Biometrics 68, 486–493
June 2012

DOI: 10.1111/j.1541-0420.2011.01675.x

Permutation Tests for Random Effects in Linear Mixed Models

Oliver E. Lee[*] **and Thomas M. Braun**[**]

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029, U.S.A.
[*]*email:* oel@umich.edu
[**]*email:* tombraun@umich.edu

Summary. Inference regarding the inclusion or exclusion of random effects in linear mixed models is challenging because the variance components are located on the boundary of their parameter space under the usual null hypothesis. As a result, the asymptotic null distribution of the Wald, score, and likelihood ratio tests will not have the typical χ^2 distribution. Although it has been proved that the correct asymptotic distribution is a mixture of χ^2 distributions, the appropriate mixture distribution is rather cumbersome and nonintuitive when the null and alternative hypotheses differ by more than one random effect. As alternatives, we present two permutation tests, one that is based on the best linear unbiased predictors and one that is based on the restricted likelihood ratio test statistic. Both methods involve weighted residuals, with the weights determined by the among- and within-subject variance components. The null permutation distributions of our statistics are computed by permuting the residuals both within and among subjects and are valid both asymptotically and in small samples. We examine the size and power of our tests via simulation under a variety of settings and apply our test to a published data set of chronic myelogenous leukemia patients.

Key words: Hypothesis testing; Longitudinal data; Variance components.

1. Introduction

Linear mixed models (LMMs) are a rich class of models containing both fixed and random effects. LMMs are often used to fit longitudinal or repeated measures data (Laird and Ware, 1982), where outcomes for a limited number of subjects are collected repeatedly over time, or with multilevel or clustered data, where random effects are used to account for the within-level or within-cluster correlations. Often, inference focuses upon the need for the inclusion of random effects. For example, subjects in a clinical trial may be recruited from a set of hospitals that are participating in the study. Homogeneity among patients from the same hospital is likely and can be accounted for through a random hospital effect in the model. However, if there is no correlation among patients from the same hospital then there would be a loss of power by estimating an unnecessary random effect variance.

The difficulty in testing for random effects lies in the fact that the variance component of the random effect is equal to 0 under the null hypothesis, a value that is on the boundary of the parameter space. As a result, the usual χ^2 asymptotic distributions of the Wald, score, and likelihood ratio test statistics do not hold. Instead, the correct null distribution for the likelihood ratio statistic has been shown to be a mixture of χ^2 distributions (Self and Liang, 1987; Stram and Lee, 1994). For example, when testing for one random effect, the null distribution becomes a 50:50 mixture of χ^2_q and χ^2_{q-1} distributions, where q is the total number of random effects in the alternative model. The score (Silvapulle and Silvapulle, 1995; Verbeke and Molenberghs, 2003) and Wald (Silvapulle, 1992) tests for variance components have been proven to have equivalent mixture χ^2 distributions. These modified tests also rely on asymptotic approximations and are not guaranteed to have nominal size with small sample sizes.

Other methods for variance component inference have been published. Öfversten (1993) developed an exact test for uncorrelated random effects in unbalanced LMMs through orthogonal transformations of the model matrix. Crainiceanu and Ruppert (2004) derived the finite sample null distribution for the likelihood ratio and restricted likelihood ratio test statistics when testing for a single variance component with no other nuisance variance components. They derived the spectral decomposition of each test statistic, and they also developed a simulation algorithm that generates the approximate finite sample null distribution via the spectral decomposition. Greven et al. (2008) extended the methods of Crainiceanu and Ruppert to test for a single variance component in the presence of multiple independent nuisance random effects and also developed an approximation to the parametric bootstrap. Kinney and Dunson (2008) used a Bayesian stochastic search variable selection method to identify nonzero random effect variances in LMMs using a modified Cholesky decomposition of the random effect covariance matrix. By reparameterizing the LMM, the stochastic search variable selection method can perform variable selection with the random effects. An alternative Bayesian method was developed by Saville and Herring (2009) in which null and alternative models are compared via Bayes factors.

Permutation tests are a viable alternative to the above methods, as permutation tests are known to have nominal size in finite samples while requiring only a few weak assumptions. Nonetheless, the only existing permutation approach for testing for random effects was presented by Fitzmaurice and Ibrahim (2007). The test was specifically designed for multilevel studies where inclusion of a single random effect to quantify the heterogeneity among the different levels may be required. They compared the likelihood ratio test statistic to an empirical null distribution generated by randomly permuting the observed level assignments among the subjects. However, their test is limited to the setting at hand and cannot be generalized to longitudinal studies and other correlated data sources if there are multiple random effects or a single continuous random effect, such as time.

Our work is a generalization to the approach of Fitzmaurice and Ibrahim and leads to a pair of permutation tests that allow for inference with any number and type of random effects in an LMM. Both test statistics are a sum of weighted squared residuals with the weights determined by the among- and within-subject variance components, and the empirical null distributions generated via permutations of the residuals. The first test statistic is based on the best linear unbiased predictions (BLUPs) (Robinson, 1991) and the second statistic is the restricted likelihood ratio test statistic assuming normality of the data. We will show that our tests have valid size and their powers are comparable to existing methods. We will also demonstrate that our likelihood ratio based permutation test can address simultaneous inference on multiple random effects. We begin with LMM notation and some background on permutation methods in Section 2. Section 3 follows with a presentation of our proposed methods. We present the results of simulations in Section 4 that demonstrate the validity and power of our methods as we vary both the numbers of subjects and the numbers of observations per subject. In Section 5, we apply our methods to data from a longitudinal study investigating the levels of adenosine deaminase (ADA) in chronic myelogenous leukemia patients. We close with a discussion of our work in Section 6.

2. Methods

2.1 *Linear Mixed Models*

Let Y_{ij} be observation j of subject or cluster i for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, n_i$. Following the Laird and Ware (1982) formulation of the LMM, we have

$$Y_{ij} = \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_{i1} z_{1ij} + \cdots + b_{iq} z_{qij} + \epsilon_{ij},$$

where $\beta_1, \ldots, \beta_p$ are the population level fixed-effect coefficients, and $b_{i1}, \ldots, b_{iq}$ are the random effects for the ith subject or cluster. The $x_{1ij}, \ldots, x_{pij}$ and $z_{1ij}, \ldots, z_{qij}$ are the observed fixed-effect covariates and random effect covariates, respectively, for observation j of subject i. Generally, x_{1ij} and z_{1ij}, are constant and equal to 1 to represent the fixed and random intercepts, respectively. The random effects, $\boldsymbol{b}_i = \{b_{i1}, b_{i2}, \ldots, b_{iq}\}$ are assumed to have a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, in which the respective variances for $b_{i1}, b_{i2}, \ldots, b_{iq}$ are denoted as $\sigma^2_{b_1}, \sigma^2_{b_2}, \ldots, \sigma^2_{b_q}$. The random errors, ϵ_{ij}, are independent, identically distributed normal random variables with mean 0 and variance σ^2_ϵ. For each j, $\boldsymbol{b}_i$ and ϵ_{ij} are assumed to be independent, although the elements of $\boldsymbol{b}_i$ are not necessarily independent of each other.

Equivalently, we can write the LMM for subject i using matrix notation, $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_p\}$, $\boldsymbol{\epsilon}_i = \{\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{in_i}\}$, and $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are subject-specific design matrices for the p fixed-effect covariates and q random effect covariates, respectively. We then combine data from all subjects so that $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N\}$ is the $\sum_i n_i$ vector of outcomes, $\boldsymbol{\epsilon} = \{\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_N\}$ is the $\sum_i n_i$ vector of errors, and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the respective design matrices for the p fixed-effect covariates and q random effect covariates formed by successively placing each subject's design matrices under each other. Furthermore, if we denote $\boldsymbol{b} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_N\}$, we have

$$Var \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \boldsymbol{G} & 0 \\ 0 & \boldsymbol{R} \end{bmatrix}$$

where $\boldsymbol{G} = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_G$ and $\boldsymbol{R} = \sigma^2_\epsilon \boldsymbol{I}_R$, in which $\otimes$ denotes the Kroenecker product, and $\boldsymbol{I}_G$ and $\boldsymbol{I}_R$ are $N \times N$ and $\sum_i n_i \times \sum_i n_i$ identity matrices, respectively.

Estimation of the elements of $\boldsymbol{\beta}$, $\boldsymbol{G}$, and $\boldsymbol{R}$ is typically done through maximum likelihood or restricted maximum likelihood (REML). Asymptotically, the maximum likelihood and REML estimators are equivalent, but for small sample sizes, the REML estimator is expected to be less biased than the maximum likelihood estimator (Ruppert, Wand, and Carroll, 2003). In addition, a comprehensive simulation study performed by Morrell (1998) found that the asymptotic likelihood ratio test based on the REML estimates are closer to nominal than test statistics utilizing the maximum likelihood estimates. Therefore, in our proposed methods we used the REML estimators. Subject-specific random effects, $b_{i1}, \ldots, b_{iq}$, can be predicted using BLUP, the results from which we denote $\tilde{\boldsymbol{b}} = \{\tilde{\boldsymbol{b}}_1, \tilde{\boldsymbol{b}}_2, \ldots, \tilde{\boldsymbol{b}}_N\}$, where $\tilde{\boldsymbol{b}}_i = \{\tilde{b}_{i1}, \tilde{b}_{i2}, \ldots, \tilde{b}_{iq}\}$. The estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}} = \{\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\}$, and $\tilde{\boldsymbol{b}}$ are solutions to the following mixed model equations given by Henderson (1950)

$$\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\tilde{\boldsymbol{b}} = \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Y},$$
$$\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} + (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\tilde{\boldsymbol{b}} = \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Y},$$

and lead to the solutions

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{Y},$$
$$\tilde{\boldsymbol{b}} = \hat{\boldsymbol{G}}\boldsymbol{Z}\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{e}}, \qquad (1)$$

where $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ are the residuals and $\hat{\boldsymbol{V}} = \boldsymbol{Z}\hat{\boldsymbol{G}}\boldsymbol{Z}^T + \hat{\boldsymbol{R}}$ is the estimated covariance matrix for $\boldsymbol{Y}$. In general, $\tilde{\boldsymbol{b}}$ can be interpreted as realized values of the random vector $\boldsymbol{b}$ (Robinson, 1991).

Our objective in this article is to compare an LMM containing p fixed effects and q random effects to a model with the same p fixed effects but only $q - r$ random effects, where $0 < r \leqslant q$. Performing this inference is equivalent to testing if the variances of the r random effects are all equal to 0. As stated before classical tests in this situation do not follow their typical χ^2_r distributions. Intuitive arguments as to why this is the case are presented by Molenberghs and Verbeke (2007).

2.2 *Permutation Tests*

A permutation test is one in which the null distribution of the test statistic is determined through permutations of the data; the test will have nominal size when the permutations are performed correctly. As an example, consider a study investigating the efficacy of a new treatment by comparing it to a placebo. The investigators wish to see if the treatment has an effect on some measured outcome of interest and randomize subjects equally to the treatment and placebo groups. Let X_i be the measured outcome for subject i in the treatment group, $i = 1, 2, \dots, n_x$, and Y_j be the outcome for subject j in the placebo group, $j = 1, 2, \dots, n_y$. The X_i are assumed to have distribution $\mathcal{F}$ with mean μ_x and variance σ^2, and the Y_i are assumed to have distribution $\mathcal{F}$ with mean μ_y and variance σ^2. Under the null hypothesis of no treatment effect, $\mu_x = \mu_y$, the two groups will have the same mean response, and more importantly, the same distribution.

Therefore, we can test our null hypothesis using the mean difference in observed response between treatment and placebo groups or $T = \bar{X} - \bar{Y}$, in which $\bar{X}$ is the observed mean response in the treatment group and $\bar{Y}$ is the observed mean in the placebo group. If $\mathcal{F}$ were a normal distribution, then T, appropriately standardized by its standard error, would have a t-distribution and the appropriate critical value would be determined from this distribution. If $\mathcal{F}$ were not a normal distribution, we could still appeal to the Central Limit Theorem and use the same t-distribution as an asymptotic approximation to the exact null distribution.

However, under the null hypothesis of no treatment effect, and conditioning on the observed outcomes of the $n_x + n_y$ subjects, the observed response of each patient would have occurred independent of group assignment. Thus, we can generate the null distribution for T by recomputing T under all $P = \binom{n_x+n_y}{n_x}$ possible permutations of group assignments. The p-value is obtained by computing the percentage of values in the permutation distribution whose magnitudes are at least as large as the magnitude of T. This permutation test is guaranteed to be nominal, meaning its size is no larger than desired (Hoeffding, 1952). More specifically, permutation tests assume that the values being permuted are exchangeable under the null hypothesis (Good, 2005). A vector, $\boldsymbol{Y}$, is exchangeable if, for any permutation of $\boldsymbol{Y}$ denoted as $\boldsymbol{Y}^*$, $\boldsymbol{Y}^*$ has the same distribution as $\boldsymbol{Y}$ (Commenges, 2003). It should be noted that exchangeability is a weaker condition than independent and identically distributed.

As the amount of data increases, so does the number of possible permutations, eventually making exact enumeration of all P permutations computationally unfeasible. Instead of calculating all possible permutations, an approximate permutation distribution can be generated through Monte Carlo sampling (Dwass, 1957). By randomly permuting the data between 100 and 1600 times (Good, 2005), an approximate permutation distribution can be generated, assuming the randomly selected permutations are drawn to sufficiently represent the tails of the exact permutation distribution.

3. Proposed Methods

3.1 *Best Linear Unbiased Predictors Based Permutation Test*

We begin by considering the hypothesis test for the inclusion or exclusion of a single random effect, $\boldsymbol{b_i} \sim N(0, \sigma^2_{b_i})$, in an LMM with no other random effects present. This is equivalent to testing if $\sigma^2_{b_i} = 0$. Thus, we are comparing the following models:

$$H_0 : Y_{ij} = \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + \epsilon_{ij}, \tag{2}$$

$$H_1 : Y_{ij} = \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_{i1} z_{1ij} + \epsilon_{ij}. \tag{3}$$

We use

$$T_1 = \sum_{i=1}^{N} \tilde{b}^2_{i1} / N, \tag{4}$$

as our test statistic, which is the sample variance of the BLUPs for the random effect, b_i. This statistic involves the sum of the squared BLUPs where the BLUPs are treated as a random sample of $b_i \sim N(0, \sigma^2_{b_i})$. Note that the denominator of the test statistic is constant for all of the permutations and does not affect the validity or power of our test.

To construct the permutation distribution with which to compare the observed test statistic, we permute the marginal errors, $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{X\beta}$. Under the null hypothesis of no random effects, the $\boldsymbol{\epsilon}$ are exchangeable, and more specifically, independent and identically normally distributed with mean 0 and variance σ^2_ϵ. By subtracting the fixed effects, $\boldsymbol{X\beta}$ from the response $\boldsymbol{Y}$, the errors have the benefit of not requiring the continuous $\boldsymbol{X}$'s to be identical among all subjects nor do the number of observations for each subject need to be the same. Therefore, we can permute the errors both within and between subjects. In practice, the errors are estimated by the residuals, $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, calculated from estimates fit from the alternative model, and Schmoyer (1994) showed that the residuals are also asymptotically exchangeable both within and among subjects under the null hypothesis.

The marginal residuals are part of the calculation for the BLUPs and lead to a straightforward permutation distribution for T_1. For each permutation $k = 1, 2, \dots, 1000$, we randomly permute the marginal residuals. Using these permuted residuals, we generate a permuted estimate $\hat{\sigma}^2_{b_i,k}$ for $\sigma^2_{b_i}$, from which we compute permuted values of the BLUPs that are used to compute T^*_{1k}, the permuted value of our test statistic T_1. These 1000 permuted values of T_1 result in an approximate empirical null distribution of T_1. The reestimation of $\sigma^2_{b_i}$ is performed because some permutations of the residuals will result in $\hat{\sigma}^2_{b_i} = 0$ and lead to the empirical null distribution having positive mass at zero. We then generate a p-value by calculating the percentage of permutations with T^*_1 greater than T_1.

Next, we extend the permutation test to test for the presence of a single random effect in a model that contains other random effects such as:

$$H_0 : Y_{ij} = \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_{i1} z_{1ij} + \epsilon_{ij}, \tag{5}$$

$$H_1 : Y_{ij} = \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_{i1} z_{1ij} + b_{i2} z_{2ij} + \epsilon_{ij}. \tag{6}$$

In this setting, the null model now contains other random effects so that all $\sum_{i=1}^{N} n_i$ errors are no longer exchangeable under the null hypothesis. Instead, the errors are normally distributed with mean $\boldsymbol{0}$ and covariance matrix, $\boldsymbol{V_0} = \sigma^2_{b_{1_0}} \boldsymbol{Z}^T \boldsymbol{Z} + \boldsymbol{R_0}$ with $\boldsymbol{R_0} = \sigma^2_{\epsilon_0} \boldsymbol{I}$. We resolve this issue by weighting the errors by the matrix $(\boldsymbol{U}_0^T)^{-1}$, where $\boldsymbol{U}_0$ is the Cholesky decomposition of $\boldsymbol{V}_0$, i.e., $\boldsymbol{V_0} = \boldsymbol{U}_0^T\boldsymbol{U}_0$. As a result, the set of weighted errors, $(\boldsymbol{U}_0^T)^{-1}(\boldsymbol{Y}-\boldsymbol{X\beta})$, are normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{I}$, and are thereby exchangeable, allowing once again for permutations both within and between subjects. We reexpress the test statistic T_1 in equation (4), to incorporate the Cholesky decomposition as:

$$T_2 = \sum_{i=1}^{N} \tilde{b}_{i2}^2/N = \sum_{i=1}^{N} \left[\boldsymbol{G}_1\boldsymbol{Z}\hat{\boldsymbol{V}}_1^{-1}\boldsymbol{U}_0^T\left(\boldsymbol{U}_0^T\right)^{-1}\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\right]^2/N. \tag{7}$$

Note that T_2 is only calculated for the single random effect being tested. For the observed data the statistic remains the sample variance of $\tilde{b}_{i2}$ because $\boldsymbol{U}_0^T(\boldsymbol{U}_0^T)^{-1}$ equals the identity for the unpermuted weighted residuals. Also, the earlier random intercept hypothesis test is a special case of this test, because the Cholesky decomposition in that scenario is equal to the identity, and (7) reduces to (4). With the appropriate weights, this BLUP-based permutation test can be used to perform inference on any single random effect of interest.

In simulation studies, this permutation test is shown to be valid and displays power comparable to the asymptotic mixture χ^2 likelihood ratio tests. The test is very intuitive and easy to perform. However, because the test is based on the BLUPs, it does have one limitation: it can only test for one random effect at a time. In the next section, we present a likelihood ratio based permutation test that allows for testing of multiple random effects and of which the BLUP permutation test is a special case.

3.2 *Likelihood Ratio Based Permutation Test*

This permutation test is based on the restricted likelihood ratio test statistic, $\lambda = -2\log(L_{H_0} - L_{H_1})$, where L_{H_0} and L_{H_1} are the restricted likelihoods under the null and alternative hypotheses, respectively. Using the same LMM notation as described previously where $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$ and $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{X\beta}$, we have $\lambda = \log\left[|\boldsymbol{V_0}|/|\boldsymbol{V_1}|\right] + \boldsymbol{\epsilon}^T(\boldsymbol{V_0}^{-1} - \boldsymbol{V_1}^{-1})\boldsymbol{\epsilon} + \log\left[|\boldsymbol{X}^T\boldsymbol{V_0}^{-1}\boldsymbol{X}|/|\boldsymbol{X}^T\boldsymbol{V_1}^{-1}\boldsymbol{X}|\right]$.

Let us test for a random intercept using the null and alternative hypotheses stated in (2) and (3). Similar to the BLUP-based permutation test, the likelihood ratio test statistic involves the marginal errors, $\boldsymbol{\epsilon}$, and we can permute $\boldsymbol{\epsilon}$ within and between the subjects under the null hypothesis. Therefore, the test statistic becomes

$$\begin{aligned} T_3 = {}& \log\left[|\hat{\boldsymbol{V}}_0|/|\hat{\boldsymbol{V}}_1|\right] + \hat{\boldsymbol{e}}_1^T(\hat{\boldsymbol{V}}_0^{-1} - \hat{\boldsymbol{V}}_1^{-1})\hat{\boldsymbol{e}}_1 \\ & + \log\left[|\boldsymbol{X}^T\hat{\boldsymbol{V}}_0^{-1}\boldsymbol{X}|/|\boldsymbol{X}^T\hat{\boldsymbol{V}}_1^{-1}\boldsymbol{X}|\right], \end{aligned} \tag{8}$$

which is λ with all parameters replaced by their estimates under the null and alternative hypotheses as denoted by their subscripts.

Similar to the BLUP-based permutation test, a new $\hat{\boldsymbol{V}}_0$ and $\hat{\boldsymbol{V}}_1$ is estimated for each permutation of $\hat{\boldsymbol{e}}_1$ and denoted as $\hat{\boldsymbol{V}}_0^*$ and $\hat{\boldsymbol{V}}_1^*$. The permuted residuals are treated as an outcome, and $\hat{\boldsymbol{V}}_0^*$ is estimated from a mixed model with a fixed intercept and random effects from the null hypothesis. We estimate $\hat{\boldsymbol{V}}_1^*$ from a mixed model with a fixed intercept and random effects from the alternative hypothesis.

Reestimation of $\hat{\boldsymbol{V}}_0^*$ and $\hat{\boldsymbol{V}}_1^*$ is necessary due to the changes that occur in the rank of $\hat{\boldsymbol{\Sigma}}$ when random effect variances are estimated to be equal to 0. If we do not reestimate $\boldsymbol{V_0}$ and $\boldsymbol{V_1}$ (including $\boldsymbol{\Sigma_0}$ and $\boldsymbol{\Sigma_1}$), the permutation distribution will be completely based on estimates from the observed data. By estimating $\boldsymbol{V_0}$ and $\boldsymbol{V_1}$ for each permutation, we allow the empirical distribution to 'mix' as the rank of $\hat{\boldsymbol{\Sigma}}$ varies, thereby generating a distribution similar to the mixture χ^2 asymptotic distribution of Stram and Lee (1994). We create the permutation distribution by calculating T_3^* for each of the random permutations and determine a *p*-value through the location of T_3 in the permutation distribution.

When testing the presence of one random effect with one or more additional random effects in the null hypothesis, things proceed similar to that of the BLUP permutation test. To be able to permute the errors, they must first be weighted by $(\boldsymbol{U}_0^T)^{-1}$. Once weighted, the errors are exchangeable and can be permuted. The permuted weighted errors are then multiplied (unweighted) by $(\boldsymbol{U}_0^T)$ to get them back on the original scale of the residuals, and for each permutation, $\hat{\boldsymbol{V}}_0^*$ and $\hat{\boldsymbol{V}}_1^*$ are reestimated using the unweighted permuted errors as described earlier. Then T_3^* is calculated, and the permutation distribution is generated for the likelihood ratio test statistic to which the observed test statistic will be compared and a *p*-value calculated.

If we wish to test for the inclusion of $0 < r \leqslant q$ random effects, we have the models:

$$\begin{aligned} H_0: Y_{ij} &= \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_{i1}z_{1ij} \\ &\qquad + \cdots + b_{i(q-r)}z_{(q-r)ij} + \epsilon_{ij}, \\ H_1: Y_{ij} &= \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_{i1}z_{1ij} + \cdots + b_{iq}z_{qij} + \epsilon_{ij}. \end{aligned}$$

The steps for this scenario are identical to those from the previous scenario where testing for one random effect in the presence of additional random effects in the null hypothesis. Nonetheless, we emphasize the importance of reestimating $\boldsymbol{\Sigma_0}$ and $\boldsymbol{\Sigma_1}$ after each permutation when testing for multiple random effects. Herein lies the largest contribution of our methods: for a general value of *r*, simulation is the only existing approach for calculating the correct mixing probabilities for the χ^2 distributions. In contrast, our permutation test based on the likelihood ratio statistic will automatically generate the correct mixing probabilities as the rank of $\hat{\boldsymbol{\Sigma}}^*$ changes from permutation to permutation.

4. Simulation Studies

4.1 *Validity*

We performed a series of simulation studies to examine the performance of our permutation tests under a number of different settings. The first study was used to evaluate the validity of our two tests under four different scenarios: (1) testing for a random intercept, (2) testing for a random slope given an independent random intercept is present in the null hypothesis, (3) testing for a random slope given a potentially correlated random intercept, and (4) simultaneously testing for both random intercept and random slope. Five hundred data sets were generated for each of the simulation scenarios using the following random intercept model:

$$Y_{ij} = \beta_1 + \beta_2 x_{2ij} + b_{i1} + \epsilon_{ij}, \tag{9}$$

with $\beta_1 = 3$, $\beta_2 = 2.75$, $\sigma_\epsilon^2 = 1$, $b_{i1} \sim N(0, \sigma_{i1}^2)$, and our fixed effect, x_{2ij}, was randomly drawn from the standard normal distribution. Then, similar to Saville and Herring (2009), x_{2ij} was centered at 0 and scaled by twice its standard error. For scenarios 1 and 4, $\sigma_{b_{i1}}^2$ was set equal to 0, while for scenarios 2 and 3, $\sigma_{b_{i1}}^2$ was set to 1. We varied both the number of subjects, $N \in \{50, 10\}$, as well as the number of observations per subject, $n \in \{10, 5\}$, and compared the size of our permutation tests to that of the asymptotic restricted likelihood ratio test with a 50:50 mixture of χ^2 distributions with 0 and 1 degrees of freedom, 1 and 2 degrees of freedom, 1 and 2 degrees of freedom, and 0, 1, and 2 degrees of freedom in a 25:50:25 ratio, for scenarios 1, 2, 3, and 4, respectively. The mixing probabilities for scenario 4 were derived from case 4 of Stram and Lee (1994) who state that when the information matrix is equal to the identity under the null hypothesis, the likelihood ratio test has an asymptotic null distribution that is a mixture of χ^2 distributions with binomial mixing probabilities. For all other situations they recommend finding the critical value through simulations.

All estimates were performed in the statistical package `R` using the `lmer()` function from the `R`-package `lme4` (Bates, Maechler, and Bolker, 2011). Unlike other LMM fitting algorithms that can only estimate extremely small values for variances, `lmer()` is able to estimate 0 for the variance components. The simulations were performed using 20 cores of an Intel Xeon X5660 2.80 GHz server with 32 gigabytes of memory.

The simulation results for validity are presented in Table 1. In all settings, both permutation tests have valid size, defined as a size contained in the interval (0.031, 0.061), the approximate 95% confidence interval for type I error rate with 500 simulations. In contrast, the asymptotic test for one random effect (scenarios 1, 2, and 3) becomes more conservative as the number of subjects or the number of observations decreases. In addition, it appears that under scenario 4, the asymptotic likelihood ratio test is liberal when $N = 10$ and $n = 5$.

4.2 *Power*

The simulations to examine the power of the tests were performed for the same four scenarios in the validity study. We generated 500 data sets using the random intercept and slope model:

$$Y_{ij} = \beta_1 + \beta_2 x_{2ij} + b_{i1} + b_{i2} z_{2ij} + \epsilon_{ij}, \tag{10}$$

with the same fixed effects from the validity simulations and with $b_{i1} \sim N(0, \sigma_{i1}^2)$, $b_{i2} \sim N(0, \sigma_{i2}^2)$, and $x_{2ij} = z_{2ij}$. We varied the variance of the random effect (or random effects under scenario 4) of interest, $k \in \{1, 2\}$, $\sigma_{ik}^2 \in \{0.15, 0.2, 0.3\}$ as well as both the number of subjects, $N \in \{50, 10\}$, and the number of observations per subject, $n \in \{10, 5\}$. For scenarios 3 and 4 the correlation of the random effects, ρ, was set equal to -0.3.

The results of the power simulations are shown in Table 1. With the exception of scenario 4, both permutation tests displayed strictly better power than the asymptotic test, even

Table 1

Size and power for the permutation tests compared to the asymptotic likelihood ratio test

			Testing scenarios										
			(1)			(2)			(3)			(4)	
N	n	σ_{i1}^2 and/or σ_{i2}^2	B	L	A	B	L	A	B	L	A	L	A
50	10	0	5.8	5.8	5.0	5.8	5.8	5.6	5.0	4.8	4.2	4.2	4.4
		0.15	99.2	99.2	98.8	40.2	41.2	38.4	45.2	42.6	38.9	98.0	98.0
		0.20	100.0	100.0	100.0	57.4	58.4	55.4	62.2	58.4	56.4	98.8	98.8
		0.30	100.0	100.0	100.0	82.4	81.0	78.8	78.8	75.4	73.6	99.8	99.8
	5	0	3.8	3.6	2.6	6.2	5.2	5.0	5.0	5.2	4.4	4.8	4.0
		0.15	80.0	80.0	77.6	16.2	18.2	16.0	21.4	22.0	18.6	56.4	55.4
		0.20	91.2	91.2	90.2	27.8	27.4	24.4	27.0	25.4	22.0	70.1	69.3
		0.30	97.6	97.6	97.6	38.4	39.0	36.8	41.6	39.6	36.6	92.6	92.6
10	10	0	5.4	5.4	4.0	5.2	4.4	3.0	6.2	4.8	3.4	4.2	5.0
		0.15	63.4	63.2	58.8	16.2	15.2	12.6	17.3	16.3	11.0	55.6	58.3
		0.20	75.2	74.6	69.6	23.6	23.6	19.8	23.1	21.7	17.3	68.1	70.1
		0.30	89.0	89.0	87.6	34.4	34.8	29.8	30.7	27.3	22.5	88.2	89.0
	5	0	4.6	4.4	3.6	5.2	3.8	2.6	5.6	5.2	3.8	5.6	7.0
		0.15	31.6	29.4	27.0	10.0	8.6	7.0	9.8	10.0	7.2	24.8	29.1
		0.20	44.6	43.4	37.6	12.6	11.4	9.2	12.6	13.0	8.8	37.5	42.1
		0.30	63.6	62.0	58.6	12.8	13.8	11.2	15.7	15.7	10.6	47.9	53.3

Results are reported in percentages.

(1): Random intercept test.

(2): Random slope test with an independent random intercept present.

(3): Random slope test with a correlated random intercept present.

(4): Simultaneous test for the random intercept and random slope.

B: BLUP-based permutation test.

L: Likelihood ratio based permutation test.

A: Asymptotic likelihood ratio test.

when the asymptotic test had nominal size. For scenario 4 the asymptotic likelihood ratio test using the 25:50:25 ratio of χ^2 distributions and the likelihood ratio based permutation test performed very similarly when $N = 50$. However, the number of rejections of the asymptotic test is higher than the permutation test for $N = 10$, and this can be explained by its inflated type I error rate. In fact, when critical values found through simulation were used instead of the 25:50:25 mixture χ^2 null distribution, the power results for the asymptotic test were almost identical to those from the permutation test for all combinations of N and n.

Given that the residuals follow known normal distributions, it is possible that residuals could be drawn directly from those distributions (bootstrapped), rather than permuting the actual residuals, to generate the empirical null distributions of T_1, T_2, and T_3. To examine this idea, we performed simulations in which we replaced permuting the residuals with instead simulating new values from the appropriate normal distributions. All other steps in the permutation tests were identical to those presented in Section 3. Both the BLUP and the restricted likelihood ratio versions were examined. N and n were set at 10, and we varied the variance of the random slope, $\sigma^2_{i2} \in \{0, 0.15, 0.2, 0.3\}$. We tested for the presence of a random slope given a potentially correlated random intercept. The results from these simulations closely mirrored the results of the permutation tests in Table 1. Both test statistics using bootstrap residuals led to valid inference. When $\sigma^2_{i2} = 0.15$ the powers were 16.1% and 16.6% for the BLUP test and the restricted likelihood ratio tests, respectively, compared with the 17.3% and 16.3% from the permutation tests. For $\sigma^2_{i2} = 0.2$ the powers for the BLUP and the restricted likelihood ratio tests were 23.1% and 20.7%, respectively, and for $\sigma^2_{i2} = 0.3$, the powers were 29.1% and 27.9%, respectively.

4.3 *Sensitivity to Nonnormality*

We also investigated the sensitivity of the permutation tests to nonnormality of the random effects and/or residuals when testing for a random slope given an independent random intercept in the model with $N = 10$ and $n = 10$. Both the null model with $\sigma^2_{i2} = 0$ and the alternative with $\sigma^2_{i2} = 0.3$ were run. Four different settings were studied: (1a) normal errors and normal random effects, (1b) logistically distributed errors and normal random effects, (1c) normal errors and logistically distributed random effects, and (1d) logistically distributed errors and logistically distributed random effects. Size and power estimates are given in Table 2. We see that under the null hypothesis, both permutation tests appear have size closer to nominal than the asymptotic test, with the asymptotic test being conservative in settings 1a, 1b, and 1c. Under the alternative hypothesis, we see that as expected, the permutation test is most powerful when the data truly are normally distributed (setting 1a), with slight losses in power when extra variation due to nonnormality exists in the data. Nonetheless, the power losses of the permutation tests are slight, and in all settings, the permutation tests display greater power than the asymptotic test.

4.4 *Comparison to Existing Methods*

In our final simulation study, we compared the permutation tests to a portion of the results published by Saville and Herring (2009) when testing for the presence of a random slope. Following their simulation settings, we generated 250 data sets from (10) with $\beta_0 = 2.75$, $\beta_1 = 3$, $n_i = n = 10$, $\sigma^2_{i1} = 1$, and $\rho = -0.3$. The standard deviation for the random slope, $\sigma_{i2} \in \{0, 0.15, 0.30, 0.45, 0.60\}$. Table 3 presents the BLUP and likelihood ratio based permutation results for $N \in \{100, 50\}$ next to the published results from Saville and Herring resulting from Bayes factors based on two different parameterizations of the model.

We see that the power for the likelihood ratio based permutation test is comparable with the approximate Bayes factors method employed by Saville and Herring. Despite some difference in results due to simulation variability, for all settings, our permutation test is as powerful or even more powerful than one or both of the tests of Saville and Herring.

Table 2

Size and power of proposed permutation tests when random effects and/or errors are nonnormally distributed

		Method		
Model	Setting	B	L	A
$\sigma^2_{i2} = 0.0$	1a	5.2	4.4	3.0
	1b	5.4	4.4	3.6
	1c	4.4	4.4	3.2
	1d	5.0	5.0	5.6
$\sigma^2_{i2} = 0.3$	1a	34.4	34.8	29.8
	1b	29.2	29.4	26.8
	1c	29.4	30.4	25.2
	1d	29.4	30.0	27.2

Results are reported in percentages.
Settings: (1a): Normal errors and normal random effects.
(1b): Logistic errors and normal random effects.
(1c): Normal errors and logistic random effects.
(1d): Logistic errors and logistic random effects.
B: BLUP-based permutation test.
L: Likelihood ratio based permutation test.
A: Asymptotic likelihood ratio test.

Table 3

Comparison of power of permutation tests to results reported by Saville and Herring when testing for the inclusion of a random slope

N	σ_{i2}	SH1	SH2	BLUP	LRT
50	0.00	8	3	3	4
	0.15	14	7	8	9
	0.30	30	17	28	22
	0.45	56	57	66	60
	0.60	75	90	94	91
100	0.00	4	4	5	4
	0.15	12	8	14	12
	0.30	38	38	44	43
	0.45	69	87	91	90
	0.60	72	99	100	100

Results are reported in percentages.
SH1: Bayes' factor as described in Saville and Herring (2009, p. 370).
SH2: Bayes' factor as described in Saville and Herring (2009, p. 371).
BLUP: BLUP-based permutation test.
LRT: Likelihood ratio based permutation test.

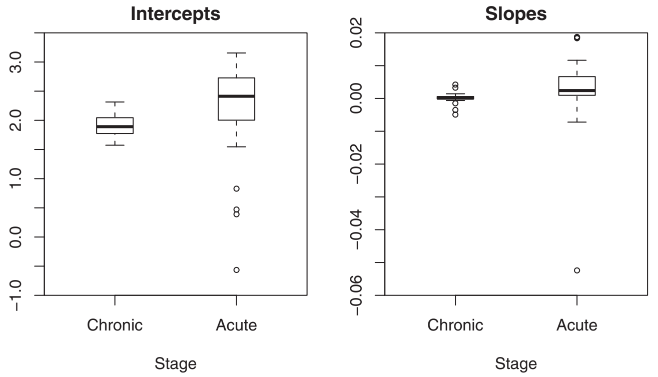

Figure 1. Boxplots stratified by stage of the patient-specific intercepts and slopes produced from a linear regression model of ADA levels over time.

Table 4

Permutation and asymptotic likelihood ratio test results for inclusion of specific random effects when modeling ADA levels in patients with chronic myelogenous leukemia

Test	Observed LRTS	Permutation p-value	Asymptotic p-value
(5) versus (4)	3.00	0.226	0.474
(4) versus (1)	234.59	< 0.001	< 0.001
(4) versus (2)	146.99	< 0.001	< 0.001
(4) versus (3)	12.88	0.019	0.004

(1): No random effects.
(2): Random intercept only model.
(3): Random intercepts for both stages model.
(4): Random intercepts for both stages and a random slope for acute stage only.
(5): Random intercepts and slopes for both stages.

5. Application

We applied our permutation test to a set of data presented in Klein, Klotz, and Grever (1984) that was collected on patients with chronic myelogenous leukemia. Chronic myelogenous leukemia is characterized by a lengthy chronic phase with little to no symptoms that eventually transitions into an accelerated phase that behaves similar to acute leukemia. The length of time until the transition from a chronic to an accelerated phase can vary greatly among patients, motivating the discovery of markers that can indicate when chronic myelogenous leukemia is about to change from a chronic to an accelerated stage. One potential marker is adenosine deaminase or ADA. This particular data set contains the ADA levels of 55 patients that were measured at various time points during their follow-up. Time is quantified as days following the initial observation date, and at each time point, investigators also recorded the phase of each patient's disease as chronic or accelerated. The frequency of the repeated measurements as well as the times of the measurements were not fixed and fluctuated greatly. Patients had anywhere from 2 to 59 measurements, and the repeated measurements took place from the initial observation date up to 1073 days following the diagnosis date.

We modeled the ADA measurements as patients progress from chronic to accelerated phases, and we were primarily interested in evaluating the level of heterogeneity among the patients to see if random effects are necessary in our model. Figure 1 contains boxplots stratified by stage of disease of the slopes and intercepts from individual linear regressions of each patient's ADA measurements on time. The figure indicates significant variation between the two stages, both in terms of mean ADA levels as well as changes over time, necessitating the inclusion of random effects.

We are also interested in investigating how the rate of change in ADA differs between chronic and accelerated phases. We applied a cubed root transformation to the ADA values so that they were approximately normally distributed, and fit an LMM with the cubed root ADA assay values regressed on disease phase, with chronic as the baseline category, number of days from the initial observation date, and interaction terms between the two to allow the time effect to differ between the two disease states. Our initial model is $ADA_{ij}^{1/3} = \beta_1 + b_{i1} + (\beta_2 + b_{i2})State_{ij} + (\beta_3 + b_{i3})Days_{ij} + (\beta_4 + b_{i4})State_{ij} * Days_{ij} + \epsilon_{ij}$. The full random effects model includes four random effects, b_{i1}, b_{i2}, b_{i3}, and b_{i4}, to allow for at most a random intercept and time effect for each of the two disease stages. We wish test if any or all of these random effects should be included.

Table 4 shows the results of our permutation tests, based on 1000 permutations, for the inclusion or exclusion of the random effects, along with results from the asymptotic likelihood ratio test. Both tests support what is seen in Figure 1: the random day effect for the chronic stage is not significant, while the other three random effects appear to be significant. As a gauge of the computation time necessary, each of these tests takes around 4 minutes to perform when using 20 cores of an Intel Xeon X5660 2.80 GHz server with 32 gigabytes of memory.

6. Discussion

In this article, we have proposed two methods for performing inference on random effects by permuting the weighted residuals both within and among subjects. In some simulations, we have found that the convergence of the solutions derived from the `lmer()` function in the statistical package `R` appears to suffer as the number of random effects increases. Our current solution is to generate more permutations to ensure that there are enough permutations to create the null distribution.

As demonstrated, the proposed permutation tests perform well even when the number of patients and the number of observations per patient is small. The tests also do not require balanced data nor do the measurements need to occur at the same points in time. As a result, our methods can be applied to the use of an LMM representation of penalized spline models (Ruppert et al., 2003) in which the smoothing parameter is a random effect. Finally, implementing these permutation tests is straightforward and can be incorporated into standard practice for analysis of LMMs using existing software; example computer code can be found at `www.sph.umich.edu/~tombraun/software.html`. Although the methods are computationally intensive, the recent rise in parallel computing through clusters and multicore processors has made it possible to greatly reduce the amount of time necessary to implement these tests.

We are currently generalizing the methods presented in this manuscript to allow for permutation-based inference in generalized linear mixed models (GLMMs). Our approach is based upon a first-order approximation of the GLMMs to make it resemble the form of an LMM, an approach that is the foundation of penalized quasi-likelihood (Breslow and Clayton, 1993) for estimation in GLMMs. We plan to present the results of our research in a forthcoming manuscript.

Acknowledgement

Funding for Oliver Lee was provided by NIH grant T32 CA083654.

References

Bates, D., Maechler, M., and Bolker, B. (2011). *lme4: Linear Mixed-Effects Models using S4 Classes.* `R` package version 0.999375–39. Vienna, Austria: CRAN.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88,** 9–25.

Commenges, D. (2003). Transformations which preserve exchangeability and application to permutation tests. *Journal of Nonparametric Statistics* **15,** 171–185.

Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B* **66,** 165–185.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* **28,** 181–187.

Fitzmaurice, G. M. and Ibrahim, J. G. (2007). A note on permutation tests for variance components in multilevel generalized linear mixed models. *Biometrics* **63,** 942–946.

Good, P. I. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, 3rd edition. New York: Springer-Verlag Inc.

Greven, S., Crainiceanu, C. M., Küchenhoff, H., and Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics* **17,** 870–891.

Henderson, C. R. (1950). Estimation of genetic parameters. *Annals of Mathematical Statistics* **21,** 309–310.

Hoeffding, W. (1952). The large-sample power of tests based on permutation of observations. *Annals of Mathematical Statistics* **23,** 169–192.

Kinney, S. K. and Dunson, D. B. (2008). Fixed and random effects selection in linear and logistic models. *Biometrics* **63,** 690–698.

Klein, J. P., Klotz, J. H., and Grever, M. R. (1984). A biological marker model for predicting disease transitions. *Biometrics* **40,** 927–936.

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

Molenberghs, G. and Verbeke, G. (2007). Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician* **61,** 22–27.

Morrell, C. H. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics* **54,** 1560–1568.

Öfversten, J. (1993). Exact tests for variance components in unbalanced mixed linear models. *Biometrics* **49,** 45–57.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science* **6,** 15–51.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge; New York: Cambridge University Press.

Saville, B. R. and Herring, A. H. (2009). Testing random effects in the linear mixed model using approximate Bayes factors. *Biometrics* **65,** 369–376.

Schmoyer, R. L. (1994). Permutation tests for correlation in regression errors. *Journal of the American Statistical Association* **89,** 1507–1516.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82,** 605–610.

Silvapulle, M. J. (1992). Robust Wald-type tests of one-sided hypotheses in the linear model. *Journal of the American Statistical Association* **87,** 156–161.

Silvapulle, M. J. and Silvapulle, P. (1995). A score test against one-sided alternatives. *Journal of the American Statistical Association* **90,** 342–349.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50,** 1171–1177.

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics* **59,** 254–262.

Received December 2010. *Revised July* 2011.
Accepted August 2011.