Journal of Biomechanics xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomechanics

journal homepage: www.elsevier.com/locate/jbiomech
www.JBiomech.com

Analysis of hierarchical biomechanical data structures using mixed-effects models

Timothy F. Tirrell [a,c,e], Alfred W. Rademaker [d], Richard L. Lieber [a,b,c,e,*]

[a] Departments of Orthopaedic Surgery, University of California, San Diego, CA, United States
[b] Departments of Bioengineering, University of California, San Diego, CA, United States
[c] Departments of Biomedical Sciences Graduate Program, University of California, San Diego, CA, United States
[d] Division of Biostatistics, Department of Preventive Medicine, Northwestern University, Chicago, IL, United States
[e] Research Service, Hines VA Medical Center, Chicago, IL, United States

ARTICLE INFO

Article history:
Accepted 8 January 2018
Available online xxxx

Keywords:
Biomechanical testing
Repeated measures
Sample size
Data analysis

ABSTRACT

Rigorous statistical analysis of biomechanical data is required to understand tissue properties. In biomechanics, samples are often obtained from multiple biopsies in the same individual, multiple samples tested per biopsy, and multiple tests performed per sample. The easiest way to analyze this hierarchical design is to simply calculate the grand mean of all samples tested. However, this may lead to incorrect inferences. In this report, three different analytical approaches are described with respect to the analysis of hierarchical data obtained from muscle biopsies. Each method was used to analyze an actual experimental data set obtained from muscle biopsies of three different muscles in the human forearm. The results illustrate the conditions under which mixed-models or simple models are acceptable for analysis of these types of data.

Published by Elsevier Ltd.

1. Introduction

Understanding tissue response to altered loading is fundamental to the fields of biomechanics, tissue engineering, and orthopaedic surgery. Measurement variation in tissue properties arises from within repeated measures of the same tissue (within-subject variability), from heterogeneity among different individuals (between-subject variability) and from experimental error. From a statistical perspective, accounting for within- and between-subject variability may require large sample sizes to accurately estimate and test parameters of interest. Sample size comprises two elements—number of subjects and number of measurements per subject—and clearly defining both depends on the specific questions being addressed. Although increasing subject number partly mitigates the effects of between-subject variability, within-subject variability can only be addressed by increasing the number of specimens tested per subject or defining a small region of interest (ROI). Unfortunately, focusing on an ROI may preclude generalizing the result to the whole subject. Calculation of sample size in simple experimental designs is fairly straightforward (Sokal and Rohlf, 1981, Dixon and Massey, 1983); however for mixed models that include within- and between-subject variability, sample size calculations may be more difficult and investigators must balance statistical requirements against available time and resources.

Random effects analysis of variance (ANOVA) models have been the subject of several previous studies, including a description of the basic random effects model (Snedecor and Cochran, 1989), and the use of random effects models in the context of estimating reliability for inter-rater designs of varying complexity (Shrout and Fleiss, 1979). Jovanovic et al. (2015) present variance component estimation in multi-level hierarchical designs and include an assessment of allocation to different levels according to the varying cost of measuring different levels. Oberfeld and Franke (2013) provide a comprehensive analysis and simulation study to assess different methods for the analysis of Type I error rate in repeated measures.

To estimate the mean value and its standard error from a group of subjects with repeated observations at a given time per subject, three analytical approaches have traditionally been used (Snedecor and Cochran, 1989): (a) the mean of means, (b) the grand mean, in which data are pooled, ignoring data structure, or (c) the random effects model. While there may be exceptions based on the study design and data variability, this paper will show that the use of the grand mean or the mean of means is either totally inappropriate or less optimal compared to the use of the random effects model. One could argue that the mean of means is appropriate

* Corresponding author at: Shirley Ryan Ability Lab, 355 E. Erie St., Chicago, IL 60611, United States.
E-mail address: rlieber@sralab.org (R.L. Lieber).

https://doi.org/10.1016/j.jbiomech.2018.01.013
0021-9290/Published by Elsevier Ltd.

because sample subdivisions are not truly independent samples. However, averaging within-subject measurements reduces the number of data points; the resulting smaller sample size and decreased confidence in the estimated population mean make this a decidedly conservative approach. Averaging individual measurements across a biopsy also ignores within-subject variability and thus, if within-subject variability is large with respect to between-subject variability, this approach underreports the true data variability. Calculating a grand mean by pooling all replicates into one sample results in a larger sample size; however, this method likely underestimates mean variability, especially when there is considerable between-subject variability. The random effects model calculates explicit values for between-subject and within-subject variability in order to calculate the standard error of the grand mean. The term "random effect" refers to random subject-specific differences that arise due to normal variability within a population, and which are quantified by a between-subject variance component σ_B^2 and a within-subject variance component σ^2. At its most basic level, the random effects model is a one-way ANOVA that takes a hierarchical dataset of k subjects with n_i measurements per subject and partitions the overall variability of this data into the two variance components σ_B^2 and σ^2. The output from a one-way ANOVA includes mean squares for between-subjects and within-subjects, which are used to estimate the two variance components, calculate the F statistic, and determine significance. While it is clear that any of the three approaches are readily available to researchers with basic statistical and computational skills, the choice of method is important, as this may determine whether an experimental result is determined to be statistically significant along with the clinical and/or biological implications of such a conclusion.

In this paper, we compare the actual and expected standard errors for each of the three methods applied to our muscle biopsy dataset. We use each method to compare muscle stiffness among muscles measured on different subjects in a defined dataset, and we make recommendations on the appropriate method to use when analyzing hierarchical data. Interestingly, we find that there is a "gradient of correctness" across the three methods, and the degree of acceptability actually depends on the data themselves.

2. Materials and methods

2.1. Experimental dataset

The experimental study measured muscle stiffness in three muscles that were biopsied during surgical procedures (Fridén and Lieber 2003, Lieber et al. 2003). The goal of the study was to compare mean stiffness measures among muscles. Ethical approval for this study was provided by Institutional Review Boards at the University of California, San Diego, and the Veterans Affairs Healthcare System, San Diego. All patients (n = 24) provided informed consent for muscle biopsies, which were obtained secondary to surgical procedures. Three muscles were biopsied—the brachioradialis (BR), the flexor carpi ulnaris (FCU) and the pronator teres (PT). In total, 34 muscle biopsies were collected from the 24 study subjects; both single muscle fibers (FB) and fiber bundles (BU) were dissected from each biopsy. In most cases, three FB and three BU were tested from each biopsy. Two muscles were biopsied in 10 subjects; one muscle was biopsied in 14 subjects.

Passive properties of muscle tissue at different size scales (fiber or bundle) were measured similarly to previous experiments (Fridén and Lieber 2003, Lieber, Runesson et al., 2003, Smith, Lee et al., 2011). Muscle fibers (FB) and bundles (BU) were dissected from biopsies, secured to a force transducer and a motor arm, and transilluminated by a 5 mW diode laser. The resultant diffraction pattern was used to calculate sarcomere length. Segments were then loaded to achieve incremental strains of ~0.25 μm/sarcomere, which were held for 180 s; resultant force and sarcomere lengths were measured during each hold. Segments were loaded until failure or slippage occurred or until sarcomere length reached 4.10 μm. The stress at the end of each 180-s hold was used to fit a stress-sarcomere length relationship, which fit well to a second order polynomial (average R^2 for fibers: 0.989; for bundles: 0.984). A representative tangent stiffness value was then calculated for each test by taking the derivative of the stress-sarcomere length relationship at a sarcomere length of 3.5 μm, providing the raw data for this analysis.

2.2. Statistical methods

Biopsies were obtained from k experimental subjects, and each biopsy was subdivided into n parts that were each tested once, yielding a total data set containing $N = k * n$ data points. Such data are commonly analyzed using one of the following three methods (Table 1):

2.2.1. Method 1: Mean of means

This method estimates the population mean based on a sample size of k. For this method, tangent stiffnesses within a single biopsy are averaged to obtain a representative tangent stiffness value for that biopsy. These values are then averaged across all biopsies to obtain a representative value for each size scale (FB or BU) and for each muscle.

2.2.2. Method 2: Grand mean

This approach considers each replicate within a biopsy as an independent data point. In the example, tangent stiffness values for all FB or BU are pooled or across all biopsies so that the sample size is $k * n$, the total number of samples from all biopsies for each muscle. When each biopsy is subdivided into n parts, the grand mean equals the mean of means. However, if the data are unbalanced in that n_i (where $i = 1, \ldots, k$) measurements are taken on subject i, the grand mean is a weighted mean of means, where each of the k means is weighted by n_i.

2.2.3. Method 3: Random effects model

This model is the most conceptually accurate and describes the value of a variable *y* for a subject, where tangent stiffness values for all fibers from a single biopsy are kept distinct. The random effects model (Snedecor and Cochran, 1989) defines

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (1)$$

where y_{ij} is the j^{th} fiber measurement of stiffness on subject i, ($i = 1, \ldots, k$ subjects), $j = 1, \ldots, n_i$ measurements per subject i (n_i = number of samples for subject i) and $N = n_1 + n_2 + \ldots + n_k$. Finally, α_i refers to the subject specific effect for subject i and ε_{ij} is the j^{th} random error term for subject i.

To perform the analysis, we assume that the α_i have a normal (Gaussian) distribution centered at 0 with variance σ_B^2, abbreviated as $\alpha_i \sim G(0, \sigma_B^2)$. Similarly, we assume that the ε_{ij} have a Gaussian distribution centered at 0 with variance σ^2, abbreviated as $\varepsilon_{ij} \sim G(0, \sigma^2)$. We also assume that there is no correlation between the α_i and the ε_{ij}; i.e., that α_i and ε_{ij} are independent.

Thus, the grand mean and random effects methods estimate the overall mean using the weighted average of subject-specific means, whereas the mean of means uses the unweighted average of subject-specific means.

For each of the three methods, the standard error of the mean (SEM) is calculated in a different way, and thus each standard error estimate has a different statistical expected value. Since standard

Please cite this article in press as: Tirrell, T.F., et al. Analysis of hierarchical biomechanical data structures using mixed-effects models. J. Biomech. (2018), https://doi.org/10.1016/j.jbiomech.2018.01.013