Received: 2 November 2019 | Revised: 19 November 2020 | Accepted: 23 November 2020

DOI: 10.1002/sim.8850

RESEARCH ARTICLE

Statistics in Medicine WILEY

Longitudinal multivariate normative comparisons

Zheng Wang[1] | **Yu Cheng**[1,2] | **Eric C. Seaberg**[3] | **Leah H. Rubin**[3,4] | **Andrew J. Levine**[5] | **James T. Becker**[6] | **Neuropsychology Working Group of the MACS**[6]

[1]Department of Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania

[2]Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania

[3]Department of Epidemiology, Johns Hopkins University, Baltimore, Maryland

[4]Department of Neurology, Johns Hopkins University, Baltimore, Maryland

[5]Department of Neurology, University of California Los Angeles, Los Angeles, California

[6]Departments of Psychiatry, Neurology, and Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania

Correspondence
Yu Cheng, Department of Statistics, University of Pittsburgh, 230 S Bouquet St., Pittsburgh, PA 15260, USA.
Email: yucheng@pitt.edu

Motivated by the Multicenter AIDS Cohort Study (MACS), we develop classification procedures for cognitive impairment based on longitudinal measures. To control family-wise error, we adapt the cross-sectional multivariate normative comparisons (MNC) method to the longitudinal setting. The cross-sectional MNC was proposed to control family-wise error by measuring the distance between multiple domain scores of a participant and the norms of healthy controls and specifically accounting for intercorrelations among all domain scores. However, in a longitudinal setting where domain scores are recorded multiple times, applying the cross-sectional MNC at each visit will still have inflated family-wise error rate due to multiple testing over repeated visits. Thus, we propose longitudinal MNC procedures that are constructed based on multivariate mixed effects models. A χ^2 test procedure is adapted from the cross-sectional MNC to classify impairment on longitudinal multivariate normal data. Meanwhile, a permutation procedure is proposed to handle skewed data. Through simulations we show that our methods can effectively control family-wise error at a predetermined level. A dataset from a neuropsychological substudy of the MACS is used to illustrate the applications of our proposed classification procedures.

KEYWORDS

cognitive impairment, false discovery rate, family-wise error rate, longitudinal analysis, multivariate mixed-effect model

1 | INTRODUCTION

Classification plays an important role in many fields of medical science. For example, identifying participants with cognitive impairment will enable clinicians to provide patients with proper treatments. As true cognitive impairment status is typically unknown, researchers often identify a group of healthy controls and measure their cognitive functioning over multiple domains to understand how performance is distributed in the healthy population. If a participant to be tested performs far below a typical healthy control, he or she is deemed to have abnormal scores/impairment, though further diagnostic tests are often carried out in clinical settings. Several methods of counting the number of domains with

Members of the Neuropsychology Working Group of the Multicenter AIDS Cohort Study include: Ned Sacktor, MD, Leah H. Rubin, Ph.D., Eric C. Seaberg, Dr.P.H., James T. Becker, Ph.D., Andrea Weinstein, Ph.D., Ann B. Ragin, Ph.D., Eileen Martin, Ph.D., Andrew J. Levine, Ph.D., Pim Browers, Ph.D., Lisa Jacobson, Dr.P.H., Cynthia Munro, Ph.D., Eithne Keelaghan, Robin Huebner, Ph.D., Carlie Williams, Ph.D.

abnormal scores[1,2] have been used in the fields of HIV and Alzheimer's Disease, despite evidence that these methods are associated with inflated family-wise error rates (FWERs). The FWER here refers to the probability of making a false cognitive impairment classification, given that all cognitive domains of a participant function normally. In order to control the FWER at a predetermined level and correct for intercorrelations among multiple cognitive domains, Huizenga et al[3] developed the so-called multivariate normative comparison (MNC) method which specifically takes the covariance of the domain scores into consideration. Let $\boldsymbol{X}_i$ denote a vector of q cognitive domain scores for participant i. The healthy control group contains n participants, and their sample mean and sample covariance matrix of the q domain scores are denoted as $\hat{\boldsymbol{\mu}}_{\boldsymbol{c}}$ and $\hat{\boldsymbol{\Psi}}_{\boldsymbol{c}}$. If each vector of q domain scores is independent and identically distributed over a multivariate normal distribution for every participant, one could build an F-statistic to classify cognitive impairment for the ith individual:

$$\frac{n(n-q)}{(n+1)(n-1)q}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_c)^T \hat{\boldsymbol{\Psi}}_{\boldsymbol{c}}^{-1}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_c) \sim F(q, n-q).$$

In principle, the MNC method can effectively control the FWER in impairment classification as long as all domain variables follow a multivariate normal distribution.[4,5] In practice, participants may visit the same clinician or institution multiple times. For example, if participants come to an Alzheimer Disease Research Center with memory complaints, they will be followed roughly annually and their cognitive functioning will be assessed repeatedly over time. In a retrospective analysis, these longitudinal scores, which are used to identify prior cognitive impairment, can provide important guidance on future treatments or help researchers identify risk factors. If the MNC is employed at each visit and a participant is repeatedly tested at a prespecified α level, the resulting FWER, which is the probability of being categorized as having prior impairment at some visits, would be greatly inflated, because the MNC fails to account for multiple testing over repeated visits. Thus, here we propose longitudinal MNC procedures that specifically take into account multiple tests over repeated measures to quantify an individual's prior impairment.

The initial step, similar to the cross-sectional MNC, is to obtain characteristics, such as mean scores over time and the covariance structure of the longitudinal measurements, from healthy controls. One typical way to approach longitudinal data is to utilize a multivariate linear mixed effects (MLME) model. Reinsel[6] established theories for multivariate longitudinal models with repeated measures when data are balanced and parameters are unrestricted. Heitjan and Sharma[7] further considered an autoregressive error structure for longitudinal data and estimated the parameters with the maximum likelihood approach. Fang et al[8] introduced a modified expectation-maximization algorithm to facilitate the estimation of unknown parameters in an MLME model with constrained intercepts. Fieuws and Verbeke[9] studied how the associations between different responses evolve over time and jointly modeled two responses by allowing a dependence structure among the random terms in the model. They further proposed modeling longitudinal outcomes in a pairwise fashion for computation efficiency when too many outcome variables are considered.[10] Verbeke et al[11] gave a rather comprehensive review of development in multivariate longitudinal analysis, and noted that joint modeling is preferred over univariate modeling to address research questions regarding associations among various outcomes over time. van den Hout et al[12] proposed a longitudinal MLME model with change-point predictors for nonlinear trends.

Here, we initially assume a multivariate normal distribution for longitudinal domain scores, and use the MLME to obtain the mean function and covariance structure of domain scores from healthy controls. As the true impairment status of an individual is generally unknown, such characteristics from healthy controls can provide a benchmark to decide if an individual's repeatedly measured domain scores are abnormally low. Similar to the way that the cross-sectional MNC tests all scores together, the proposed extended longitudinal multivariate normative comparison (LMNC) is developed to test all scores across visits simultaneously. Under multivariate normality, a testing procedure based on χ^2 is then proposed to classify cognitive status for each participant. However, if the dependency structure is not sufficiently specified or the data fail to follow a multivariate normal distribution, the χ^2 procedure may still have an inflated FWER. Therefore, we propose a permutation test for our proposed test statistic which is robust against distribution assumptions.

The structure of the remaining article is as follows. First, we detail modeling and testing procedures in Section 2. Next, we present results from simulation studies when the multivariate normal distribution is satisfied and when the assumption is not satisfied (Section 3). Third, we illustrate in Section 4 how to implement the MLME and the χ^2 and permutation tests for neuropsychological (NP) data collected in the Multicenter AIDS Cohort Study (MACS). Finally, we conclude by discussing some advantages and disadvantages of the MNC method compared with prior methods.

2 | LONGITUDINAL MULTIVARIATE NORMATIVE COMPARISONS

2.1 | Testing procedure based on χ^2

Assume there are n participants enrolled in a healthy group which is used as the reference, and each participant has q cognitive domains tested over m_i total visits during the study. Domain test scores are usually normalized so that a multivariate normal distribution holds for each visit. Let $Y_{ijk}, i = 1, \ldots, n; j = 1, \ldots, q; k = 1, \ldots, m_i$ denote the tested score of participant i for domain j over the kth visit. Considering that scores of a single domain assessed across m_i visits are correlated with each other, and scores of two different domains from the same participant are correlated, we model Y_{ijk} using an MLME model:

$$Y_{ijk} = \beta_{j0} + \beta_{j1} t_{ik} + \beta_{j2} t_{ik}^2 + \beta_{j3} t_{ik}^3 + \nu_{ij} + \delta_{ik} + \epsilon_{ijk}. \tag{1}$$

Here, we use q polynomial functions of degree 3 to describe the changes in the mean domain scores over time, and can add higher order terms if necessary. Alternatively, the B-spline technique can be used to approximate the true mean domain scores over time.[13-17] ϵ_{ijk} is assumed to be independent and identically distributed (i.i.d.) normal $N(0, \sigma^2)$, which is specific to each observation or measurement. Similarly, δ_{ik}, which represents the visit-specific effect, is also assumed to be i.i.d. normal $N(0, \theta^2)$. Given different domain functions tend to be correlated with each other for the same participant, $\boldsymbol{\nu}_i = (\nu_{i1}, .., \nu_{iq})^\intercal$ is assumed to be $N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = [\rho_{sr}], s, r = 1, \ldots, q$. Generally, the symmetric matrix $\boldsymbol{\Sigma}$ could be left unspecified, or assumed to have the structure of autoregression or compound symmetry.

All unknown parameters can be estimated from an MLME model,[8,9] which are denoted as $\hat{\beta}_{j0}, \hat{\beta}_{j1}, \hat{\beta}_{j2}, \hat{\beta}_{j3}, j = 1, \ldots, q, \hat{\rho}_{sr}, s, r = 1, \ldots, q, \hat{\theta}^2$, and $\hat{\sigma}^2$. For participant d to be tested, we take all q domain scores observed over m_d visits, and stack them into a single vector

$$\boldsymbol{U}_d = (Y_{d11}, \ldots, Y_{dq1}, Y_{d12}, \ldots, Y_{dq2}, \ldots, Y_{d1m_d}, \ldots, Y_{dqm_d})^\intercal. \tag{2}$$

From the linear mixed effects model in (1), the estimated mean vector of $\boldsymbol{U}_d$ is written as $\hat{\boldsymbol{\mu}}_d = (\hat{\beta}_{10} + \hat{\beta}_{11} t_{d1} + \hat{\beta}_{12} t_{d1}^2 + \hat{\beta}_{13} t_{d1}^3, \hat{\beta}_{20} + \hat{\beta}_{21} t_{d1} + \hat{\beta}_{22} t_{d1}^2 + \hat{\beta}_{23} t_{d1}^3, \ldots, \hat{\beta}_{q0} + \hat{\beta}_{q1} t_{d1} + \hat{\beta}_{q2} t_{d1}^2 + \hat{\beta}_{q3} t_{d1}^3, \ldots, \hat{\beta}_{10} + \hat{\beta}_{11} t_{dm_d} + \hat{\beta}_{12} t_{dm_d}^2 + \hat{\beta}_{13} t_{dm_d}^3, \ldots, \hat{\beta}_{q0} + \hat{\beta}_{q1} t_{dm_d} + \hat{\beta}_{q2} t_{dm_d}^2 + \hat{\beta}_{q3} t_{dm_d}^3)^\intercal$, which is of length qm_d. Furthermore, based on the covariance matrix structured in this model, we can estimate the covariance matrix for $\boldsymbol{U}_d$ as $\hat{\boldsymbol{\Psi}}_d = [\tau_{sr}], s, r = 1, \ldots, qm_d$. Each element in $\boldsymbol{\Psi}_d$ corresponds to the covariance between a pair $Y_{dj_1k_1}$ and $Y_{dj_2k_2}$, which can be estimated as $\hat{\rho}_{j_1j_2} + \hat{\theta}^2 \mathbb{I}\{k_1 = k_2\} + \hat{\sigma}^2 \mathbb{I}\{j_1 = j_2, k_1 = k_2\}$, with domain indexes $1 \le j_1, j_2 \le q$, visit indexes $1 \le k_1, k_2 \le m_d$, and $\mathbb{I}\{\cdot\}$ being an indicator function.

Under the assumption of multivariate normal distribution for all observations measured over time, we now propose an extended LMNC statistic for testing whether the dth participant has impaired cognition:

$$T_d = (\boldsymbol{U}_d - \hat{\boldsymbol{\mu}}_d)^\intercal \hat{\boldsymbol{\Psi}}_d^{-1} (\boldsymbol{U}_d - \hat{\boldsymbol{\mu}}_d) \sim \chi^2_{qm_d}, \tag{3}$$

which can be modified to an F test when the number of participants is small in the healthy control group. For participant d, if we are concerned that this participant's performance is either too high or too low, we will use $(1 - \alpha)$ quantile of $\chi^2_{qm_d}$ as the threshold for the significance level α. In practice, clinicians are typically more interested in screening for cognitive impairment with extremely low scores. One can conduct a statistical test considering the direction of domain scores by rejecting the null hypothesis if participant d's measured distance T_d exceeds the $(1 - 2\alpha)$ quantile of $\chi^2_{qm_d}$ and $\boldsymbol{U}_d' \mathbf{1}_{qm_d} < \hat{\boldsymbol{\mu}}_d' \mathbf{1}_{qm_d}$, where $\mathbf{1}_{qm_d}$ is the qm_d-vector of ones.

2.2 | Permutation testing

In practice, multivariate normality may not hold for the recorded measurements, and the test statistic in (3) might not follow an χ^2 distribution. In such case, the T_d statistic in (3) can still serve as a distance measure of individual scores to the norm. However, we need to develop a new method to find the critical value for the test statistic without relying on a particular parametric distribution. We propose the innovative use of a permutation test to find such a critical value for each participant. In order that a test statistic from a permuted sample is comparable to the one from the original

data, the permutation should retain the covariance structure of $\boldsymbol{\nu}_i$. For example, the covariance structure in Model (1), $\boldsymbol{\Sigma} = [\rho_{sr}], s, r = 1, \ldots, q$, is set to be compound symmetric, where $\rho_{ss} = \rho_{rr}$ for $s, r = 1, \ldots, q$, and $\rho_{sr} = \rho_{ut}$ for $s, r, u, t = 1, \ldots, q$ and $s \neq r, u \neq t$. The compound symmetry is a reasonable covariance structure when all cognitive domain scores in the reference group have been standardized and their errors can be assumed to follow an identical distribution. Meanwhile, as the test statistic depends on the number of total visits m_d completed by the dth participant, the permutation test should be done in a way specific to m_d.

Suppose there are M distinct number of visits in the testing group. We take M bootstrap samples, one for each unique number of visits. The following procedure details how permutation tests should be done for all of the participants in the testing group who have m total number of visits. We first take a bootstrap sample of the desired number N of participants with replacement (say 5000) from the healthy control group. Then, we remove the time effect (ie, $\hat{\beta}_{j0} + \hat{\beta}_{j1} t_{ik} + \hat{\beta}_{j2} t_{ik}^2 + \hat{\beta}_{j3} t_{ik}^3$ from model (1)) to obtain participant-specific errors over time for participant i from the bootstrap sample, $1 \leq i \leq N$. Next, to carry out the permutation test for each participant in the bootstrap sample, we consider errors of each domain function across all visits as a whole column. As a result, the multivariate longitudinal measures can be organized into a matrix of q-domain columns and m_i-visit rows. Then, we permute these q columns within the same participant so that this compound symmetric covariance structure will be sustained after each permutation.

For each participant i in the bootstrap sample, we then sample m visits with replacement to represent the bootstrapped sample errors with the number of visits matching with that of those participants to be tested. The bootstrapped sample errors from the m visits can be stacked in a similar way as in Equation (2) to a vector $\boldsymbol{V}_i = (E_{i11}, \ldots, E_{iq1}, E_{i12}, \ldots, E_{iq2}, \ldots, E_{i1m}, \ldots, E_{iqm})^\intercal$. Then, the bootstrap test statistic is calculated for the rearranged error sample from participant i as $T_i = (\boldsymbol{V}_i)^\intercal \hat{\boldsymbol{\Phi}}_{m_i}^{-1} \boldsymbol{V}_i$. However, the covariance structure $\boldsymbol{\Phi}_{m_i}$ used here is not the same as $\boldsymbol{\Psi}$ from Equation (3), given that we draw errors with replacement for m times at the visit level within participant i. $\boldsymbol{\Phi}_{m_i}$ is a $mq \times mq$ matrix. For domain indexes $1 \leq j_1, j_2 \leq q$ and visit indexes $1 \leq k_1, k_2 \leq m$, its element can be estimated as $\hat{\rho}_{j_1 j_2} + \hat{\theta}^2(\mathbb{I}\{k_1 = k_2\} + m_i^{-1}\mathbb{I}\{k_1 \neq k_2\}) + \hat{\sigma}^2(\mathbb{I}\{j_1 = j_2, k_1 = k_2\} + m_i^{-1}\mathbb{I}\{j_1 = j_2, k_1 \neq k_2\})$, where $\mathbb{I}$ is an indicator function. This covariance matrix $\boldsymbol{\Phi}_{m_i}$ cannot be inverted when $m > 1$ and the participant that was bootstrapped has only one visit ($m_i = 1$). As a result, we will exclude participants with only one visit from the healthy control (reference) group when the permutation test is administered after longitudinal modeling. It is worth noting that for a testing participant with a specific number of visits, we can use all participants in the control group to create permutation samples, as long as those individuals have at least two visits. In subsequent sections, we demonstrate in our simulations that the permutation test performs well when the number of remaining participants is 100 or above after excluding those individuals with only one visit. This exclusion seems to have a minimal impact on the MACS sample that we use. With a sufficient number of permutation tests conducted, the $(1 - \alpha)$ quantile specific to m visits can be found among all $T_i \mathbb{I}\{(\boldsymbol{V}_i)^\intercal \mathbf{1}_{qm} < 0\}, i = 1, \ldots, N$, to serve as the critical value. Thus, we relax the assumptions that the test statistic follows a χ^2 distribution and that the upper tails and the lower tails of the domain scores are symmetric. Participant d with total $m_d = m$ visits will be classified as cognitively impaired if their test statistic exceeds this critical value and $\boldsymbol{U}_d^\intercal \mathbf{1}_{qm} < \hat{\boldsymbol{\mu}}^\intercal \mathbf{1}_{qm}$.

3 | SIMULATION ANALYSIS

We ran a series of simulation studies to evaluate the performance of the proposed procedures. Given that the MACS data analyses in Section 4 involve six cognitive domains, we also considered $q = 6$ hypothetical domains in the simulation studies. We first generated longitudinal multivariate data following the multivariate normal distribution with several forms of polynomial mean functions over time. The testing procedure based on χ^2 was evaluated by FWER over different levels of α. Then, we considered data that do not follow multivariate normality to evaluate the performance of the proposed permutation test. Two forms of data were examined. The first form was generated from multivariate t distributions with symmetric but heavier tails than normal distributions. The second form was generated by transforming Gamma distributions to achieve negative skewness.

We carried out 1000 simulations for each scenario. For each simulation, we generated longitudinal scores for 1000 participants supposedly from the healthy control group, and generated longitudinal scores for another 1000 participants independently as the test group. For each participant, we simulated survival time from an exponential distribution with mean 30 years and censored at 15 years. Since participants in the MACS were tested semiannually (around 0.5 year

between any consecutive two visits) or biannually on their cognitive performance,[18,19] the time between any consecutive two visits was assumed to follow independent and uniform (0,1) distribution with the first visit at time 0. We continued to simulate visits until the accumulated visit times exceeded the censored survival time for the ith participant. The number of visits at the last visit before the boundary was recorded as m_i.

In practice, one might be interested in determining whether cognitive functions are significantly better in one group compared with another. Thus, we also examined and compared various testing groups with different visit frequencies and mean functions under alternatives and under the null. Finally, we studied the performance of the proposed tests under various sample sizes and percentage of subjects with only one visit. Detailed simulation specification and results are described below.

3.1 | Multivariate normal distribution

After the set of visits m_i was generated for participant i, six domain scores were simulated from the multivariate normal distribution at each visit. The covariance matrix for $\boldsymbol{U}_i$ was specified as following. We set $\sigma^2 = 30$, $\theta^2 = 10$, $\rho_{sr} = 20$, for $s = r$, and $\rho_{sr} = 60$, for $s \neq r$ with $s, r = 1, \ldots, 6$. Each element for covariance matrix can then be computed. Diagonal elements are $\sigma^2 + \theta^2 + \rho_{11} = 100$. Covariance of different cognitive domains at the same visit is $\theta^2 + \rho_{12} = 30$. Covariance of the same cognitive domains at different visits is $\rho_{11} = 60$. The remaining elements are $\rho_{12} = 20$.

We considered four types of polynomial mean functions over time. For the constant trend, all six cognitive domains were assumed to have mean of 50 at any given t. For the linear trend, the first three cognitive domains were set to have means of $50 - 0.3t$, and the other three to have means of $50 - 0.5t$. For the quadratic trend, the first three cognitive domains were set to have means of $50 - 0.02t^2 + 0.1t$, and the other three to have means of $50 - 0.15t^2 + 0.2t$. Finally, for the cubic trend, the first three were set to have means of $50 - 0.001t^3 + 0.05t^2 + 0.3t$ and the remaining to have means of $50 - 0.0015t^3 + 0.07t^2 + 0.6t$.

The `mvrnorm` from the R library `MASS` was then used to generate longitudinal cognitive errors following the multivariate normal distribution with means set to 0 and the covariance matrix as described above. The mean polynomial functions with the four forms (see above) were added to the errors to represent the simulated longitudinal cognitive scores. For the healthy control group, the `lmer` from the library `lme4` was used to implement model (1). Without assuming any prior knowledge of the true longitudinal mean trend, cubic polynomial functions were used to describe the mean functions for all four sets of data. For each type of mean functions, the test statistics were then computed for 1000 testing participants using the sample mean and covariance matrix obtained from the corresponding healthy control group. The χ^2 tests were conducted for each simulated dataset at different levels of α (from 0.001 to 0.1), and the average FWER was computed based on 1000 simulations for each type of mean functions. Figure 1 illustrates the obtained FWERs of the LMNC χ^2 test across all α levels under the cubic mean trend. The results for the other three mean trends are almost identical to those for the cubic trend and thus are not shown here. The estimated FWERs are denoted by the black solid lines, and the nominal α levels are denoted by gray dash lines. The two lines are almost identical under the four mean trends. The LMNC χ^2 test seems to have exact FWER when domain scores follow multivariate normal distributions and the underlying means and covariance structure are correctly specified.

3.2 | Multivariate *t* and Gamma distributions

Real data often do not follow multivariate normal distributions. Skewness and heavy tails are commonly observed. In this simulation setting, we considered the same four mean functions described in Section 3.1 but with nonnormal errors. One set of errors had symmetric heavy tails from multivariate t distributions, and the other set had negative skewness transformed from correlated Gamma distributions.

We generated longitudinal random errors from multivariate t distributions with 5, 25, and 50 degrees of freedom. The covariance matrix for the error terms was assumed to follow the same structure as described in Section 3.1, and the means of the errors were set at 0. The `rmt` from the library `csampling` was used for multivariate t random error generation. Then we added four polynomial mean trends to the simulated random errors to represent observed longitudinal scores with heavy symmetric tails.

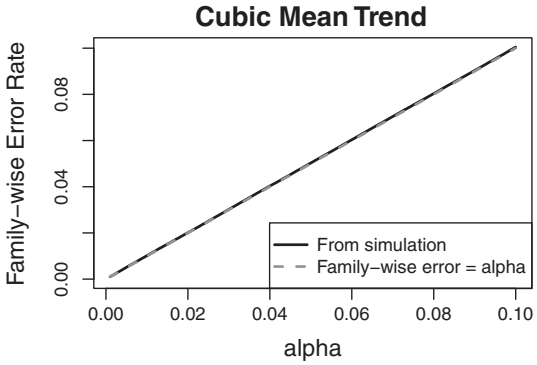

FIGURE 1 The longitudinal multivariate normative comparison χ^2 test when data follow multivariate normal distribution

Next, a gamma distribution was utilized to simulate data with negative skewness. In order to comply with certain covariance structure, that is, compound symmetric, we first generated longitudinal multivariate normal errors $\varsigma_{ijk}, j = 1, \ldots, 6, k = 1, \ldots, m_i$ for participant i with the means of zero. The covariance matrix from Section 3.1 divided by 100 was used here. Then we considered three different gamma distribution designs. For the first one, we calculated $70 - \Gamma^{-1}(\Phi(\varsigma_{ijk}))$ as our negative skewed errors, where Γ is the cumulative distribution function (CDF) of the gamma distribution with shape of 4 and scale of 5 and Φ is the CDF of the standard normal distribution. For the second design, we calculated $100 - \Gamma^{-1}(\Phi(\varsigma_{ijk}))$ as our negative skewed errors, where we assumed shape of 25 and scale of 2 for the gamma distribution. For the third design, we used $150 - \Gamma^{-1}(\Phi(\varsigma_{ijk}))$ as our negative skewed errors, where the gamma distribution has shape of 100 and scale of 1. The same longitudinal mean functions from Section 3.1 were again added to the simulated errors to obtain observed longitudinal cognitive domain scores with negative skewness. All three designs have baseline scores with mean 50 and variance 100.

For each scenario we generated longitudinal cognitive domain scores for 1000 participants from the healthy control group and scores for the other 1000 as the test group. Other simulation setups were the same as those from Section 3.1. To implement the permutation test, we first fit an MLME with cubic polynomial terms to data from the healthy control group as specified in Model (1) and obtained the estimates for the mean trends and the covariance matrix. Then, for each unique number of visits M observed in the test group, we bootstrapped 5000 participants with replacement (N = 5000). For each participant, we subtracted the estimated mean trend from their longitudinal scores. The resulting errors were rearranged randomly by columns as illustrated in Section 2.2 and then sampled by rows with replacement for M visits. The $(1 - \alpha)$th quantile was found among those 5000 test statistics whose average mean values are negative to serve as the threshold for cognitive impairment classification in the test group. After 1000 simulations, summarized FWERs at various levels of α are shown in the upper panel of Figure 2 for data generated from multivariate t distributions and in the bottom panel of Figure 2 for data generated from gamma distributions, both with cubic mean trends. The results under other mean trends are similar and not shown here. For comparison, we also carried out the testing procedure based on χ^2 to examine how FWERs are controlled relative to different α levels. Their FWERs at various levels of α are also shown in Figure 2 along with those from the permutation test.

When multivariate normality does not hold, the FWER based on the χ^2 procedure can be greatly inflated, as shown in Figure 2 where the three black curves denoting the FWERs from the χ^2 test are way above the empirical α levels denoted by the gray broken dash lines. Moreover, the inflation appeared more drastic at smaller levels of α. Conversely, the permutation test successfully maintained FWER at or below any predetermined level as shown in Figure 2. Since the permutation test was applied to the error terms, this suggests that Model 1 is adequate in capturing the mean functions even when the data do not follow multivariate normal. Another interesting phenomenon about the permutation test that we observed from the plots is that FWERs were smaller compared with α when the multivariate t distribution had less heavy tailedness and the gamma distribution had less skew. Simply, when the data move closer to normality, the permutation test becomes more conservative. Though the conservativeness of permutation tests has been observed previously,[20] our permutation test is more complicated and the dependency on the skewness of the data requires further investigation. Therefore, it remains important to check the normality of the data before determining whether the χ^2 or permutation test should be used when applying the LMNC for classification.

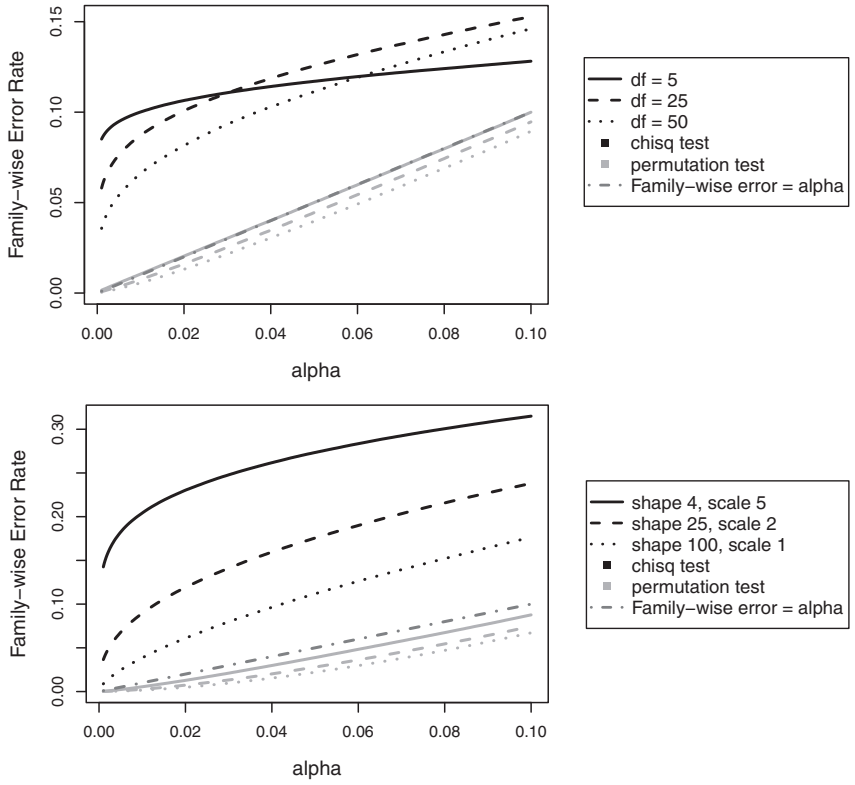

FIGURE 2 The longitudinal multivariate normative comparison χ^2 and permutation tests; the upper panel is when data follow multivariate t distributions (permutation test when $df = 5$ overlapped with the nominal α line); the bottom panel is when data are transformed from Gamma distributions

3.3 | Comparing groups under different visit frequencies

In this section, we examined the power and the FWER of the proposed tests under different settings of visit frequency for the test group. The MACS study, which inspired us to develop the LMNC method, followed men living with and without HIV at roughly the same frequency. Thus, the two comparison groups have similar distributions for the number of visits as shown later in Section 4. However, this may not hold when a new study with certain treatment/condition is tested against an old study, because various factors can contribute to significant differences in visit frequencies. Even within the same study, participants from different cohorts may have different follow-up visits. Therefore, we carried out the following numerical studies to examine how different visit frequencies affect FWER as well as power if comparisons between groups are desired. Four different designs were considered for the test group by changing mean survival time and censoring time:

1. Survival time follows exponential distribution with mean 30 years and is censored at 15 years (median visit number 28);
2. Survival time follows exponential distribution with mean 50 years and is censored at 15 years (median visit number 29);
3. Survival time follows exponential distribution with mean 30 years and is censored at 10 years (median visit number 19);
4. Survival time follows exponential distribution with mean 30 years and is censored at 25 years (median visit number 42).

Here for the healthy control group, we adopted the same multivariate t setting with 5 degrees of freedom and the same quadratic mean trend from Section 3.1. Under the first design the test and control groups had an identical visiting frequency. To evaluate the FWER under the null, we generated 1000 participants following the same mean trend and covariance structure as the test group at each simulation. The only difference was the observed survival time and the subsequent visit frequency. To examine power under alternatives, we assumed the first, third and fifth cognitive domains of the test group to have mean trends of $50 - 0.02t^2$ and the remaining cognitive domains to have mean trends of $30 - 0.04t^2$. As the two sets of domain scores had different means, we lowered the dependence in our covariance structure by setting

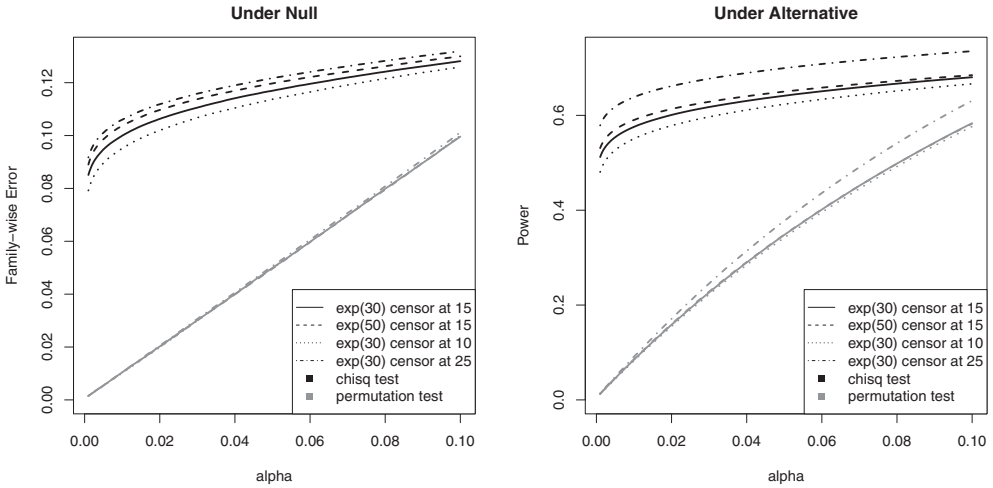

FIGURE 3 The longitudinal multivariate normative comparison χ^2 test and permutation test using multivariate t data when testing group has different visit frequencies

$\sigma^2 = 60$, $\theta^2 = 10$, $\rho_{sr} = 5$, for $s = r$, and $\rho_{sr} = 25$, for $s \neq r$ with $s, r = 1, \ldots, 6$. Therefore, the covariance of different cognitive domains at the same visit was $\theta^2 + \rho_{12} = 15$. Covariance of the same cognitive domains at different visits was $\rho_{11} = 25$. The remaining elements were $\rho_{12} = 5$. Again, we considered the four different designs of visit frequencies for the test group and simulated 1000 participants for each design at every simulation run.

At each simulation for every participant, both the χ^2 test and the permutation test based on 5000 permutations were used for cognitive impairment classification. One thousand simulations were implemented and summarized in Figure 3. From the graph under the null hypothesis, we can see that the frequency-specific permutation test can effectively control FWER at a per-determined α level for all different survival time designs. Because data do not follow a multivariate normal distribution, using the χ^2 test will inflate FWER. The inflation in cognitive impairment differs with various survival time designs and visit frequencies and seems more noticeable when the testing group has many more visits than the control group. Not surprisingly, the χ^2 test has more power than the permutation test under alternative hypotheses. Although close, the testing group with many more visits tends to have higher power in both the χ^2 and permutation tests. Considering both the FWER and power, it is important to make sure that the visit number distributions are comparable when comparisons between two groups are desired.

3.4 | Impact of effective sample size

Under the multivariate normal assumption, the inference is based on an asymptotic χ^2 distribution. With few participants in the healthy control group, the χ^2 test may yield inflated FWERs. When the permutation test is used, a small sample size may limit its ability to obtain a proper permutation distribution. Moreover, the permutation test only uses participants with more than one visit to establish a permutation distribution, because repeated measures are considered in the covariance matrix. Thus, if more people from the healthy control groups are lost to follow-up after the initial measurement, we will have less people left to conduct the permutation test, which translates into a smaller effective sample size. Considering all these, we are interested in evaluating how different effective sample sizes of healthy controls impact the LMNC method in terms of FWER.

To examine how the performance of our proposed χ^2 test was affected, we simulated multivariate normal data using the same setup as in Section 3.1 for the quadratic mean trends but with different effective sample sizes. The number of participants enrolled in the healthy control group was set at 25, 50, 100, 200, 500, and 1000. The size of the testing group was the same as the healthy control group. Average FWERs computed based on 10 000 simulations from both the χ^2 and permutation tests are summarized in Figure 4 with $\alpha = 0.05$ for each sample size considered. To evaluate how different effective sample sizes of healthy controls impact the performance of our proposed permutation test, we used the same multivariate t errors as from Section 3.2. Changes were made to the survival time generation, the time between any consecutive two visits, and the number of participants enrolled in healthy and testing groups. The survival times for both groups follow an exponential distribution with mean 10 years and are censored at 10 years. For the time between any two consecutive visits, uniform (0,1), (0,6), (0,11.5), (0,18), and (0,41) were used to create about 5%, 25%, 45%, 65%, and 85% of participants with only one visit (lost to follow-up) in both healthy control and testing groups. Meanwhile, we

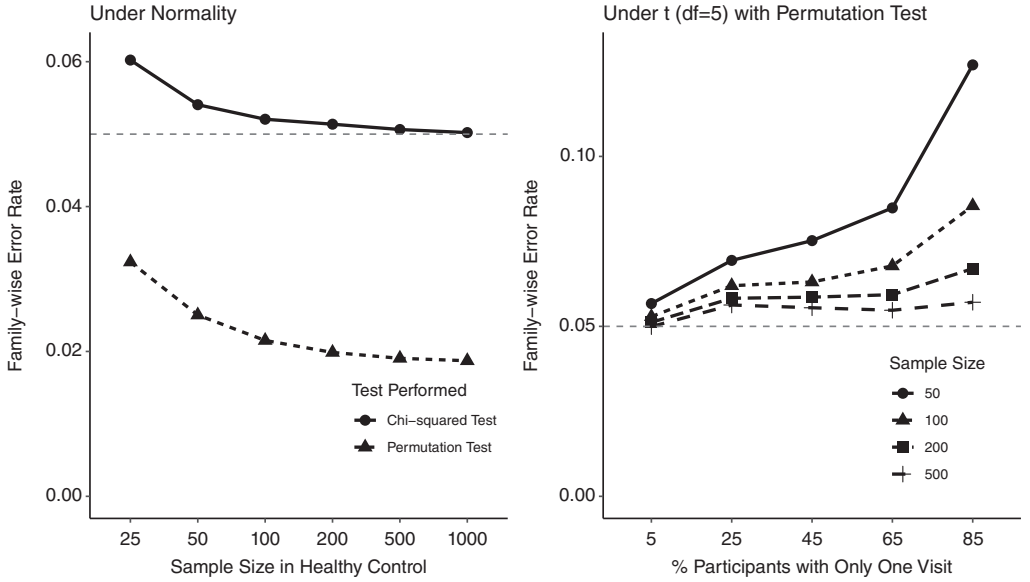

FIGURE 4 The longitudinal multivariate normative comparison (LMNC) χ^2 test and permutation test with different sample sizes under multivariate normality (left); the LMNC permutation test with different sample sizes and visit frequencies using multivariate t data (right); $\alpha = 0.05$ represented as the horizontal dashed line

examined different sample sizes of 50, 100, 200, and 500. Results from 10 000 simulations are summarized in Figure 4 as average FWERs at $\alpha = 0.05$.

From the left plot of Figure 4, we can see that a small number of participants in the healthy control group yields an inflated FWER. When sample size reaches 100, the χ^2 test has FWERs very close to the α level. On the other hand, the permutation test can always control the FWER below the α level, though it can be quite conservative when the sample size is large and errors follow a multivariate normal distribution. When data do not follow multivariate normality, the permutation test also inflates FWERs when the number of participants is small in the healthy control group. At a fixed sample size, a large proportion of people lost to follow-up after initial measurement can also inflate FWER for the LMNC permutation test. Based on the simulations here, when we have more than 100 participants with more than one visit, or the effective sample size is greater than 100, the FWER from the permutation test is very close to the α level.

4 | APPLICATION TO THE MACS

We applied the proposed LMNC to the NP data that were collected from an ongoing MACS. The MACS study has been administered by the University of Pittsburgh, Johns Hopkins University, Northwestern University, and the University of California at Los Angeles.[21,22] Since its first enrollment in 1984, the MACS has recruited more than 7000 men who have sex with men (MSM), either infected with HIV or at risk for infection at study entry. Participants have been regularly interviewed and examined semiannually about a broad range of variables including their age, depressive symptoms, sexual activity, substance use, cognitive functioning, and physical measurements. HIV infection negatively impacts patients' brain, and the effect of HIV on brain functioning was found to be less drastic after the highly active antiretroviral therapy (HAART) became available in early 1990s. In a MACS NP substudy, participants have been repeatedly tested on a NP test battery assessing six cognitive domains which included learning, motor speed and coordination, speed of information processing, memory, working memory and attention, and executive functioning.[23,24] As of October 2017, some participants had more than 20 years of longitudinal NP data. This provides a unique opportunity to examine how cognitive impairment compares between those infected with HIV and those not infected in the HAART era.

At each NP visit, the battery of tests was administrated, and these test scores were summarized by T-scores which were calculated from regression models adjusting for education, race/ethnicity, age, and the number of tests administrated, and standardized to have a mean of 50 and standard deviation of 10. Then, summary T-scores were obtained from taking the arithmetic mean of all T-scores in each domain, except for motor speed and coordination domain, where the lowest T score

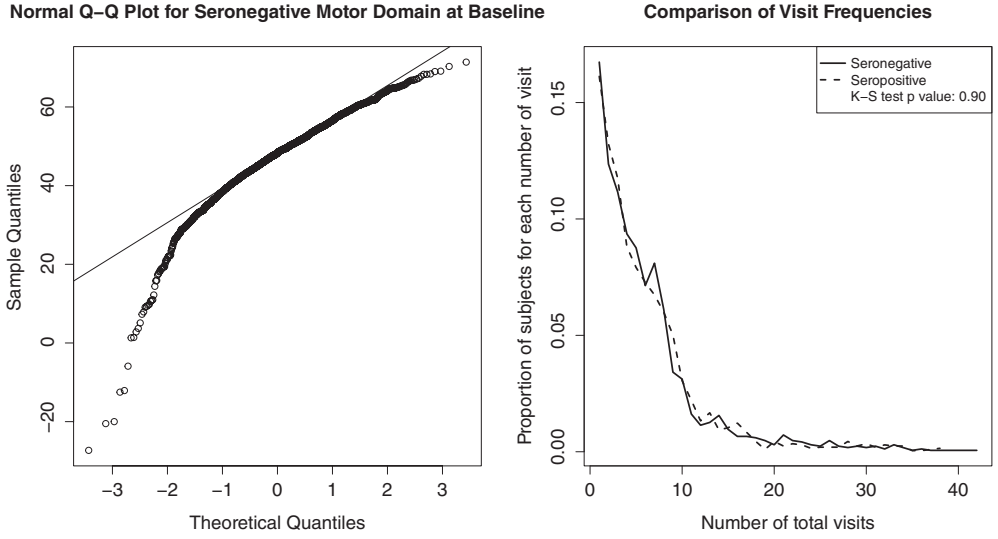

FIGURE 5 Q-Q plot of baseline motor score in the seronegative group and visit frequencies of two serostatus groups

is used. In this analysis, we focus on visits where participants had all six cognitive domain scores available, and include 3701 participants who have at least one such visit. Among participants included in this analysis, 1667 were seronegative (279 having one visit), while 2034 were infected with HIV (328 having one visit) at the study entry. Those not infected with HIV serve as the "healthy" control group, representing HIV-uninfected MSM. Because the motor speed and coordination domain used the lowest T score instead of the average, we can see from Figure 5 that baseline motor domain scores for seronegative participants failed to follow a normal distribution. The LMNC using the χ^2-test may be of concern and the permutation test should be considered. For both seropositive and seronegative groups, we calculated the number of participants for each total visit frequency and plotted them by group in Figure 5. The Kolmogorov-Smirnov test shows that the visit number distributions do not differ ($P = .90$), and the visit frequencies are comparable between the two groups. At the same time, the number of participants with more than one visit is large enough to construct the permutation distribution. Thus, the LMNC permutation test is expected to work well in this application.

Specifically, we first fit the model described in (1) with cubic mean trends in the healthy control group. After estimates were obtained, both the χ^2 test and the permutation test were applied to data from the healthy control group across different levels of α. For both tests, fivefold cross validation was used to test cognitive impairment among those not infected with HIV. The results are shown in Figure 6. The first thing we can see is that the permutation test (N = 100 000) can effectively control FWER at predetermined α levels. By contrast, the χ^2 test would have inflated the family-wise error when the data fail to follow a multivariate normal distribution but the model is sufficiently specified. We also applied both the permutation test and the χ^2 test to data from seropositive men. The results are also shown in Figure 6. The permutation test identified about the same proportion of seropositive men with cognitive impairment as in the seronegative group across αs. Meanwhile, the χ^2 test identified a much higher proportion of cognitively impaired men in the seropositive group than in the seronegative group. The standard Chi-square test for the association between serostatus and cognitive impairment that is identified by our χ^2 test yields a P value less than .0001, and that for cognitive impairment identified by our proposed permutation test results in a P value of .07, suggesting different levels of associations. This may have subsequent clinical and research implications. Not only would the χ^2 test identify more people with cognitive impairment in seronegative and seropositive groups, but also different conclusions might be drawn about the relationship between serostatus and cognitive impairment during the HAART era. By contrast, the permutation test shows that the association between cognitive impairment and HIV infection is rather weak, leading to the conclusion that people infected with HIV seem to enjoy relatively healthy cognitive functioning after being properly treated with HAART.

Both the χ^2 and permutation tests are based on the fact that the seronegative group (not infected with HIV) is treated as the reference group of healthy controls. Unlike the simulation study, where we know all the participants tested are under the null distribution when evaluating FWER, the true cognitive impairment status from the MACS seronegative group is actually unknown but their functioning scores are assumed to follow a normal aging process. The impairment rate and the above conclusion may differ had we used another reference group. To further validate the comparisons between the two tests, Table 1 shows the mean scores of all six cognitive domains for both seronegative and seropositive groups at the first visit (100% participants), the forth visit (50% participants), and the tenth visit (15% participants). We can see that,

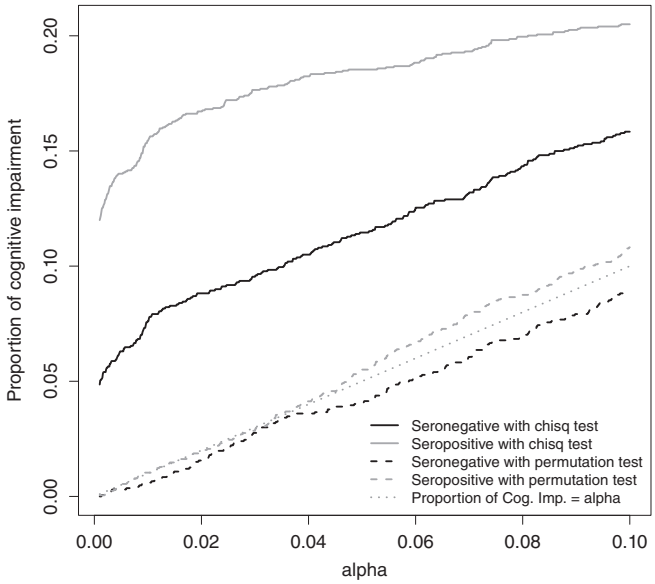

FIGURE 6 Comparing proportion of cognitive impairment in seronegative and seropositive groups in the Multicenter AIDS Cohort Study

TABLE 1 Mean scores of six cognitive domains for seronegative and seropositive groups at different visit

	Cognitive domain	Motor	Executive	Speed	Learning	Memory	Working memory
At visit 1	Seronegative	47.12	49.81	49.92	49.67	49.90	49.64
	Seropositive	46.73	49.77	49.25	49.71	49.98	49.40
At visit 4	Seronegative	45.86	50.17	50.45	49.33	49.04	48.99
	Seropositive	45.79	49.31	49.33	48.96	49.04	48.36
At visit 10	Seronegative	48.14	53.14	51.25	50.94	50.74	51.86
	Seropositive	48.26	51.93	51.00	52.41	52.38	51.51

relative to standard deviation of 10, the score differences are very small between the two groups. A wrong conclusion would be drawn if a method failing to control family-wise error, like the χ^2 test in this case, is used otherwise.

5 | DISCUSSION

Our work demonstrated that the proposed LMNC method can effectively control FWER. Multivariate normality is a key assumption in using the χ^2 test for cognitive impairment classification. When such an assumption is not satisfied by data or the model in use does not fully address random effects, the permutation test can still guard FWER at a predetermined level.

The MNC method specifically takes intercorrelations among domain scores into account, and may lead to different results as some existing methods that are used in AIDS research. As an example, we only consider two cognitive domains at a single visit. Suppose that the variance of two domain scores is 1 and the correlation is 0.5, and both mean cognitive scores are zero. The participant having cognitive scores of $(-1, -2)$ will have a larger P value than the one with scores $(0, -2)$. This is contrary to the intuition that the first participant seems to have more extreme scores. However, the correlation between two domains is high. Thus, the scores $(0, -2)$ from the second participant is more unusual than $(-1, -2)$ under the strong positive correlation, and consequently, the second participant has a longer "distance" from the means, after inversely weighted by the covariance matrix. If the correlation between two domains is set to be zero, then the first participant will have a smaller P value. Therefore, the MNC results may not be consistent with some existing ad hoc diagnoses methods such as counting the number of domains with scores 1 or 1.5 standard deviations below the means.[1,2]

This paradox also exists in a longitudinal setting. For illustration purposes, let us assume that only one cognitive domain is tested, with a mean of 0 and variance of 1. The correlation between any two visits is 0.5. One participant with the

domain score tested at two visits as $(0, -2)$ will have a larger P value than another participant with the domain score tested at three visits as $(0, 0, -2)$. This is also against the intuition as the first participant seems to have worse cognition earlier. However, the second participant has longer records of being "normal," so the "distance" from the means is also larger after weighted by the inverse covariance matrix. Consequently, the second participant has a smaller P value. If domain scores are independent among all visits, the P value for the second participant would be larger, because of more visits and a larger degree of freedom when performing the χ^2 test. This may serve as an explanation to why we observed greater power under a higher visit frequency design, even though they follow the same mean trends. To generalize our proposed method to groups with very different visit number distributions, further efforts should be made to improve our proposed permutation test. Moreover, our current MNC analysis only uses visits at which all domain scores are available, while naive methods can tolerate one or two missing domains. Nevertheless, our proposed LMNC method provides insights into how "abnormal" domain scores may be, which could be missed by naive methods ignoring intercorrelations among domain scores and repeated visits. How to extend our LMNC method to missing domain data will be an interesting future research topic.

ACKNOWLEDGEMENTS

The work is partially supported by the NSF DMS -1916001 to Cheng and the University of Pittsburgh Center for Research Computing through the resources provided. We are grateful to the associate editor and two anonymous reviewers for their constructive comments and suggestions that led to a substantially improved article. Data in this article were collected by the MACS with centers at Baltimore (U01-AI35042): The Johns Hopkins University Bloomberg School of Public Health: Joseph B. Margolick (PI), Todd Brown (PI), Jay Bream, Adrian Dobs, Michelle Estrella, W. David Hardy, Lisette Johnson-Hill, Sean Leng, Anne Monroe, Cynthia Munro, Michael W. Plankey, Wendy Post, Ned Sacktor, Jennifer Schrack, Chloe Thio; Chicago (U01-AI35039): Feinberg School of Medicine, Northwestern University, and Cook County Bureau of Health Services: Steven M. Wolinsky (PI), Sheila Badri, Dana Gabuzda, Frank J. Palella, Jr., Sudhir Penugonda, John P. Phair, Susheel Reddy, Matthew Stephens, Linda Teplin; Los Angeles (U01-AI35040): University of California, UCLA Schools of Public Health and Medicine: Roger Detels (PI), Otoniel Martínez-Maza (PI), Otto Yang (Co-PI), Peter Anton, Robert Bolan, Elizabeth Breen, Anthony Butch, Shehnaz Hussain, Beth Jamieson, John Oishi, Harry Vinters, Dorothy Wiley, Mallory Witt, Stephen Young, Zuo Feng Zhang; Pittsburgh (U01-AI35041): University of Pittsburgh, Graduate School of Public Health: Charles R. Rinaldo (PI), Lawrence A. Kingsley (PI), Jeremy J. Martinson (PI), James T. Becker, Phalguni Gupta, Kenneth Ho, Susan Koletar, John W. Mellors, Anthony J. Silvestre, Ronald D. Stall; Data Coordinating Center (UM1-AI35043): The Johns Hopkins University Bloomberg School of Public Health: Lisa P. Jacobson (PI), Gypsyamber D'Souza (PI), Alison Abraham, Keri Althoff, Michael Collaco, Priya Duggal, Sabina Haberlen, Eithne Keelaghan, Heather McKay, Alvaro Muñoz , Derek Ng, Anne Rostich, Eric C. Seaberg, Sol Su, Pamela Surkan, Nicholas Wada. Institute of Allergy and Infectious Diseases: Robin E. Huebner; National Cancer Institute: Geraldina Dominguez. The MACS is funded primarily by the National Institute of Allergy and Infectious Diseases (NIAID), with additional cofunding from the National Cancer Institute (NCI), the National Institute on Drug Abuse (NIDA), and the National Institute of Mental Health (NIMH). Targeted supplemental funding for specific projects was also provided by the National Heart, Lung, and Blood Institute (NHLBI), and the National Institute on Deafness and Communication Disorders (NIDCD). MACS data collection is also supported by UL1-TR001079 (JHU ICTR) from the National Center for Advancing Translational Sciences (NCATS) a component of the NIH, and NIH Roadmap for Medical Research. Additional support was provided by the Johns Hopkins University Center for AIDS Research (P30AI094189). The contents of this publication are solely the responsibility of the authors and do not represent the official views of the National Institutes of Health (NIH), Johns Hopkins ICTR, or NCATS.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The MACS website is located at http://aidscohortstudy.org/. Investigators interested in doing research in the MACS submit a concept sheet (https://mwccs.org/) which is reviewed by the relevant working groups and then sent to the MACs executive committee for a decision. Approved concept sheets have assigned MACS coinvestigators from participating sites.

ORCID

Yu Cheng https://orcid.org/0000-0003-4805-203X

REFERENCES

1. Antinori A, Arendt G, Becker JT, et al. Updated research nosology for HIV-associated neurocognitive disorders. *Neurology*. 2007;69(18):1789-1799.
2. Gisslén M, Price RW, Nilsson S. The definition of HIV-associated neurocognitive disorders: are we overestimating the real prevalence? *BMC Infect Dis*. 2011;11(1):356.
3. Huizenga HM, Smeding H, Grasman RPPP, Schmand B. Multivariate normative comparisons. *Neuropsychologia*. 2007;45(11):2534-2542.
4. Su T, Schouten J, Geurtsen GJ, et al. Multivariate normative comparison, a novel method for more reliably detecting cognitive impairment in HIV infection. *AIDS*. 2015;29(5):547-557.
5. Wang Z, Molsberry S, Cheng Y, et al. Cross-sectional analysis of cognitive function using multivariate normative comparisons in men with HIV disease. *AIDS*. 2019;33(14):2115-2124.
6. Reinsel G. Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *J Am Stat Assoc*. 1982;77(377):190-195.
7. Heitjan DF, Sharma D. Modeling repeated-series longitudinal data. *Stat Med*. 1997;16(4):347-355.
8. Fang H, Tian G, Xiong X, Tan M. A multivariate random-effects model with restricted parameters: application to assessing radiation therapy for brain tumours. *Stat Med*. 2006;25(11):1948-1959.
9. Fieuws S, Verbeke G. Joint modelling of multivariate longitudinal profiles: pitfalls of the random-effects approach. *Stat Med*. 2004;23(20):3093-3104.
10. Fieuws S, Verbeke G. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*. 2006;62(2):424-431.
11. Verbeke G, Fieuws S, Molenberghs G, Davidian M. The analysis of multivariate longitudinal data: a review. *Stat Methods Med Res*. 2014;23(1):42-59.
12. van den Hout A, Fox J, Muniz-Terrera G. Longitudinal mixed-effects models for latent cognitive function. *Stat Model* 2015; 15(4): 366-387.
13. Bloxom B. A constrained spline estimator of a hazard function. *Psychometrika*. 1985;50(3):301-321.
14. De Boor C. *A Practical Guide to Splines*. New York, NY: Springer; 2001.
15. Shumaker L. *Spline Functions: Basic Theory*. Cambridge, MA: Cambridge University Press; 2007.
16. Rutherford MJ, Crowther MJ, Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *J Stat Comput Simul*. 2015;85(4):777-793.
17. Harrell FE Jr. *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. New York, NY: Springer; 2015.
18. Miller EN, Seines OA, McArthur JC, et al. Neuropsychological performance in HIV-1-infected homosexual men. *Neurology*. 1990;40(2):197-197.
19. Becker J, Kingsley L, Molsberry S, et al. Cohort profile: recruitment cohorts in the neuropsychological substudy of the multicenter AIDS cohort study. *Int J Epidemiol*. 2014;44(5):1506-1516.
20. Berger VW. Pros and cons of permutation tests in clinical trials. *Stat Med*. 2000;19(10):1319-1328.
21. Kingsley L, Kaslow R, Rinaldo CJR, Detre K, Odaka N, Vanraden M. Risk factors for seroconversion to human immunodeficiency virus among male homosexuals. *Lancet*. 1987;329(8529):P345-P349.
22. Kaslow R, Ostrow D, Detels R, Phair J, Polk F, Rinaldo J. The multicenter AIDS cohort study: rationale, organization, and selected characteristics of the participants. *Am J Epidemiol*. 1987;126(2):310-318.
23. Farinpour R, Miller E, S P, et al. Psychosocial risk factors of HIV morbidity and mortality: findings from the multicenter AIDS cohort study (MACS). *J Clin Exp Neuropsychol*. 2003;25(5):654-670.
24. Popov M, Molsberry S, Lecci F, et al. Brain structural correlates of trajectories to cognitive impairment in men with and without HIV disease. *Brain Imag Behav*. 2020;14(3):821-829.

How to cite this article: Wang Z, Cheng Y, Seaberg EC, et al. Longitudinal multivariate normative comparisons. *Statistics in Medicine*. 2021;40:1440–1452. https://doi.org/10.1002/sim.8850