HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2024 January 15.

Published in final edited form as:

Stat Med. 2023 January 15; 42(1): 52–67. doi:10.1002/sim.9601.

Dynamic Impairment Classification Through Arrayed Comparisons

Zheng Wang[1], **Zi Wang**[1], **Lingyun Lyu**[2], **Yu Cheng**[*,1,2], **Eric C. Seaberg**[3], **Samantha A. Molsberry**[4], **Ann Ragin**[5], **James T. Becker**[6]

[1]Department of Statistics, University of Pittsburgh, PA, USA

[2]Department of Biostatistics, University of Pittsburgh, PA, USA

[3]Department of Epidemiology, Johns Hopkins University, MD, USA

[4]Population Health Sciences, Harvard University, MA, USA

[5]Department of Radiology, Northwestern University, IL, USA

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

Summary

The Multivariate Normative Comparison (MNC) method has been used for identifying cognitive impairment. When participants' cognitive brain domains are evaluated regularly, the longitudinal MNC (LMNC) has been introduced to correct for the inter-correlation among repeated assessments of multiple cognitive domains in the same participant. However, it may not be practical to wait until the end of study for diagnosis. For example, in participants of the Multicenter AIDS Cohort Study (MACS), cognitive functioning has been evaluated repeatedly for more than 35 years. Therefore, it is optimal to identify cognitive impairment at each assessment, while the family-wise error rate is controlled with unknown number of assessments in future. In this work, we propose to use the difference of consecutive LMNC test statistics to construct independent tests. Frequency modeling can help predict how many assessments each participant will have, so Bonferroni-type correction can be easily adapted. A Chi-squared test is used under the assumption of multivariate normality, and permutation test is proposed where this assumption is violated. We showed through simulation and the MACS data that our method controlled family-wise error rate below a pre-determined level.

Keywords

Cognitive impairment; Dynamic classification; Family-wise error rate; Frequency Modeling; Multivariate mixed-effect model; Sequential test

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

1 | INTRODUCTION

This article concerns dynamic classification of cognitive impairment over time, which is motivated by the Multi-center AIDS Cohort Study (MACS) and of considerable general interest to neuroscience researchers in psychiatry and brain science fields. This approach to classification can further guide treatments to effectively prolong life expectancy and promote quality of life in persons living with HIV. Given normalized neuropsychological scores collected from a battery of tests across several cognitive brain domains, existing popular methods use t tests on each domain separately or count number of domains with abnormal scores. [1,2] These approaches fail to properly control family-wise error rate (FWER) by not taking inter-correlations among all brain domains into account. Here the type I error is defined in reference to the deviations from "normal" cognitive functions gathered from a large reference group of healthy controls.

Huizenga et al [3] introduced a method called the Multivariate Normative Comparison (MNC). The MNC statistic is a function of an overall distance between a subject's multiple domain scores and the norms of the reference population. Then a test analogous to Hotelling's multivariate T test is carried out to flag any abnormally large distance from normative scores. Under the assumption that cognitive brain domain scores follow a multivariate normal distribution, the MNC method will control FWER effectively at a pre-determined level. Su et al [4] and Wang et al [5] have studied and demonstrated the effectiveness of the MNC with cross-sectional neuropsychological data collected among persons infected with HIV. Wang et al [6] further extended the MNC method from cross-sectional data to historically collected longitudinal data and proposed a Longitudinal Multivariate Normative Comparison (LMNC) method by modeling multiple neuropsychological scores with a multivariate linear mixed effects (MLME) model. Correlations among different brain domains and across all assessments within the same participant are explicitly considered within the model. A permutation procedure is also available when the model cannot sufficiently explain the data or the data are not normally distributed.

For research purposes, we often carry out analyses after all necessary information has been collected and then examine the data retrospectively. Thus, the static longitudinal MNC method in Wang et al [6] is adequate for providing classification of prior cognitive impairment with proper control of family-wise error. In clinical practice, this is probably not sufficient, particularly when treatment should be prescribed in the early stage of cognitive decline. Ideally, classification of impaired cognition should be done fluidly at each assessment. This will have practical use in Just-in-Time Adaptive Interventions (JITAIs) which track health status on mobile device. [7,8] Proper identification of departures from normal health can lead to effective treatments.

For an ongoing longitudinal study, the history of data provide useful insight into participants' behaviors, such as how their cognitive functions change over time, how frequently they visit research centers, and how long they might survive. Assuming that we have collected sufficient data from the same or a similar cohort for an ongoing study, we can predict how many assessments each participant will have, using survival analysis and frequency modeling techniques such as Cox proportional hazard regression and Poisson

regression. Analogous to retrospective longitudinal data, MLME models [9] can be applied to prospectively collected data to capture dynamic changes in domain scores over time for each individual, while accounting for inter-correlations among various cognitive domains and repeated measures for the same participant. Longitudinal multivariate normative comparison (LMNC) statistics can be built from MLME estimations for each participant at every assessment. The proposed dynamic arrayed comparison (DAC) test is based on differences in two consecutive LMNC statistics, which are shown to be independent across assessments. This is key in constructing our test procedures for normally distributed data. When data do not follow multivariate normal distributions, we flexibly adapt permutation tests for dynamic classification, and then apply our proposed classification procedures to control family-wise error.

The proposed methodology will be detailed in Section 2. We will first formulate frequency prediction and MLME models, and then construct test statistics for DAC with the Bonferroni-type adaptive procedure. In Section 3, numerical studies will be carried out to examine the performance of the proposed method. Applications of DAC to a neuropsychological substudy in the MACS are shown in Section 4. We conclude with discussion of DAC performance and future directions.

2 | FAMILY-WISE ERROR CONTROLLING PROCEDURES

2.1 | Dynamic Arrayed Comparisons Based on χ^2

Before identifying cognitive impairment prospectively, we assume that relevant population normative data are available. Following the notation from Wang et al, [6] n participants enrolled in a healthy reference group and were evaluated on q cognitive domains over m_i assessments. Cognitive domain j is measured as Y_{ijk}, $i = 1, \ldots, n$; $j = 1, \ldots, q$; $k = 1, \ldots, m_i$ for participant i at k-th assessment. A multivariate normal distribution is assumed on cognitive scores since they are generally normalized in practice. Within the same participant, the MLME model is used for cognitive functions from different domains over time with a covariance structure capturing dependence among cognitive domains and repeated measures at different assessments. Thus we have:

$$\boldsymbol{Y}_{ijk} = \beta_{j0} + \beta_{j1} t_{ik} + \beta_{j2} t_{ik}^2 + \beta_{j3} t_{ik}^3 + \upsilon_{ij} + \delta_{ik} + \epsilon_{ijk}\,. \tag{1}$$

Here we use q polynomial functions of degree 3 to characterize the mean cognitive scores over time. If desired, polynomials with a higher degree can be added. The B-spline technique, besides polynomial, can also approximate the true average cognitive scores over time. [10,11,12,13,14] We assume ϵ_{ijk}, representing random error from each observation, to be independent and identically distributed following normal $N(0, \sigma^2)$. We also assume δ_{ik}, which are errors specific for each assessment, to follow independent and identical normal $N(0, \theta^2)$, since the variances and covariances are generally stable over time in MACS. Due to the inter-correlations among various cognitive domains for the same participant, $\boldsymbol{\upsilon}_i = \left(\upsilon_{i1}, \ldots, \upsilon_{iq}\right)^\top$ is assumed to follow multivariate normal $N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = [\rho_{sr}]$, $s, r = 1, \ldots, q$. For the MACS domain data, we assume a compound symmetric structure, since the covariances between two different domains from two different assessments are

almost always around 20, except for three outliers (two 5's and one 50). The structure of covariance matrix depends on research designs and data collected, and can be unspecified, auto-regressive, compound symmetric, or follow other complicated structures such as the Damped Exponential Correlation. [15] If the outcomes are viral loads, CD4 counts, and other highly variable variables, more flexible models [15] should be considered.

Fang et al [16] and Fieuws [17] provided estimation procedures used to estimate unknown parameters from the MLME model. Estimated parameters 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 $\hat{\sigma}^2$. Assuming participant d is tested, we take all q cognitive scores observed over m_d assessments, and stack them into m_d vectors

$$\boldsymbol{U}_{\omega}^{d} = \left(\boldsymbol{Y}_{d11}, \ldots, \boldsymbol{Y}_{dq1}, \boldsymbol{Y}_{d12}, \ldots, \boldsymbol{Y}_{dq2}, \ldots, \boldsymbol{Y}_{d1\omega}, \ldots, \boldsymbol{Y}_{dq\omega}\right)^{\top}, \omega = 1, \ldots, m_d\,. \tag{2}$$

From the MLME model in (1), the estimated mean vector of $\boldsymbol{U}_{\omega}^{d}$ is written as

$\hat{\boldsymbol{\mu}}_{\omega}^{d} = \left(\hat{\beta}_{10} + \hat{\beta}_{11}t_{d1} + \hat{\beta}_{12}t_{d1}^2 + \hat{\beta}_{13}t_{d1}^3, \hat{\beta}_{20} + \hat{\beta}_{21}t_{d1}\right.$
$\left. + \hat{\beta}_{22}t_{d1}^2 + \hat{\beta}_{23}t_{d1}^3, \ldots, \hat{\beta}_{q0} + \hat{\beta}_{q1}t_{d1} + \hat{\beta}_{q2}t_{d1}^2 + \hat{\beta}_{q3}t_{d1}^3, \ldots, \hat{\beta}_{10} + \hat{\beta}_{11}t_{d\omega} + \hat{\beta}_{12}t_{d\omega}^2 + \hat{\beta}_{13}t_{d\omega}^3, \ldots, \hat{\beta}_{q0} + \hat{\beta}_{q1}t_{d\omega} + \hat{\beta}_{q2}t_{d\omega}^2 + \hat{\beta}_{q3}t\right.$
of length $q\omega$. Moreover, from the covariance matrix specified in this model, we can estimate the covariance matrix for $\boldsymbol{U}_{\omega}^{d}$ as $\hat{\boldsymbol{\Psi}}_{\omega}^{d} = [\tau_{sr}]$, $s, r = 1, \ldots, q\omega$. Each element in $\boldsymbol{\Psi}_{\omega}^{d}$ corresponds to the covariance between a pair $\boldsymbol{Y}_{dj_1k_1}$ and $\boldsymbol{Y}_{dj_2k_2}$, 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 indices $1 \le j_1, j_2 \le q$, assessment indices $1 \le k_1, k_2 \le \omega$ and indicator functions $\mathbb{I}\{\cdot\}$.

Under the multivariate normal assumption on the q longitudinal cognitive functioning scores, Wang et al[6] proposed a LMNC test statistic for participant d at assessment ω as

$$\boldsymbol{G}_{\omega}^{d} = \left(\boldsymbol{U}_{\omega}^{d} - \hat{\boldsymbol{\mu}}_{\omega}^{d}\right)^{\top}\left(\hat{\boldsymbol{\Psi}}_{\omega}^{d}\right)^{-1}\left(\boldsymbol{U}_{\omega}^{d} - \hat{\boldsymbol{\mu}}_{\omega}^{d}\right) \sim \chi_{q\omega}^{2}, \omega = 1, \ldots, m_d, \tag{3}$$

and used it to classify prior impairment status on retrospectively collected data. In order to identify cognitive impairment at each assessment ω, we create DAC test statistics as the difference between two consecutive LMNC test statistics and set $S_{\omega}^{d} = \boldsymbol{G}_{\omega}^{d} - \boldsymbol{G}_{\omega-1}^{d}$ for $2 \le \omega \le m_d$ and $S_1^d = \boldsymbol{G}_1^d$. We can show below that they are independent from each other. Without loss of generality denote $\boldsymbol{X}_{\omega} = \boldsymbol{U}_{\omega} - \mu_{\omega} = \left(\boldsymbol{X}_{\omega-1}^{\top}, W_{\omega}^{\top}\right)^{\top}$ for $\omega \ge 2$ and $\boldsymbol{X}_1 = W_1$. The covariance matrix of $\boldsymbol{X}_{\omega}$ is

$$\boldsymbol{\Psi}_{\omega} = \begin{pmatrix} \boldsymbol{\Psi}_{\omega-1} & \boldsymbol{\Delta}_{\omega} \\ \boldsymbol{\Delta}_{\omega}^{\mathrm{T}} & \boldsymbol{\Phi}_{\omega} \end{pmatrix}, \omega \ge 2,$$

with $\boldsymbol{\Psi}_1 = \boldsymbol{\Phi}_1$. Then for $\omega \ge 2$,

$$\begin{aligned} S_\omega &= G_\omega - G_{\omega-1} \\ &= X_\omega^\top \Psi_\omega^{-1} X_\omega - X_{\omega-1}^\top \Psi_{\omega-1}^{-1} X_{\omega-1} \\ &= X_{\omega-1}^\top \Psi_{\omega-1}^{-1} \Delta_\omega \Theta_\omega^{-1} \Delta_\omega^\top \Psi_{\omega-1}^{-1} X_{\omega-1} - 2 X_{\omega-1}^\top \Psi_{\omega-1}^{-1} \Delta_\omega \Theta_\omega^{-1} W_\omega + W_\omega^\top \Theta_\omega^{-1} W_\omega, \end{aligned}$$

where $\Theta_\omega = \Phi_\omega - \Delta_\omega^\top \Psi_{\omega-1}^{-1} \Delta_\omega$. We also know from properties of multivariate normal distributions that $X_{\omega-1}$, with covariance $\Psi_{\omega-1}$, and $W_\omega - \Delta_\omega^\top \Psi_{\omega-1}^{-1} X_{\omega-1}$, with covariance Θ_ω, are independent. Therefore,

$$\begin{aligned} S_\omega &= X_{\omega-1}^\top \Psi_{\omega-1}^{-1} \Delta_\omega \Theta_\omega^{-1} \Delta_\omega^\top \Psi_{\omega-1}^{-1} X_{\omega-1} - 2 X_{\omega-1}^\top \Psi_{\omega-1}^{-1} \Delta_\omega \Theta_\omega^{-1} W_\omega + W_\omega^\top \Theta_\omega^{-1} W_\omega \\ &= \left(W_\omega - \Delta_\omega^\top \Psi_{\omega-1}^{-1} X_{\omega-1}\right)^\top \Theta_\omega^{-1} \left(W_\omega - \Delta_\omega^\top \Psi_{\omega-1}^{-1} X_{\omega-1}\right) \sim \chi_q^2. \end{aligned}$$

As a result, we can claim DAC test statistics within $\left\{S_\omega^d, \omega = 1, \ldots, m_d\right\}$ are independent.

If m_d is known, we can construct visit-by-visit testing procedures easily and apply Bonferroni or other procedures to control FWER. In order to identify cognitive impairment for participant d at assessment ω, we will use $\left(1 - 2\alpha_\omega^d\right)$ quantile of χ_q^2 as the threshold to control the overall significance level α while $\mathbf{1}_{q\omega}^\top \hat{X}_\omega^d < \mathbf{1}_{q(\omega-1)}^\top \hat{X}_{\omega-1}^d$ when $\omega > 1$ or $\mathbf{1}_q^\top \hat{X}_1^d < 0$ when $\omega = 1$, since we are interested in screening impairment, i.e., cognitive scores that are largely lower than the means. However, m_d is generally unknown for a prospective study. We will address it in the following section.

2.2 | Frequency Prediction

Given that we have observed some data with respect to our population of interest in an ongoing study, the difficulty of unknown m_d can be addressed by using some frequency model to estimate the expected number of assessments each patient will have based on their baseline characteristics and historical data H. One may first use a survival modeling approach to predict the mean residual lifetime ($\hat{T}$) of each patient and then use Poisson regression to predict the frequency of visiting during a unit of time ($\hat{\Lambda}$). For example, we assume that the survival time follows a proportional hazards model

$$\lambda(t \mid Z_1) = \lambda_0(t) \exp\left(\beta^\top Z_1\right), \tag{4}$$

and that the number of assessments per time unit follows a log-linear model

$$\log(\Lambda \mid Z_2) = \gamma_0 + \gamma_1^\top Z_2, \tag{5}$$

though other proper models may be used, where covariate vectors Z_1 and Z_2 can be completely different or have overlapping predictors. Commonly used techniques like (partial) maximum likelihood estimation can be used to obtain the estimates $\hat{\beta}$, $\hat{\gamma}_0$ and $\hat{\gamma}_1$, and the Nelson-Aalen type estimator can be computed for $\hat{\lambda}_0(t)$. As a result, the product of the expected survival time

$$\hat{T}_i = \int \exp\left\{-\hat{\Lambda}_0(t)e^{\hat{\beta}^\top Z_{1i}}\right\}dt$$

and the expected frequency at a unit of time

$$\hat{\Lambda}_i = \exp\left(\hat{\gamma}_0 + \hat{\gamma}_1^\top Z_{2i}\right)$$

can be used to predict the number of assessments, i.e., $\widehat{N}_i = \hat{T}_i\hat{\Lambda}_i$ for participant i. We assume $\widehat{N}_i$ is an unbiased estimator for participant's expected number of future visits. The nearest integer greater than $\widehat{N}_i$ is used if the prediction is a decimal number and it is denoted by $\lceil \widehat{N}_i \rceil$ Based on the estimated expected number of assessments, we propose a Bonferroni-type adaptive procedure on controlling family-wise errors.

2.3 | Bonferroni-type Adaptive Procedure

Bonferroni procedure seems to be a natural choice in controlling family-wise error with expected number of assessments known. [18,19,20] If a uniform $\frac{1}{\lceil \widehat{N} \rceil_i}\alpha$ is applied to every assessment, the participants who have fewer assessments than expected will have a much smaller family wise error than those who have more assessments than expected. Therefore, we propose a new procedure to control FWER, where participants who visit the clinic more than expected will be subject to more stringent testing. We define $\boldsymbol{M}_i$ to be the actual number of assessments for participant i, which is unknown at the beginning of the study. For participant i at any assessment $1 \le j \le M_i$, if $j \le \lceil \widehat{N}_i \rceil$ the p value at this visit is compared with $\frac{C\left(\lceil \widehat{N}_i \rceil - j + 1\right)}{\lceil \widehat{N}_i \rceil^2}\alpha$ When $j > \lceil \widehat{N}_i \rceil$, the p value is compared with $\frac{C}{\lceil \widehat{N}_i \rceil^2}\alpha$. C is chosen to be $\frac{2c\lceil \widehat{N}_i \rceil^2}{\lceil \widehat{N}_i \rceil\left(\lceil \widehat{N}_i \rceil + 1\right)}$. When $c = 1$ and under the null hypothesis $\boldsymbol{H}_0$ that the participant being tested is not impaired cognitively at any visit, we have

$$
\begin{aligned}
&P(\text{ classified cognitive impairment during the study } \mid Z_{1i}, Z_{2i}, H_0, \boldsymbol{H}_S) \\
&= 1 - \prod_{j=1}^{\infty} \{1 - P(\text{ classified cognitive impairment at visit } j \mid Z_{1i}, Z_{2i}, H_0, \boldsymbol{H}_S)\} \\
&= 1 - \mathbb{E}\left(\sum_{j=1}^{M_i} \left[1 - \alpha \left\{ \frac{C\left(\lceil \widehat{N}_i \rceil - j + 1 \right)}{\lceil \widehat{N}_i \rceil^2} \mathbb{I}\left(j \leq \lceil \widehat{N}_i \rceil \right) + \frac{C}{\lceil \widehat{N}_i \rceil^2} \mathbb{I}\left(j > \lceil \widehat{N}_i \rceil \right) \right\} \right] \mid Z_{1i}, Z_{2i}, H_0, \boldsymbol{H}_S \right) \\
&\leq \mathbb{E}\left(\sum_{j=1}^{M_i} \left[\alpha \left\{ \frac{C\left(\lceil \widehat{N}_i \rceil - j + 1 \right)}{\lceil \widehat{N}_i \rceil^2} \mathbb{I}\left(j \leq \lceil \widehat{N}_i \rceil \right) + \frac{C}{\lceil \widehat{N}_i \rceil^2} \mathbb{I}\left(j > \lceil \widehat{N}_i \rceil \right) \right\} \right] \mid Z_{1i}, Z_{2i}, H_0, \boldsymbol{H}_S \right) \\
&\leq \alpha \mathbb{E}\left[\left\{ \frac{C \lceil \widehat{N}_i \rceil \left(\lceil \widehat{N}_i \rceil + 1 \right)}{2 \lceil \widehat{N}_i \rceil^2} + \frac{C\left(M_i - \lceil \widehat{N}_i \rceil \right)}{\lceil \widehat{N}_i \rceil^2} I\left(M_i > \lceil \widehat{N}_i \rceil \right) \right\} \mid Z_{1i}, Z_{2i}, H_0, \boldsymbol{H}_S \right] \\
&\leq \alpha,
\end{aligned}
$$

where $\mathbb{I}(\cdot)$ is an indicator function. The first equation is valid because of the independence of DAC test statistics. The second equality holds since subject i only visits M_i times and the probability of committing a type I error at visit j is determined by the proposed adaptive procedure depending on whether j falls before or after the predicted number of visits $\lceil \widehat{N}_i \rceil$. The following inequality holds because of Jensen's inequality and the last inequality holds since C is chosen to be the reciprocal of $\sum_{j=1}^{\lceil \widehat{N}_i \rceil} \frac{\lceil \widehat{N}_i \rceil - j + 1}{\lceil \widehat{N}_i \rceil^2} = \frac{\lceil \widehat{N}_i \rceil \left(\lceil \widehat{N}_i \rceil + 1 \right)}{2 \lceil \widehat{N}_i \rceil^2}$, and $\lceil \widehat{N}_i \rceil$ is assumed to be unbiased, i.e., $\mathbb{E}\left[\lceil \widehat{N}_i \rceil - M_i \right] = 0$. We can choose c adaptively to be a number greater than 1 to improve power while keeping FWER strictly controlled. However, if the number of visits is small, the loss of power under independent tests is expected to be small even with $c = 1$. Alternatively, we can also replace expected mean survival $\widehat{T}_i$ with median survival time, which is supposedly smaller as a survival time distribution is often right skewed. The median survival time is easier to estimate and will tend to have less stringent family-wise error while controlled at a pre-determined level.

2.4 | Permutation Test

Generally, it can be hard to justify a multivariate normal distribution for recorded data and the testing procedure may fail to follow a χ^2 distribution. Intuitively, the test statistic still measures the incremental departure from the norm at the current visit as compared to the previous visit, though there is no known distribution that we can use to find thresholds for abnormality. Here, we propose an innovative use of bootstrapping and permutation to obtain a series of critical values for the test statistics over time, when we cannot assume multivariate normality.

First, we bootstrap $\boldsymbol{B}$ (i.e. 10,000) participants from the study with replacement. For the b-th bootstrapped participant with M_b visits, we will remove the time effect obtained from model (3) (i.e. $\hat{\beta}_{j0} + \hat{\beta}_{j1} t_{bk} + \hat{\beta}_{j2} t_{bk}^2 + \hat{\beta}_{j3} t_{bk}^3$) to obtain participant-specific errors over M_b visits.

Under the assumption that the covariance matrix $\boldsymbol{\Sigma} = [\rho_{sr}]$, $s, r = 1, \ldots, q$ characterizing cognitive domains follows a compound symmetry structure, we can permute the errors in the following way without disrupting the covariance structure. We first rearrange the errors as a matrix with M_b rows representing all visits and q columns representing all cognitive domains. We then permute the columns in whole. If the compound symmetry is not proper for the data of interest, this column permutation step could be skipped. Next, we permute the rows in whole to preserve the assumed structure.

Then, based on the permuted errors, we obtain an error vector as

$$\boldsymbol{V}_b = \left(\boldsymbol{E}_{b11}, \ldots, \boldsymbol{E}_{bq1}, \boldsymbol{E}_{b12}, \ldots, \boldsymbol{E}_{bq2}, \ldots, \boldsymbol{E}_{b1\boldsymbol{M}_b}, \ldots, \boldsymbol{E}_{bq\boldsymbol{M}_b}\right)^\top.$$

The DAC test statistics can be calculated as $\left\{S_\omega^b\left\{\left(\boldsymbol{E}_{b1\omega}, \ldots, \boldsymbol{E}_{bq\omega}\right)^\top \mathbf{1}_{qm} < 0\right\}, \omega = 1, \ldots, \boldsymbol{M}_b\right\}$ without assuming any specific error distributions. Pooling them together after $\boldsymbol{B}$ bootstraps, we can calculate any empirical proportion of impairment α_0 based on the predicted number of visits from models (4) and (5). The corresponding $(1 - \alpha_0)$ quantile may serve as a critical value. Participant d at the ω-th visit will be identified as cognitively impaired if this visit-specific test statistic exceeds this critical value while $\mathbf{1}_{q\omega}^\top \widehat{\boldsymbol{X}}_\omega^d < \mathbf{1}_{q(\omega-1)}^\top \widehat{\boldsymbol{X}}_{\omega-1}^d$ when $\omega > 1$ or $\mathbf{1}_q^\top \widehat{\boldsymbol{X}}_1^d < 0$.

As we have described in Section 2.3, the factor c, which is used to control how much we can spend α, can be set at 1 under multivariate normal distributions, because the independence of the tests guarantees the loss of power is small. However, this independence becomes questionable when the multivariate normal distribution cannot be justified. As a result, c should be adjusted to preserve power while controlling family-wise errors. Cross-validation (CV) can be used here to determine an appropriate c value. We first randomly divide the healthy reference group by several CV sets. For each set, we apply the MLME model to the remaining participants and calculate DAC test statistics for this fold. Then permutation test statistics are built from the healthy reference group leaving out the fold to be tested. Putting the DAC test statistics together with their corresponding permutation test statistics, an iterative process is used to determine an appropriate c value such that the FWER is controlled just around the pre-determined α level.

3 | NUMERICAL STUDIES

Here, we ran some simulation studies to assess how the proposed DAC method works under various distribution assumptions. As described in Section 4, the MACS study regularly evaluates six cognitive domains, we also considered $q = 6$ cognitive domains in our simulations. First, longitudinal multivariate data were generated with various forms of mean score functions over time. When a multivariate normal distribution was assumed, the DAC testing procedure based on χ^2 was evaluated for its FWER over various levels of α. Otherwise, the permutation testing procedure was used to evaluate DAC. Multivariate t and Gamma distributions were considered for non-normal errors with heavier tails or skew, similar to Wang et al. [6]

Specifically, for these simulations we assumed 1,000 participants with cognitive functioning tested in the past would serve as historical healthy controls. At the same time, we generated longitudinal scores for 1,000 additional participants as the testing group assuming they would enroll in future. For each participant in the healthy control and testing groups, we first simulated their enrollment time uniformly over 20 years. Their survival time follows a Weibull distribution with five covariates,

$$\log(T) = \beta_0 + \sum_{i=1}^{5} \beta_i z_i + \sigma W,$$

where $\boldsymbol{W}$ was generated from the standard extreme value distribution, and covariates were independently generated with z_1-z_4 from the standard normal distribution and z_5 from a uniform (0,1) distribution. Participants were assumed to be censored at year 20 for both the historical and newer cohorts. Based on the simulated duration in the study, each subject's visit times were generated such that the time between two assessments follows an independent exponential distribution with the first visit happening at time 0. The hazard rate of the exponential was determined by Poisson regression from equation (5). We used the same five covariates and set $\gamma_0 = 0$, $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0.1$, and $\gamma_5 = -0.1$, yielding a median of 8 visits. At all the visit times, 6 cognitive domain scores were generated from different multivariate distributions as detailed in the subsections. One thousand simulations were carried out for each scenario.

In practice, the study duration and visit frequency of an ongoing study may differ from its historical cohort. Therefore, we examined various scenarios where the duration of the prospective study changes or when participants visit the study more and less frequently. FWER and power were examined under the null and the alternative, respectively. Meanwhile, as mentioned in Section 2.4, it's important to determine an appropriate c value to relax the thresholds and to preserve power. The relationship between c and the FWER and power was also investigated. The details of each simulation and results are described below.

3.1 | Multivariate Normal Distribution

After obtaining the number of assessments m_i for participant i based on their duration and visit times, we generated six domain scores from the multivariate normal distribution at each visit. For the covariance matrix $\boldsymbol{U}^i_{m_i}$, we set $\sigma^2 = 30$, $\theta^2 = 10$, $\rho_{sr} = 60$, for $s = r$, and $\rho_{sr} = 15$, for $s \neq r$ with $s, r = 1, \ldots, 6$. That is, covariance of different cognitive domains at the same visit is $\theta^2 + \rho_{12} = 25$, covariance of the same cognitive domains at different visits is $\rho_{11} = 60$, and the rest elements are $\rho_{12} = 15$.

Four forms of polynomial mean trends were considered. 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 $50 - 0.6t$, and the other three had means of $50 - 0.8t$. For the quadratic trend, the first three cognitive domains were set to have means of $50 - 0.08t^2 + 0.2t$, and the other three to have means of $50 - 0.06t^2 + 0.1t$. Lastly, for the cubic trend, the first three were set to have means of $50 - 0.008t^3$

$+ 0.08t^2 + 0.55t$ and the rest to have means $50 - 0.007t^3 + 0.06t^2 + 0.55t$, similar to the settings in Wang et al. [6] The `mvrnorm` from the R library MASS was then used to simulate longitudinal cognitive errors following the multivariate normal distribution with means 0 and the covariance matrix $\boldsymbol{U}^i_{m_i}$. The mean polynomial functions mentioned above were then added to the errors to represent the generated longitudinal cognitive scores.

The `lmer` from the library `lme4`, the coxph from the library `survival`, and the `glm` were used to implement models (1), (4) and (5). We adopt cubic polynomial functions for mean scores in the MLME, pretending that the true polynomial functions are unknown. At various levels of α (from 0.001 to 0.1), χ^2 tests were conducted for each simulated participant in the newer cohort. For the purpose of comparison, we also implemented the permutation test here. Results based on 1,000 simulations are summarized in Figure 1. As the results for all four mean trends are almost identical, we only include the result under the cubic mean trend. The estimated FWER from the χ^2 test is denoted by the black solid line, the one from the permutation test is given by the black dash line, and the nominal α level is denoted by the gray broken line. The DAC χ^2 test and the permutation test seem to perform equally well, both successfully controlling FWERs below pre-determined levels for all mean trend functions, when domain scores follow multivariate normal distributions and we can correctly model the visit frequency and multiple longitudinal domain scores. As expected, the departures of the FWERs from the nominal levels are small. The FWERs of four mean functions are all around 0.046 when $\alpha = 0.05$.

3.2 | Multivariate *t* and Gamma Distribution

In practice, real data may present skewness and heavy tails, which do not follow multivariate normal distributions. Thus, we considered two sets of non-normal errors here. One set follows multivariate t distributions for heavy tails, and the other presents negative skewness from correlated Gamma distributions. The same four mean trends from Section 3.1 are used here.

To simulate longitudinal scores with heavy tails, multivariate t distributions with 5, 25 and 50 degrees of freedom were used. The `rmt` from the library `csampling` was used for multivariate t random error generation. Means of the random errors were set to 0, and the covariance matrix used here is the same as $\boldsymbol{U}^i_{m_i}$ from Section 3.1. The four polynomial score trends were then added to the generated errors as the observed longitudinal scores.

To simulate longitudinal scores with negative skew, gamma distributions were used. Again a compound symmetric covariance structure was considered by transforming longitudinal multivariate normal errors to get correlated gamma errors. We first simulated multivariate standard normal errors $\varsigma_{ijk}, j = 1, \ldots, 6, k = 1, \ldots, m_i$ with means 0 and covariance $\boldsymbol{U}^i_{m_i}/100$ from Section 3.1. Three gamma distribution designs were considered. The first one took transformation of $70 - \boldsymbol{\Gamma}^{-1}(\boldsymbol{\Phi}(\varsigma_{ijk}))$, where $\boldsymbol{\Gamma}$ is the cumulative distribution function (CDF) of the gamma distribution with shape of 4 and scale of 5 and $\boldsymbol{\Phi}$ is the CDF of the standard normal distribution. For the second design, we used $100 - \boldsymbol{\Gamma}^{-1}(\boldsymbol{\Phi}(\varsigma_{ijk}))$ as our negative

skewed errors, where Γ has shape of 25 and scale of 2. We calculated $150 - \Gamma^{-1}(\Phi(\varsigma_{ijk}))$ for the third design, where Γ has shape of 100 and scale of 1. Each score error has mean of 0 and variance of 100. The four polynomial mean trends were again added to the generated scores to represent measured longitudinal domain scores with negative skew.

For each setting, survival time and visit frequency were generated in the same way as mentioned at the beginning of Section 3. We assumed 1,000 participants had been measured in the historical healthy control group, while 1,000 more participants were going to enroll in the new study and get tested. As described in Section 2.4, the proposed permutation test was used here. We bootstrapped 10,000 participants with replacement from the historical healthy control group. For each selected participant, we obtained longitudinal errors by subtracting the estimated means from the original scores. Then we rearranged the errors, added back the longitudinal mean scores, and computed a series of DAC test statistics. After repeating 10,000 times, visit-by-visit classification of cognitive impairment is performed for the newer cohort with corresponding quantiles from these permutation test statistics as the thresholds. The results of FWER at various α levels after 1,000 simulations from multivariate t and correlated gamma distributions are summarized in Figure 2. As a comparison, results from the DAC χ^2 tests are also shown in the figures. Figure 2 is under the cubic mean trend, other three mean trends are omitted as they are similar.

As shown in Figure 2, the FWERs from the DAC χ^2 test can be greatly inflated when multivariate normality does not hold. Illustrated by the black curves, FWER inflation is smaller when the multivariate t distribution has less heavier tails or when the gamma distribution is less skewed. When the permutation test is used, FWERs are controlled at any pre-specified α level and noticeably smaller at larger α values. This conservativeness has also been observed by other work. [21] We can relax the thresholds by increasing the factor c through CV as described in Section 2.4, so that FWER is adjusted around the pre-determined level. In Section 3.4, we will discuss how we can adjust the value of c when performing power analysis.

3.3 | Impact of mis-specified visit prediction on FWER

Both the χ^2 and permutation tests rely on the predicted number of visits. In this section, we evaluate the impact of mis-specified visit prediction on FWER. Caudill and Mixon [22] proposed a censored Poisson regression model (CPR) for right-censored count data, and Famoye and Wang [23] proposed a censored generalized Poisson regression model (CGPR) with an extra dispersion parameter. The performance of these two single models was assessed relative to the two-model approach discussed in Section 2.2. We first compared the predicted number of assessments using these three methods and then evaluated the resulting FWER from the χ^2 and permutation tests based on the three predicted numbers.

Specifically, we calculated the mean differences between the estimated numbers of assessments using the three methods and the true observed numbers in each simulated dataset of the 1000 simulations, adopting the same setting as described in Section For the two Poisson models, we used the logarithm of of the simulated event time as an offset. The

results are given in Figure 3 and Table 1, which show that the two-model method has the best performance among the three methods.

Next, under the same settings as in Section 3.1, we calculated FWER for the χ^2 test and the permutation test at a range of α's, using the predicted number of visits from the censored Poisson regression model and the two-model method. The results from the generalized Poisson model are similar and not illustrated here. Figure 4 displays the resulting FWER and shows that the one calculated using the two-model method is slightly closer to α compared to that using the censored Poisson regression model.

We then evaluated how the DAC χ^2 and permutation tests perform when the proportional hazards (PH) survival model or the Poisson frequency model or both are mis-specified in predicting the number of visits. The simulations setting is the same as that in Section 3.1 expect for the generation of visit numbers for each individual. In this new setting, we generated the survival times from a log-logistic model: $\log(T) = \beta_0 + \sum_{i=1}^{5} \beta_i z_i + \sigma W$, where W follows the standard logistic distribution. As this model has the proportional odds interpretation, we refer to it as the PO model. We generated the gap times between two adjacent assessments for subject i from a gamma-exponential distribution $Exp(r_i\Lambda_i)$ to account for overdispersion, where $r_i \sim Gamma(\delta^{-1}, \delta^{-1})$. We set $\delta = 1$ and $\Lambda_i = \exp(\gamma_0 + \gamma_1^T Z_{2i})$. The resulting number of visits by time t follows a Negative-Binomial process: $NB\left(\delta^{-1}, \frac{\Lambda_i t}{\delta^{-1} + \Lambda_i t}\right)$, according to McShane et al. [24] Since $E(Y_i(t)) = \Lambda_i t$ and $V_{ar}(Y_i(t)) = \Lambda_i t + \delta(\Lambda_i t)^2$, δ can be interpreted as the overdispersion parameter.

We then calculated the differences between the estimated numbers of assessments using the proposed two-model approach and the true simulated numbers under four different scenarios: PH survival and Poisson frequency models, PH survival and NB frequency models, PO survival and Poisson frequency models, and PO survival and NB frequency models. The results are summarized in Table 2. When the survival time is generated from a PO model while a PH model is used in prediction, the differences seem to be close to those when the numbers of visits are generated from a Weibull (PH) survival model and a Poisson frequency model. In contrast, when the frequency follows a NB model but is mis-fit with a Poisson model, we observe larger differences between the predicted numbers and the true simulated numbers. When both survival and frequency models are mis-specified, the effect seems to be dominated by the mis-specification of the frequency model. Figure 5 displays the FWER from the χ^2 test at a range of α's under these four scenarios. Consistent with what we have observed in Table 2, the mis-specification of survival models has a minimal effect on FWER, and ignoring overdispersion in the frequency data leads to somewhat more conservative FWER. Not shown here, the mis-specification in the prediction models has a similar effect on the permutation test. Overall, the proposed procedures appear robust against the mis-specification of survival models and frequency models. We thus use the two-model approach in all the subsequent analyses.

3.4 | Different Number of Visits and Power Analysis

In practice, multivariate normality is often violated with collected measurements. In addition duration and visit frequency of a future study may differ from historical ones. For example, in the MACS study, there were several enrollment waves over time and the MACS proposed some study design changes in mid-2000 so participants started to visit the centers less frequently than in 1990's. Thus, in this subsection, we examined how the FWER and power of the proposed DAC methods may be affected by different study duration and visit frequencies in the newer testing cohort. Five different designs were evaluated for the testing group by changing the study duration (survival distributions remained the same as in the previous two subsections) and the visit frequencies through Poisson regression parameters. For the historical healthy control group, the survival time and visit frequency remain the same as in the previous two subsections. The distributions of five covariates remain the same as well. Below detail the five settings for the newer cohort:

1. **Original:** In the Weibull model used for survival time, we set $\beta_0 = 3$, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.2$, $\beta_5 = -0.2$ and $\sigma = 0.1$. Participants were set to enroll uniformly until the study ended in 20 years. In the Poisson model used for assessment intensity Λ, we set $\gamma_0 = 0$, $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0.1$, and $\gamma_5 = -0.1$. The median number of visits is 8.
2. **Shorter Study:** Same as setting (1), except that we ended the study in 10 years. The median number of visits is 4.
3. **Longer Study:** Same as setting (1), except that we ended the study in 30 years. The median number of visits is 10.
4. **Less Visits:** Same as setting (1), except that we divided the visit intensity Λ by 2. The median number of visits is 4.
5. **More Visits:** Same as setting (1), except that we multiplied the visit intensity Λ by 2. The median number of visits is 16.

The multivariate t with 5 degrees of freedom was used here to generate longitudinal scores in the historical healthy control group and in the testing group from the prospective study. For both groups under the null, the setup is the same as in Section 3.1. We first simulated the number of visits based on each scenario of survival and assessment frequency. Then we generated multivariate t errors and added with quadratic mean trends from Section 3.1. For the testing group under the alternative, the setup is the same except for mean trends. We specified first three cognitive domains to have means $20 - 0.08t^2$, and the other three to have means $50 - 0.1t^2$. One thousand participants were assumed to have enrolled in the historical study, and 1,000 participants were expected to enroll in the new study. As mentioned in Section 2.4 and shown in Figure 2, the FWER is conservative with the permutation test, and CV can be used to determine an appropriate factor c to increase the thresholds. After 5-fold CV on 1,000 simulations, we determined that $c = 2.0$ for the new study under the original study design. In predicting the number of assessments within the newer study, we assumed that we had known policy changes such as when the study ended and how the frequency of visits changed. The results of FWER and power from the permutation test (10,000 times) after 1,000 simulations under five cases are shown in Figure 6.

As illustrated in Figure 6, different study duration or assessment frequencies have an impact on FWER and power. Fewer visits in the newer study, from a shorter study period or lower visit requirements, seem to have inflated FWER and larger power. However, the FWER inflation is in general small and cases (2)-(5) are quite different from case (1). Thus, it remains important to make sure the visit frequency does not deviate too much from the historical study. At the same time, the factor c used to relax the thresholds can be increased or reduced slightly based on researchers' understanding of how study policy changes the assessment frequency for the future study.

4 | APPLICATION TO THE MULTICENTER AIDS COHORT STUDY

Here, the proposed DAC method was applied to the neuropsychological (NP) sub-study data collected through 2017 from the Multicenter AIDS Cohort Study (MACS) which began in 1984 and merged with the WIHS to become the MWCCS on 4/1/2019. The MACS has been administered by Johns Hopkins University, Northwestern University, The University of California at Los Angeles, and the University of Pittsburgh. [25,26] More than 7,000 men who have sex with men (MSM) have been recruited in the study. Participants were either infected with HIV or at risk for infection at enrollment. They have been regularly interviewed and examined on a wide range of variables, such as drug use, depressive symptoms, age, sexual disorder, cognitive functioning, and physical measurements. Participants' cognitive functioning is negatively impacted by HIV infection. However, highly active antiretroviral therapy (HAART) was found to have positive effects on cognitive functioning among people infected with HIV since its first availability in early 1990's. Participants in the NP substudy have been regularly evaluated with a NP test battery for six cognitive domains, including motor speed & coordination, speed of information processing, executive functioning, learning, memory, and working memory & attention. [27,28] These test scores provide a rare opportunity to assess, in the HAART era, how those infected with HIV and those without the infection differ in cognitive decline over time.

The MACS has four enrollment waves starting at 1984, 1987, 2001 and 2010. We took the first two pre-HAART cohorts as our historical data, and used data from the two more recent enrollment cohorts prospectively to examine how cognitive impairment is developing over time after HARRT. During each NP visit, a battery of tests was carried out, and collected scores were summarized by T-scores, which were computed from regression models adjusting for ethnicity, education, age, and the number of tests administered. The T-scores have means of 50 and standard deviations of 10. For motor speed & coordination, the lowest T score is used for summary, while the other five use the arithmetic means of all T-scores in each specific domain as summary T-scores. The multivariate normal distribution assumption on the NP data from the MACS is of concern, because the summary T-score in motor speed & coordination is very skewed. Consequently, the permutation test should work better for identifying cognitive impairment with FWER controlled at a pre-specified level.

For this analysis, only participants with complete scores on six cognitive domains were included. We used participants prior to 2001 as the historical cohort for visit-by-visit cognitive impairment classification starting 2001. Prior to 2001, 1,231 men were infected with HIV, while 870 were not. Five-fold CV on the 870 participants without HIV infection

was used to find an appropriate factor c to relax the thresholds on permutation testing. For each CV set, we used Cox proportional hazard regression to model survival and Poisson regression to model visit frequencies with all historical cohort participants except for the hold-out set. For survival modeling, participants were censored at 4 years past the last NP visit or 2001, whichever came earlier, if death was not observed or death happened beyond 4 years past the last NP visit or 2001. Covariates in Cox regression included CD4 cell count along with its quadratic transformation, age at the first NP visit, Center for Epidemiologic Studies Depression (CESD) score, hepatitis C status, testing centers, and HIV serostatus. We controlled for CD4 cell count and its quadratic transformation, age, CESD, and HIV serostatus in Poisson regression.

For each CV set, the MLME was applied to the rest of historical healthy controls (i.e., participants without HIV infection in the earlier cohort) to estimate the mean trends and covariance structure. Based on the frequency prediction and MLME results, we treated each fold as if they were a newer cohort and conducted visit-by-visit classification of cognitive impairment using the DAC permutation test. After summarizing the rates from the five folds of healthy controls from the historical study, we found that we can relax the factor c to 1.4 while keeping FWER around the pre-specified level. The results from the historical controls are illustrated in Figure 7 using this cross-validated c of 1.4.

Then, we applied the DAC to those participants enrolled from 2001, where 803 were infected with HIV and 796 were not. However, due to study policy changes, participants on average halved their frequencies to the NP substudy since 2001. Given this knowledge, we also halved the predicted frequencies for these participants. We truncated the predicted survival times by 2017 of the data freeze. The results for the newer cohorts are also shown in Figure 7 using the permutation test and the factor c of 1.4 from CV. First, we can observe that the newer cohorts have significantly more people with cognitive impairment identified as compared to historical healthy controls. Second, since 2001, the impairment rates between seronegative and seropositive groups are not significantly different if we set $\alpha = 0.05$. This is consistent with the findings from Wang et al. [5,6] Table 3 shows the mean scores for all six cognitive domains at visits when participants were evaluated around the same time. Because participants were measured at a roughly half frequency in the newer cohorts due to policy changes since 2001, the visits in Table 3 were selected from the old and new cohorts to have comparable duration since the first visit. Across three comparison points, we can see that domains scores from the newer cohorts are generally lower than those from the historical seronegative group, especially in the motor speed & coordination domain for which the difference becomes more profound as time progresses. It has also been noted by other studies that seronegative participants in the newer cohorts are in general less healthy than the historical controls. [29] However, seronegative and seropositive participants in the newer cohorts do not seem to differ that much, which is consistent with what we have observed in Figure 7.

5 | DISCUSSION

The LMNC proposed by Wang et al [6] considered a research setting, where data have been collected and we only need one classification of prior impairment for each patient

by looking at retrospective data all together. As a comparison, the proposed DAC method has greater utility by affording visit-by-visit diagnosis. To facilitate the implementation of our methods, we have posted our code at https://github.com/tlwangzi123/Dynamic-Arrayed-Comparison and developed an R Shiny app (https://lingyun-shiny.shinyapps.io/DACShinnyApp/) to display individual scores over time and to illustrate dynamic classification using simulated domain scores, as the MACS NP data are not in the public domain. It is worth mentioning that the DAC method, along with the LMNC method, has broad applicability for classification, as long as measurements collected are longitudinal in nature and dynamic classification is of interest. The app can be easily adapted for other applications.

The proposed DAC method can effectively control FWER. Multivariate normality is an important assumption for the χ^2 test, although FWER is slightly lower than the pre-determined level. Since the χ^2 test is much easier to compute, we suggest to use the χ^2 test when the normality assumption is valid. When such an assumption is violated, permutation test can also control FWER as shown in the simulation studies. It remains critical to select an appropriate factor to relax the thresholds while keeping FWER under control. CV is an effective way to utilize the data collected from historical healthy controls to choose this factor. However, it does not generalize well to a new study when study duration or visit frequency are markedly different from the historical controls. Researchers must decide whether to adjust such factors and how much to adjust before conducting the DAC method.

The independence and the distribution of the DAC test statistics at visit ω have been established assuming that domain scores are available at all visits up to ω. This assumption can be relaxed for the test procedures based on the χ^2 distribution. The MLME can easily handle missingness in domain scores, the independence proof does not rely on the dimensions of $\boldsymbol{X}_{\omega-1}$ and $\boldsymbol{W}_{\omega}$, and the χ^2 distribution still holds, though the degree of freedom changes with the length of available domain scores. However, it is more challenging to extend our proposed methods to handle incomplete domain scores when data do not follow a multivariate normal distribution. The permutation procedure fails to work with missing values because different scores at different visits become missing and we cannot build a permutation test statistic with missing elements in the vector. Other inference procedures need to be developed instead. As missing domains are common in HIV research, the extension to handle incomplete data is of interest for future research.

In this article we focus on dynamic classification of impairment of individual participants during the course of a longitudinal study. Some participants may return to normal cognitive functioning due to random variation, regression to the mean, or treatment and they will remain under monitoring for future cognitive decline. With ongoing testing, intuitively it is reasonable to add some “reward” to the significance level, or the α wealth, that has been spent at the first detection, analogous to the ideas of α investment in Foster and Stine [30] and Aharoni and Rosset. [31] In reality, more severe patients with extreme domain scores tend to be followed more frequently. In our proposed method, the number of assessments is predicted using baseline covariates, which are all determined at the start of a study. To take into account the domain scores, one would need to predict the number of assessments dynamically, which makes the adjustment of the testing procedures (to control FWER) much

more complicated. These are beyond the scope of this article and remain interesting future research topics.

ACKNOWLEDGMENTS

The authors thank the Associate Editor and two referees for their helpful comments which lead to an improved manuscript. The work was partially supported by the NSF DMS −1916001 to Cheng and the University of Pittsburgh Center for Research Computing through the resources provided. Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS), now the MACS/WIHS Combined Cohort Study (MWCCS). 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). MWCCS (Principal Investigators): Atlanta CRS (Ighovwerha Ofotokun, Anandi Sheth, and Gina Wingood), U01-HL146241; Baltimore CRS (Todd Brown and Joseph Margolick), U01-HL146201; Bronx CRS (Kathryn Anastos and Anjali Sharma), U01-HL146204; Brooklyn CRS (Deborah Gustafson and Tracey Wilson), U01-HL146202; Data Analysis and Coordination Center (Gypsyamber D'Souza, Stephen Gange and Elizabeth Golub), U01-HL146193; Chicago-Cook County CRS (Mardge Cohen and Audrey French), U01-HL146245; Chicago-Northwestern CRS (Steven Wolinsky), U01-HL146240; Northern California CRS (Bradley Aouizerat, Jennifer Price, and Phyllis Tien), U01-HL146242; Los Angeles CRS (Roger Detels and Matthew Mimiaga), U01-HL146333; Metropolitan Washington CRS (Seble Kassaye and Daniel Merenstein), U01-HL146205; Miami CRS (Maria Alcaide, Margaret Fischl, and Deborah Jones), U01-HL146203; Pittsburgh CRS (Jeremy Martinson and Charles Rinaldo), U01-HL146208; UAB-MS CRS (Mirjam-Colette Kempf, Jodie Dionne-Odom, and Deborah Konkle-Parker), U01-HL146192; UNC CRS (Adaora Adimora), U01-HL146194. The MWCCS is funded primarily by the National Heart, Lung, and Blood Institute (NHLBI), with additional co-funding from the Eunice Kennedy Shriver National Institute Of Child Health & Human Development (NICHD), National Institute On Aging (NIA), National Institute Of Dental & Craniofacial Research (NIDCR), National Institute Of Allergy And Infectious Diseases (NIAID), National Institute Of Neurological Disorders And Stroke (NINDS), National Institute Of Mental Health (NIMH), National Institute On Drug Abuse (NIDA), National Institute Of Nursing Research (NINR), National Cancer Institute (NCI), National Institute on Alcohol Abuse and Alcoholism (NIAAA), National Institute on Deafness and Other Communication Disorders (NIDCD), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute on Minority Health and Health Disparities (NIMHD), and in coordination and alignment with the research priorities of the National Institutes of Health, Office of AIDS Research (OAR). MWCCS data collection is also supported by UL1-TR000004 (UCSF CTSA), UL1-TR003098 (JHU ICTR), UL1-TR001881 (UCLA CTSI), P30-AI-050409 (Atlanta CFAR), P30-AI-073961 (Miami CFAR), P30-AI-050410 (UNC CFAR), P30-AI-027767 (UAB CFAR), and P30-MH-116867 (Miami CHARM).

The authors gratefully acknowledge the contributions of the study participants and dedication of the staff at the MWCCS sites.

References

1. Antinori A, Arendt G, Becker JT, et al. Updated research nosology for HIV-associated neurocognitive disorders. Neurology 2007; 69(18): 1789–1799. [PubMed: 17914061]
2. Gisslén M, Price RW, Nilsson S. The definition of HIV-associated neurocognitive disorders: are we overestimating the real prevalence?. BMC Infectious Diseases 2011; 11(1): 356. [PubMed: 22204557]
3. Huizenga HM, Smeding H, Grasman RPPP, Schmand B. Multivariate normative comparisons. Neuropsychologia 2007; 45(11): 2534–2542. [PubMed: 17451757]
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. [PubMed: 25587908]
5. Wang Z, Molsberry SA, 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. [PubMed: 31335803]
6. Wang Z, Cheng Y, Seaberg E, et al. Longitudinal Multivariate Normative Comparisons. Statistics in Medicine 2021; 40(6): 1440–1452. [PubMed: 33296952]
7. Klasnja P, Hekler EB, Shiffman S, et al. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. Health Psychology 2015; 34(S): 1220–1228.

8. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Annals of Behavioral Medicine 2018; 52(6): 446–462. [PubMed: 27663578]
9. Verbeke G, Fieuws S, Molenberghs G, Davidian M. The analysis of multivariate longitudinal data: A review. Statistical Methods in Medical Research 2014; 23(1): 42–59. [PubMed: 22523185]
10. Bloxom B A constrained spline estimator of a hazard function. Psychometrika 1985; 50(3): 301–321.
11. De Boor C A practical guide to splines New York, NY: Springer. 2001.
12. Shumaker L Spline Functions: Basic Theory Cambridge: Cambridge University Press. 2007.
13. 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. Journal of Statistical Computation and Simulation 2015; 85(4): 777–793.
14. Harrell FE Jr. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis New York, NY: Springer. 2015.
15. Mattos TB, Avila Matos L, Lachos VH. A semiparametric mixed-effects model for censored longitudinal data. Statistical Methods in Medical Research 2021; 30(12): 2582–2603. [PubMed: 34661487]
16. Fang H, Tian G, Xiong X, Tan M. A multivariate random-effects model with restricted parameters: application to assessing radiation therapy for brain tumours. Statistics in Medicine 2006; 25(11): 1948–1959. [PubMed: 16220474]
17. Fieuws S, Verbeke G. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. Biometrics 2006; 62(2): 424–431. [PubMed: 16918906]
18. Bonferroni C. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze 1936; 8: 3–62.
19. Dunn OJ. Estimation of the medians for dependent variables. The Annals of Mathematical Statistics 1959: 192–197.
20. Dunn OJ. Multiple comparisons among means. Journal of the American statistical association 1961; 56(293): 52–64.
21. Berger VW. Pros and cons of permutation tests in clinical trials. Statistics in Medicine 2000; 19(10): 1319–1328. [PubMed: 10814980]
22. Caudill SB, Mixon FG. Modeling Household Fertility Decisions: Estimation and Testing of Censored Regression Models for Count Data. Empirical Economics 1995; 20: 183–196. [PubMed: 12320575]
23. Famoye F, Wang W. Censored generalized Poisson regression model. Computational Statistics & Data Analysis 2004; 46(3): 547–560.
24. McShane B, Adrian M, Bradlow ET, Fader PS. Count models based on Weibull interarrival times. Journal of Business & Economic Statistics 2008; 26(3): 369–378.
25. Kingsley L, Kaslow R, Rinaldo CJR, Detre K, Odaka N, Vanraden M. Risk factors for seroconversion to human immunodeficiency virus among male homosexuals. The Lancet 1987; 329: 345–349.
26. 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. American Journal of Epidemiology 1987; 126(2): 310–318. [PubMed: 3300281]
27. Farinpour R, Miller E, S P, et al. Psychosocial Risk Factors of HIV Morbidity and Mortality: Findings from the Multicenter AIDS Cohort Study (MACS). Journal of Clinical and Experimental Neuropsychology 2003; 25(5): 654–670. [PubMed: 12815503]
28. Popov M, Molsberry S, Lecci F, et al. Brain structural correlates of trajectories to cognitive impairment in men with and without HIV disease. Brain Imaging and Behavior 2020; 14(3): 821–829. [PubMed: 30623289]
29. Becker J, Kingsley L, Molsberry S, et al. Cohort Profile: Recruitment cohorts in the neuropsychological substudy of the Multicenter AIDS Cohort Study. International Journal of Epidemiology 2014; 44(5): 1506–1516. [PubMed: 24771276]

30. Foster DP, Stine RA. α-investing: a procedure for sequential control of expected false discoveries. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2008; 70(2): 429–444.

31. Aharoni E, Rosset S. Generalized α-investing: definitions, optimality results and application to public databases. Journal of the Royal Statistical Society: Series B: Statistical Methodology 2014: 771–794.

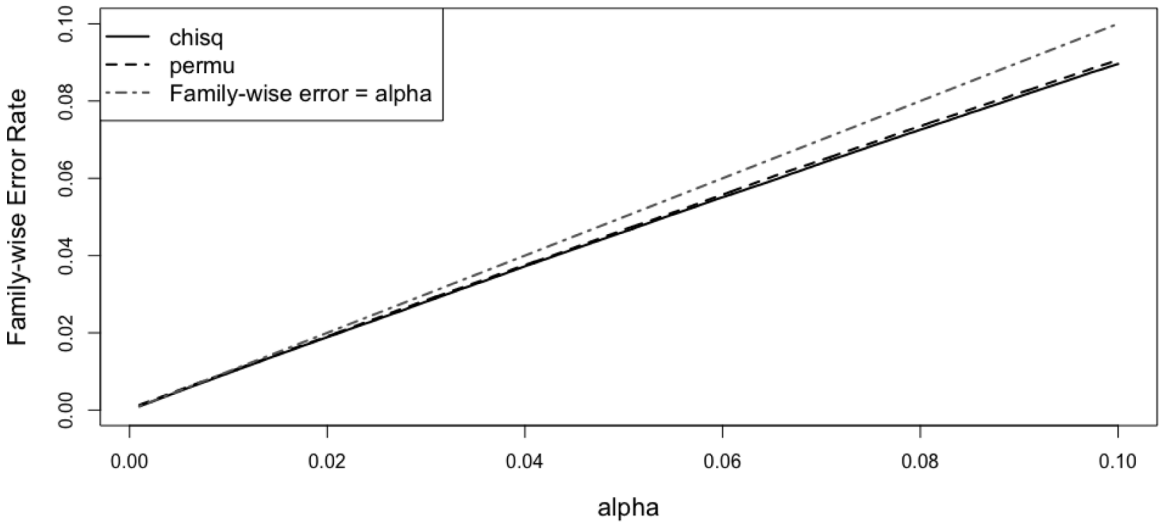

FIGURE 1.
The DAC χ^2 and permutation tests when data follow multivariate normal distributions

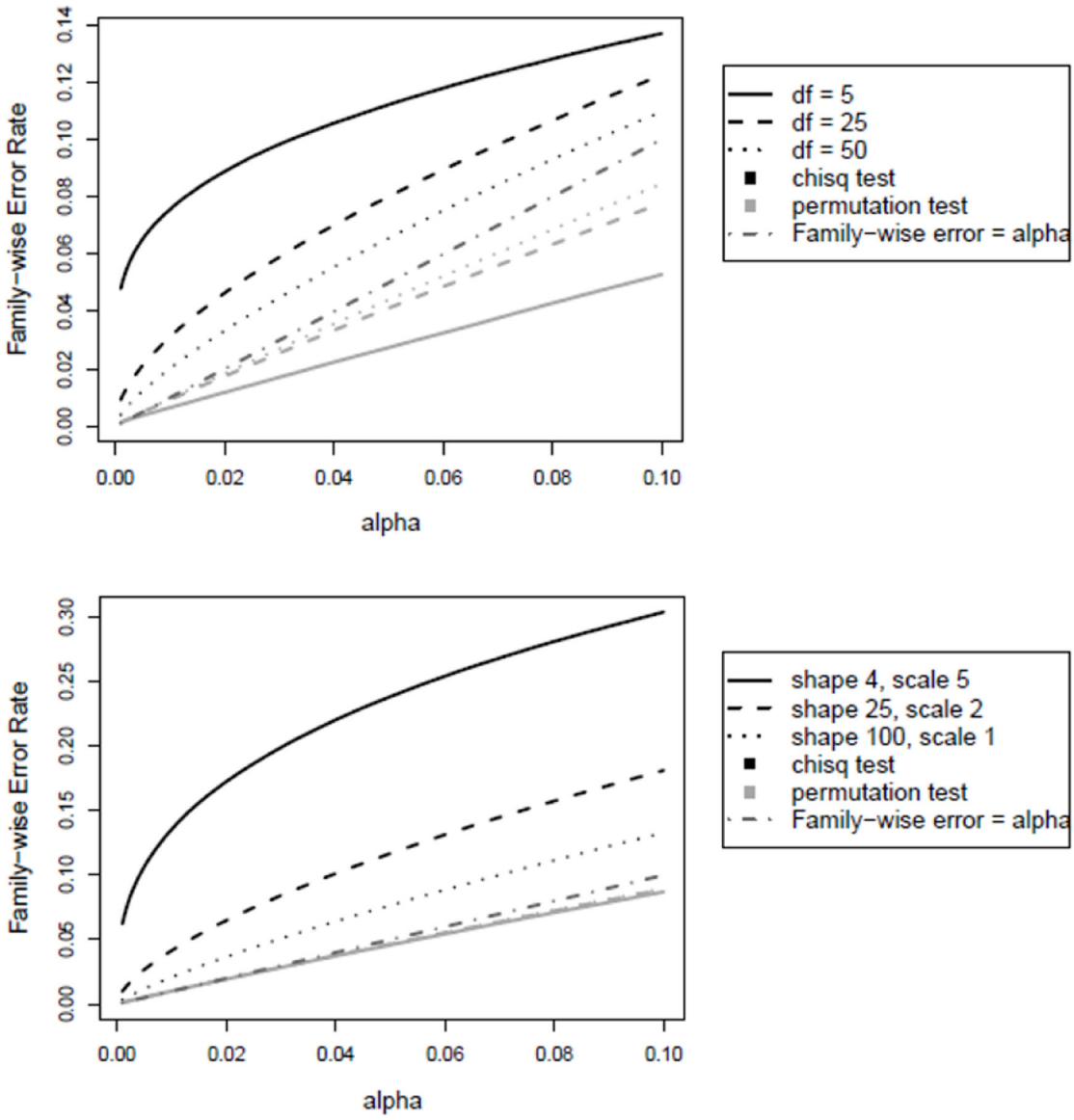

FIGURE 2.

The DAC χ^2 and permutation tests; the upper panel is when data follow multivariate t distributions; the bottom panel is when data are transformed from *Gamma* distributions

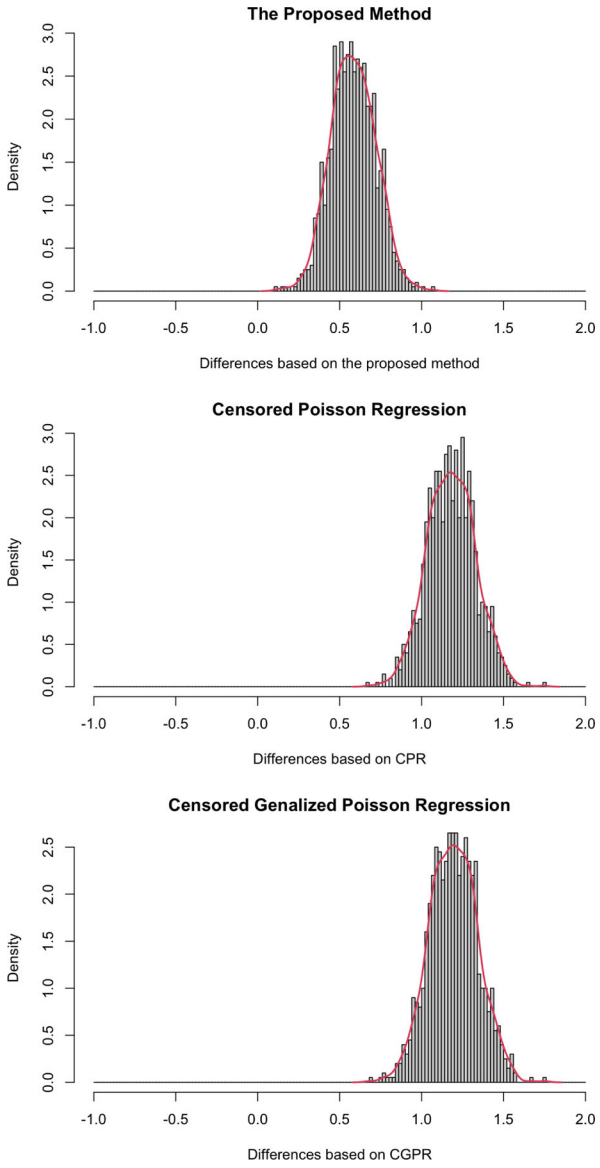

FIGURE 3.
Histograms of the **mean** differences in simulations using the three methods

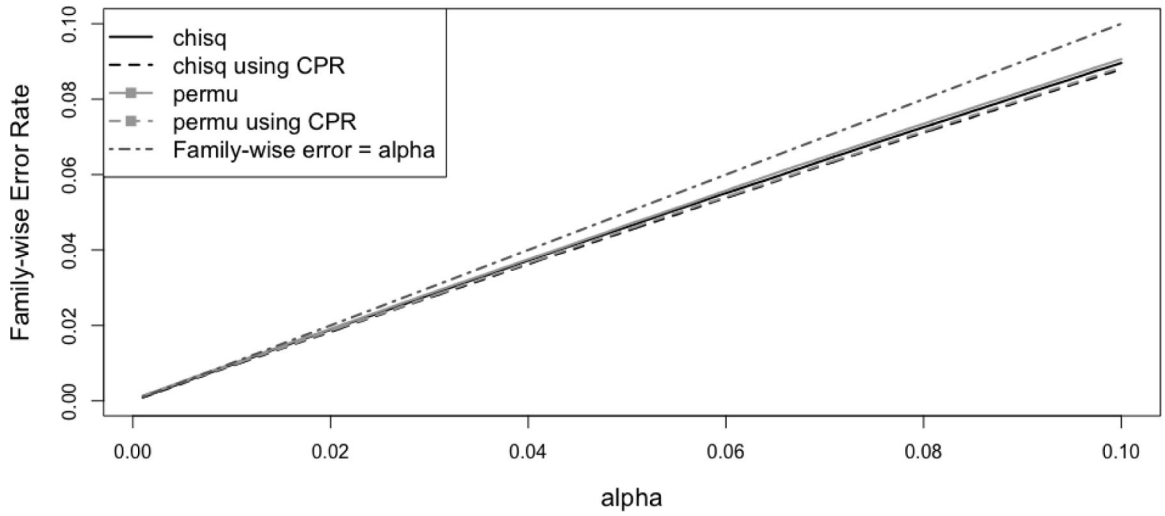

FIGURE 4.

The DAC χ^2 and permutation tests based on the predicted numbers of visits from the two-model approach and the CPR

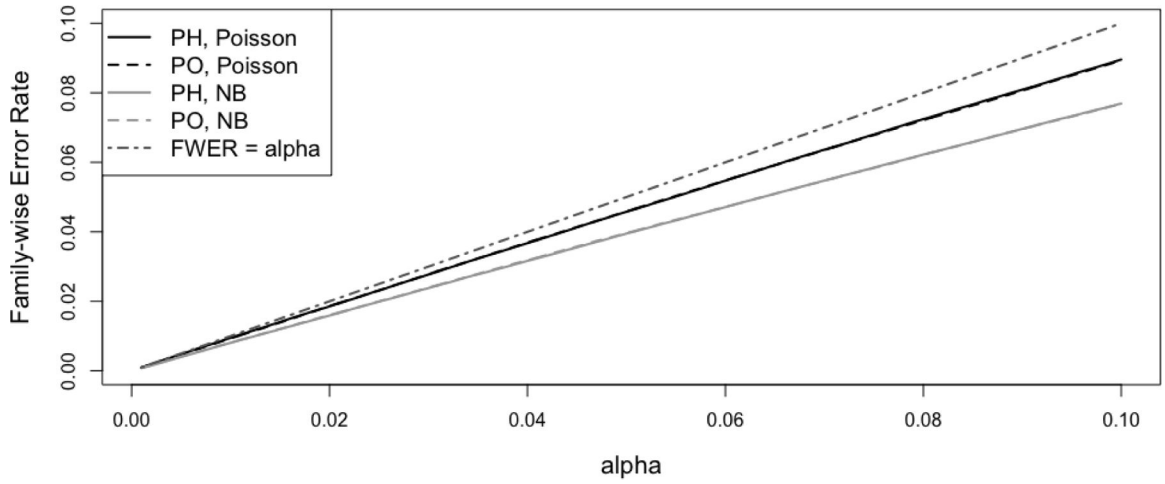

FIGURE 5.

The DAC χ^2 test based on the predicted number of visits from the PH-Poisson models when true survival follows a PH or PO model and frequency follows a Possion or NB model

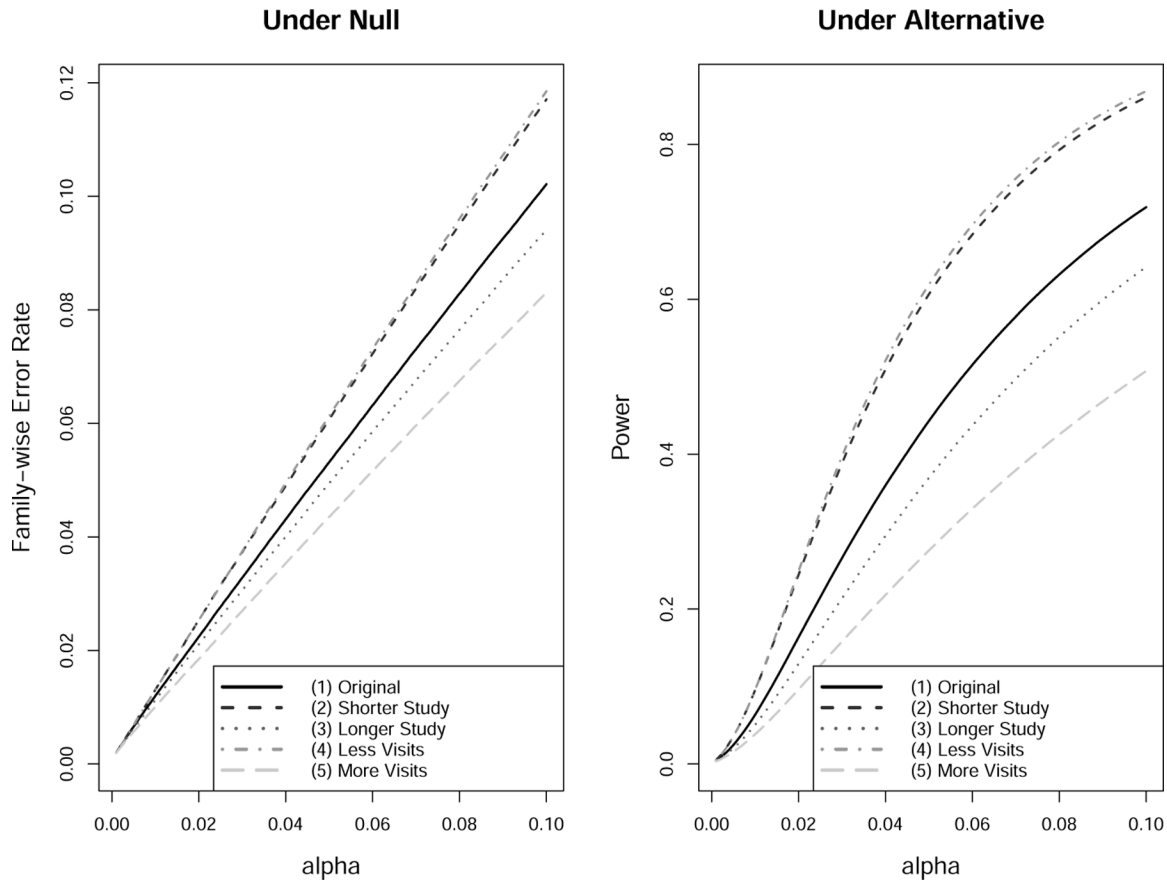

FIGURE 6.

FWER and power of the DAC permutation tests when data follow multivariate *t* distributions and newer cohort has different study period or visit frequency (Case (2) and Case (4) are close to each other)

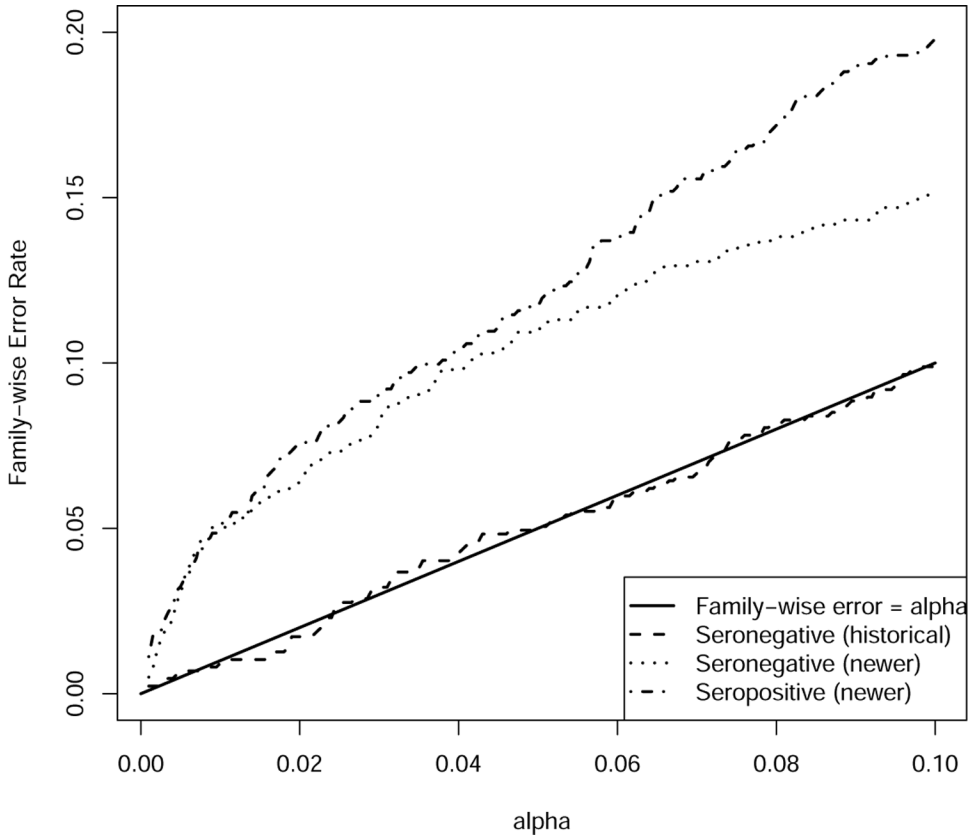

FIGURE 7.
Comparing proportion of cognitive impairment in the MACS

TABLE 1

Summaries of the **mean** differences in simulations using the three methods

Quantities	The Proposed Method	CPR	CGPR
Estimated Mean	0.582	1.180	1.194
Empirical SD	0.136	0.146	0.147

TABLE 2

Summaries of the differences between predicted and simulated numbers under different time and frequency models with each quantity averaged over 1000 simulations

Quantities	PH, Poisson	PO, Poisson	PH, NB	PO, NB
Min	−12.912	−12.913	−110.471	−97.051
Mean	0.585	0.613	0.627	0.596
Median	0.982	0.992	2.277	2.049
Max	11.845	11.985	27.574	26.737
SD	3.097	3.136	12.397	11.381

TABLE 3

Mean scores of six cognitive domains for seronegative and seropositive groups at comparable times

Group (visit, % available)	Motor	Executive	Speed	Learning	Memory	Memory Memory
Seronegative (historical, visit 1, 100%)	48.06	50.18	50.45	51.08	51.31	49.28
Seronegative (newer, visit visit 1, 100%)	46.10	49.41	49.34	48.12	48.35	50.04
Seropositive (newer, visit visit 1, 100%)	46.11	49.17	48.42	48.78	48.78	50.00
Seronegative (historical, visit 5, 46%)	50.34	52.12	51.11	50.98	51.02	51.08
Seronegative (newer, visit3.68%)	44.04	49.42	49.54	48.39	48.54	49.10
Seropositive (newer, visit 3, 72%)	44.21	47.99	48.31	48.57	48.21	48.74
Seronegative (historical, visit 9, 16%)	50.51	54.72	52.21	52.29	51.45	52.55
Seronegative (newer, visit 5, 44%)	42.81	50.06	50.14	48.68	48.85	47.94
Seropositive (newer, visit 5, 53%)	42.78	48.57	48.40	48.41	48.64	46.92