Neuron

CellPress

Primer

Beyond t test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research

Zhaoxia Yu,[1,7,*] Michele Guindani,[1] Steven F. Grieco,[2] Lujia Chen,[2] Todd C. Holmes,[3,7] and Xiangmin Xu[2,4,5,6,7,*]
[1]Department of Statistics, Donald Bren School of Information and Computer Sciences, University of California, Irvine, Irvine, CA 92697-3425, USA
[2]Department of Anatomy and Neurobiology, School of Medicine, University of California, Irvine, Irvine, CA 92697-1275, USA
[3]Department of Physiology and Biophysics, School of Medicine, University of California, Irvine, Irvine, CA 92697- 4560, USA
[4]Department of Biomedical Engineering, University of California, Irvine, Irvine, CA 92697-2715, USA
[5]Department of Microbiology and Molecular Genetics, University of California, Irvine, Irvine, CA 92697-4025, USA
[6]Department of Computer Science, University of California, Irvine, Irvine, CA 92697-3435, USA
[7]The Center for Neural Circuit Mapping, University of California, Irvine, Irvine, CA 92697, USA
*Correspondence: zhaoxia@ics.uci.edu (Z.Y.), xiangmix@uci.edu (X.X.)
https://doi.org/10.1016/j.neuron.2021.10.030

SUMMARY

In basic neuroscience research, data are often clustered or collected with repeated measures, hence correlated. The most widely used methods such as t test and ANOVA do not take data dependence into account and thus are often misused. This Primer introduces linear and generalized mixed-effects models that consider data dependence and provides clear instruction on how to recognize when they are needed and how to apply them. The appropriate use of mixed-effects models will help researchers improve their experimental design and will lead to data analyses with greater validity and higher reproducibility of the experimental findings.

OVERVIEW

The importance of using appropriate statistical methods for experimental design and data analysis is well recognized across scientific disciplines. The growing concern over reproducibility in biomedical research is often referred to as a "problem of inadequate rigor" (Kilkenny et al., 2009; Prinz et al., 2011). The reproducibility crisis has been attributed to various factors that include lack of adherence to good scientific practices, underdeveloped experimental designs, and the misuse of statistical methods (Landis et al., 2012; Steward and Balice-Gordon, 2014). Further compounding these challenges, we are in the midst of an ever-expanding biomedical research revolution. "Big Data" are being produced at an unprecedented rate (Margolis et al., 2014). The proper analysis of Big Data requires up-to-date statistical methodologies that take complex features of data such as explicit and implicit data dependencies into consideration. Better matching of statistical models that take data characteristics into account will allow for better interpretation of data outcomes. It will also boost the confidence in biomedical research of all stakeholders in the scientific enterprise, including industry and the taxpaying public (Alberts et al., 2014; Freedman et al., 2015; Macleod et al., 2014). Despite recent advances in statistical methods, current neuroscience research is often conducted using a limited set of well-known statistical tools. Many models and tests assume that the observations are independent of one another. Failure to account for this dependency in the data often leads to an increased number of false positives, a major cause of the irreproducibility crisis (Aarts et al., 2014).

The t test and analysis of variance (ANOVA) are familiar methods to all neuroscience researchers. Both methods assume that individual observations are independent of one another. For example, data measurements from multiple mice observed under different conditions (e.g., different mouse genetic models) are taken to be unique. However, this assumption of independence is false for animals clustered into cages or litters and for neuroanatomical and neurophysiological studies that rely on large-scale longitudinal recordings and involve repeated measurements over time of the same neurons and/or animals (Aarts et al., 2014; Galbraith et al., 2010; Wilson et al., 2017). In those cases, data are structured as clusters of multiple measurements collected from single units of analyses (neurons and/or animals), leading to natural dependence and correlation between the observations (Figure 1).

A quick examination of recently published articles indicates that reported results in basic neuroscience research often use inappropriate statistical methods for which the experimental designs and the ensuing/resulting data dependencies are not taken into account (Aarts et al., 2014; Boisgontier and Cheval, 2016). Our conclusion is supported by our survey of the studies published in prestigious journals over the past few years. In total, we identified >100 articles in which recordings of individual neurons from multiple animals were pooled for statistical testing. Alarmingly, only ~50% of these articles accounted for data dependencies in any meaningful way. Our finding agrees with an

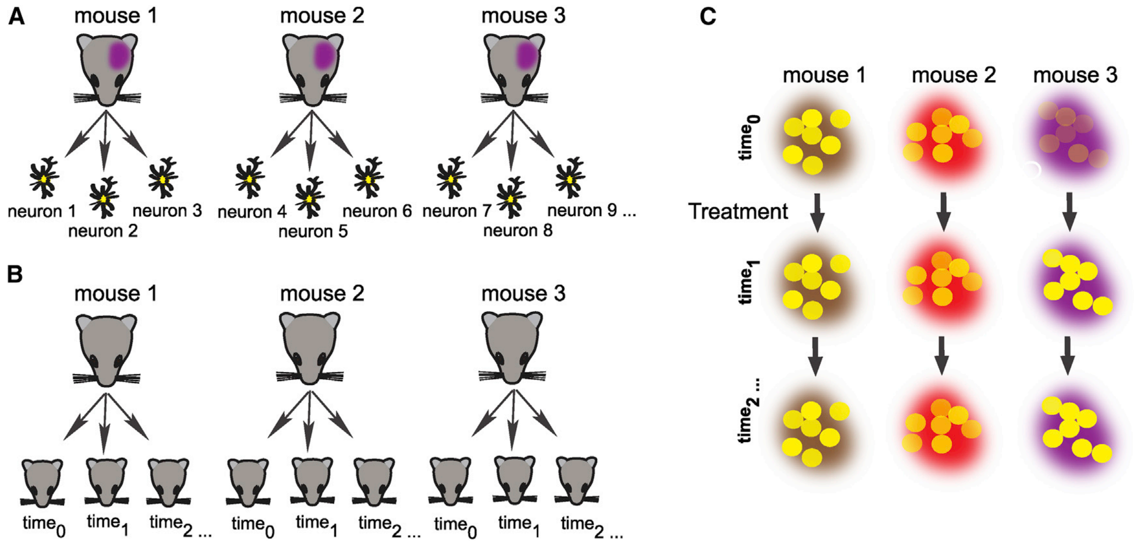

Figure 1. Sources of correlation

A graphical representation shows potential sources of correlated data.

(A) The data are correlated because neurons from the same animal tend to be more similar to one another than neurons from different animals.

(B) The observations are dependent when they are taken from the same animal temporally, while the data from different animals are independent.

(C) Correlation arises from 2 sources: individual observations are made from neurons from 3 different mice before and after drug treatment.

investigation published a few years ago (Aarts et al., 2014), which found that 53% of neuroscience articles failed to account for the dependence in their data. Representative descriptions of the inappropriate analyses read, "t(28656) = 314 with $p<10^{-10}$ over a total of n=28657 neurons pooled across six mice," "n = 377 neurons from four mice, two-sided Wilcoxon signed rank test," "610 A cells, 987 B cells and 2584 C cells from 10 mice, one-way ANOVA and Kruskal–Wallis test," "two-sided paired t test, n=1597 neurons from 11 animals, d.f. = 1596," among numerous others. Such analyses can lead to astonishingly high type I error (false positive) rates (see below). Even in cases for which multi-level data dependencies are obvious, investigators continue to use repeated ANOVA, paired t test, or their nonparametric versions. In many cases, errors due to the use of inappropriate statistics affect the main conclusion of the article (Fiedler, 2011).

Statisticians have developed effective methods to analyze correlated data. Several widely used statistical tools that take data dependencies into account are the linear and generalized mixed-effects (ME) models, which include t test and ANOVA as special cases. Although the value of analyzing correlated data has been increasingly recognized in many scientific disciplines, including clinical research, genetics, psychological science studies, ME models have been underutilized in basic neuroscience research.

The purpose of our article is to provide a readable primer to neuroscience experimentalists, who do not have extensive training in statistics. We illustrate and discuss what features of the experimental questions require an appropriate consideration of adequate design and data structure, and how the proper use of ME models will lead to more rigorous analysis, reproducibility, and richer conclusions. We provide concrete data examples on how to properly use ME models. In addition to providing an improved perspective on appropriate statistical analyses, we provide easy-to-follow instructions for the implementation of ME models, with access to code and practice datasets to all interested users. See Glossary Box 1 for a useful glossary related to this Primer.

INTRODUCTION TO LINEAR AND GENERALIZED LINEAR ME (LME, GLMM) MODELS

Important concepts and definitions related to statistical testing

To understand the practical issues of ME models in the context of neuroscience research, we introduce several important concepts and definitions using real-world data illustrations. Considering 5,000 cells measured from 5 mice, what is the effective sample size (n_{eff}) in this study? Is it 5,000 or 5? Perhaps it is neither. The number of biological units, experimental units, and observational units can be quite distinct from one another. A detailed discussion of sample size in cell cultures and animal experiments is provided by an earlier paper (Lazic et al., 2018). Here, we use an example dataset collected from our laboratory to illustrate the concept and definition of intra-class-correlation (ICC), which is a metric to quantify the degree of correlation due to clustering. We also introduce the concepts of design effect (D_{eff}) and n_{eff} and discuss why conventional methods such as t test and ANOVA are not appropriate for this example.

ICC is a widely used metric to quantify the degree to which measurements from the same group are correlated. Depending on the specific settings that are concerned, different definitions have been proposed. For simplicity, let us consider the simple 1-way ANOVA setting, in which each animal is considered as a class. The total variance of data can be partitioned into the between- (inter-) and within- (intra-) class variances. The population

Box 1. Glossary

Clustered data: In neuroscience research, the data from a study are often obtained from a number of different experimental units (referred to as clusters). The key feature of clustered data is that observations from the same cluster tend to be correlated with each other.
Dependent versus independent: For dependent samples, the selection of subjects for consideration (e.g., neurons, animals) in one sample is affected by the selection of subjects in the other samples. For independent samples, the selection of subjects for consideration (e.g., neurons, animals) is not affected by the selection of subjects in the other sample.
Effect size: An effect size is a numerical quantity for the magnitude of a certain relationship such as the difference between population means or the association between two quantitative variables.
Fixed versus random effects: Fixed effects often refer to fixed but unknown population parameters such as coefficients in the traditional linear model (LM). Random effects often refer to effects at the individual or subject level that are included in the model to take into account the heterogeneity/variability of individual observations but are usually not of direct interest.
Frequentist versus Bayesian approaches in mixed-effects models: In frequentist analysis, a fixed effect is a fixed but unknown population parameter, whereas a random effect is a value drawn from a distribution to capture individual variability. In Bayesian analysis, both fixed and random effects are random variables drawn from distributions (priors); the inference is conducted by computing the posterior distribution for the fixed effects and the variance-covariance of the random effects. The posterior distribution updates the prior information using the observed data.
Hypothesis testing: A hypothesis is a statement about a parameter (or a set of parameters) of interest. Statistical hypothesis testing is formalized to make a decision between rejecting or not rejecting a null hypothesis on the basis of a set of experimental observations and measurements. Two types of errors can result from any decision rule (test): (1) rejecting the null hypothesis when it is true (a Type I error, ‘‘false positive’’) and (2) failing to reject the null hypothesis when it is false (a Type II error, ‘‘false negative’’).
Independently and identically distributed: A set of random variables are independently and identically distributed (i.i.d.) if they are mutually independent and each of them follows the same distribution.
Linear regression model (or linear model): A linear regression model is an approach to model the linear relationship between a response variable and one or more explanatory variables.
Linear mixed-effects model (LME) and generalized linear mixed model (GLMM): The LME is an extension of the linear regression model to consider both fixed and random effects. It is particularly useful when the data are clustered or have repeated measurements. The GLMM is an extension to the generalized linear model, in which the linear predictor contains random effects in addition to the fixed effects.
Parameters: Parameters are the characteristic values of an entire population, such as the mean and standard deviation of a normal distribution. Samples can be used to estimate population parameters.
Parametric versus nonparametric tests: A parametric test assumes that the data follow an underlying statistical distribution. A nonparametric test does not impose a specific distribution on data. Nonparametric tests are more robust than parametric tests as they are valid over a broader range of situations.

ICC (Fischer, 1944) is defined as the ratio of the between-class variance to the total variance:

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2},$$

where σ_b^2 denotes the between-class variance and σ_e^2 denotes the within-class variance. For naturally occurring clusters, ICC often falls between 0 and 1. If ICC = 0, then the data can be treated as uncorrelated; if ICC = 1, then all of the observations in each cluster are perfectly correlated.

In our study of ketamine effects on neuroplasticity (example 1, see below), we measured phosphorylated cyclic AMP response element binding protein (pCREB) immunoreactivity of 1,200 putative excitatory neurons of mouse visual cortex at different time points: collected at baseline (saline), 24, 48, and 72 h and 1 week following ketamine treatment from 24 mice (Figure 2). The original data and full description of the experiments can be found in Grieco et al. (2020). For this example, a large ICC suggests that neurons from the same mouse tend to be more similar to one another than neurons from different mice. For larger values of ICC, there is greater homogeneity within clusters and greater heterogeneity between clusters. As shown in Figure 2, the pCREB values of the 357 neurons in the saline group tend to cluster into groups indexed by the 7 mice. The estimated ICC (Wolak and Wolak, 2015) is 0.61, which implies that the 357 observations should not be treated as independent data points.

To understand why conventional methods (t test, ANOVA) fail when data dependencies are not taken into account, it is helpful to quantify the magnitude of clustering of an experiment using the D_{eff} (Kish, 1965), which is defined as

$$D_{eff} = 1 + (M - 1)ICC$$

where M denotes the average cluster size of an experiment design. It is a useful metric to recalibrate the standard error of an estimate in the presence of clustering or adjusting sample size when designing an experiment. For the saline group, with $n = 357$ and $ICC = 0.61$, the D_{eff} is 32 (i.e., on average, 32 neurons under the current design are equivalent to 1 uncorrelated neuron). This experimental design may call for more measurements, but how many should be made?

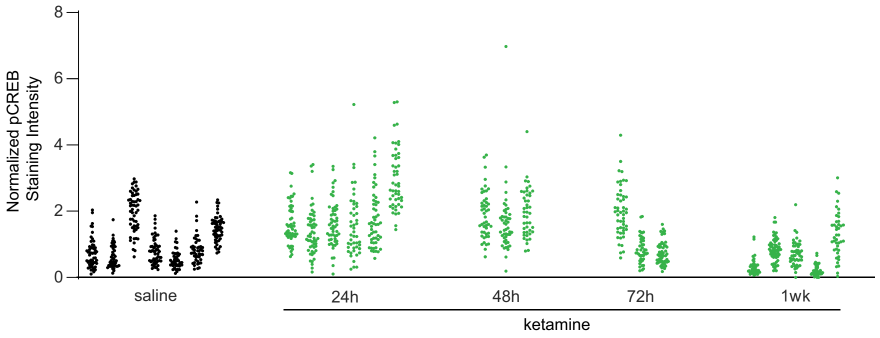

Figure 2. Avoiding false positives that arise from correlated measurements taken from the same animals

Normalized pCREB staining intensity values from 1,200 neurons (example 1). The values in each cluster were from 1 animal. In total, pCREB values were measured for 1,200 neurons from 24 mice at 5 conditions: saline (7 mice, ICC = 0.61), 24 h (6 mice, ICC = 0.33), 48 h (3 mice, ICC = 0.02), 72 h (3 mice, ICC = 0.63), and 1 week (5 mice, ICC = 0.54) after treatment. According to ICC, observations at 48 and 72 h show the smallest and largest ICCs, respectively.

Another closely related concept that helps answer this question is the n_{eff}, which is the equivalent sample size if there is no clustering/correlation. It is defined as $n_{\text{eff}} = n/D_{eff}$, where n is the total sample size (number of observations). This definition is also an interpolation of the 2 extreme cases of ICC = 0 or 1, with ICC = 0 leading to $n_{\text{eff}} = n$ (no correlation) and ICC = 1 leading to $n_{\text{eff}} = n/M$ (complete correlation). In sample size calculations, the D_{eff} can be interpreted as a multiplying factor to obtain the desired sample size under the assumption of independence. With $D_{eff} = 32$ in the saline example, the n_{eff} based on the 357 neurons is $n_{\text{eff}} = 357/32 \approx 11$, which is only ~50% more than the number of mice. The ICC, D_{eff}s, and n_{eff}s for the 5 groups are shown in Table 1. The results indicate that there is substantial dependence in data. Unfortunately, when researchers analyze data under such circumstances, the methods they choose often make the wrong assumption that all of the observations are independent from one another. One well-known consequence of ignoring correlations in data is an increased number of false positives, which is discussed below.

Failing to account for data dependence leads to high type I error (false positive) rates

When dependence is ignored in the data analysis, null hypotheses can be erroneously rejected, and confidence intervals do not have enough coverage. In the statistical literature, the action of erroneously rejecting a null hypothesis (see Glossary Box 1 and supplemental information) is called a "false positive." For a given test, its size, or type I error rate, is defined as the probability that the null hypothesis is erroneously rejected. We say that a test has an inflated type I error rate when its type I error rate is greater than its significance level, which is often denoted as α. To evaluate the severity of inflated type I error rates due to failure to consider data dependencies in realistic scenarios, we simulated data using the dependence structure of example 1. The number of neurons from each of the 24 animals, the number of animals from each of the 5 groups, and the ICCs from example 1, illustrated in Figure 2 and Table 1, were used to generate simulated data. To ensure that the data were simulated under the null hypothesis, the responses in each of the 5 groups were simulated from a multivariate normal distribution, with mean 0 and correlation structure based on the ICC of that group. Thus, the a priori known ground truth is that the 5 groups (baseline [saline], 24, 48, and 72 h and 1 week) share the same population mean.

We simulated 10,000 datasets, each of which was analyzed using the linear model by pooling all of the neurons, or was analyzed using the LME model, to test equal population means of the 5 groups. The histogram of linear model p values indicates that most of the p values are small (Figure 3A, left panel); the type I error rate is ~90% when $\alpha = 0.05$ is used. Thus, with no difference between the 5 groups, the probability that the linear model will reject the null hypothesis is 90%. This strikingly large type I error rate of the linear model confirms that when substantial data dependency exists, the cost of failure to take data dependency into account is very serious due to the higher probability of false positives.

In comparison, the histogram of LME p values is approximately uniform between 0 and 1 (Figure 3B, right panel); if the significance level is chosen at $\alpha = 0.05$, then the estimated type I error rate is 8.6%, which indicates that the LME test is effective in accounting for data dependency. This convincingly illustrates the need for use of the LME in neuroscience research. Next, we provide some background and describe the method of the LME model.

LME model

The word "mixed" in LME means that the model consists of both fixed and random effects. Fixed effects refer to fixed but unknown coefficients for the variables of interest and the explanatory covariates, as identified in the traditional linear model developed by Francis Galton more than a century ago. Random effects, first proposed by Fisher (1919), refer to variables that are not of direct interest; however, they may lead to correlated outcomes. A major difference between fixed and random effects is that the fixed effects are considered unknown parameters, whereas the random effects are considered random variables drawn from a distribution (e.g., a normal distribution). LME was pioneered by C.R. Henderson in his series on animal breeding (Henderson, 1949). It is now widely accepted and has been successfully applied in various scientific disciplines such as economics, genetics, psychology, medicine, and sociology (Fitzmaurice et al., 2012; Jiang and Nguyen, 2021; Laird and Ware, 1982). Depending on the disciplines and application domains, alternative names have been used for LME, including random-effects model, multi-level model, hierarchical model, and variance component model. To apply LME, it is necessary to understand its assumptions and representation in sufficient detail, especially with respect to simpler methods. We start by reviewing the

Table 1. ICC, design effect, and effective sample size for the 5 groups in example 1

	Saline (7 mice)	24 h (6 mice)	48 h (3 mice)	72 h (3 mice)	1 week (5 mice)
No. cells	357	209	139	150	245
ICC	0.61	0.33	0.02	0.63	0.54
Design effect	32.0	17.7	1.8	31.8	26.8
Effective sample size	11.1	17.5	76.9	4.7	9.1

ICC and the design effect were the lowest at 48 h, when the data were relatively homogeneous across animals. At baseline and 72 h, the data were noticeably heterogeneous across animals, leading to high ICC.

2-sample t test, 1-way ANOVA, and the linear model, and then introduce the LME model.

Background: 2-sample t test, 1-way ANOVA, and linear model

We start from the familiar 2-sample case with n_0 observations (Y_1, …, Y_{n0}) from a control group and n_1 observations from a treatment group (Y_{n0+1}, …, Y_{n0+n1}). Under independence and normality assumptions, the t test statistic, which standardizes the difference of the sample means by its standard error, follows a *t* distribution. Equivalently, one can use a simple linear model to model the difference between treatment and control.

Let x_i denote a covariate (predictor) variable such that $x_i = 1$ if the observed outcome Y_i is from a subject assigned to the treatment group and $x_i = 0$ otherwise. Then, we can assume a linear relationship between the outcome and the treatment assignment as follows:

$$Y_i = \beta_0 + x_i \, \mathrm{x}\beta_1 + \varepsilon_i, \; i = 1, \; \ldots, n_0, n_0 + 1, \; \ldots, n_0 + n_1 \quad \text{(Equation 1)}$$

In this model, β_0 is the mean of the control group and ($\beta_0 + \beta_1$) is the mean of the treatment group. The null hypothesis of no effect of the treatment versus control is expressed as H_0: $\beta_1 = 0$ and the test statistic of the well-known t test is identical to the least-squares estimate of the coefficient β_1 divided by its standard error. The ε_i is the random error term. The generalization from 1 treatment to *p* treatments is straightforward since it is possible to use *p* indicator variables, also known as dummy variables, for each of the treatment labels:

$$Y_i = \beta_0 + x_{i,1}\mathrm{x}\beta_1 + \; \ldots \; + x_{i,p}\mathrm{x}\beta_p + \varepsilon_i, i = 1, \; \ldots, \; n, \quad \text{(Equation 2)}$$

where *n* is the total number of observations. In the above multiple linear regression, β_0 indicates the population mean of the reference group (which is often just the control group). Then, each coefficient β_k is the difference in population means between the *k*th treatment and the reference group, since $x_{i,k} = 1$ if observation *i* belongs to the *k*th treatment group and $x_{i,k} = 0$ otherwise. Most often, we are interested in whether there is any difference in population means among all of the ($p + 1$) groups (i.e., H_0: $\beta_1 = \ldots = \beta_p = 0$). If the random errors (ε_i) are independently and identically distributed (i.i.d.) from a normal distribution, then we can use an F-test to assess the null hypothesis H_0. The same F-test is probably more familiar to practitioners from the 1-way ANOVA. The idea is to decompose the total variance of the data into different sources. The 2 sources modeled in the multiple linear regression are the variation due to different treatments and the variation due to randomness. The F statistic used in the F-test characterizes the variation due to treatments relative to the variation due to randomness. Thus, ANOVA, in a broad sense, is a method of understanding the contributions of different factors to an outcome variable by studying the proportion of variance explained by each factor (Gelman, 2005).

Unfortunately, ANOVA is frequently misused in neuroanatomical and neurophysiological studies due to a failure of the practitioner to account for the collection of multiple observations from the same animal. Many investigators tend to use the default setup in statistical software or packages, and they may not be familiar with more advanced regression frameworks. ME models are a generalization of the previous methods (t test, ANOVA, linear model) and provide researchers with an effective strategy to analyze correlated data by taking dependence into account.

A practical guidance to the LME model

We consider the data in example 1. The data consist of 1,200 observed pCREB immunoreactivity values from 24 mice under 5 groups, which include the baseline group (7 mice) and 24 h (6 mice), 48 h (3 mice), 72 h (3 mice), and 1 week after ketamine treatment (5 mice), as shown in Table 1 and Figure 2. Here, the data are recorded as multiple measurements from each mouse, which represents a single unit (cluster) of analysis. Let Y_{ij} indicate the *j*th observation of the *i*th mouse, and ($x_{ij,1}$, …, $x_{ij,4}$) are the dummy variables for the treatment labels, with $x_{ij,1} = 1$ for 24 h, $x_{ij,2} = 1$ for 48 h, $x_{ij,3} = 1$ for 72 h, and $x_{ij,4} = 1$ for 1 week after ketamine treatments, respectively. Because there are multiple observations from the same animal, the data are naturally clustered by animal. We account for the resulting dependence by adding an animal-specific effect to the regression framework discussed in the previous section, as follows:

$$Y_{ij} = \beta_0 + x_{ij,1}\mathrm{x}\beta_1 + \; \ldots \; + x_{ij,4}\mathrm{x}\beta_4 + u_i + \varepsilon_{ij}, i = 1, \; \ldots, \; 24; j = 1, \; \ldots, \; n_i, \quad \text{(Equation 3)}$$

where n_i is the number of observations from the *i*th mouse, u_i indicates the deviance between the overall intercept β_0 and the mean specific to the *i*th mouse, and ε_{ij} represents the deviation in pCREB immunoreactivity of observation (cell) *j* in mouse *i* from the mean pCREB immunoreactivity of mouse *i*. Among the coefficients, the coefficients of the fixed-effects component, (β_0, β_1, β_2, β_3, β_4), are assumed to be fixed but unknown, whereas (u_1, …, u_{24}) are treated as independent and identically distributed random variables from a normal distribution with mean 0 and a variance parameter that reflects the variation across animals. It is important to notice that the cluster/animal-specific means are more generally referred to as random intercepts in an LME. Equivalently, one could write the previous equation by using a vector ($z_{ij,1}$, …, $z_{ij,24}$) of dummy variables for the cluster/animal IDs such that $z_{ij,k} = 1$ for $i = k$ and 0, otherwise:

$$Y_{ij} = \beta_0 + x_{ij,1}\mathrm{x}\beta_1 + \; \ldots \; + x_{ij,4}\mathrm{x}\beta_4 + z_{ij,1}u_1 + \; \ldots \; + z_{ij,24}u_{24} + \varepsilon_{ij},$$

$$I = 1 \;\; 24; j = 1, \; \ldots, n_i \quad \text{(Equation 4)}$$

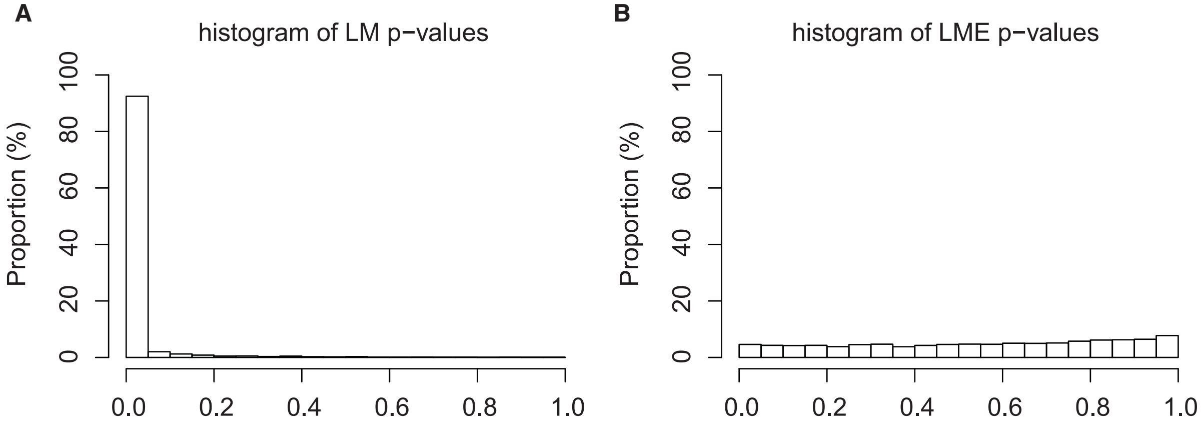

Figure 3. Histograms of p values using simulated data that assume (1) no treatment effects and (2) the same sample sizes and correlation structure with example 1

(A) Histogram of the p values from the inappropriate method (linear model) shows that ignoring the correlation structure of the data led to a surprisingly high type I error rate (90%) at significance level $\alpha = 0.05$.

(B) Histogram of the p values from LME.

In the model above, Y_{ij} is modeled by 4 components: the overall intercept β_0, which is the population mean of the reference group in this example, the fixed effects from the covariates ($x_{ij,1}, \ldots, x_{ij,4}$), the random effects due to the clustering ($z_{ij,1}, \ldots, z_{ij,24}$), and the random errors ε_{ij}'s, assumed to be i.i.d. from a normal distribution with mean 0.

In the application of these methods, one practical issue is to determine which effects should be treated as fixed and which should be considered random. A number of definitions of fixed effects and random effects have been given (Gelman and Hill, 2006). It is generally agreed that a fixed effect captures a parameter at the population level; as such, it should be a constant across subjects/clusters. Population-level treatment effects, which are often of direct scientific interest, are included in the fixed effects. When scientifically relevant, predictors (e.g., age, gender) whose effects are not expected to change across subjects should also be treated as fixed effects. In contrast, a random effect captures cluster-specific effects (e.g., due to the animal or the cell considered), which are only relevant for capturing the dependence among observations and are typically of no direct relevance for assessing scientific hypotheses. The mice in a study are a sample from a large population and they are randomly chosen among all of the possible mice. Thus, the animal-specific effects are often not of primary interest; hence, they are added to the random-effects component. In example 1, the mean in pCREB immunoreactivity from a particular mouse is not relevant for the final analysis; however, including the mouse-specific means accounts for the correlation between observations from the same animal.

In addition to cluster-specific means, an LME model may include additional terms that describe the variability observed within a cluster (e.g., animal, cell). Most often, this is the case when measurements are taken at different times from within the same animal and cell, and it may be important to account for possibly different cluster-specific trajectories over time. We discuss this in more detail as it pertains to example 3 below.

The LME in a matrix format

It is often convenient to write the LME in a very general matrix form, which was first derived in Henderson et al. (1959). This format gives a compact expression of the LME model, as follows:

$$Y = \mathbf{1}\beta_0 + X\beta + Zu + \varepsilon, \qquad \text{(Equation 5)}$$

where Y is an $n \times 1$ vector of individual observations; $\mathbf{1}$ is the $n \times 1$ vector of ones; the columns of X are predictors whose coefficients β, a $p \times 1$ vector, are assumed to be fixed but unknown; the columns of Z are the variables whose coefficients u, a $q \times 1$ vector, are random variables drawn from a distribution, with mean 0 and a partially or completely unknown covariance matrix; and ε is the residual random error.

In addition to being compact, the matrix form is convenient from a data analysis perspective, since many software packages for LMEs often require that the data are organized according to the "long format"—each row of the dataset contains only the values for 1 observation. For example, using the long format, the data in example 1 can be stored in a matrix with 1,200 rows; the dummy variables introduced above the Supplemental Information for the treatment labels and the cluster/animal identification numbers are used as the columns for X and Z, respectively. Because many software packages such as MATLAB and R can take categorical variables and convert them to dummy variables automatically in their internal computation, the data for example 1 can be stored in a 1,200 × 3 matrix, with the first column being the pCREB immunoreactivity values, the second column being the treatment labels, and the last column being the animal identification numbers (see the supplemental information).

Since the LME model consists of both fixed and random effects, it is highly versatile and includes the traditional linear regression model (linear model), random-effects model, t test, paired t test, ANOVA, and repeated ANOVA as special cases. In fact, software implementing the LME model can also be

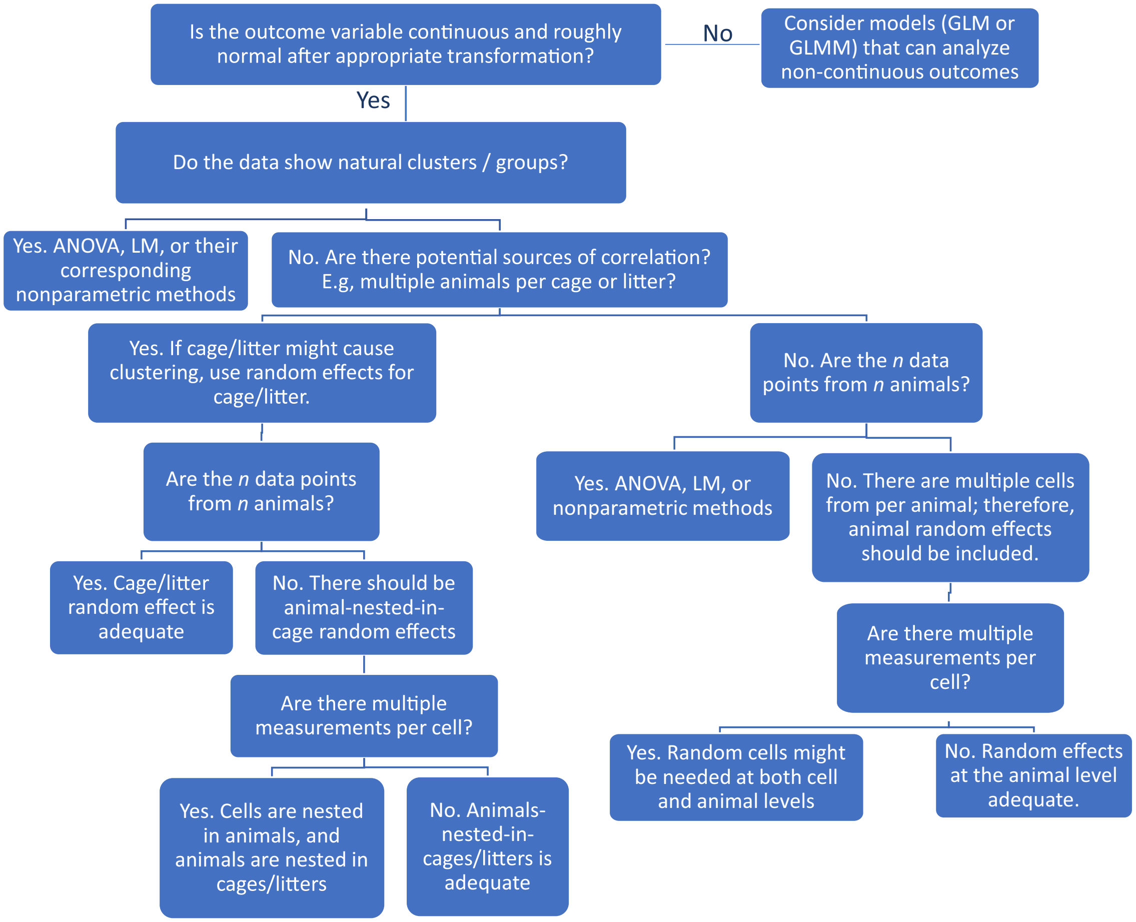
Is the outcome variable continuous and roughly normal after appropriate transformation?
No
Consider models (GLM or GLMM) that can analyze non-continuous outcomes
Yes
Do the data show natural clusters / groups?
Yes. ANOVA, LM, or their corresponding nonparametric methods
No. Are there potential sources of correlation? E.g, multiple animals per cage or litter?
Yes. If cage/litter might cause clustering, use random effects for cage/litter.
No. Are the n data points from n animals?
Are the n data points from n animals?
Yes. ANOVA, LM, or nonparametric methods
No. There are multiple cells from per animal; therefore, animal random effects should be included.
Yes. Cage/litter random effect is adequate
No. There should be animal-nested-in-cage random effects
Are there multiple measurements per cell?
Are there multiple measurements per cell?
Yes. Random cells might be needed at both cell and animal levels
No. Random effects at the animal level adequate.
Yes. Cells are nested in animals, and animals are nested in cages/litters
No. Animals-nested-in-cages/litters is adequate

Figure 4. A decision chart for setting up ME model analysis
This basic decision chart shows in a stepwise fashion how to identify the ME application scenarios and random effects.

used to implement the linear model, ANOVA, 2-sample t test, paired t test, and other methods. To determine whether and which LME model should be used, one needs to understand the sources of correlation. Data visualization, as depicted in Figure 2, is the first step we recommend to gain a good understanding of the data. It is helpful to have a visual inspection of model assumptions, especially regarding whether there is any data dependency due to factors that should be modeled. The decision chart in Figure 4 provides a user-friendly guide to determine whether some variables should be included in the matrix Z to model the correlation in animal experiments appropriately. (Please also refer to the Practical applications of the LME and GLMM implementation details below.)

GLMM

In this section, we discuss how to model data dependency for a broader range of outcome types. Traditional linear models and the LME are designed to model a continuous outcome variable with a fundamental assumption that its variance does not change with its mean. This assumption can be violated for commonly collected outcome variables, such as the choice made in a 2-alternative forced choice task (binary data), the proportion of neurons activated (proportional data), the number of neural spikes in a given time window, and the number of behavioral freezes in each session (count data). For example, a natural choice of distribution for count data is the Poisson distribution, for which its mean and variance are equal. This violates the homoscedasticity (meaning "constant variance") assumption that is a fundamental assumption of a standard linear regression model. In addition, negative predictive values may occur in a linear model, which is undesirable for count or proportional data. These issues can be addressed by the generalized linear model (GLM) framework, which is an important extension of the linear model.

We present a unified framework to analyze various outcome types, known as the GLM (McCullagh and Nelder, 2019; Nelder and Wedderburn, 1972). It includes the conventional linear regression (for continuous variables), logistic regression (for binary outcomes), and Poisson regression (for count data) as special cases. Let Y_i be the ith outcome variable and $X_i = (X_{i,1}, \ldots, X_{i,p})$ be the corresponding covariates. The critical operation of GLM is to link the expected value of Y_i and a linear predictor (i.e., a linear combination of the covariates) through a "link" function g:

$$g(E(Y_i|X_i)) = \beta_0 + X_{i,1}x\beta_1 + \ldots + X_{i,p}x\beta_p \qquad \text{(Equation 6)}$$

The link function g connects the expected mean of the outcome variable to a linear predictor. An equivalent expression is $E(Y_i \mid X_i) = g^{-1}(\beta_0 + X_{i,1} \text{ x } \beta_1 + \ldots + X_{i,p} \text{ x } \beta_p)$, where g^{-1} denotes the inverse function of g. For example, the link function of the linear regression model is the identity function, which implies that

$$E(Y_i|X_i) = \beta_0 + X_{i,1} x\beta_1 + \ldots + X_{i,p} x\beta_p \qquad \text{(Equation 7)}$$

To further help delineate the link function g, we then consider the situation in which the outcome variable is binary, which is often modeled using a logistic regression. Note that a logistic regression is a special GLM, with the link function g being the *logit* function; in other words, we model the *logit*-transformed success probability using a linear combination of the covariates:

$$\log it(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + X_{i1} \times \beta_1 + \cdots + X_{ip} \times \beta_p, \qquad \text{(Equation 8)}$$

where the success probability $\pi_i = E(Y_i \mid X_i) = \Pr(Y_i = 1 \mid X_i)$, with the latter equation due to the fact that Y_i is either 0 or 1. The *logit* function ensures that the estimated success probabilities are always between 0 and 1, thus preventing negative predictive values or predictive values >1. To complete the specification of the model, a data-generating mechanism for the outcomes is needed. One natural choice is the Bernoulli distribution:

$$Y_i|\pi_i \sim Bernoulli(\pi_i) \qquad \text{(Equation 9)}$$

The corresponding likelihood function can then be used to make inferences about parameters using the maximum likelihood. The distributional assumptions can be relaxed by specifying the relationship between mean and variance, rather than the full distribution, which is expected to have good robustness. This approach is known as the quasi-likelihood method. We refer interested readers to Wedderburn (1974). The GLM generalizes the conventional LM for various types of outcomes by using appropriate link functions and by distributional assumptions of the outcomes. Like the conventional linear model, all of the coefficients in the GLM are assumed to be unknown but fixed parameters. Next, we further extend GLM to GLMMs so that the data dependence due to the underlying experimental design can be appropriately accounted for by including random effects.

To account for data dependency, the GLM has been extended to the GLMM (Breslow and Clayton, 1993; Liang and Zeger, 1986; Stiratelli et al., 1984; Wolfinger and O'connell, 1993; Zeger and Karim, 1991; Zeger and Liang, 1986):

$$\begin{aligned} g(E(Y_{ij}|X_i)) &= \beta_0 + X_{ij,1}x\beta_1 + \ldots + X_{ij,p}x\beta_p \\ &\quad + Z_{ij,1}xu_1 + \ldots + Z_{ij,q}xu_q \end{aligned} \qquad \text{(Equation 10)}$$

The random-effects terms in LME (Equation 4) and GLMM (Equation 10) play the same role; they explicitly model the dependence structure by specifying subject-specific or other relevant random effects and their joint distribution. With appropriate assumptions on the distribution of the outcome variables Y_{ij} and the mean assumption specified in Equation 10, likelihood-based approaches are often used for parameter estimation. Compared to LME, the computation involved in GLMM with non-normal data is substantially more challenging, both in computational speed and stability. As a result, several strategies have been developed to approximate the likelihood (Bolker et al., 2009).

A robust alternative is the generalized estimating equation (GEE) (Zeger et al., 1988) approach. GEE makes assumptions based on the first 2 moments rather than imposing explicit distributional assumptions. The idea of GEE is to estimate coefficients using a "working" correlation structure, which does not have to be identical to the unknown underlying true correlation. An incorrect correlation structure, while it would not bias the estimates, would affect the estimate of the variance. Thus, a correction approach is applied to obtain consistent estimates of variance and covariance. However, caution is merited, as GEE and GLMM may lead to different estimates and interpretations (Fitzmaurice et al., 2012). Moreover, the correction procedure in GEE relies on aggregated information across subject-level data, but for cases of animal studies that only use a few animals in an experiment, the accuracy of GEE results may be questionable.

Bayesian analysis

In the LME and GLMM framework, the random-effects coefficients are drawn from a given distribution (typically Gaussian). Therefore, Bayesian analysis provides a natural alternative for analyzing the data considered in this Primer. One inherent advantage of Bayesian analysis is that it is easy to incorporate prior information on all of the parameters in the model, including both the fixed-effects coefficients and the parameters involved in the variance-covariance matrices. In particular, the Bayesian framework allows practitioners to consider distributions of the random effects that are far from Gaussian, or to consider more flexible covariance structures needed to characterize the underlying data-generating process. In the frequentist framework (see Glossary Box 1 and the supplemental information), computational algorithms can become formidably complex and prohibitive in those cases. The Bayesian framework obtains inferences on the parameters of interest by means of the posterior distribution, which results from combining the prior information with the data using the Bayes' theorem. Therefore, Bayesian inference does not rely on asymptotic approximations that may be invalid with limited sample sizes.

To describe how Bayesian analysis works for ME model, consider again the model (Equation 4) in LME model:

$$Y_{ij} = \beta_0 + X_{ij,1}x\beta_1 + \ldots + X_{ij,p}x\beta_p + Z_{ij,1}xu_1 + \ldots + Z_{ij,q}xu_q + \varepsilon_{ij} \qquad \text{(Equation 11)}$$

For simplicity of presentation and to avoid advanced statistical and mathematical details required for more general models, we assume i.i.d. random effects (i.e., the random effects are i.i.d. from $N(0, \sigma^2_u))$. We also assume the errors are i.i.d. from $N(0, \sigma^2)$. While we focus here for simplicity on the linear model (Equation 4) from "LME model", our discussion can also be extended to the generalized linear framework of "GLMM." Using the Bayes' theorem, the posterior distribution,

Table 2. p values for comparing pCREB immunoreactivity at each time point (24 h, 48 h, 72 h, and 1 week) after ketamine treatment to the baseline (saline)

	Overall	24 h	48 h	72 h	1 week
Linear model (ANOVA)	1.2×10^{-78}	6.0×10^{-38}	6.8×10^{-26}	0.0291	1.1×10^{-8}
LME	0.0029	0.0049	0.0164	0.5601	0.2525

The "Overall" column corresponds to the null hypothesis of no difference among the 5 groups (example 1). The LME p values are based upon the *lme* function in the *nlme* package, in which the denominator degrees of freedom are determined by the animal grouping level (Pinheiro et al., 2007). The methods for obtaining more accurate p values with adjustments for multiple comparisons can be found in the supplemental information.

$f(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2_u, \sigma^2|Y)$, is proportional to the product of the likelihood function $f(Y|\beta_0,\ \beta_1\ldots,\beta_p,\sigma_u^2,\sigma^2)$ and the prior distribution $\pi(\beta_0,\beta_1,\ldots,\beta_p,\sigma_u^2,\sigma^2)$ (summarizing the available knowledge on the parameters):

$$f(\beta_0,\ \beta_1,\ldots,\beta_p,\sigma_u^2,\sigma^2|Y) = \frac{f(Y|\beta_0,\ \beta_1\ldots,\beta_p,\sigma_u^2,\sigma^2)\,\pi(\beta_0,\beta_1,\ldots,\beta_p,\sigma_u^2,\sigma^2)}{f(Y)}, \qquad \text{(Equation 12)}$$

where *f*(*Y*) is a constant that depends only on the observed data but does not depend on the model parameters. If possible, the prior distribution $\pi(\beta_0,\beta_1,\ldots,\beta_p,\sigma_u^2,\sigma^2)$ should be chosen to reflect the beliefs or information that investigators may have about the parameters. In the absence of prior knowledge about the parameters, uninformative prior distributions are often used. These types of priors are also known as flat, weak, objective, vague, or diffuse priors. For example, a uniform distribution over a wide range or a normal distribution with a very large variance can be regarded as a weak prior for the fixed-effects coefficients.

Once the likelihood and the priors have been specified, Bayesian inference often requires the use of sophisticated sampling methods to obtain quantities from the posterior distribution, generally denoted as Markov chain Monte Carlo (MCMC) algorithms such as the Gibbs sampling (Gelfand and Smith, 1990), the Metropolis-Hastings algorithm (Casella and George, 1992; Hastings, 1970; Metropolis et al., 1953), and the Hamiltonian Monte Carlo algorithm (Betancourt, 2017; Duane et al., 1987; Hoffman and Gelman, 2014; Neal, 2011; Shahbaba et al., 2014). However, in practical applications, it is possible to use existing software packages to conduct Bayesian analyses of ME models without the necessity of in-depth knowledge of the underlying computational details (Bürkner, 2017, 2018; Fong et al., 2010; Hadfield, 2010). Inference on a parameter can then be conducted using its marginal posterior distribution. For example, one can consider the mean of the posterior distribution as a point estimate of the unknown parameter as well as a 95% credible interval to obtain the Bayesian counterpart of a confidence interval in frequentist analysis. In a Bayesian framework, the 95% credible interval is an uncertainty estimate that identifies the shortest interval containing 95% of the posterior distribution of the parameter of interest (highest posterior density interval). Hypothesis testing on the parameters of the ME models can be conducted by comparing the marginal likelihoods under 2 competing models, via the so-called Bayes factor. The use of a Bayesian approach and Bayes factors has been sometimes advocated as an alternative to p values since the Bayes factor represents a direct measure of the evidence of one model versus the other (Benjamin and Berger, 2019; Held and Ott, 2018; Kass and Raftery, 1995).

PRACTICAL APPLICATIONS OF THE LME AND GLMM

We provide practical examples to demonstrate why conventional linear models, including t test and ANOVA, fail for the analysis of correlated data, and why LME should be used instead, with its advantages in each practical example explained.

Example 1

As described in "Important concepts and definitions related to statistical testing", we measured pCREB immunoreactivity of 1,200 putative excitatory neurons in the mouse visual cortex at different time points: collected at baseline (saline), 24, 48, and 72 h and 1 week following ketamine treatment, collected from 24 mice (Figure 2). If we use ANOVA or a linear model to compare each time point to the baseline (saline), as shown in Table 1, we find that the p values of all of the comparisons are <0.05 and the overall difference between the 5 groups is highly significant ($p = 1.2 \times 10^{-78}$). However, recall that the 1,200 neurons are clustered in 24 mice. The ICC, D_{eff}, and n_{eff}s (Table 1) indicate that the dependency due to clustering is substantial. Therefore, the 1,200 neurons should not be treated as 1,200 independent cells. The lesson from this example is that the number of observational units is much larger than the number of experimental units (see Lazic et al. [2018] for helpful discussion). We used an LME with animal-specific random effects to handle the dependency due to clustering. The p values are much larger than those from the linear model; thus, they are less likely to reach the threshold of significance (Table 2). Note that the difference between saline and 72 h or 1 week by LME analysis is not significant after accounting for the dependency of the data.

Example 2

Data were derived from an experiment designed to determine how *in vivo* calcium (Ca^{2+}) activity of PV cells (measured longitudinally) changes over time after ketamine treatment (Grieco et al., 2020). Ca^{2+} event frequencies were measured from the brain cells of 4 mice at 24, 48, and 72 h and 1 week after ketamine treatment; Ca^{2+} event frequencies at 24 h were compared to the other 3 time points. In total, Ca^{2+} event frequencies of 1,724 neurons were measured. The boxplot in Figure 5A and the linear model (or ANOVA, t test) analysis results in Table 3 indicate significantly reduced Ca^{2+} activity at 48 h relative to 24 h with $p = 4.8 \times 10^{-6}$, and significantly increased Ca^{2+} event frequency at 1 week compared to 24 h with $p = 2.4 \times 10^{-3}$. However, if we account for repeated measures due to cells clustered in mice using LME with random intercepts (the model is similar to Equation 4), most of the p values are >0.05 and

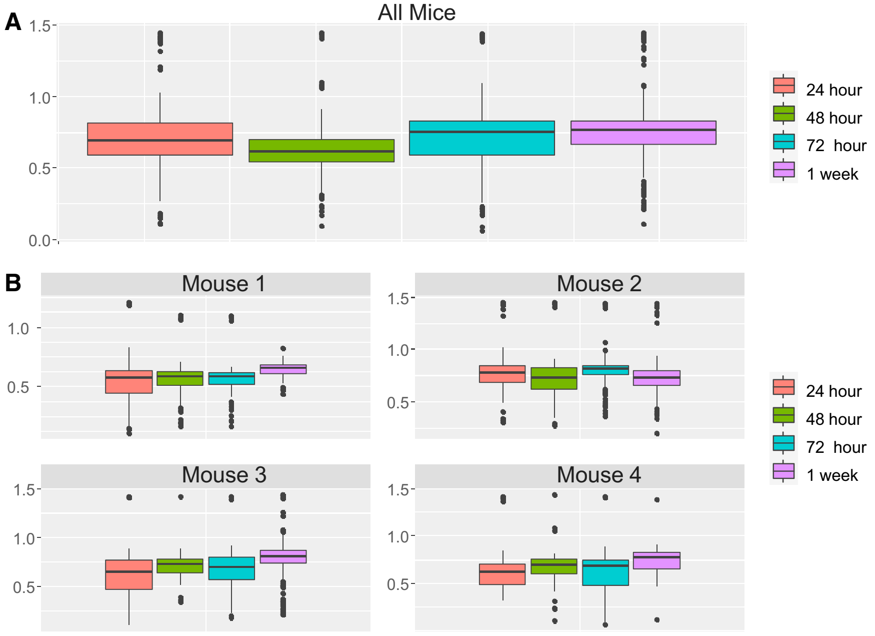

Figure 5. Weighting effects from single animals

When data from different animals are naively pooled, the result can be dominated by the data from a single animal (example 2). To illustrate this point, we present the boxplots of Ca^{2+} event frequencies measured at 4 time points using 2 different ways. The median (horizontal bar) and 25th and 75th quartile values (lower and upper boundaries of the box) of each data set are represented in each box.

(A) Boxplot of Ca^{2+} event frequencies using the pooled neurons from 4 mice. ANOVA or t test showed that Ca^{2+} activity was significantly reduced at 48 h relative to 24 h, with $p = 4.8 \times 10^{-6}$, and significantly increased Ca^{2+} activity at 1 week compared to 24 h with $p = 2.4 \times 10^{-3}$.

(B) However, when looking at boxplots of Ca^{2+} event frequencies stratified by individual mice, these changes occur only in mouse 2. This is because mouse 2 contributed 43% of the cells, which likely explains why the pooled data are more similar to mouse 2 than to other mice. Note that the comparisons are not significant if we account for repeated measures due to cells clustered in mice using LME, thus avoiding an erroneous conclusion.

thus fail to reach significance, except that the overall p value is 0.04.

To understand the discrepancy between the results from the linear and LME models, we created boxplots for the pooled data and for each mouse (Figure 5B). Although the pooled data (Figure 5A) and the corresponding p value from the linear model show a significant reduction in Ca^{2+} activities from 24 to 48 h, we noticed that the only mouse showing a noticeable reduction was mouse 2. In fact, a close examination of Figure 5B suggests that there may be small increases in the other 3 mice. To examine why the pooled data follow the pattern of mouse 2 and not that of other mice, we checked the number of neurons in each of the mouse × time combinations (Table 4). The last column of Table 4 shows that mouse 2 contributed 43% of all cells, which likely explains why the pooled data are more similar to mouse 2 than to the other mice. The lesson from this example is that naively pooling data from different animals is a potentially dangerous practice, as the results can be dominated by a single animal that can misrepresent a substantial proportion of the measured data. Investigators limited to using the linear model often notice outlier data of a single animal, and they may agonize about whether they are justified in "tossing that animal" from their analysis, sometimes by applying "overly creative post-hoc exclusion criteria." The other way out of this thorny problem is the brute force approach of repeating the experiment with a much larger sample size—a more honest, but expensive solution. The application of LME solves this troubling potential problem as it takes dependency and weighting into account.

In this example, there are only 4 mice. This number may be smaller than the one recommended for using random-effects models. However, as discussed in Gelman and Hill (2006), using a random-effects model in this situation will not provide much gain versus simpler analyses, but it probably will not do much harm either. An alternative would be to include the animal identification variable as a factor with fixed animal effects in the conventional linear regression. However, a recent study suggests that clusters should be modeled using random effects as long as the software does not incur any computational issue such as flags due to convergence (Oberpriller et al., 2021). Note that neither of the 2 analyses is the same as fitting a linear model to the pooled cells together, which erroneously ignores the between-animal heterogeneity and fails to account for the data dependency due to the within-animal similarity. In a more extreme case, for an experiment using only 2 monkeys, for example, naively pooling the neurons from the 2 animals incurs the risk of drawing conclusions mainly from 1 animal and unrealistic homogeneous assumptions across animals, as discussed above. A more appropriate approach is to analyze the animals separately and check whether the results from these 2 animals "replicate" each other. Exploratory analysis such as data visualization is highly recommended to identify potential issues.

Example 3

In this experiment, Ca^{2+} event-integrated amplitudes are compared between baseline (saline) and 24 h after ketamine treatment (Grieco et al., 2020). A total of 1,248 cells were sampled from 11 mice, and each cell was measured twice (baseline and after ketamine treatment). As a result, correlation arises from both cells and animals, which creates a 3-level structure: repeated measurements (baseline and after treatment) within cells and cells within animals. It is clear that the ketamine treatment should be included as a fixed effect. The choice of the random effects deserves more careful consideration. The hierarchical structure (i.e., 2 observations per cell and multiple cells per animal) suggests that the random effects of the cells should be nested within individual mice. We

Table 3. The results (estimates ± SE and p values) for the Ca^{2+} event frequency data using linear model and LME (example 2)

	48 h	72 h	1 week
Linear model (est)	−0.078 ± 0.017	0.009 ± 0.017	0.050 ± 0.016
Linear model (p)	4.8×10^{-6}	0.595	2.4×10^{-3}
LME (est)	−0.011 ± 0.014	0.020 ± 0.014	0.025 ± 0.014
LME (p)	0.424	0.150	0.069

Table 4. Number of neurons by mouse and time in example 2

	24 h	48 h	72 h	1 week	Total (%)
Mouse 1	81	254	88	43	466 (27)
Mouse 2	206	101	210	222	739 (43)
Mouse 3	33	18	51	207	309 (18)
Mouse 4	63	52	58	37	210 (12)
Total	383	425	407	509	1,724 (100)

In total, Ca^{2+} event frequencies at 1,718 neurons were measured. When splitting the number by mouse, mouse 2 has the largest number of measured neurons (43%). Thus, when pooling the cells naively, the overall results would be dominated by the results observed in mouse 2.

consider a basic model that includes random intercepts at both cell and animal levels:

$$Y_{ijk} = \beta_0 + x_{ijk} x \beta_1 + u_i + u_{ij} + \varepsilon_{ijk}, i = 1, \ldots, 11; j = 1, \ldots, n_i; k = 0, 1, \quad \text{(Equation 13)}$$

where the indices i, j, and k stand for the ith mouse, the jth cell, and the kth measurement of neuron j from mouse i. Similarly, $x_{ijk} = 1$ if the measurement is taken after treatment and 0 if it is taken at baseline. By including the cell variable in the random effect, we implicitly capture the change from "before" to "after" treatment for each cell. This is similar to how paired data are handled in a paired t test. Moreover, by specifying that the cells are nested within individual mice, we essentially model the correlations within both mouse and cell levels. As explained in the supplemental information, part II, example 3, when the cell identifications are not unique, specifying nested random effects is necessary; otherwise, 2 cells with the same cell identification from 2 different mice will be considered as sharing a cell-specific effect (known as crossed random effects, in comparison to nested random effects), which does not make sense. We recommend that users use unique cell identification numbers across animals to avoid confusion and mistakes in the model specification.

For the treatment effect, LME and the linear model produce similar estimates; however, the standard error of the linear model was larger. Thus, the p value based on LME was smaller (0.0036 for the linear model versus 0.0001 for LME). In this example, since the 2 measures from each cell are positively correlated (Figure 6), the variance of the differences is smaller when treating the data as paired than as independent. As a result, the more rigorous practice of using cell effects as random effects leads to lower but more accurate p values. The lesson in this example is that the LME can actually yield lower p values than conventional approaches. This is opposite to example 1 and example 2 and dispels the potential notion that LME incurs a "cost" by always leading to greater p values. Rigorous statistical analysis is not a hunt for the smallest p value (commonly known as p-hacking or significance chasing); the objective of the experimenter should be always to use the most appropriate and thorough analysis method.

In this example, the random effects involve >1 level, and the LME model we fit includes neuron-specific and animal-specific random intercepts. Sometimes, models incorporating additional random effects may be appropriate to account for additional sources of variability (Barr et al., 2013; Ferron et al., 2002; Heisig and Schaeffer, 2019; Kwok et al., 2007; Matuschek et al., 2017). For example, both the overall mean levels and the treatment effects may vary across animals and neurons. A mouse may have a higher (lower) treatment response than the average population response, for example, due to unobserved individual physiology. The plausibility of including extra random effects can often be assessed visually by linearly interpolating the observed response over the values of the predictor of interest in each cluster (e.g., all of the recorded Ca^{2+} event integrated amplitudes pre- and post-treatment within a specific animal); that is, by conducting a linear model regression within each cluster. Suppose the interpolation suggests that the slopes of the regression differ across clusters/animals along with their intercepts. In that case, the LME may incorporate both random intercepts and random slopes to capture how each mouse responds differently to the treatment. It may also be helpful to allow correlations between the different random-effects components. In the example considered here, there is a nested structure of clusters: cells within animals. Therefore, it is possible to conceive 3 other models with additional random effects: a model that includes random slopes only at the neuron level, a model with random slopes only at the animal level, and a model with random slopes for both neurons and animals. By conducting likelihood ratio tests to compare these models, we find that including random slopes at the neuron level leads to substantial improvement in the likelihood. However, random slopes at the animal level seem unnecessary. More detailed analyses and technical remarks are provided in our accompanying supplemental information. It should be noted that the modeling decisions should not be based on tests and p values alone, as the result may be significant even with a very small effect size if the sample size is large enough or be insignificant with a moderate or large effect size for small sample sizes. Rather, the modeling decision should always be guided by the combined information provided by the study design, scientific reasoning, and previous evidence. For example, different animals are expected to have different mean levels on outcome variables; thus, it is reasonable to model the variation due to animals by considering animal-specific random effects. A similar argument is the inclusion of baseline covariates such as age in many biomedical studies, even when they are not significant. Also, when random slopes are included, it is typically recommended to include the corresponding random intercepts. If random slopes (for treatment) are included at the animal level, then it is sensible to also include the animal-specific random intercepts.

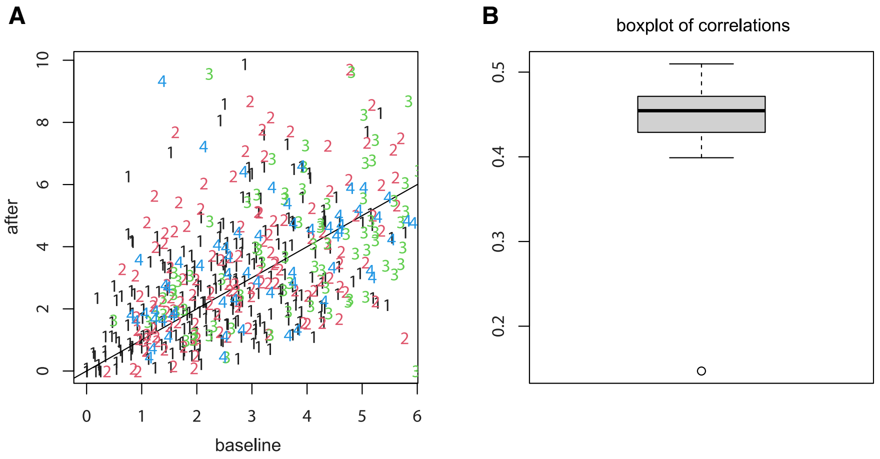

Figure 6. LME does not always lead to larger p values than methods that ignore data dependencies

(A) The scatterplot of the Ca^{2+} event integrated amplitude at baseline versus 24 h after treatment for the neurons from 4 example mice (labeled 1, 2, 3, and 4) indicates that the baseline and after-treatment measures are positively correlated.

(B) Boxplot of the baseline and after-treatment correlations of all 11 mice. Due to the positive correlations shown in the data, the variance of differences is smaller when treating the data as paired than as independent. As a result, LME produced a smaller p value than the t test. The median (horizontal bar) and 25th and 75th quartile values (lower and upper boundaries of the box) of the data set are represented.

Example 4

In this example, we illustrate how to use both frequentist and Bayesian GLMM approaches to analyze binary outcomes. The dataset analyzed here is simulated based on a published study (Wei et al., 2020), in which 8 mice were trained in a tactile delayed response task to use their whiskers to predict the location (left or right) of a water pole and report it with directional licking (lick left or lick right). The behavioral outcome we are interested in is whether the animals made the correct predictions. Therefore, we code correct left or right licks as 1 and erroneous licks as 0. In total, 512 trials were generated in our simulation, which includes 216 correct trials and 296 wrong trials. One question we would like to answer is whether a particular neuron is associated with the prediction. For this purpose, we analyze the prediction outcome and mean neural activity levels (measured by neuronal calcium signal changes, dF/F) from the 512 trials using a GLMM. The importance of modeling correlated data by introducing random effects has been shown in examples 1–3. In this example, we focus on how to interpret results from a GLMM model for the mouse behavioral and imaging experiment.

The result from a frequentist approach shows that with the increase of 1% of mean calcium intensity (dF/F), the odds that the mice will make a correct prediction will increase by 6.4% (95% confidence interval: 2.4%–10.7%) and the corresponding p value is 0.0016 based on the large-sample Wald test. The large-sample likelihood ratio test and a parametric bootstrap test give similar p values.

The Bayesian analysis requires the specification of the prior distributions for the model parameters. Due to the lack of prior information, we select priors that are relatively non-informative (i.e., those have large variances around their means). More specifically, we use a normal prior with mean 0 and large standard deviation 10 for the fixed-effect coefficients. For the variances of the random intercept and the errors, we imposed a half-Cauchy distribution with a scale parameter of 5. The results showed that the odds that the mice will make a correct prediction increase by 6.2% (95% credible interval: 2.0%–10.6%) with every 1% increase in dF/F. The Bayes factor of the model with dF/F versus the null model is 5.02; in other words, the posterior odds of the model with dF/F to the null model is 5 times that of the prior odds, suggesting a moderate association of dF/F with prediction (Held and Ott, 2018; Kass and Raftery, 1995). These results are comparable to those from the frequentist GLMM in the preceding paragraph.

RESOURCES

We provide effective and easy-to-follow instructions for the implementation of LME and GLMM with access to the R code, with practice datasets to help with such analysis and results interpretation in the supplemental information. We choose R because it is a free and open source software (CRAN) (R Development Core Team, 2020), widely adopted by the data science community. One major advantage of R over other open source or commercial software is that R has a rich collection of user-contributed packages (>15,000), greatly facilitating a programming environment for developers and the access to cutting-edge statistical methods. There are many statistical packages. A selected (but not complete) list of packages that provide statistical inference and tools for ME models is summarized in Table 5. Our sample code, explanations, and interpretations of results from *lme4* (Bates et al., 2014), *nlme* (Pinheiro et al., 2007), *icc* (Wolak and Wolak, 2015), *pbkrtest* (Halekoh and Højsgaard, 2014), *brms* (Bürkner, 2017; Bürkner, 2018), *lmerTest* (Kuznetsova et al., 2017), *emmeans* (Lenth et al., 2019), *car* (Fox and Weisberg, 2018), and *sjPlot* (Lüdecke, 2018) are provided in the supplemental information.

DISCUSSION AND CONCLUSIONS

Our goal was to raise awareness of the widespread issue in correlated data analysis by t test and ANOVA and to introduce effective solutions and provide clear guidance on how to analyze data that are clustered or have repeated measurements. We note that the issues raised in our article should be considered ideally in the first steps of experimental design, rather than as post hoc applications. Prior knowledge based on direct experience, information from published literature, or pilot studies on the possible ranges of ICC are useful for optimizing statistical power with fixed available resources. For repeated measurements

Table 5. Selected R packages and functions for mixed-effects modeling and statistical inference

Package name	Functions related to mixed-effect modeling
nlme	*lme*: fit a linear mixed-effects model
lme4	*lmer*: fit a linear mixed-effects model *glmm*: fit a generalized linear mixed-effects model
brms	It can conduct Bayesian mixed-effects modeling
lmerTest	It can perform hypothesis testing on fixed and random effects based on models from lme4::lmer
emmeans	It can provide adjusted p values for pairwise and treatments versus control comparisons
pbkrtest	It can perform the F-test (Kenward-Roger and Satterthwaite type) and parametric bootstrap test
car	car::Anova provides large-ample Wald test or F-test with Kenward-Roger denominator degrees of freedom
sjPlot	It can provide visualization and create manuscript-style tables

involving a single level of clusters, formulas to obtain the optimal number of clusters (e.g., animals) and the number of observations per cluster (e.g., cells) can be determined (Aarts et al., 2014). For more complicated scenarios, simulation-based methods seem to be more suitable for accurate power analysis and sample size calculations (Green and MacLeod, 2016).

One may be tempted to use summary statistics such as cluster means to remove correlations due to animal effects. These approaches are not applicable to all experimental designs, such as those involving crossed random effects (Baayen et al., 2008). When methods based on summary statistics work, they give correct type I error rates, but they often have lower power than LME (Aarts et al., 2014; Galbraith et al., 2010). Compared to LME, the paired t test and repeated ANOVA are far more familiar to most researchers. For simple designs such as paired samples or balanced designs, they are still valuable tools; however, they can be less efficient in the presence of missing data. For example, repeated ANOVA implements list-wise deletion (i.e., the entire list or case will be deleted if a single measure is missing). Since an incomplete case still provides information about the parameters we are interested in, deleting the entire case does not make full use of data. As a comparison, by using a likelihood approach, LME is still able to capture information provided by incomplete cases.

As generalizations of linear models, ME models (LME and GLMM) also share many of the same challenges: model selection and diagnostics, heterogeneous variances, and adjustments for multiple comparisons. What if the outcome data are severely skewed? How will one jointly analyze multiple features? Statisticians have developed methods to address these challenges. For example, resampling methods have been proposed as robust alternatives to LME (Halekoh and Højsgaard, 2014; Zeger et al., 1988). To relax the Gaussian assumption of random errors, statisticians have proposed semiparametric methods in which treatment effects remain parametric and the distributions of random effects are estimated using nonparametric methods (Datta and Satten, 2005; Dutta and Datta, 2016; Rosner et al., 2006; Rosner and Grove, 1999). In addition, it is important to conduct model diagnostics on the random effects when conducting LME. Due to the limited space, it is overambitious to cover all of the practical issues one may encounter in handling dependent data, including the issue of multiple testing and the misuse and misinterpretation of p values. We refer the interested reader to specialized research articles (Aickin and Gensler, 1996; Altman and Bland, 1995; Benjamin and Berger, 2019; Benjamini and Hochberg, 1995; Gelman and Stern, 2006; Goodman, 2008; Holm, 1979; McHugh, 2011; Storey, 2002; Wasserstein and Lazar, 2016) or to consult with experienced statisticians.

We believe that the proper use of LME and GLMM will help neuroscience researchers to improve their experimental design and leverage the advantages of more recently developed statistical methodologies. The recommended statistical approach introduced in this article will lead to data analyses with greater validity and will enable accurate and informative interpretation of results toward higher reproducibility of experimental findings in the neurosciences.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2021.10.030.

ACKNOWLEDGMENTS

This work was supported by US National Institutes of Health (NIH) grants (R01EY028212, R01MH105427, and R01NS104897). T.C.H. was supported by NIH grant R35GM127102. M.G. was supported by NSF grant SES 1659921.

AUTHOR CONTRIBUTIONS

Z.Y., S.F.G., M.G., L.C., T.C.H., and X.X. prepared the figures and wrote the manuscript. X.X. conceived and oversaw the work.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Published: November 15, 2021

REFERENCES

Aarts, E., Verhage, M., Veenvliet, J.V., Dolan, C.V., and van der Sluis, S. (2014). A solution to dependency: using multilevel analysis to accommodate nested data. Nat. Neurosci. *17*, 491–496.

Aickin, M., and Gensler, H. (1996). Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. American Journal of Public Health *86*, 726–728, PMID: 8629727.

Altman, D.G., and Bland, J.M. (1995). Statistics notes: Absence of evidence is not evidence of absence. BMJ *311*, 485, PMID: 7647644.

Alberts, B., Kirschner, M.W., Tilghman, S., and Varmus, H. (2014). Rescuing US biomedical research from its systemic flaws. Proc. Natl. Acad. Sci. USA *111*, 5773–5777.

Baayen, R.H., Davidson, D.J., and Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. *59*, 390–412.

Barr, D.J., Levy, R., Scheepers, C., and Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. J. Mem. Lang. *68*, 255–278.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. J. Stat. Softw. *67*, 1–48.

Benjamin, D.J., and Berger, J.O. (2019). Three recommendations for improving the use of p-values. Am. Stat. *73*, 186–191.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. arXiv 1701.02434v2, http://arxiv.org/abs/1701.02434v2.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.

Boisgontier, M.P., and Cheval, B. (2016). The anova to mixed model transition. Neurosci. Biobehav. Rev. *68*, 1004–1005.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., and White, J.-S.S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. *24*, 127–135.

Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. *88*, 9–25.

Bürkner, P.-C. (2017). brms: an R package for Bayesian multilevel models using Stan. J. Stat. Softw. *80*, 1–28.

Bürkner, P. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. R J. *10*, 395.

Casella, G., and George, E.I. (1992). Explaining the Gibbs sampler. Am. Stat. *46*, 167–174.

Datta, S., and Satten, G.A. (2005). Rank-sum tests for clustered data. J. Am. Stat. Assoc. *100*, 908–915.

Duane, S., Kennedy, A.D., Pendleton, B.J., and Roweth, D. (1987). Hybrid monte carlo. Phys. Lett. B *195*, 216–222.

Dutta, S., and Datta, S. (2016). A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative. Biometrics *72*, 432–440.

Ferron, J., Dailey, R., and Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. Multivariate Behav. Res. *37*, 379–403.

Fiedler, K. (2011). Voodoo Correlations Are Everywhere-Not Only in Neuroscience. Perspect. Psychol. Sci. *6*, 163–171.

Fischer, R. (1944). Statistical Methods for Research Workers, 1925 (Oliver Boyd), p. 518.

Fisher, R.A. (1919). XV.—The correlation between relatives on the supposition of Mendelian inheritance. http://l.academicdirect.org/Horticulture/GAs/Refs/Fisher_1918_Correlation.pdf.

Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2012). Applied Longitudinal Analysis*Volume 998* (John Wiley & Sons).

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. Biostatistics *11*, 397–412.

Fox, J., and Weisberg, S. (2018). An R Companion to Applied Regression (Sage Publications).

Freedman, L.P., Cockburn, I.M., and Simcoe, T.S. (2015). The Economics of Reproducibility in Preclinical Research. PLoS Biol. *13*, e1002165.

Galbraith, S., Daniel, J.A., and Vissel, B. (2010). A study of clustered data and approaches to its analysis. J. Neurosci. *30*, 10601–10608.

Gelfand, A.E., and Smith, A.F. (1990). Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. *85*, 398–409.

Gelman, A. (2005). Analysis of variance—why it is more important than ever. Ann. Stat. *33*, 1–53.

Gelman, A., and Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models (Cambridge University Press).

Gelman, A., and Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. The American Statistician *60*, 328–331.

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. In Seminars in Hematology 45 (WB Saunders), pp. 135–140.

Green, P., and MacLeod, C.J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. Methods Ecol. Evol. *7*, 493–498.

Grieco, S.F., Qiao, X., Zheng, X., Liu, Y., Chen, L., Zhang, H., Yu, Z., Gavornik, J.P., Lai, C., Gandhi, S.P., et al. (2020). Subanesthetic Ketamine Reactivates Adult Cortical Plasticity to Restore Vision from Amblyopia. Curr. Biol. *30*, 3591–3603.e8.

Hadfield, J.D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. J. Stat. Softw. *33*, 1–22.

Halekoh, U., and Højsgaard, S. (2014). A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest. J. Stat. Softw. *59*, 1–30.

Hastings, W.K. (1970). Monte-Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika *57*, 97–109.

Heisig, J.P., and Schaeffer, M. (2019). Why you should always include a random slope for the lower-level variable involved in a cross-level interaction. Eur. Sociol. Rev. *35*, 258–279.

Held, L., and Ott, M. (2018). On p-Values and Bayes Factors. Annu. Rev. Stat. Appl. *5*, 393–419.

Henderson, C.R. (1949). Estimation of changes in herd environment. J. Dairy Sci. *32*, 706.

Henderson, C.R., Kempthorne, O., Searle, S.R., and Von Krosigk, C. (1959). The estimation of environmental and genetic trends from records subject to culling. Biometrics *15*, 192–218.

Hoffman, M.D., and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res. *15*, 1593–1623.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 65–70.

Jiang, J., and Nguyen, T. (2021). Linear and Generalized Linear Mixed Models and Their Applications, Second Edition (Springer).

Kass, R.E., and Raftery, A.E. (1995). Bayes factors. J. Am. Stat. Assoc. *90*, 773–795.

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M.F., Cuthill, I.C., Fry, D., Hutton, J., and Altman, D.G. (2009). Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoS ONE *4*, e7824.

Kish, L. (1965). Survey Sampling (Wiley).

Kuznetsova, A., Brockhoff, P.B., and Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. J. Stat. Softw. *82*, 1–26.

Kwok, O.-m., West, S.G., and Green, S.B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: a Monte Carlo study. Multivariate Behav. Res. *42*, 557–592.

Laird, N.M., and Ware, J.H. (1982). Random-effects models for longitudinal data. Biometrics *38*, 963–974.

Landis, S.C., Amara, S.G., Asadullah, K., Austin, C.P., Blumenstein, R., Bradley, E.W., Crystal, R.G., Darnell, R.B., Ferrante, R.J., Fillit, H., et al. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. Nature *490*, 187–191.

Lazic, S.E., Clarke-Williams, C.J., and Munafò, M.R. (2018). What exactly is ‘N’ in cell culture and animal experiments? PLoS Biol. *16*, e2005282.

Lenth, R., Singmann, H., Love, J., Buerkner, P., and Herve, M. (2019). Estimated marginal means, aka least-squares means. R package version 1.3.2. https://rdrr.io/cran/emmeans/.

Liang, K.-Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. Biometrika *73*, 13–22.

Lüdecke, D. (2018). sjPlot: data visualization for statistics in social science. R package version 2. https://zenodo.org/record/2400856#.YXF4vhrMKM8.

Macleod, M.R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J.P.A., Al-Shahi Salman, R., Chan, A.W., and Glasziou, P. (2014). Biomedical research: increasing value, reducing waste. Lancet *383*, 101–104.

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., Guyer, M., and Green, E.D. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J. Am. Med. Inform. Assoc. *21*, 957–958.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing type I error and power in linear mixed models. J. Mem. Lang. *94*, 305–315.

McCullagh, P., and Nelder, J.A. (2019). Generalized Linear Models (Routledge).

McHugh, M.L. (2011). Multiple comparison analysis testing in ANOVA. Biochemia Medica *21*, 203–209, PMID: 22420233.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equation of state calculations by fast computing machines. J. Chem. Phys. *21*, 1087–1092.

Neal, R.M. (2011). MCMC using Hamiltonian dynamics. In Handbook of Markov Chain Monte Carlo, S. Brooks, A. Gelman, G.L. Jones, and X.-L. Meng, eds. (Chapman & Hall/CRC), pp. 113–162.

Nelder, J.A., and Wedderburn, R.W. (1972). Generalized linear models. J. R. Stat. Soc. [Ser A] *135*, 370–384.

Oberpriller, J., de Souza Leite, M., and Pichler, M. (2021). Fixed or random? On the reliability of mixed-effect models for a small number of levels in grouping variables. bioRxiv. https://doi.org/10.1101/2021.05.03.442487.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., EISPACK Authors, Heisterkamp, S., Van Willigen, B., and Ranke, J.; R Core Development Team (2007). nlme: linear and nonlinear mixed effects models. R package version 3. https://rdrr.io/cran/nlme/.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? Nat. Rev. Drug Discov. *10*, 712.

R Development Core Team (2020). R: A language and environment for statistical computing (R Foundation for Statistical Computing).

Rosner, B., and Grove, D. (1999). Use of the Mann-Whitney U-test for clustered data. Stat. Med. *18*, 1387–1400.

Rosner, B., Glynn, R.J., and Lee, M.L.T. (2006). Extension of the rank sum test for clustered data: two-group comparisons with group membership defined at the subunit level. Biometrics *62*, 1251–1259.

Shahbaba, B., Lan, S., Johnson, W.O., and Neal, R.M. (2014). Split hamiltonian monte carlo. Stat. Comput. *24*, 339–349.

Steward, O., and Balice-Gordon, R. (2014). Rigor or mortis: best practices for preclinical research in neuroscience. Neuron *84*, 572–581.

Stiratelli, R., Laird, N., and Ware, J.H. (1984). Random-effects models for serial observations with binary response. Biometrics *40*, 961–971.

Storey, J.D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *64*, 479–498.

Wasserstein, R.L., & Lazar, N.A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose, The American Statistician 70, 129-133. https://doi.org/10.1080/00031305.2016.1154108.

Wedderburn, R.W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. Biometrika *61*, 439–447.

Wei, Z., Lin, B.-J., Chen, T.-W., Daie, K., Svoboda, K., and Druckmann, S. (2020). A comparison of neuronal population dynamics measured with calcium imaging and electrophysiology. PLoS Comput. Biol. *16*, e1008198.

Wilson, M.D., Sethi, S., Lein, P.J., and Keil, K.P. (2017). Valid statistical approaches for analyzing sholl data: Mixed effects versus simple linear models. J. Neurosci. Methods *279*, 33–43.

Wolak, M., and Wolak, M. (2015). R Package ''ICC.'' Facilitating estimation of the intraclass correlation coefficient (R Documentation).

Wolfinger, R., and O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. J. Stat. Comput. Simul. *48*, 233–243.

Zeger, S.L., and Karim, M.R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. J. Am. Stat. Assoc. *86*, 79–86.

Zeger, S.L., and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. Biometrics *42*, 121–130.

Zeger, S.L., Liang, K.-Y., and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. Biometrics *44*, 1049–1060.