Published in final edited form as:

J Biomech. 2018 March 01; 69: 34–39. doi:10.1016/j.jbiomech.2018.01.013.

Analysis of hierarchical biomechanical data structures using mixed-effects models

Timothy F. Tirrell, BS^{1,3,5}, Alfred W. Rademaker, PhD⁴, and Richard L. Lieber, PhD^{1,2,3,5}

¹Department of Orthopaedic Surgery, University of California, San Diego, CA

²Department of Bioengineering, University of California, San Diego, CA

³Department of Biomedical Sciences Graduate Program, University of California, San Diego, CA

⁴Division of Biostatistics, Department of Preventive Medicine, Northwestern University, Chicago, IL

⁵Research Service, Hines VA Medical Center, Chicago, IL

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.

Keywords

Biomechanical testing; repeated measures; sample size; data analysis

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

Address for correspondence: Richard L. Lieber, PhD, Shirley Ryan AbilityLab, 355 E. Erie St., Chicago, IL 60611, Tel.: +1 312 238 6260; fax: +1 312 238 1392., rlieber@srnlab.org (R.L. Lieber).

Conflict of Interest Statement

The authors declare no conflict of interest.

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 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.

Materials and Methods

Experimental Dataset

The experimental study measured muscle stiffness in three muscles that were biopsied during surgical procedures (Fridén and Lieber 2003, Lieber, Runesson 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 $\sim 0.25 \mu\text{m}/\text{sarcomere}$, which were held for 180 seconds; resultant force and sarcomere lengths were measured during each hold. Segments were loaded until failure or slippage occurred or until sarcomere length reached $4.10 \mu\text{m}$. The stress at the end of each 180-second 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 $3.5 \mu\text{m}$, providing the raw data for this analysis.

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):

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.

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 $k \cdot n$, 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 $i = 1, \dots, k$) 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 .

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 , ($i = 1, \dots, k$ subjects), $j = 1, \dots, n_i$ measurements per subject i (n_i = number of samples for subject i) and $N = n_1 + n_2 + \dots + n_k$. 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 error is used to calculate confidence intervals, variations in SEM can affect calculations of statistical significance. The SEM is calculated as follows for the three methods (Table 2):

- a. For the mean of means, the SEM is the standard deviation of the k means divided by the square root of k (Snedecor and Cochran 1989). Its expected value is the square root of $\sigma_B^2/k + \sigma^2/kn_h$, where n_h is the harmonic mean of the n_i , i.e. $n_h = k/(1/n_1 + 1/n_2 + \dots, 1/n_k)$;

- b. For the grand mean, the SEM is the standard deviation of the N measurements divided by the square root of N and has an expected value that is the square root of $(\sigma_B^2 + \sigma^2)/N = \sigma_B^2/N + \sigma^2/N$;
- c. For the random effects model, the SEM is the square root of the ANOVA mean square value between subjects divided by the square root of N with an expected value of the square root of $\sigma_B^2/k + \sigma^2/N$ (Snedecor and Cochran 1989).

To illustrate the considerations for choosing the most appropriate approach for hierarchical studies, we present expected and actual means and standard errors for each muscle and sample scale (fiber, bundle), using each method described above. Pairs of muscles are then compared using each approach. It will be seen that the accuracy of the approach actually depends on the data themselves.

Results

The raw data, comprising each replicate data point from each person, are provided in the Supplemental Table. The sample size, mean and standard error of the stiffness measures by type of measurement (fiber or bundle), muscle, and analysis method are presented in Table 1. A summary of the random effects variance components and the standard errors for the three methods are provided for fiber stiffness (Table 2) and bundle stiffness (Table 3). These tables include expected values for each standard error, and then apply the random effects estimates of the within- and between-subjects variance to these expected values. Based on these results, we make the following five observations:

1. For the random effects model, the actual value agrees closely with the target value for all cases. This is because the variance components were calculated explicitly using the random effects model.
2. The standard error for the random effects model always estimates a larger quantity ($\sigma_B^2/k + \sigma^2/N$) than the standard error of the overall grand mean ($\sigma_B^2/N + \sigma^2/N$) since k is always less than N.
3. The standard error for the random effects model always estimates a smaller quantity ($\sigma_B^2/k + \sigma^2/N$) than the standard error of the mean of means ($\sigma_B^2/k + \sigma^2/kn_h$) since N is always larger than kn_h .
4. Using a method other than the random effects model will thus either under- or over-estimate the standard error and result in inaccurate p-values. This is important because random effects models are quite rare in biomechanical studies and thus, as a community, we are at increased risk for committing type 1 and type 2 errors when using these inappropriate tests. Using a statistical test with a standard error smaller than it should be (grand mean vs. random effects) leads to an over-commitment of a Type I error (more tests are called significant when there are actually no significant group differences), an underestimation of a Type II error (fewer tests are called non-significant when there are actually significant group differences) and an overestimation of power (more tests are called significant when there are group differences). Using a statistical test with a standard error larger than it should be (mean of means vs random effects) leads

to an under-commitment of a Type I error, an overestimation of a Type II error and an underestimation of power.

5. The extent of the over- or under-estimation of standard error using the mean of means or the grand mean depends on the ratio σ_B^2 / σ^2 as well as the variability in sample size across subjects. The larger the σ_B^2 / σ^2 ratio, *the greater the under estimation using the grand mean and the more compelling the reason to use the random effects model*. This is the key point that allows the investigator to determine whether the mixed model approach is required. This is clearly seen in Tables 2 and 3, where the ratio is relatively small for FCU fiber, PT fiber, and BR bundle (<0.10) and in these cases, underestimation of the grand mean is modest. For FCU bundle and PT bundle, the ratio is larger (>0.60) resulting in a greater underestimation of the grand mean SEM.
6. The mean of means method is preferable to grand mean if it is necessary to use a method other than the random effects method if, for example, a random effects statistical package is not available. The mean of means standard error is similar in expectation to the random effects mean, while the grand mean standard error dampens the effect of the between-subject variability by inflating the effect of within-subject variability, especially if the number of replicates within subjects is large. Moreover, the grand mean inappropriately uses residual degrees of freedom for the statistical test which exceed the between subject degrees of freedom used by the random effects and mean of means model.

The over-estimation of the SEM using the mean of means also depends on the extent to which the sample sizes across subjects differ. The more different they are, the more the harmonic mean differs from the arithmetic mean of the subject-specific sample sizes. These effects, together with the variation in the ratio of between- to within-subject variance, are illustrated in Table 4. This table assumes 5 subjects, a target of 3 observations per subject and a within-subject variance of 10. Under reasonable variation in sample size per subject, the harmonic mean does not differ appreciably from the arithmetic mean so the two standard errors will be similar. Underestimation of SEM using the grand mean depends only on the ratio of the between- to within-subject variance, and the underestimation increases (numbers in grand mean row get smaller) as this ratio increases. Overestimation of SEM using the mean of means also depends on the harmonic mean. There is no overestimation when the subject-specific sample sizes are equal (harmonic mean=3), but overestimation increases (standard error of the mean of means gets larger) as the inequality increases and the harmonic mean gets smaller.

All three methods were used to compare stiffness among biopsied muscles where different subjects were used for the different muscles. The comparisons between BR and FCU, or between BR and PT involved different subjects leading to unmatched independent comparisons. Pairwise comparisons are given for fiber stiffness (Table 5) and for bundle stiffness (Table 6). We report the mean difference, the SEM, and the p-value as determined from a mixed effects model with a random effects model within each muscle as described above and a fixed effect for muscle. In addition, the expected value of the standard error is given, together with an estimate of this expected value based on the data. The standard error

for the mean of means overestimates the expected value because it does not take the dampening effect of replicates into account. The grand mean estimates what it is supposed to estimate, but overcompensates for replicates, thus the standard error is smaller than that of either the mean of means or the random effects methods. While the random effects method is the method of choice for these analyses, the mean of means, not the grand mean, will give results that are more comparable to the random effects model.

The comparisons of fiber stiffness between muscles indicated that the BR muscle was similar to the FCU and PT muscles using any of the three methods. The comparisons of bundle stiffness indicated no difference between the muscles for most of the comparisons done. However, the bundle stiffness analysis draws attention to the inappropriate use of the grand mean analysis, which is significant at $p < 0.05$ for the BR vs FCU comparison. The other methods have non-significant results at $p > 0.05$. The between-to-within variance ratio is 1.10 (Table 3) indicating that the grand mean severely underestimates the standard error for the FCU muscle, and consequently for the BR vs FCU comparison. In this case, it leads to a significant result when the other methods do not. This example illustrates Type I error inflation. If twice the standard error of the mean difference is used as a critical value for significance testing, then that critical value is ~ 29 (2×14.65) for the random effects model, while it is only ~ 20 (2×10.36) for the grand mean model. Using a critical value of 20 for the random effects model increases the Type I error from 5% to 16%. Overall, in this experiment, there were no significant differences between the muscles for either fiber or bundle stiffness.

Discussion

The purpose of this report was to compare analytical methods for calculation of error for experimental designs in which determining the within-subject and between-subject variability are relevant. These types of experiment are extremely common in biomechanical studies. The primary advantage of using averaging methods such as the mean of means or the grand mean is simplicity and universality. These methods do not require sophisticated software, interpreting the results is a straightforward process, and the results are most likely to be understood by a broad audience. Because readers naturally gravitate towards methods with which they are familiar, using commonly understood analysis procedures may be advantageous for broad dissemination of findings. However, the primary limitation of these methods depends on the experimental design. If the data are hierarchical in nature, simple averaging does not adequately represent the different sources of variability that exist in the data and does not allow certain statistical comparisons to be made. Because p-values and confidence intervals depend on variability, misrepresenting variability affects significance calculations for differences between groups. If the goal of an experiment is to determine whether a null hypothesis should be rejected, accurate analysis of differences between groups is of the utmost importance. Mixed-effects models most accurately represent the variability in a hierarchical dataset and can subset this variability into its appropriate levels.

Statistical software currently makes techniques such as mixed-effects modeling broadly accessible with a minimal investment of time and money. However, easy access to such statistical tools does not imply that they are always superior to simpler methods. Although

implementation of advanced statistical methods may be technically simple, doing so correctly and interpreting output quality are nontrivial. Implementation of these methods is most accessible to the general population through an open source software environment, which allows users great latitude. However, this can be a positive or negative feature: while flexibility allows complete customization of input parameters and generates computationally correct outputs, appropriate use in a given situation depends on the user's understanding and programming acumen. User interfaces in open source packages tend to be less user-friendly than commercially available software.

In the data presented, the residual (within-subject) variance is larger than the variance associated with subject-specific random effects (between-subject variability) at both the fiber and bundle size scales. While between-subject variability may be much smaller than within-subject variability, it is the ratio of these two parameters that determines whether a random effects model is necessary. Our data show that use of the random effects model is warranted when the between- to within-subject variance ratio is greater than 0.10. In these cases, using simple averaging methods to analyze these data masks the variability within a biopsy. Using the mixed model allows accurate partitioning of sources of variability and most accurately quantifies the significance of the main effects in the data. In addition, as within-subject sample sizes increase in variability, the mean of means becomes overly conservative due to the increased standard error.

We attempted to generalize these findings through a sensitivity analysis (Table 4), by comparing the ratio of the SEM of each of the grand mean and mean of mean methods to the SEM for the random effects method. The grand mean SEM does not depend on the difference in sample size across subjects, and it becomes smaller, thereby less conservative, as the ratio of the between- to within-subject variance increases. The first row of Table 4 indicates that, as the between-subject variance changes from 10% to 90% of the within-subject variance, the reduction in the SEM ranges between 8% and 28%. The mean of means becomes more conservative as variability in subject sample size increases. In the sensitivity analysis with 5 measures per subject, if all 5 subjects had 3 measures, then the mean of means is as good as the random effects model. If 1 subject had 5 measures, 1 subject had 4 measures and 3 subjects had 2 measures (so that the harmonic mean is 2.5 with a mean of 3), then the SEM is overestimated by between 2.7% and 7.4%, depending on the between- to within-subject variance ratio. If 2 subjects had 5 measures, 2 subjects had 2 measures and 1 subject had 1 measure (so that harmonic mean is 2.1 while the mean is 3), then the SEM is overestimated by between 5.6% and 15.3%.

Based on this study, our recommendations are as follows:

1. Use a one-way analysis of variance within each group to estimate the between- and within-subject variance. Determine the ratio of these variances and calculate the harmonic mean of the subject sample sizes.
2. The grand mean, pooling all the data across subjects and ignoring the data structure, is to be avoided since it routinely underestimates the SEM for a group, especially as the between to within variance ratio increases, leading to unwarranted significant conclusions (type 1 error).

3. The mean of means is the preferred method if a random effects model is not used. However, if sample sizes vary considerably across subjects so that the harmonic mean of the sample sizes is less than the arithmetic mean, then this method is overly conservative and leads to false negative results (type 2 error).
4. The random effects model is the preferred method both to estimate the mean and standard error for each group being analyzed, but also to perform comparisons between groups.

Appropriate statistical approaches to the analysis of biomechanical experiments will increase the fidelity and utility of experiments in this field.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Department of Veterans Affairs and the National Institutes of Health (R24 HD050837 and R01 AR057393).

References

- Dixon, WJ., Massey, FJ. Introduction to Statistical Analysis. New York: McGraw Hill; 1983.
- Fridén J, Lieber RL. Spastic muscle cells are shorter and stiffer than normal cells. *Muscle Nerve*. 2003; 27(2):157–164. [PubMed: 12548522]
- Jovanovic BD, Subramanian H, Helenowski IB, Roy HK, Backman V. Clinical trial laboratory data nested with in subject: components of variance, sample size and cost. *Biomedical and Biostatistical International Journal*. 2015; 2:1–7.
- Lieber RL, Runesson E, Einarsson F, Fridén J. Inferior mechanical properties of spastic muscle bundles due to hypertrophic but compromised extracellular matrix material. *Muscle & Nerve*. 2003; 28:464–471. [PubMed: 14506719]
- Oberfeld D, Franke T. Evaluating the robustness of repeated measures analysis: The case of small sample sizes and nonnormal data. *Behav Res*. 2013; 45:792–812.
- Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*. 1979; 86:420–428. [PubMed: 18839484]
- Smith LR, Lee KS, Ward SR, Chambers HG, Lieber RL. Hamstring contractures in children with spastic cerebral palsy result from a stiffer extracellular matrix and increased in vivo sarcomere length. *J Physiol*. 2011; 589(Pt 10):2625–2639. [PubMed: 21486759]
- Snedecor, GW., Cochran, WG. Statistical methods. Ames, Iowa: Iowa State Univ Press; 1989.
- Sokal, RR., Rohlf, FJ. Biometry. San Francisco: W.H. Freeman and Company; 1981.

Sample size, mean and standard error of the stiffness measures by type of measurement (fiber or bundle), muscle and analysis method

Table 1

Muscle identity	Analysis type	Fiber stiffness (kPa/ μ m)		Bundle stiffness (kPa/ μ m)	
		n	mean $\pm$ sem	n	mean $\pm$ sem
BR	Mean of means	10	15.78 $\pm$ 1.74	10	33.56 $\pm$ 3.96
	Grand mean	29	15.91 $\pm$ 1.55	28	32.18 $\pm$ 3.42
	Random effects	29	15.88 $\pm$ 1.77	28	32.30 $\pm$ 3.69
FCU	Mean of means	14	15.91 $\pm$ 2.39	14	55.63 $\pm$ 14.19
	Grand mean	40	15.58 $\pm$ 2.26	42	55.63 $\pm$ 9.78
	Random effects	40	15.63 $\pm$ 2.44	42	55.62 $\pm$ 14.19
PT	Mean of means	13	15.32 $\pm$ 1.54	13	25.44 $\pm$ 4.05
	Grand mean	36	14.92 $\pm$ 1.45	38	25.31 $\pm$ 3.07
	Random effects	36	14.96 $\pm$ 1.55	38	25.39 $\pm$ 4.09

Table 2

Fiber stiffness variance components and standard errors for the three methods described in the text

Muscle	Analysis type	Fiber stiffness (kPa/ μm)		
BR	Between subject	σ_B^2	10.3	$\sigma_B^2 / \sigma^2 = 0.17$
	Within subject	σ^2	60.3	
		k, N	10, 29	
		n_h	2.857	
		Expected sem		Actual sem
	Mean of means	$\sigma_B^2/k + \sigma^2/kn_h$	1.772	1.744
	Grand mean	$\sigma_B^2/N + \sigma^2/N$	1.560	1.550
	Random effects	$\sigma_B^2/k + \sigma^2/N$	1.763	1.770
FCU	Between subject	σ_B^2	17.1	$\sigma_B^2 / \sigma^2 = 0.09$
	Within subject	σ^2	187.9	
		k, N	14, 40	
		n_h	2.800	
		Expected sem		Actual sem
	Mean of means	$\sigma_B^2/k + \sigma^2/kn_h$	2.453	2.388
	Grand mean	$\sigma_B^2/N + \sigma^2/N$	2.264	2.300
	Random effects	$\sigma_B^2/k + \sigma^2/N$	2.433	2.440
PT	Between subject	σ_B^2	5.7	$\sigma_B^2 / \sigma^2 = 0.08$
	Within subject	σ^2	70.5	
		k, N	13, 36	
		n_h	2.516	
		Expected sem		Actual sem
	Mean of means	$\sigma_B^2/k + \sigma^2/kn_h$	1.611	1.543
	Grand mean	$\sigma_B^2/N + \sigma^2/N$	1.455	1.500
	Random effects	$\sigma_B^2/k + \sigma^2/N$	1.548	1.550

Table 3

Bundle stiffness variance components and standard errors for the three methods described in the text

Muscle	Analysis type	Bundle stiffness (kPa/ μm)		
BR	Between subject	σ_B^2	27.1	$\sigma_B^2 / \sigma^2 = 0.09$
	Within subject	σ^2	301.9	
		k, N	10	
		n_h	2.500	
		Expected sem		Actual sem
	Mean of means	$\sigma_B^2/k + \sigma^2/kn_h$	3.845	3.965
	Grand mean	$\sigma_B^2/N + \sigma^2/N$	3.428	3.420
	Random effects	$\sigma_B^2/k + \sigma^2/N$	3.673	3.690
FCU	Between subject	σ_B^2	2164.4	$\sigma_B^2 / \sigma^2 = 1.10$
	Within subject	σ^2	1959.3	
		k, N	14	
		n_h	3.000	
		Expected sem		Actual sem
	Mean of means	$\sigma_B^2/k + \sigma^2/kn_h$	14.186	14.186
	Grand mean	$\sigma_B^2/N + \sigma^2/N$	9.909	9.800
	Random effects	$\sigma_B^2/k + \sigma^2/N$	14.186	14.190
PT	Between subject	σ_B^2	140.3	$\sigma_B^2 / \sigma^2 = 0.63$
	Within subject	σ^2	223.8	
		k, N	13	
		n_h	2.889	
		Expected sem		Actual sem
	Mean of means	$\sigma_B^2/k + \sigma^2/kn_h$	4.093	4.047
	Grand mean	$\sigma_B^2/N + \sigma^2/N$	3.095	3.100
	Random effects	$\sigma_B^2/k + \sigma^2/N$	4.084	4.090

Table 4

Effects on within group standard errors of the mean (SEM) by using the grand mean or the mean of means in place of the random effects mean.

		Ratio of between to within subject variance (within subject variance = 10)							
		1	3	5	7	9			
Under-estimation of grand mean SEM (ratio of SEM compared to random effects model)		0.920	0.827	0.775	0.741	0.717			
Over-estimation of mean of means SEM (ratio of SEM compared to random effects model)	Harmonic mean								
	3.0	1.000	1.000	1.000	1.000	1.000			1.000
	2.9	1.013	1.009	1.007	1.006	1.005			1.005
	2.7	1.042	1.029	1.022	1.018	1.015			1.015
	2.5	1.074	1.051	1.039	1.032	1.027			1.027
	2.3	1.111	1.077	1.059	1.048	1.040			1.040
	2.1	1.153	1.107	1.082	1.067	1.056			1.056

Table 5

Comparison of mean fiber stiffness between pairs of muscles using the three methods described in the text.

Analysis type	BR vs FCU		p-value	Expected value of standard error	Calculated expected value of standard error
	Mean difference	Standard error of mean difference			
Mean of means	0.13	2.96	0.97	$2(\sigma_B^2/k + \sigma^2/kn_b)$	3.01
Grand mean	0.33	2.74	0.90	$2(\sigma_B^2/N + \sigma^2/N)$	2.75
Random effects	0.25	3.01	0.94	$2(\sigma_B^2/k + \sigma^2/N)$	3.00
Analysis type	BR vs PT				
	Mean difference	Standard error of mean difference	p-value	Expected value of standard error	Calculated expected value of standard error
Mean of means	0.46	2.33	0.85	$2(\sigma_B^2/k + \sigma^2/kn_b)$	2.35
Grand mean	1.00	2.13	0.64	$2(\sigma_B^2/N + \sigma^2/N)$	2.13
Random effects	0.92	2.35	0.70	$2(\sigma_B^2/k + \sigma^2/N)$	2.35

Table 6

Comparison of mean bundle stiffness between pairs of muscles using the three methods.

Analysis type	BR vs FCU		p-value	Expected value of standard error	Calculated expected value of standard error
	Mean difference	Standard error of mean difference			
Mean of means	22.07	14.73	0.15	$2(\sigma_B^2/k + \sigma^2/kn_b)$	14.65
Grand mean	23.44	10.36	0.027	$2\sigma_B^2/N + \sigma^2/N$	10.48
Random effects	23.32	14.65	0.13	$2\sigma_B^2/k + \sigma^2/N$	14.65
Analysis type	BR vs PT				
	Mean difference	Standard error of mean difference	p-value	Expected value of standard error	Calculated expected value of standard error
Mean of means	8.11	5.67	0.17	$2(\sigma_B^2/k + \sigma^2/kn_b)$	5.49
Grand mean	6.87	4.60	0.14	$2\sigma_B^2/N + \sigma^2/N$	4.62
Random effects	6.92	5.51	0.22	$2\sigma_B^2/k + \sigma^2/N$	5.49