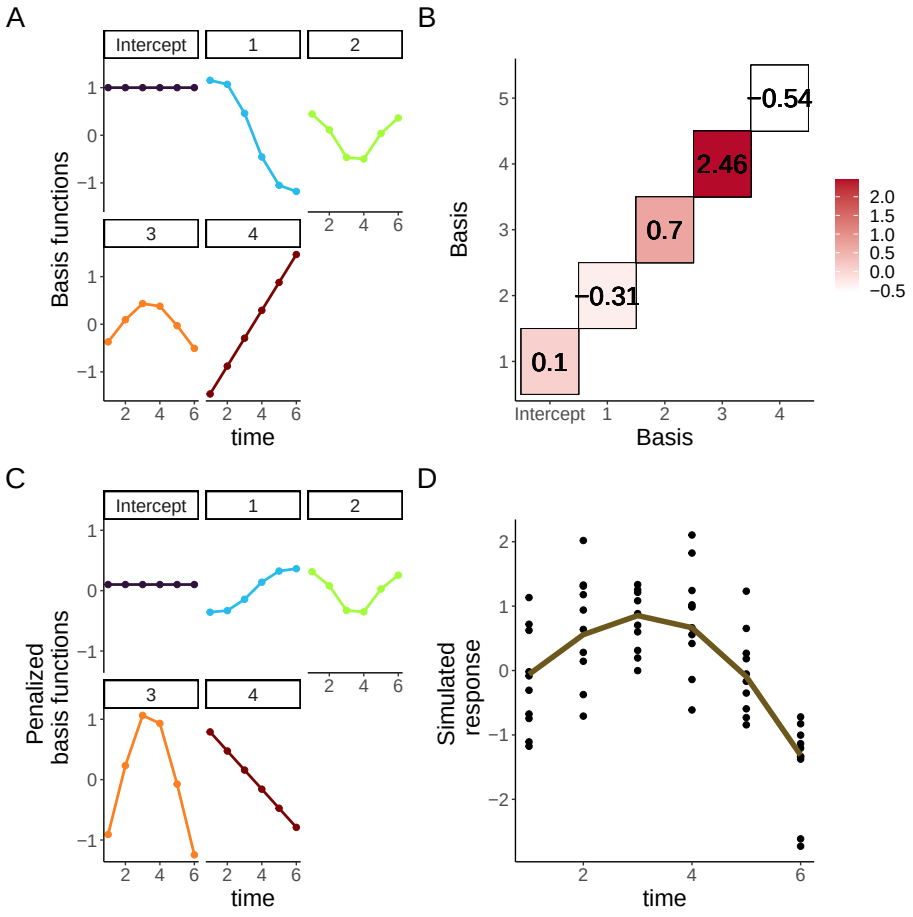

Figure 2: Basis functions for a single smoother for time. A: Basis functions for a single smoother for time for the simulated data of Group 1 from Figure 2. B: Matrix of basis function weights. Each basis function is multiplied by a coefficient which can be positive or negative. The coefficient determines the overall effect of each basis in the final smoother. C: Weighted basis functions. Each of the four basis functions 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 the group shown as points.

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 prior distribution as in the "fully Bayesian" case, which can be fitted using JAGS, Stan, or other probabilistic programming language).[57] Therefore, the confidence intervals (CIs) calculated for the smooth terms using *mgcv* are considered empirical Bayesian posterior credible intervals,[33] which have good "frequentist" coverage (point-wise coverage or "single point" coverage), and *across the function* coverage.[37]

To understand this last part, it is worth reminding 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 a "point-wise" CI for non-linear data, the "true" function will lie outside of the CI in certain regions at a given frequency (because the data does not follow a linear trend). This means that if a point-wise CI is obtained for 100 random samples, the total number of CIs that contain the true function *entirely* is much less than the nominal value of the CI (95%, usually).

Contrary to this, an *across the function* CI (like those estimated for GAMs using *mgcv*) contains the true function through the entire range of the covariates (which would be time in the case of longitudinal data). For GAMs this means that if repeated samples are taken and a GAM and corresponding CIs are calculated, the percentage of CIs entirely containing the true function over the entire time period would be close to the nominal value (i.e., if 100 random samples are obtained and a GAM and CI is calculated for each one of them, it would be expected that 95 out of the 100 fitted CIs entirely contain the true function), thereby allowing more robust estimates from the model. In-depth theory of the Bayesian interpretation of GAMs is beyond the scope of this paper, but can be found in[34,37,57] and.[58] With this brief introduction to the Bayesian interpretation of GAMs, we henceforth refer to the confidence intervals for the smooths in GAMs as "empirical Bayesian" through the rest of this paper.

6 The analyisis of longitudinal biomedical data using GAMs

The previous sections provided the basic framework to understand 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`. The code for the simulated data and figures, and a brief guide for model selection and diagnostics appear in the Appendix.

6.1 Simulated data

The simulated data is based on the reported longitudinal changes in oxygen saturation (StO_2) in subcutaneous tumors that appear in Figure 3C in.[16] In the paper, 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 simulated and real data appear in Figure 3A and the inset, respectively.

6.2 An interaction GAM for longitudinal data

An interaction effect is typically the main interest in longitudinal biomedical data, as it 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 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 basis functions and correlations can be found in.[53]

For the data in Figure 3, A the main effect of interest is how StO_2 changes over time for each treatment. To estimate this, the model incorporates separate smooths for *Group* and *Day*, respectively. 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 the Appendix) 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 the 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). The parametric term `Group` is added to quantify overall mean differences in the effect of treatment between groups. 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.

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*.[59] A guide for model selection and diagnostics is in the Appendix, and an in-depth analysis can be found in[37] and.[60]

One question that might arise at this point is "what is the fit that an rm-ANOVA model produces for the simulated data?" The 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 missingness exists. The rationale behind this is that GAMs are able to pick the trend in the data even when some observations are missing. However, this usually causes the resulting smooths to have wider confidence intervals and less ability to pick certain trends. Consider the simulated StO_2 values from Figure 3B. If 40% of the observations are randomly deleted and the same interaction GAM fitted for the complete dataset is used, the resulting smooths are still able to show a different trend for each group, but because the empirical Bayesian credible intervals for the smooths overlap during the first 3 days with fewer data points, the trend is less pronounced than in the full dataset (Figure 3D). Although the confidence intervals have increased for both smooths, the model still shows different trends with as few as 4 observations per group at certain time points.

6.3 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. Whereas in rm-ANOVA a *post-hoc* analysis is required to answer such question by calculating some *p-values* after multiple comparisons, GAMs can use a different approach to estimate significance. In essence, the idea behind the estimation of significance in GAMs across different treatment groups is that if the *difference* between the empirical Bayesian confidence intervals of the fitted smooths for such groups is non-zero, then a significant difference exists at that time point(s). The absence of a *p-value* in this case might seem odd, but the empirical Bayesian confidence interval interpretation can be conceptualized in the following manner: Different trends in each group are an indication of an effect by the treatment. This is what happens for the simulated data 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

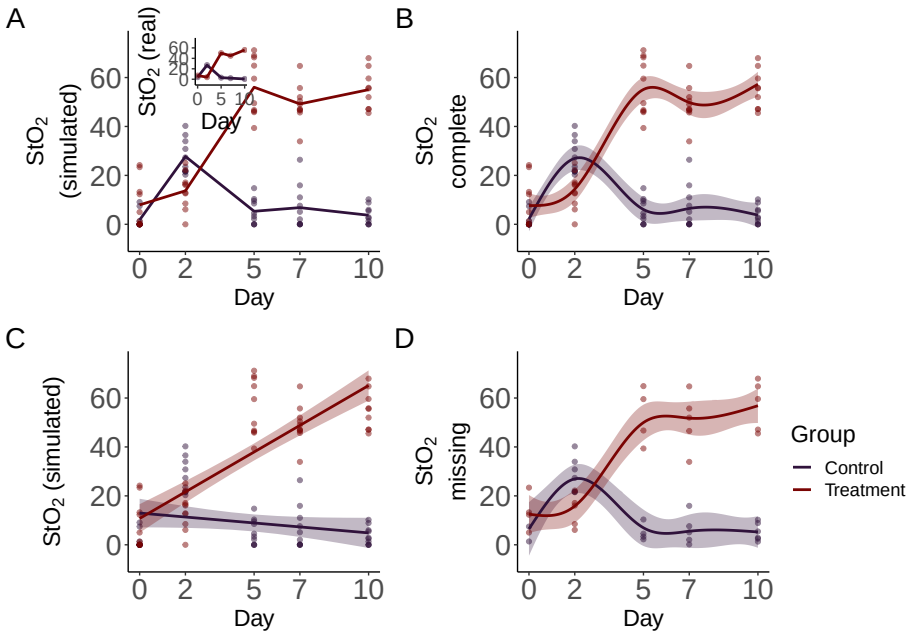

Figure 3: Simulated data and smooths for oxygen saturation in tumors. A: Simulated data that follows previously reported trends (inset) in tumors under chemotherapy (Treatment) or saline (Control) treatment. Simulated data is from a normal distribution with standard deviation of 10% with 10 observations per time point. Lines indicate mean oxygen saturation 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% 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 40% of its observations missing. Lines represent trends for each group, shaded regions are 95% empirical Bayesian confidence intervals.

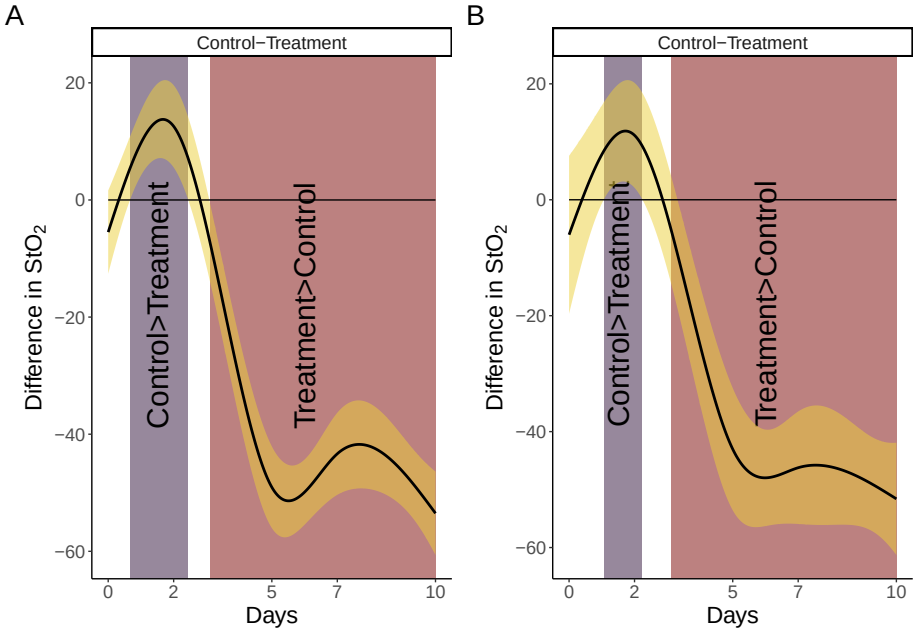

Figure 4: Pairwise comparisons for smooth terms. A: Pairwise comparisons for the full dataset. B: Pairwise comparisons for the dataset with missing observations. Significant differences exist where the 95% empirical Bayesian credible interval does not cover 0. In both cases the effect of treatment is significant after day 3.

the observed difference is significant. The difference between groups with similar trends is likely to yield 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).

Consider the calculation of pairwise differences for the smooths in Figure 3B and Figure 3D. Figure 4 shows the comparison between each treatment group for the full and missing datasets. Here, the "Control" group is used as the reference to which "Treatment" group is being compared. Of notice, the pairwise comparison has been set on the response scale (see Appendix for code details), because otherwise the comparison appears shifted and is not intuitively easy to relate to the original data.

With this correction in mind, the shaded regions over the confidence interval (where it does not cover 0) indicate the time interval where each group has a higher effect than the other. Notice that the shaded region between days 0 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 ≈ 3 day it is the "Treatment" group the one that 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 empirical Bayesian credible intervals of the smooths show the same trend of the full dataset. Consequently, the smooth pairwise comparison (Figure 4B) shows that the Control Group has higher StO_2 before day 3, and is able to estimate the change on day 3 where the Treatment Group becomes significant as the full dataset smooth pairwise comparison.

In a sense, the pairwise smooth comparison is more informative than a *post-hoc p-value*. For biomedical studies, the smooth comparison is able to provide an estimate of *when* and by *how much* a biological process becomes significant. This 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.

7 Discussion

Biomedical longitudinal data is particularly challenging to analyze due to the likelihood of missing observations and different correlation structures in the data, which limit the use of rm-ANOVA. Although LMEMs can handle missing 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, where it is clear that these models do not capture the non-linear trend of the data, thereby causing 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. It could be argued that in Section 3.5 a LMEM with quadratic effects could have been used, but because in reality the true function in the data is not known, using polynomial effects ends up causing more questions than answers in this case (e.g., what is the biological interpretation of a quadratic, cubic, or other effect of even higher order?). Guidelines for statistical reporting in biomedical journals exist (the SAMPL guidelines)[61] 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, instead of merely relying on a R^2 value.

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 (among which are palaeoecology, geochemistry, and ecology)[33,53] they are not routinely used in biomedical research despite an early publication from Hastie and Tibshirani that demonstrated their use in medical research.[62] This is possibly due to the fact that the theory behind GAMs can seem very different from that of rm-ANOVA and LMEMs, but the purpose of Section 4 is to demonstrate that at its core the principle 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 following non-linear trends in the data.

However, 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 6, while a basic workflow for model selection is in the Appendix. One of the features of GAMs is that their Bayesian interpretation allows to indicate differences between groups without the need of a *p-value*, and in turn provide 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 model is therefore able to identify changes between the groups at time points were data was not directly measured even with missing data exists ($\approx$ day 3 in Figure 4 A, B), which 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 not only provides a fully developed environment to fit GAMs, but also eases 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 reproducible. In this regard, reproducibility is still an issue in biomedical research,[63,64] but it is becoming apparent that what other disciplines have experienced in this aspect is likely to impact sooner rather than later this field. 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.

8 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.

9 Acknowledgements

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

10 Declaration of Conflicting Interests

The Authors declare that there is no conflict of interest.

Supplementary Materials

An Appendix which contains all the code used to create this manuscript, along with a basic workflow to implement GAMs in R is available as Supplementary Material in PDF. 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.

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

52. 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

53. 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

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

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

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

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

58. 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

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

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

61. 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

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

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

64. 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