

# A Distribution-Free Test of Constant Mean in Linear Mixed Effects Models

Johan Lim<sup>1</sup>, Xinlei Wang<sup>2\*,†</sup>, Seokho Lee<sup>3</sup> and Sin-Ho Jung<sup>4</sup>

<sup>1</sup> *Department Statistics, Seoul National University, Seoul, Korea*

<sup>2</sup> *Department of Statistical Science, Southern Methodist University, Dallas, TX, USA*

<sup>3</sup> *Department of Statistics, Texas A&M University, College Station, TX, USA*

<sup>4</sup> *Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA*

## SUMMARY

We propose a distribution free procedure, an analogy of the DIP test in nonparametric regression, to test whether means of responses are constant over time in repeated measures data. Unlike existing tests, the proposed procedure requires very minimal assumptions to the distributions of both random effects and errors. We study the asymptotic reference distribution of the test statistic analytically, and propose a permutation procedure to approximate the finite-sample reference distribution. The size and power of the proposed test are illustrated and compared with competitors through several simulation studies. We find that it performs well for data of small sizes, regardless of model specification. Finally, we apply our test to a data example to compare the effect of fatigue in two different methods used for cardiopulmonary resuscitation. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: constant mean; DIP test; hypothesis testing; mixed effects model; permutation test

## 1. INTRODUCTION

Repeated measures data are now common in many fields, and they are frequently used to examine changes in the response variable (say  $y$ ) over time or other factors (or covariates). The goal of this paper is to test the existence of such changes.

The application that motivates this paper is a randomized crossover medical study, in which people are interested in comparing the effect of fatigue in two different methods used for cardiopulmonary resuscitation (CPR), continuous chest compression CPR (CCC-CPR) and standard CPR (STD-CPR). The study involves 57 medical students; each performs two sessions of 9-minute exercise separated by at least two days, one using CCC-CPR and the other using STD-CPR. The number of adequate chest compressions (defined as at least 38 millimeters of compression depth) delivered per minute is recorded in each session. Due to fatigue, the mean number of adequate compressions tends to decline over time, no matter what method is used. The problem is to test whether the fatigue has a greater effect in one

---

\*Correspondence to: Xinlei Wang, Department of Statistical Science, 3225 Daniel Avenue, PO Box 750332, Dallas, Texas 75275-0332, USA

†E-mail: swang@smu.edu

method than that in the other. To address this, we need to compare the slopes of those mean numbers between the two methods. This can be done by considering the difference in the number of adequate compressions between CCC-CPR and STD-CPR at the same time point and within a subject, and then test whether those differences change over time.

For repeated measures data, methods for testing the existence of any change in  $y$  over  $t$  are primarily based on mixed effects models. Let  $y_{ij}$  and  $t_{ij}$  be response and observation time of the  $j$ th repetition of the  $i$ th subject for  $i = 1, \dots, n$  and  $j = 1, \dots, m$ . The mixed effects model we consider is

$$y_{ij} = \mu(t_{ij}) + z_i^T \alpha_i + \epsilon_{ij}, \quad (1)$$

where  $\mu(\cdot)$  is the mean function of fixed effects,  $z_i$  is the vector of covariates for the  $i$ th subject,  $\alpha_i$  is the corresponding vector of random effects with mean 0 and covariance matrix  $\Sigma_\alpha$ , and  $\epsilon_{ij}$ 's are independent errors with mean 0 and variance  $\sigma_\epsilon^2$ .

In the literature, likelihood ratio tests and F-tests are available for testing the constancy of the mean function. A simple description is provided below.

The likelihood ratio tests can be constructed based on a general multivariate normal assumption to  $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{im})^T$ , where the model

$$y_{ij} = \mu + \beta_j + z_i^T \alpha_i + \epsilon_{ij} \quad (2)$$

with normally distributed  $\alpha_i$  and  $\epsilon_{ij}$ s is assumed. The test statistic (2 times negative log-likelihood ratio) has the form of  $\log(\det(S_1)/\det(S_0))$ , where  $\det(A)$  is the determinant of the matrix  $A$ , and  $S_0/S_1$  is the covariance matrix estimate with/without the constant mean assumption (*i.e.*,  $H_0 : \beta_1 = \dots = \beta_m$ ). The likelihood ratio test under the model (2) is often conjectured to be asymptotically the most powerful, whose asymptotic reference distribution is the Chi-square distribution with  $m - 1$  degrees of freedom. For a full discussion, see [1]. Also, an alternative to the likelihood ratio test is developed in [2].

The F-tests for testing the constancy of means are also based on the normality assumption for both the random effects and errors. The test statistic is often chosen as a ratio of mean squares that has an F distribution under  $H_0$ . For a general balanced ANOVA setup with mixed effects, the standard F-tests have been shown to be optimal (UMPU, UMPI and UMPIU) [3, 4, 5, 6]. Here, we focus on the usual F-tests based on two basic models, the linear trend mixed effects model (LTMM) and the nonparametric trend mixed effects model (NTMM). In the LTMM, the linear trend over time is assumed so (1) becomes

$$y_{ij} = \mu + \beta t_{ij} + z_i^T \alpha_i + \epsilon_{ij}, \quad (3)$$

while in the NTMM, no parametric form is assumed so that (1) is actually (2). As a result, the F-test based on the NTMM (F-NT) is much more flexible than the one based on the LTMM (F-LT). But fitting the NTMM is very intensive in computation, especially when  $m$  is large, due to large-scale matrix calculations. Both F-LT and F-NT are supported in standard statistical packages such as R or SAS. In this paper, we implement the F-LT using R and the F-NT using the formula presented in [7].

We should also note that besides the simple structures of the fixed effects specified in (2) and (3), several other approaches are made to model  $\mu(\cdot)$  non-parametrically. For example, [8], [9] and [10] all use splines to model the fixed effects, although they differ in modeling the dependence of observations of each subject. [11] and [12] use kernel methods to model the time trend. However, such work usually concentrates on estimation of the nonparametric fixed effects, not on significance testing.

As discussed above, the previous work on testing the mean function  $\mu(\cdot)$  often requires the normality assumption. Several authors point out possible bias and inefficiency, resulted from mis-specified

distributions of random effects or errors, and suggest improved methods such as a sandwich-type correction for estimating the standard errors and the Box-Cox transformation [13, 14, 15].

In this paper, we propose a non-parametric procedure to test the constancy of the mean, which does not require any distributional assumption to random effects or errors beyond zero mean and finite variance. We show that the proposed procedure is more powerful than the standard F-tests when the sample size  $n$  is small. Even for large  $n$ , it performs reasonably well, compared with F-tests based on correctly specified models. In addition, our procedure is, by no means, restricted to the simple structures in the models (2) and (3).

The remainder of the paper is organized as follows. In Section 2, we introduce the DIP test and study the proposed test statistic in details. A permutation procedure is presented to evaluate the finite-sample reference distribution of the test statistic. Section 3 and 4 report simulation results about the size of the DIP test and the power in different models. Note that the numerical comparisons are done between the DIP test and the F-tests (F-LT and F-NT). This is because (i) the F-tests have optimal properties in theory under the normality assumption and (ii) the F-tests, due to their availability through standard software procedures, are more commonly used in practice than the likelihood ratio tests. In Section 5, we employed the DIP test to study the chest compression data. Section 6 concludes the paper.

## 2. THE DIP TEST

The main theme of this paper is testing whether the mean function  $\mu(t)$  is constant (or uniform) over  $t$ , especially for large  $m$ . Without loss of generality, assume  $t_{i1} \leq t_{i2} \leq \dots \leq t_{im}$  for each  $i$ . The null hypothesis we consider is that  $y_{i1}, y_{i2}, \dots, y_{im}$  have the same mean over  $j$ .

We begin with the residual process

$$e_{ij} = y_{ij} - \bar{y}_i, \quad \text{for } i = 1, 2, \dots, n, \quad j = 1, 2, \dots, m$$

where  $\bar{y}_i = \sum_{j=1}^m y_{ij}/m$ . Note that each component in the above residual process has zero mean under  $H_0$ . So if the process deviates much from 0, then the uniformity of the mean function is disproved.

There are several ways to measure the deviation of the residual process from 0, which include the  $\mathcal{L}_p$  distance or the Mahalanobis distance from the zero vector. In this paper, we suggest to use the maximum of the partial sum process as our test statistic

$$\mathbf{T}_n = \max_{k=1}^m \frac{1}{\sqrt{nm}} \left| \sum_{j=1}^k \sum_{i=1}^n (y_{ij} - \bar{y}_i) \right| \quad \text{or} \quad \mathbf{T}'_n = \mathbf{T}_n / s_n, \quad (4)$$

where  $s_n^2$  is any consistent estimate of  $\sigma_\epsilon^2$  under both  $H_0$  and  $H_a$ . This is similar in spirit to [16], who use the maximum difference of cumulative distribution functions to test the equality between two distributions. Hence, we call our testing procedure the DIP test (or simply DIP), as in [16]. In what follows, we study the test statistic  $\mathbf{T}_n$  in details.

First, we answer the question of what  $\mathbf{T}_n$  measures via a simple example. Suppose we have five repeated measures from one subject, say  $(y_1, y_2, y_3, y_4, y_5) = (1, 2, 3, 4, 5)$ . Let  $\bar{y}$  be the average of  $y_1, \dots, y_5$ . Then  $\mathbf{T}_n$  is the maximum discrepancy between  $\sum_{j=1}^k y_j$  and  $k \bar{y}$ . In Figure 1,  $k \bar{y}$ , the upper straight line, is the partial process under  $H_0$ ; and  $\sum_{j=1}^k y_j$ , the lower convex curve, is the observed partial sum processes. Then each  $d_k = \sum_{j=1}^k (y_j - \bar{y})$  measures the deviation of the observed process from the expected under  $H_0$  at each time point. Clearly,  $\mathbf{T}_n$  is the maximum of  $|d_1|, \dots, |d_4|$ .

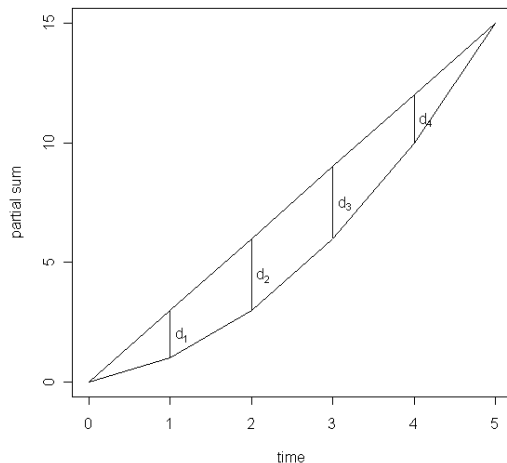


Figure 1. An illustration of the components in  $\mathbf{T}_n$ .

It can be verified that the asymptotic reference distribution of the test statistic is in the form of maxima of the Brownian bridge for each of the following cases; (i)  $n \rightarrow \infty$  with a fixed  $m$ , and (ii)  $m \rightarrow \infty$  with a fixed  $n$ . For details, see Appendix.

In practice, often the size of the data set is not large enough to apply the asymptotic distributions given in the appendix or the standard deviation  $\sigma_\epsilon$  is hard to estimate consistently in both  $H_0$  and  $H_a$ . To overcome such difficulties, we propose a permutation procedure to approximate the finite-sample reference distribution. It is based on the fact that under  $H_0$ , the joint distribution of  $y_{i1}, y_{i2}, \dots, y_{im}$  is the same as that of  $y_{i\pi_i(1)}, \dots, y_{i\pi_i(m)}$  for any permutation  $\pi_i = (\pi_i(1), \dots, \pi_i(m))$  on  $\{1, 2, \dots, m\}$  for each  $i$ . Using this invariance property under  $H_0$ , the procedure can be set up by the following steps:

1. Select a random permutation  $\pi_i$  on  $\{1, 2, \dots, m\}$
2. Let  $y_{ij}^{\pi_i} = y_{i\pi_i(j)}$  and  $t_{ij}^{\pi_i} = t_{ij}$ , i.e., we do not permute the covariate.
3. Compute the test statistic  $\mathbf{T}_n$  with permuted responses.
4. Iterate the above steps  $B$  times and let the computed test statistic be  $\mathbf{T}_n^{(1)}, \dots, \mathbf{T}_n^{(B)}$ .
5. Finally, the critical value can be approximated by the quantile of the empirical distribution of  $\mathbf{T}_n$ .

### 3. SIZE OF TESTS

This section examines the unbiasedness of F-tests and the DIP test from simulation studies. We show that DIP is less biased than F-LT when the sample size  $n$  is small.

Here, we consider three types of error distributions, normal, uniform and double exponential, all with zero mean and unit variance. The random effects  $\alpha_i$ s are simulated from the normal distribution with mean 0 and variance  $\sigma_\alpha^2 = 1$ . The sample size ( $n$ ) and the number of repetitions ( $m$ ) are chosen as  $n = 5, 10, 30$  and  $50$ , and  $m = 5, 10$  and  $20$ . For each possible choice of  $(m, n)$ , we generate 1000

data sets from the model (1) with  $\mu(x) = 0$ . And for each data set, the size (*i.e.*, the p-value) of the DIP test is computed based on 400 permuted samples, the size of the F–LT is computed using the routine provided in R-archive [17, 18, 19], and the size of the F–NT is computed using the formula provided in [7].

Table I. Computed sizes of the F–tests and DIP test under the null hypothesis. The nominal sizes of tests are fixed as  $\alpha = 0.05$ .

$(m, n)$	Normal			Uniform			Double Exponential		
	F-LT	F-NT	DIP	F-LT	F-NT	DIP	F-LT	F-NT	DIP
(5,5)	0.178	0.043	0.053	0.129	0.053	0.090	0.168	0.050	0.053
(5,10)	0.079	0.049	0.075	0.090	0.062	0.074	0.094	0.048	0.078
(5,30)	0.068	0.064	0.068	0.074	0.051	0.075	0.070	0.035	0.017
(5,50)	0.062	0.042	0.051	0.058	0.042	0.070	0.057	0.044	0.065
(10,5)	0.149	0.054	0.049	0.149	0.055	0.073	0.165	0.042	0.050
(10,10)	0.111	0.040	0.056	0.114	0.061	0.057	0.122	0.042	0.063
(10,30)	0.027	0.046	0.060	0.062	0.063	0.058	0.061	0.042	0.057
(10,50)	0.055	0.042	0.077	0.061	0.060	0.067	0.057	0.051	0.063
(20,5)	0.195	0.060	0.060	0.214	0.048	0.093	0.178	0.049	0.050
(20,10)	0.116	0.054	0.050	0.124	0.054	0.057	0.117	0.036	0.059
(20,30)	0.073	0.047	0.062	0.073	0.060	0.056	0.062	0.040	0.047
(20,50)	0.071	0.042	0.044	0.065	0.054	0.049	0.059	0.063	0.061

Table 1 reports the computed sizes of the F–tests and the DIP test performed at the significance level  $\alpha = 0.05$  under each setting. We find that, even when the errors are from normal distributions, F–LT is much biased for small  $n$ , since its computed sizes are consistently larger than 0.05. In contrast, the sizes of F–NT and DIP are close to 0.05 in every case we consider. Similar results are also observed when tests are performed at  $\alpha = 0.01$  and 0.10.

#### 4. POWER OF THE DIP TEST

In this section we investigate the power of the DIP test under various situations through simulation: (a) normal random effects and normal errors; (b) normal random effects and non-normal errors; (c) non-normal random effects and normal errors; and (d) data with missing values. To do this, we consider two non-normal distributions, uniform and double exponential, for either random effects or errors. In each of the above cases, we generate data from two non-constant mean models; the first is LTMM (3) with slope  $\beta_j = 0.5/m$ , and the other is NTMM (2) with  $\beta_j = 0.5$  for  $j = m - 1, m$ , and 0 elsewhere. As in the previous section, we set  $m = (5, 10, 20)$ ,  $n = (5, 10, 30, 50)$ , and generate 1000 data sets for each setting. And for each data set, three tests (DIP, F–LT and F–NT) are employed to test the non-constant means, where the power of each test is computed in the same way as computing the corresponding p-value in Section 3.

#### 4.1. Normal random effects and errors

Table 2 summarizes the powers of the F-tests and the DIP test under the case of normal random effects and errors. Since F-LT is biased in size for small  $n$ , we also report the size-adjusted powers in parentheses for the cases of  $n = 5$  and 10, which are computed as follows. First, find an adjusted critical value  $c$  to make the size of F-LT equal that of DIP. Next, compute the adjusted power of F-LT based on this critical value. Although computing the size-adjusted powers appears to be ad-hoc, it can help us to make a fair comparison.

In this normal setup, one would expect that under LTMM, F-LT performs best in detecting non-constant means and under NTMM, F-NT performs best. Amazingly, Table 2 shows that the DIP test often outperforms both the F-tests when  $n$  is small, no matter which models the data are generated from. For large  $n$ , the DIP test performs reasonably well compared with F-LT (F-NT) for data from LTMM (NTMM); but it appears to be more powerful than F-NT (F-LT).

Table II. The case of normal random effects and normal errors: powers of the F-tests and DIP test under models with linear/nonlinear trend

$(m, n)$	LTMM			NTMM		
	F-LT	F-NT	DIP	F-LT	F-NT	DIP
(5,5)	0.064 (0.002)	0.065	0.092	0.124 (0.009)	0.109	0.170
(5,10)	0.128 (0.123)	0.099	0.150	0.249 (0.243)	0.199	0.307
(5,30)	0.361	0.229	0.348	0.681	0.643	0.783
(5,50)	0.579	0.398	0.534	0.888	0.867	0.935
(5,100)	0.871	0.719	0.822	0.992	0.998	0.999
(10,5)	0.124 (0.016)	0.072	0.107	0.102 (0.014)	0.126	0.105
(10,10)	0.261 (0.137)	0.123	0.205	0.193 (0.106)	0.175	0.207
(10,30)	0.658	0.337	0.582	0.606	0.631	0.681
(10,50)	0.882	0.570	0.808	0.841	0.902	0.941
(10,100)	0.996	0.909	0.981	0.990	0.997	1.000
(20,5)	0.233 (0.028)	0.087	0.180	0.089 (0.007)	0.106	0.060
(20,10)	0.476 (0.269)	0.159	0.381	0.139 (0.048)	0.173	0.095
(20,30)	0.940	0.520	0.869	0.420	0.535	0.409
(20,50)	0.993	0.801	0.983	0.661	0.820	0.745
(20,100)	1.000	0.990	1.000	0.922	0.996	0.990

#### 4.2. Non-normal errors

For the case of normal random effects but non-normal errors, Tables 3 and 4 report the computed powers for testing the linear trend in LTMM (3) and the nonlinear trend in NTMM (2), respectively. As in Section 3, the means and variances of the uniform and double exponential distributions are set to be 0 and 1, respectively.

In terms of overall comparisons, the relative performances of the three tests in the case of non-normal errors are similar to what we observe in the normal case. Usually, DIP does better than the F-tests for small  $n$ . For large  $n$ , it appears that F-LT works best under LTMM and F-NT works best under NTMM. This is reasonable since the asymptotic normality would finally take the effect when

$n \rightarrow +\infty$  under the correct model specification. However, the F-tests are sensitive to model misspecification, but DIP is not; the F-LT (F-NT) becomes less powerful to test NTMM (LTMM) while DIP often achieves quite close or even better performance to the better of the two F-tests.

Table III. The case of normal random effects and non-normal errors: powers of the F-tests and DIP test under models with linear trend

$(m, n)$	LTMM with non-normal errors					
	Uniform			Double Exponential		
	F-LT	F-NT	DIP	F-LT	F-NT	DIP
(5,5)	0.118 (0.050)	0.110	0.144	0.051 (0.020)	0.060	0.065
(5,10)	0.143 (0.122)	0.106	0.156	0.096 (0.083)	0.085	0.130
(5,30)	0.376	0.242	0.360	0.207	0.158	0.237
(5,50)	0.602	0.418	0.549	0.322	0.224	0.327
(5,100)	0.876	0.748	0.839	0.604	0.400	0.544
(10,5)	0.114 (0.024)	0.076	0.123	0.069 (0.010)	0.057	0.080
(10,10)	0.240 (0.110)	0.102	0.225	0.117 (0.053)	0.082	0.115
(10,30)	0.663	0.345	0.564	0.391	0.158	0.308
(10,50)	0.876	0.561	0.800	0.598	0.268	0.483
(10,100)	0.992	0.902	0.980	0.874	0.589	0.796
(20,5)	0.212 (0.010)	0.090	0.163	0.126 (0.008)	0.069	0.100
(20,10)	0.433 (0.253)	0.132	0.322	0.257 (0.124)	0.082	0.199
(20,30)	0.924	0.500	0.839	0.679	0.249	0.535
(20,50)	0.993	0.801	0.980	0.868	0.427	0.757
(20,100)	1.000	0.990	1.000	0.996	0.787	0.978

#### 4.3. Non-normal random effects

The results for the case of non-normal random effects and normal errors are reported in Tables 5 and 6. Again, DIP does consistently better than the F-tests for small  $n$ ; for large  $n$ , F-LT (F-NT) appears to be the best under LTMM (NTMM), but DIP provides much closer performance to that of F-LT (F-NT) than F-NT (F-LT).

#### 4.4. Missing observations

We suggest a simple modification of our DIP test for data with missing observations, and then implement a simulation study to investigate its performance. We assume missing observations occur at completely random [20]. It should be remarked that several adjusted F-tests have been proposed in the previous literature *e.g.*, [21, 22].

Let  $\delta_{ij}$  be 0 or 1 according to whether  $y_{ij}$  is missing or not, for  $i = 1, 2, \dots, n$ , and  $j = 1, 2, \dots, m$ . We suggest to compute the test statistics only with the observed data points so that

$$\mathbf{T}_n = \max_{k=1}^m \frac{1}{\sqrt{nm}} \left| \sum_{j=1}^k \sum_{i=1}^n \delta_{ij} (y_{ij} - \bar{y}_i) \right|,$$

Table IV. The case of normal random effects and non-normal errors: powers of the F-tests and DIP test under models with non-linear trend

$(m, n)$	NTMM with non-normal errors					
	Uniform			Double Exponential		
	F-LT	F-NT	DIP	F-LT	F-NT	DIP
(5,5)	0.117 (0.050)	0.109	0.143	0.075 (0.029)	0.069	0.103
(5,10)	0.254 (0.219)	0.231	0.333	0.154 (0.133)	0.126	0.193
(5,30)	0.715	0.634	0.799	0.405	0.342	0.485
(5,50)	0.910	0.897	0.963	0.625	0.569	0.708
(5,100)	0.997	1.000	1.000	0.929	0.883	0.954
(10,5)	0.122 (0.025)	0.108	0.116	0.070 (0.006)	0.084	0.074
(10,10)	0.235 (0.113)	0.203	0.236	0.120 (0.053)	0.119	0.119
(10,30)	0.627	0.649	0.700	0.350	0.319	0.364
(10,50)	0.848	0.896	0.929	0.584	0.531	0.618
(10,100)	0.988	0.996	0.999	0.862	0.886	0.934
(20,5)	0.083 (0.001)	0.085	0.074	0.055 (0.002)	0.070	0.053
(20,10)	0.140 (0.052)	0.150	0.082	0.095 (0.032)	0.089	0.055
(20,30)	0.449	0.547	0.397	0.220	0.249	0.174
(20,50)	0.657	0.838	0.740	0.388	0.444	0.338
(20,100)	0.936	0.996	0.991	0.678	0.824	0.740

where  $\bar{y}_{i.} = \sum_{j=1}^m \delta_{ij} y_{ij} / \sum_{j=1}^m \delta_{ij}$ . It is equivalent to imputing all missing observations of the  $i$ th subject with  $\bar{y}_{i.}$ . The permutation procedure introduced in Section 2 remains unchanged here.

Table 7 reports the results of the DIP test when we have 10% and 30% missing observations at randomly chosen time points for the case of normal random effects and errors only. The table can be read with Tables 1– 2. Easily to see, missing even a notable portion of observations would not affect either the size or power of the DIP test much.

## 5. EXAMPLE: CHEST COMPRESSION RATE

This section provides an empirical example of testing the trend over time using repeated measures data.

Continuous chest compression cardiopulmonary resuscitation (CPR), which consists of 15 compressions at a rate of 100 per minute, followed by 2 mouth-to-mouth rescue breaths repeated for 9 minutes, has been advocated as an alternative to standard CPR. Studies have shown that continuous chest compression CPR (CCC-CPR) delivers substantially more chest compressions per minute, and is easier to perform than standard CPR (STD-CPR). However, one concern regarding CCC-CPR is that, due to fatigue, the rescuer may not be able to maintain adequate compressions rate or depth throughout a realistic EMS response time. The aim of this section is to compare the effect of fatigue on the performance of CCC-CPR and STD-CPR in a mannequin model, using the data set described in the introduction.

Let  $x_{ij}^c$  and  $x_{ij}^s$  be the number of adequate compressions of the  $i$ th subject at time  $j$  for CCC-CPR and STD-CPR, respectively. To achieve our goal, we test the constancy of the mean of  $y_{ij} = x_{ij}^c - x_{ij}^s$ .

Table V. The case of non-normal random effects and normal errors: powers of the F-tests and DIP test under models with linear trend

$(m, n)$	LTMM with non-normal random effects					
	Uniform			Double Exponential		
	F-LT	F-NT	DIP	F-LT	F-NT	DIP
(5,5)	0.081 (0.050)	0.090	0.106	0.070 (0.035)	0.076	0.090
(5,10)	0.145 (0.121)	0.116	0.156	0.133 (0.114)	0.084	0.136
(5,30)	0.385	0.244	0.359	0.378	0.223	0.345
(5,50)	0.585	0.399	0.547	0.576	0.365	0.517
(5,100)	0.873	0.715	0.835	0.885	0.678	0.821
(10,5)	0.123 (0.020)	0.077	0.119	0.132 (0.020)	0.082	0.123
(10,10)	0.248 (0.116)	0.116	0.202	0.256 (0.131)	0.127	0.231
(10,30)	0.676	0.308	0.558	0.687	0.346	0.586
(10,50)	0.883	0.555	0.786	0.898	0.584	0.800
(10,100)	0.997	0.900	0.979	0.993	0.919	0.987
(20,5)	0.223 (0.020)	0.099	0.198	0.235 (0.019)	0.095	0.211
(20,10)	0.447 (0.279)	0.176	0.374	0.463 (0.289)	0.155	0.382
(20,30)	0.921	0.496	0.848	0.919	0.502	0.860
(20,50)	0.997	0.769	0.974	0.991	0.799	0.976
(20,100)	1.000	0.996	1.000	1.000	0.970	0.999

over time  $j$ , as pointed out in the introduction.

Table VIII shows the mean numbers (standard deviations) of adequate compressions for each group over time and p-values computed from the two-sample t-tests. The mean number of adequate compressions is significantly greater for CCC-CPR than that for STD-CPR during the first 2 minutes. From the third minute of CPR, the difference is no longer significant and in the 9th minute the absolute number of effective compressions is higher in the STD-CPR group though this difference is not statistically significant. The number of adequate compressions per minute declines over time in both the CCC-CPR and STD-CPR groups but it declines more rapidly in CCC-CPR than STD-CPR. Figure 2 shows the difference in adequate compressions between CC-CPR and STD-CPR at each minute for each subject.

Table IX reports the result from the DIP test, as well as those from the F-tests based on the LTMM and NTMM. We can see that, F-NT does not show any statistical significance on the difference in slope, while F-LT just stands on the margin. The residual normal plots in Figure 3 from fitting the LTMM and NTMM reveal strong non-normality of the data. Thus, the suggested DIP could be a good alternative, which shows a significant difference in the slopes between those two groups.

## 6. CONCLUSION

In this paper, we have proposed the DIP test to test the constancy of the mean function for repeated measures data, which is virtually distribution free. The asymptotic reference distribution of the proposed test statistic has been analytically established. A permutation procedure has been described

Table VI. The case of non-normal random effects and normal errors: powers of the F-tests and DIP test under models with non-linear trend

$(m, n)$	NTMM with non-normal random effects					
	Uniform			Double Exponential		
	F-LT	F-NT	DIP	F-LT	F-NT	DIP
(5,5)	0.087 (0.040)	0.081	0.096	0.078 (0.026)	0.090	0.104
(5,10)	0.111 (0.092)	0.153	0.159	0.094 (0.079)	0.138	0.156
(5,30)	0.226	0.375	0.361	0.215	0.369	0.346
(5,50)	0.400	0.585	0.535	0.368	0.584	0.521
(5,100)	0.713	0.875	0.828	0.701	0.876	0.821
(10,5)	0.076 (0.013)	0.122	0.114	0.097 (0.020)	0.109	0.118
(10,10)	0.124 (0.046)	0.260	0.202	0.144 (0.061)	0.230	0.216
(10,30)	0.347	0.691	0.570	0.328	0.648	0.535
(10,50)	0.595	0.883	0.812	0.547	0.869	0.788
(10,100)	0.900	0.990	0.976	0.902	0.994	0.977
(20,5)	0.079 (0.001)	0.254	0.194	0.095 (0.002)	0.232	0.182
(20,10)	0.163 (0.077)	0.476	0.385	0.152 (0.078)	0.438	0.345
(20,30)	0.949	0.518	0.869	0.522	0.937	0.847
(20,50)	0.994	0.809	0.983	0.795	0.996	0.979
(20,100)	1.000	0.995	1.000	0.990	1.000	1.000

Table VII. The sizes and powers of the DIP test for data with missing observations

$(m, n)$	Size		Power			
	$\beta = 0$		LTMM		NTMM	
	10%	30 %	10%	30%	10%	30%
(5,5)	0.054	0.050	0.085	0.083	0.138	0.108
(5,10)	0.061	0.069	0.124	0.108	0.264	0.179
(5,30)	0.074	0.072	0.275	0.200	0.681	0.491
(5,50)	0.060	0.063	0.444	0.315	0.886	0.721
(5,100)	0.060	0.059	0.735	0.551	0.997	0.960
(10,5)	0.040	0.044	0.103	0.078	0.085	0.070
(10,10)	0.040	0.049	0.172	0.125	0.171	0.117
(10,30)	0.042	0.049	0.476	0.330	0.563	0.369
(10,50)	0.049	0.060	0.708	0.518	0.854	0.642
(10,100)	0.046	0.053	0.963	0.825	0.998	0.956
(20,5)	0.043	0.054	0.129	0.116	0.065	0.069
(20,10)	0.048	0.040	0.316	0.213	0.090	0.083
(20,30)	0.047	0.048	0.770	0.573	0.327	0.199
(20,50)	0.039	0.044	0.958	0.831	0.602	0.377
(20,100)	0.052	0.069	0.999	0.981	0.969	0.799

Table VIII. Mean numbers (standard deviations) of adequate compressions for CCC-CPR and STD-CPR by minute. The p-values are computed from the two-sample t-tests.

Minute	STD-CPR	CCC-CPR	P-value
1	33.0 (3.6)	48.0 (7.0)	0.0066
2	29.6 (3.6)	40.9 (7.1)	0.0402
3	28.8 (3.7)	37.8 (6.9)	0.0813
4	25.7 (3.4)	31.3 (6.4)	0.4831
5	26.3 (3.6)	29.9 (6.4)	0.6671
6	24.8 (3.6)	29.0 (6.3)	0.5849
7	23.8 (3.6)	29.2 (6.5)	0.4176
8	24.8 (3.6)	29.9 (6.6)	0.5461
9	25.3 (3.7)	24.6 (6.0)	0.8186

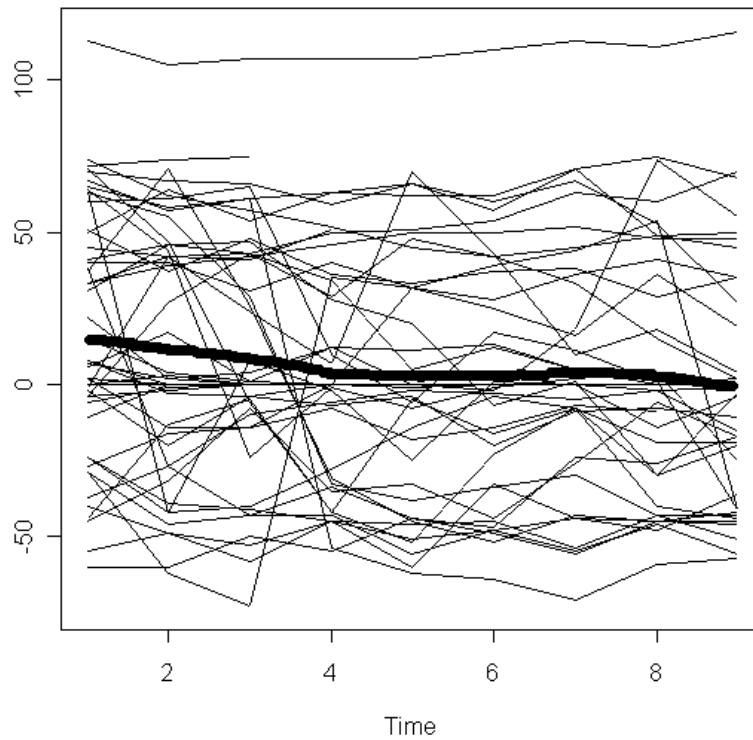


Figure 2. An empirical example: trajectories of the difference of the adequate compressions between CCC-CPR and STD-CPR. The bold line is the mean curve.

Table IX. The p-values of testing the slope in the chest compression data

	F-LT	F-NT	DIP
P-value	0.0493	0.1575	0.0000

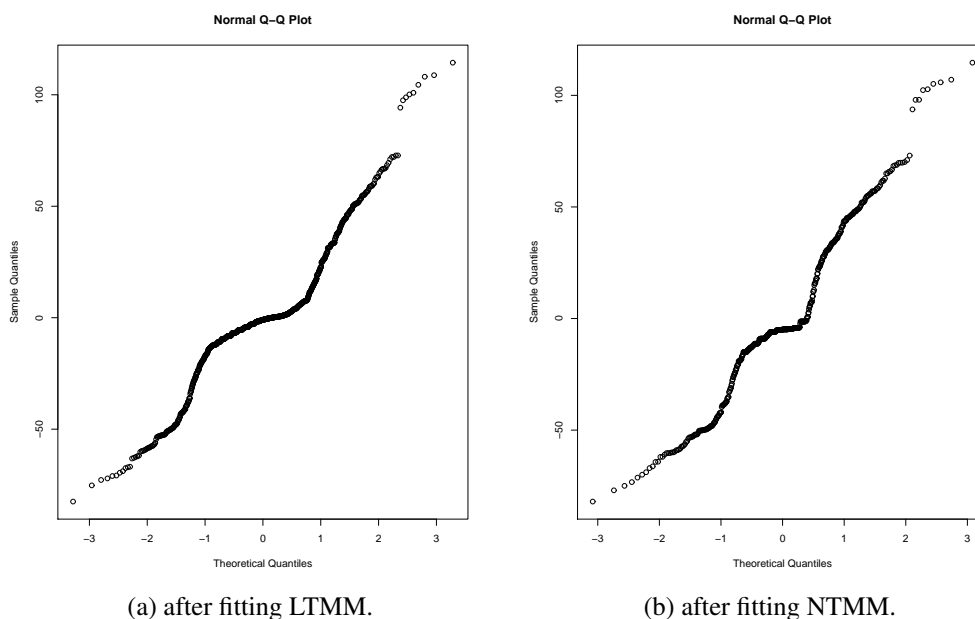


Figure 3. An empirical example: QQ-plots after fitting the linear mixed effects models

to approximate the finite-sample reference distribution.

The proposed test was illustrated with several simulation studies and a data example. Our experience in simulation suggests the DIP test performs reasonably well, especially for data of small sample size, regardless of the underlying model. In contrast, the F-tests are often sensitive to model misspecification; the F-LT (F-NT) becomes less powerful for testing the model NTMM (LTMM). We conjecture that the benefit of the DIP test is due to the proposed test statistic itself, rather than the use of bootstrapping to get the p-value. This is based on our limited numerical evidence (not reported here) that permutation-based F-tests (i.e., calculating the F-statistic, and then bootstrapping to get the p-value) performed worse than the proposed DIP test.

We conclude the paper with a discussion of unequal measurement times over subjects (i.e.,  $t_{ij} \neq t_j$ ). Although all the simulations in this paper are based on common measurement times, the proposed DIP test does not assume  $t_{ij} = t_j$  for each  $j$ . Additional simulation not reported in the paper shows that DIP is powerful enough to detect a very small departure from the constant function from data with unequal measurement times.

## APPENDIX

We derive the asymptotic distribution of testing statistic  $\mathbf{T}_n$  under  $H_0$ , constant mean.

**Theorem 1.** *Suppose the observations are from the mixed effects model (1) where the mean function  $\mu(t)$  is constant over  $t$ ,  $\alpha_i$  is a vector of random effects with mean 0 and finite covariance matrix, and  $\epsilon_{ij}$  is independently and identically distributed with zero mean and finite variance, say  $\sigma_\epsilon^2$ . Then, when  $n$  increases but  $m$  is fixed,  $\mathbf{T}_n$  converges in distribution to  $\sigma_\epsilon \max_{k=1}^m |B_0(k/m)|$ , where  $B_0(t)$  is the Brownian bridge on  $[0, 1]$ . On the other hand, when  $m$  increases but  $n$  is fixed,  $\mathbf{T}_n$  converges in distribution to  $\sigma_\epsilon \sup_{t \in [0,1]} |B_0(t)|$  whose density function is known as*

$$f(x) = \frac{2}{\sqrt{2\pi}} \exp(-x^2/2), \quad \text{for } x > 0.$$

**Proof 1.** *Both results in the above theorem are from the functional central limit theorem and the continuous mapping theorem. Since their proofs are similar, we only present the one in the case when  $n$  increases for a fixed  $m$ .*

*Under the constancy of the mean function  $\mu(x)$ , the joint distribution of  $y_{ij} - \bar{y}_i$ 's is the same as that of  $\epsilon_{ij} - \bar{\epsilon}_i$ , with  $\bar{\epsilon}_i$  defined similarly to  $\bar{y}_i$ . Then from the functional central limit theorem, the joint distribution of*

$$\frac{1}{\sqrt{m}n} \left[ \sum_{j=1}^1 \sum_{i=1}^n \epsilon_{ij}, \dots, \sum_{j=1}^m \sum_{i=1}^n \epsilon_{ij} \right] \quad (5)$$

*converges to  $[B(1/m), \dots, B(m/m)]$  where  $\{B(t), t \in [0, 1]\}$  is the Brownian motion on  $[0, 1]$ . Thus, the partial sum process*

$$\begin{aligned} & \frac{1}{\sqrt{nm}} \left[ \sum_{j=1}^1 \sum_{i=1}^n (\epsilon_{ij} - \bar{\epsilon}_i), \dots, \sum_{j=1}^m \sum_{i=1}^n (\epsilon_{ij} - \bar{\epsilon}_i) \right] \\ &= \frac{1}{\sqrt{nm}} \left[ \sum_{j=1}^1 \sum_{i=1}^n \epsilon_{ij}, \dots, \sum_{j=1}^m \sum_{i=1}^n \epsilon_{ij} \right] - \frac{1}{\sqrt{nm}} \sum_{j=1}^m \sum_{i=1}^n \epsilon_{ij} \cdot \left[ \frac{1}{m}, \dots, \frac{m}{m} \right] \end{aligned}$$

*converges to  $[B_0(1/m), \dots, B_0(m/m)]$  where  $B_0(t) = B(t) - t \cdot B(1)$  is the Brownian bridge,  $t \in [0, 1]$ . Next, from the continuity of the function  $g(x_1, x_2, \dots, x_m) = \max_{k=1}^m |x_m|$ , it can be shown that*

$$\max_{k=1}^m \frac{1}{\sqrt{nm}} \left| \sum_{j=1}^k \sum_{i=1}^n (y_{ij} - \bar{y}_i) \right|$$

*converges to  $\max_{j=1}^k |B_0(k/m)|$  as  $n$  increases for every fixed  $m$ . The density function of the limiting distribution is from p 96 in [23].*

## ACKNOWLEDGEMENTS

We are grateful to Dr. Hasan Rajab at Scott and White Hospital for providing the chest compression data and many suggestions.

## REFERENCES

1. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. Springer, 2000.
2. Demidenko, E. *Mixed Models: Theory and Applications*. Wiley, 2004.
3. Herbach LH. Properties of model-II type analysis of variance tests, A: Optimum nature of the F-test in the balanced case. *Annals of Mathematical Statistics* 1959; **30**:939–959.
4. Spjøtvoll E. Optimum invariant tests in unbalanced variance components models. *Annals of Mathematical Statistics* 1967; **38**:422–428.
5. Seifert B. Explicit formulas of exact tests in mixed balanced ANOVA models. *Biometrical Journal* 1981; **23**:535–550.
6. Mathew T, Sinha BK. Optimum tests for fixed effects and variance components in balanced models. *Journal of American Statistical Association* 1988; **83**:133–135.
7. Montgomery D. *Designs and Analysis of Experiments*. John Wiley & Sons: New York, 1997.
8. Rice JA, Silverman BW. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. B.* 1991; **53**:233–243.
9. Wang Y, Taylor JMG. Inference for smoothing curves in longitudinal data with application to an AIDS clinical trial. *Statistics in Medicine* 1995; **14**:1205–1218. DOI:10.1002/sim.4780141106.
10. Wang Y. Mixed-effects smoothing spline ANOVA. *Journal of the Royal Statistical Society. B.* 1998; **60**:159–174.
11. Hart JD, Wehrly TE. Kernel regression estimation using repeated measurements data. *Journal of American Statistical Association* 1986; **81**:1080–1088.
12. Zeger SL, Diggle PJ. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* 1994; **50**:689–699.
13. Verbeke G, Lesaffre E. A linear mixed-effects model with heterogeneity in the random effects population. *Journal of American Statistical Association* 1996; **91**:217–221.
14. Verbeke G, Lesaffre E. The effects of misspecifying the random effects distributions in linear mixed effects model for longitudinal data. *Computational Statistics and Data Analysis* 1997; **23**:541–556.
15. Gurka MJ, Edwards LJ, Muller KE, Kupper LL. Extending the Box-Cox transformation to the linear mixed effects model. *Journal of the Royal Statistical Society. A.* 2006; **169**:273–288.
16. Hartiganm JA, Hartigan PM. The Dip test for unimodality. *Annals of Statistics* 1985; **13**:70–84.
17. Lindstrom MJ, Bates DM. Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. *Journal of the American Statistical Association* 1988; **83**:1014–1022.
18. Lindstrom MJ, Bates DM. Nonlinear mixed effects models for repeated measures data. *Biometrics* 1990; **46**:673–687.
19. Pinheiro JC, Bates DM. *Mixed effects models in S and S-Plus*. Springer, 1996.
20. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley & Sons: New York, 1987.
21. Barton CN, Cramer EC. Hypothesis testing in multivariate linear mixed effects model with randomly missing data. *Communication in Statistics-Simulations* 1989; **18**:875–895.
22. Catellier DJ, Muller KE. Tests for Gaussian repeated measures with missing data in small samples. *Statistics in Medicine* 2000; **19**:1101–1114. DOI:10.1002/(SICI)1097-0258(20000430)19:8<1101::AID-SIM415>3.0.CO;2-H.
23. Karatzas I, Shreve SE. *Brownian Motion and Stochastic Calculus*. Springer: New York, 1991.