Pathway testing for longitudinal metabolomics

Ebrahimpoor, M.; Spitali, P.; Goeman, J.J.; Tsonaka, R.

Received: 26 November 2019 | Revised: 19 February 2021 | Accepted: 4 March 2021

DOI: 10.1002/sim.8957

RESEARCH ARTICLE

Statistics in Medicine WILEY

Pathway testing for longitudinal metabolomics

Mitra Ebrahimpoor[1] | **Pietro Spitali**[2] | **Jelle J. Goeman**[1] | **Roula Tsonaka**[1]

[1]Medical Statistics, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

[2]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

Correspondence
Mitra Ebrahimpoor, Medical Statistics, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden 2333 ZA, The Netherlands.
Email: m.ebrahimpoor@lumc.nl

Funding information
FP7 Ideas: European Research Council, Grant/Award Number: 305121; Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: 639.072.412

We propose a top-down approach for pathway analysis of longitudinal metabolite data. We apply a score test based on a shared latent process mixed model which can identify pathways with differentially progressing metabolites. The strength of our approach is that it can handle unbalanced designs, deals with potential missing values in the longitudinal markers, and gives valid results even with small sample sizes. Contrary to bottom-up approaches, correlations between metabolites are explicitly modeled leveraging power gains. For large pathway sizes, a computationally efficient solution is proposed based on pseudo-likelihood methodology. We demonstrate the advantages of the proposed method in identification of differentially expressed pathways through simulation studies. Finally, longitudinal metabolite data from a mice experiment is analyzed to demonstrate our methodology.

KEYWORDS
global test, joint latent process, longitudinal analysis, mixed model, pseudo likelihood

1 | INTRODUCTION

Assessing metabolite changes in healthy and diseased subjects over a follow-up period can improve our understanding of disease progression and severity. Linear mixed effects models, multivariate regression models or the generalized estimation equation method are common approaches to model the dynamics of each metabolite separately.[1,2] Based on such models we can identify which specific metabolites exhibit differential progression between experimental conditions or groups of interest. An alternative approach to analyze the high-dimensional metabolite data is pathway analysis. Pathways are a priori-defined sets of genomic features that are involved in a chemical reaction leading to a certain product or change.[3] Joint analysis of the metabolites in a pathway facilitates the biological interpretation of the findings and increases statistical power.[4] Statistical methods for pathway analysis can be classified into two general categories: bottom-up and top-down.[5] The bottom-up methods start with modeling the features individually and then aggregate the individual *P*-values to get a pathway-level *P*-value (using various methods such as Fisher's exact test, or Simes test). These approaches do not take into account the correlation structure within the pathway and also lose information by only using the *P*-value and not the full measured data. In addition, some of these approaches dichotomize the *P*-value (ie, significant/nonsignificant) before testing the association leading to loss of information. By contrast, top-down methods use all recorded measurements jointly to get pathway-level tests, for example, global test and gene-sets net correlations analysis.[6,7] Global test is a score-test which can assess the association between a set of features with a clinical outcome,

for example, binary or survival. In particular, based on an appropriate regression model, it tests the hypothesis that at least one of the features in the set is associated with the clinical outcome. Global test has desirable power properties when features are correlated which is the case for pathways.

The aforementioned top-down pathway methods have been developed for cross-sectional designs, and extending them to the longitudinal setting is not straightforward. Complications arise when the longitudinal features are missing at some time-points or when measurements are collected at irregular points in time. In addition, in longitudinal designs the goal is to test for changes over time at the pathway level, which requires explicit modeling of the progression of all the features in the pathway. This is also the case in our motivating longitudinal experiment designed to study the dynamics of Duchenne muscular dystrophy (DMD). DMD is a severe pediatric neuromuscular disorder caused by a lack of dystrophin. In this experiment, 132 metabolites are measured at five time-points in a 7-month period from 21 dystrophin-lacking and five wild type (WT) mice. These metabolites are classified in 11 pathways from the Wikipathways database and our goal is to identify the biological pathways exhibiting differential progression over time between the mice groups. Pathway testing, in this case, is complicated by the within-subject correlations, missing values, small sample size and the high number of metabolites in some pathway. To our knowledge, there are currently no pathway analysis approaches that collectively model the progression of metabolic pathways as a single dynamic system.

We follow a top-down approach and propose a pathway-level test which extends the global test to the longitudinal setting. In particular, we propose a two-step procedure. In the first step, a shared latent process is used to model the longitudinal progression and between-metabolite correlations *at the same time* at pathway level.[8] The longitudinal progression of the pathway for each individual is then summarized in terms of metabolite- and subject-specific shared random-effects. In the second step, the estimated random-effects are used as covariates in the global test. We call this method shared latent process mixed effects modeling within global test (SLaPMEG). In practice, the size of the pathway can vary from 1 to even a hundred metabolites. For larger pathways, estimation of the joint model at step 1 can be challenging. For such cases, an additional novel contribution is a computationally efficient solution we propose based on pseudo-likelihood methodology. In particular, we propose a fast approach to derive the predicted values for the random-effects and avoid the numerical challenge of optimizing the joint model of all metabolites in the pathway.

The strengths of our proposed method are that (1) it exploits the within and between metabolites correlations, (2) it can be applied in unbalanced designs potentially with missing values, (3) it can be used to analyze very large pathways, (4) it provides an insight into the most influential metabolites in the pathway by inspecting the plots of the estimated random-effects, (5) it can potentially handle any clinical outcome in the generalized linear model family and survival outcomes, and (6) it is implemented in the R package *SLaPMEG* available online.

The rest of the article is organized as follows. In Section 2, we formulate the problem of testing the association of a longitudinal pathway with a clinical binary outcome and introduce the SLaPMEG method. In Section 3, we present a two-step likelihood estimation procedure for the parameters of the model. In Section 4, we propose a computationally efficient solution based on pseudo-likelihood, to enhance the analysis of pathways with many features. In Section 5, we empirically evaluate the performance of our method in terms of type I error and power under different simulation settings. We also compare the performance of our proposed method to the naive approach, which models separately the metabolites, ignoring the between-metabolite correlation. Finally, in Section 6, we illustrate the application of our method using the motivating longitudinal experiment explained above, and end with some concluding remarks.

2 | MODEL SPECIFICATION

Let y_{kij} be the measurement for kth ($k=1, \dots, K$) longitudinal feature of the ith subject ($i=1, \dots, n$) at time point j. We assume here that the K features belong to a common biological pathway. We denote measurement occasion with ($j=1, \dots, n_{ki}$), and the actual measurement time with t. Let also w be the $n\times 1$ phenotypic vector (eg, a binary disease status indicator), which is assumed to be linked to the longitudinal features (eg, metabolites).

We can assume that there is a true unobserved biological mechanism, M_i, which drives the association of the features y with the phenotype w and induces correlation between the features in the pathway. To model the association between the clinical outcome and the longitudinal metabolites, we assume that under a conditional independence assumption, for each subject i the joint density is written as

$$f(y_i, w_i) = \int f(y_i|M_i) f(w_i|M_i) dF(M_i), \tag{1}$$

where F is the cumulative distribution function of M_i and the other two terms in the integrand are specified below.

Shared latent process model $f(y_i \mid M_i)$**:** The progression and correlation structure of the K longitudinal features in the pathway is modeled by

$$y_{kij} = x_i(t_{kij})^T\beta + z_i(t_{kij})^T u_i + b_{ki} + \epsilon_{kij}, \tag{2}$$

where the first two terms, $x_i(t_{kij})^T\beta + z_i(t_{kij})^T u_i$, represent the shared time-dependent latent trajectory. In addition, $x_i(t)$ and $z_i(t)$ are the time-dependent vectors of covariates for fixed and random-effects of the sizes n_β and n_u, corresponding to parameters β and u, respectively. In addition, u_i is the subject-specific random term $u_i \sim \text{MVN}(\mu, \Sigma_u)$, where μ is a vector of size n_u and Σ_u is the corresponding variance matrix. For identifiability reasons, we will assume that $\mu_{u_{i0}}$ is zero and $\text{var}(u_{i0}) = 1$. Finally, the b_{ki}'s are feature-specific random-effects $b_{ki} \sim N(0, \sigma^2_{b_k})$ and the ϵ_{kij}'s are error terms with $\epsilon_{kij} \sim N(0, \sigma^2_{\epsilon_K})$ The random terms b_{ki} and u_i convey the within-pathway and within-subject correlations and together represent the unknown biological mechanism M_i.

Phenotypic model $f(w_i \mid M_i)$**:** Given that the biological mechanism M_i is captured by the random-effects b_{ki} and u_i, we can model the association between the longitudinal metabolites and the phenotype via a generalized linear model. For instance, in case of a binary outcome with $\pi_i = \Pr\{w_i = 1\}$, we have

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_0 + \sum_{q=1}^{n_u}\alpha_{1q}u_{iq} + \sum_{k=1}^{K}\alpha_{2k}b_{ki}. \tag{3}$$

Therefore, testing for association between clinical outcome and longitudinal metabolites corresponds to testing the null hypothesis:

$$H_0 : \alpha_{11} = \cdots = \alpha_{1n_u} = \alpha_{21} = \cdots = \alpha_{2k} = 0. \tag{4}$$

Testing this hypothesis in a high-dimensional setting as we have here is further discussed in the following section.

3 | ESTIMATION AND INFERENCE

Let Ω be the vector of size $n_\beta + (n_u \times (n_u+1)/2) - 1 + 2K$, including all parameters in the joint distribution in (1).

By combining models (2) and (3), we can formulate the individual contribution L_i to the likelihood for subject i as

$$L_i(\Omega) = \int\int f(u_i)f(w_i|u_i, b_{ki})\prod_{k=1}^{K}\left\{f(b_{ki})\prod_{j=1}^{n_{ki}}f(y_{kij}|u_i, b_{ki})\right\}du_i db_{ki}, \tag{5}$$

where the conditional distribution of vector y_i of size $n_i = \sum_k n_{ki}$ given the random effects is

$$f(y_i|u_i, b_{ki}) = (2\pi)^{-\frac{n_i}{2}}\,|\Sigma^c_{y_i}|^{-\frac{1}{2}}\,\exp\left\{-\frac{1}{2}(y_i - \mu_{y_i})^T(\Sigma^c_{y_i})^{-1}(y_i - \mu_{y_i})\right\}, \tag{6}$$

where X_i and Z_i are matrices of size $n_\beta \times n_i$ and $n_u \times n_i$, respectively. In addition, $\mu_{y_i} = X_i(t_{kij})^T\beta + Z_i(t_{kij})^T u_i + b_{ki}$ is the mean vector and, $\Sigma^c_{y_i} = I_K \otimes \sigma^2_{\epsilon_k} I_{n_{ki}}$ is the variance matrix of the conditional distribution of y_i. Here, I is the identity matrix with the specified dimension.

The conditional distribution of the w_i given the random effects is

$$f(w_i|u_i, b_{ki}) = \left(\frac{\exp(\alpha_0 + \sum_{q=1}^{n_u}\alpha_{1q}u_{iq} + \sum_{k=1}^{K}\alpha_{2k}b_{ki})}{1 + \exp(\alpha_0 + \sum_{q=1}^{n_u}\alpha_{1q}u_{iq} + \sum_{k=1}^{K}\alpha_{2k}b_{ki})}\right)^{w_i} \times \left(\frac{1}{1 + \exp(\alpha_0 + \sum_{q=1}^{n_u}\alpha_{1q}u_{iq} + \sum_{k=1}^{K}\alpha_{2k}b_{ki})}\right)^{(1-w_i)},$$

and the random effects u_i and b_{ki} follow normal distribution as defined previously.

To make inferences about the association of longitudinal omics y_i and the phenotype w_i, we need to estimate the vector of parameters Ω by maximizing the likelihood function in (5). However, the number of parameters to be estimated (Ω) and dimensionality of the random effects increases with pathway size, rendering our approach computationally intensive. To address this issue we follow the two-step procedure described in Tsonaka et al.[9] In our case, under the assumption that $f(M \mid y, w)$ depends only on y and that the parameter spaces for $y_i \mid M_i$ and $w_i \mid M_i$ are separable, we can show that

$$\log f(y_i, w_i) = \log f(w_i|M_i) + \log f(y_i). \tag{7}$$

Derivation of equality (7) is provided as supplementary material. Equation (7) implies that the observed marginal likelihood for the longitudinal features $f(y_i)$ can be fitted separately from the phenotypic model $f(w_i \mid M_i)$, hence reducing the dimensionality of the problem. Consequently, we adopt a two-step estimation procedure, where in step I, M_i is estimated in terms of the shared random process model (2). Then, in step II, the associations between pathway and phenotype are estimated using the estimated random-effects from the previous step. A detailed explanation of each step is as follows.

Step I. Let Θ be the vector of unknown parameters for the shared latent process model (2), the normality of y, u and b implies that the marginal densities of y_i are normal with mean vector $E_i = X_i\beta$ and variance matrix $\Sigma_{y_i} = Z_i\Sigma_u Z_i^t + I_K \otimes \sigma^2_{b_k}\mathbb{I}_{n_{ki}} + I_K \otimes \sigma^2_{\epsilon_k} I_{n_{ki}}$, where Z_i is the random effects matrix $(Z_{i1}^T(t), \ldots, Z_{ik}^T(t), \ldots, Z_{iK}^T(t))^T$ of dimension $n_i \times n_u$, where $\mathbb{I}$ is a square matrix of all 1's with the specified dimension. The marginal likelihood function is $\mathcal{L}(\Theta) = \sum_i L_i(\Theta)$, where L_i denotes the individual contributions to the likelihood. Then, estimation of Θ is possible by optimizing $\mathcal{L}(\Theta)$ using a modified Marquardt algorithm presented by Proust et al.[10]

Based on the fitted model, using some matrix algebra, we also derive the random effects b_{ki} and u_i which are obtained as the modes of the posterior distribution of u and b given y. These are the best linear unbiased predictors for (u_i, b_{ki}) and are given by

$$E(\hat{u}_i, \hat{b}_{ki}|y_i) = \left[Z_i\Sigma_u \quad I_K\sigma^2_{b_k} \otimes \mathbb{I}_{(n_{ki}\times 1)}\right]^T \Sigma^{-1}_{y_i}(y_i - \beta X_i), \tag{8}$$

The unknown parameters in (8) can be replaced by their estimates.

Step II. To further proceed with testing the association of pathway with the clinical outcome, we need to test the null hypothesis in (4), based on the model presented in (3). In low-dimensional settings, such a hypothesis may be tested using the multivariate Wald test or likelihood ratio test. However, in high-dimensional settings which is common in omics research, the score test is preferred. Assuming that the coefficients α from (3) are samples from a common distribution with expectation zero and variance τ^2, the null hypothesis in (4) is equivalent to the null hypothesis $H_0 : \tau^2 = 0$. As suggested by Goeman et al,[6] this hypothesis can be tested using a score test and the test statistic

$$Q = \frac{(w - \mu_w)^T RR^T (w - \mu_w)}{(n_u + K)(w - \mu_w)^T(w - \mu_w)},$$

where μ_w is the expectation of w under H_0 and R is the matrix of all estimated random-effects (u_i and b_{ki}). Under the null hypothesis, the Q statistic from global test is distributed as a weighted sum of χ^2 variables.[11] However, small sample sizes are quite common in genomic studies where the asymptotic distribution may be a poor approximation. Therefore, for small sample sizes the P-value is computed based on the empirical distribution of Q which is calculated by phenotype permutation. It is possible to implement this two-step procedure in R using the *lcmm* package in step I and the *globaltest* package in step II. The R package *SLaPMEG* combines the functions from these two packages with the additional option of pairwise method for larger pathways. The R package *SLaPMEG* is available via CRAN.

4 | PAIRWISE FITTING FOR LARGE PATHWAYS

The first step of the procedure above becomes computationally intensive as the size of pathway, hence the dimension of the vector $\hat{\Theta}$ increases. For instance, optimizing the joint marginal likelihood for a pathway of size 20 (with 30 subjects measured at five occasions) would require estimation of 43 parameters which takes about 4 minutes on a standard desktop computer. In addition, a joint model with a large number of parameters might not necessarily converge. In this section, we offer a solution for these issues based on a pairwise modeling approach. This computational solution is an additional novelty in our work.

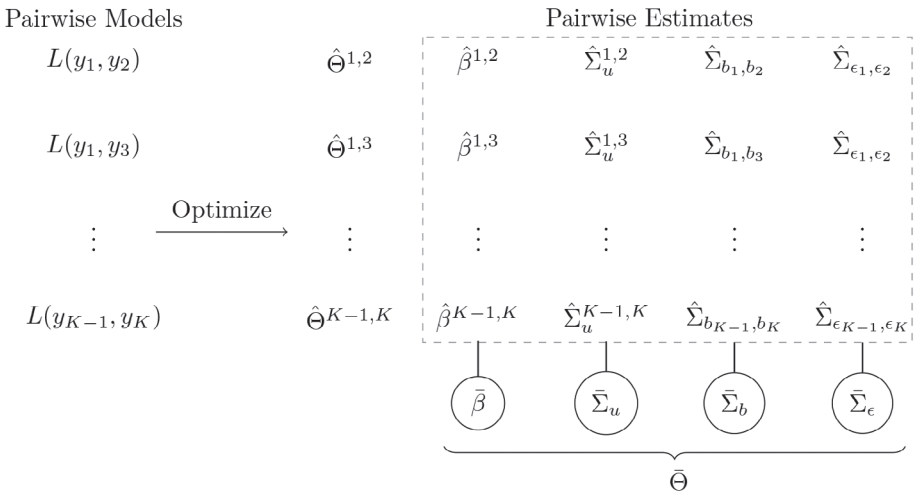

FIGURE 1 Pairwise fitting approach—A separate model is fitted on each pair of longitudinal outcomes (y_r, y_s) and the bivariate $L(y_r, y_s)$'s are optimized giving rise to the MLE $\hat{\Theta}^{r,s}$ where $\hat{\Theta}^{r,s} = (\beta_r, \sigma_r, \beta_s, \sigma_s)$. Parameters that are estimated from more than one pair, for example, β are averaged, giving rise to $\overline{\beta}$, $\overline{\Sigma}_u$, $\overline{\Sigma}_b$, and $\overline{\Sigma}_\epsilon$. These estimates are then used in Equation (8) to derive the Empirical Bayes estimates from step I based on the full joint model

We employ a pseudo-likelihood approach where a pseudo-likelihood function (L_{PL}) replaces the original likelihood function by a product over more tractable components.[12,13] For the joint likelihood presented in (1), we can write the pseudo-likelihood function as

$$L_{\text{PL}}(\Theta) = L(y_1, y_2|\Theta_{1,2}) + L(y_1, y_3|\Theta_{1,3}) + \ \ldots \ + L(y_{K-1}, y_K|\Theta_{K-1,K}), \tag{9}$$

where $L(y_r, y_s|\Theta_{r,s})$ is the likelihood function for the rth and sth metabolites in the pathway and $\Theta^{r,s}$ is the vector of parameters as defined above which is estimated based on this pair. Equality (9) implies that we can avoid optimizing the full joint likelihood presented in (5) and instead, fit the bivariate models separately. This greatly reduces the dimensionality of the joint model and hence the computation time for optimizing it.

In practice, as depicted in Figure 1, first all 2^K possible bivariate models are fitted and optimized, obviously certain parameters are estimated more than once as they appear in multiple pairs, we combine the estimated values from these bivariate models by averaging them to obtain estimates for the unknown parameters in (8). We chose to average, as averaging preserves the properties of the original estimates so the final averaged estimates are also consistent and asymptotically normally distributed.

To illustrate how the procedure works, take a pathway of three features as an example, the three models (y_1, y_2), (y_1, y_3) and (y_2, y_3) are fitted. Then, the corresponding $L(y_1, y_2)$, $L(y_1, y_3)$ and $L(y_2, y_3)$ are optimized, giving rise to the MLE's $\hat{\Theta}^{1,2} = (\hat{\beta}^{1,2}, \hat{\Sigma}_u^{1,2}, \hat{\Sigma}_b^{1,2}, \hat{\Sigma}_\epsilon^{1,2})$, $\hat{\Theta}^{1,3} = (\hat{\beta}^{1,3}, \hat{\Sigma}_u^{1,3}, \hat{\Sigma}_b^{1,3}, \hat{\Sigma}_\epsilon^{1,3})$ and $\hat{\Theta}^{2,3} = (\hat{\beta}^{2,3}, \hat{\Sigma}_u^{2,3}, \hat{\Sigma}_b^{2,3}, \hat{\Sigma}_\epsilon^{2,3})$. By averaging across the parameters that are estimated more than once, such as β, we get $\overline{\Theta} = (\overline{\beta}, \overline{\Sigma}_u, \overline{\Sigma}_b, \overline{\Sigma}_\epsilon)$. Now we can obtain valid predictions of the random-effects by substituting the unknown parameters in Equation (8) using corresponding values from $\overline{\Theta}$.

5 | SIMULATION STUDY

5.1 | Simulation study design

We designed a simulation study to evaluate the SLaPMEG approach in terms of type I error and power across several scenarios with increasing sample size, pathway size, effect size and strength of correlation between metabolites and/or repeated measures.

5.1.1 | Data generation under joint model

Longitudinal metabolite data were simulated for 26 and 100 subjects which were randomly assigned to two phenotypic groups of equal sizes. We considered a short and longer follow-up, that is, with three and five repeated measures, respectively. Four correlation scenarios were assumed; in the high-correlation scenario both the metabolites and repeated

measures are highly correlated and the low-correlation structure is the opposite. There are also two intermediate scenarios in which either the metabolites or the repeated measurements are highly correlated. To study the effect of pathway size, we included pathways of size 3, 20, and 40.

The data were simulated based on the joint model presented earlier, with a grouping variable as phenotype. The model includes a fixed effect of group (case/control status), fixed effect of time, a shared random intercept and slope u_i and the metabolite-specific random-effects b_{ki}. We will specify the model in (1) as

$$y_{kij} = \beta_0 + \beta_1 \cdot t_{kij} + \beta_{2k} \cdot \text{group}_i + u_{0i} + u_{1i} \cdot t_{kij} + b_{ki} + \epsilon_{kij}, \tag{10}$$

where,

$$\begin{aligned} u_{0i} &\sim N(0,1), \quad u_{1i} \sim N(0,\sigma_u), \\ b_{ki} &\sim N(0,\sigma_{b_k}), \quad \sigma_{b_k} \sim N(\mu_b, 0.5), \\ \epsilon_{kij} &\sim N(0,\sigma_{\epsilon_k}), \quad \sigma_{\epsilon_k} \sim N(2, 0.5). \end{aligned}$$

The fixed effect of group (β_{2k}) was set to zero for all metabolites to study type I error and it was either 2, 3, or 4 to study power, where either 1/3 or a 2/3 of the metabolites in the pathway carried the effect of interest. Therefore, for a pathway with three metabolites, either 1 (1/3) or 2 (2/3) metabolite(s) carried the group effect, and the group effect was zero for the other metabolites. For simplicity, the fixed effect of time was set to 2. For high and low-correlation scenarios, σ_u was set to 4 and 1 and μ_b was set to 4 and 2, respectively. In total, 244 scenarios were considered with 1000 repetitions for each scenario.

5.1.2 | Analysis of simulated data

SlaPMEG approach was adopted to analyze the data. The *P*-values of the global test at the second step were calculated using 10 000 permutations, type I error and power were calculated by considering a *P*-value $< .05$ as significant. For comparison, simulated data were also analyzed based on a naive approach which ignores the potential correlation between metabolites in a pathway and estimates the metabolite-specific random-effects based on separate LMMs per metabolite. To test the group effect at a pathway level, we will perform a global test using all the random-effects estimated via the LMMS for individual metabolite. Furthermore, we evaluated the performance of the pairwise approach with the same data. As explained in Section 4, in this case, the potential correlation between all metabolites in a pathway is estimated via the pairwise shared latent process models. Thus, to assess whether loss of efficiency is induced, we compared the standard SLaPMEG, which applies the shared latent process on all metabolites simultaneously, with the pairwise approach in pathways of small/moderate size under high-/low-correlation scenarios. For the case of very large pathways (40 metabolites), only the pairwise method and the naive method were applied.

5.1.3 | Robustness and misspecification

We have studied the power of SLaPMEG to detect a difference in progression of the metabolites via additional simulations. For this case, we have simulated data under a model with time-dependent group effect. So instead of $\beta_2 \cdot \text{group}_i$, we have added $\beta_2 \cdot t_{kij} \times \text{group}_i$ to (10). To make sure the results arise only from the slope effect, we set $\beta_1 = 0$ and β_2 was either 2, 3, or 4. The data are analyzed as before with one additional test. We adopted a variation of SLaPMEG, where in Step II only the random slope is tested via the global test. This last variation is to test if the group-dependent trend is well captured by the random slope via SLaPMEG framework. For simplicity, we have conducted these simulations under a moderate correlation structure, where the variances parameters are set to the average of high- and low-correlation settings from previous simulations.

Furthermore, to explore robustness to misspecification of the correlation structure, we simulated longitudinal metabolites under independence and applied SLaPMEG. Thus, the data were simulated based on separate LMM's from the model above with the b_{ki} term removed. The same parameters were adopted and the results of the naive approach and SLaPMEG were compared.

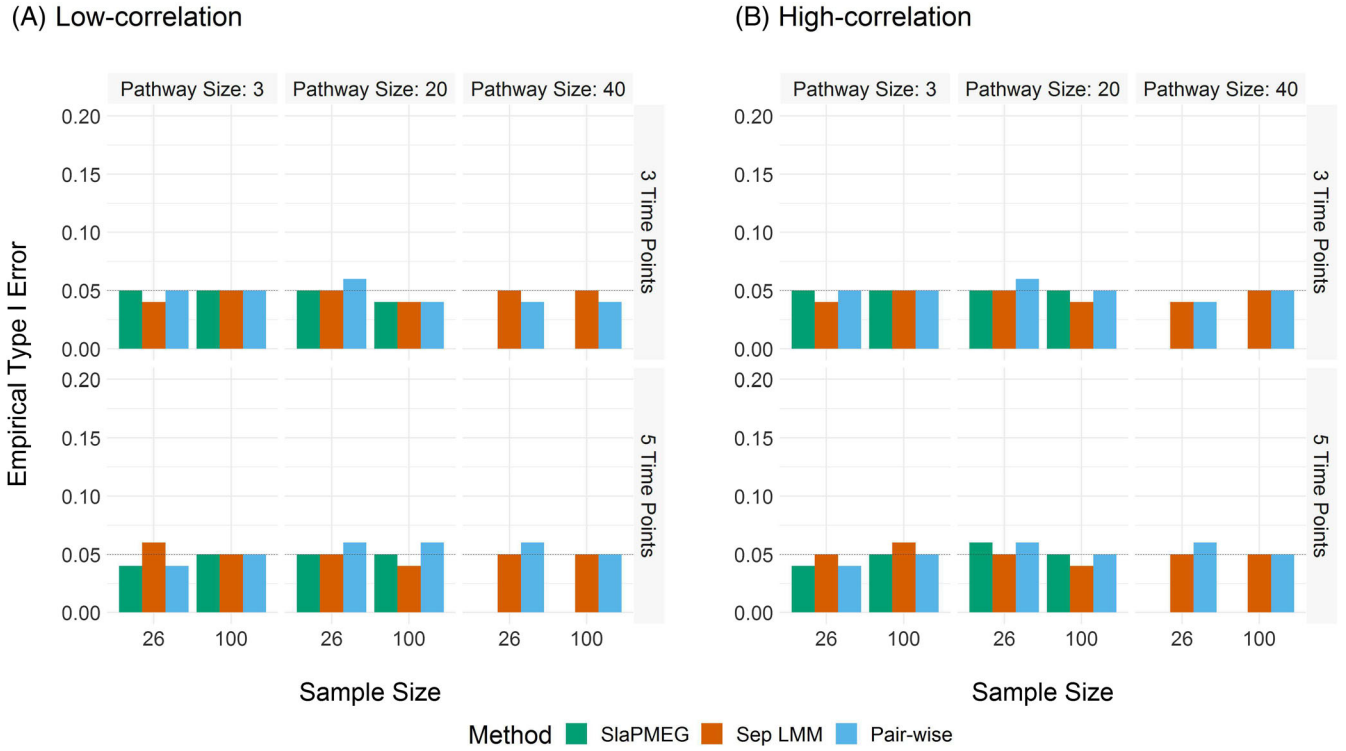

FIGURE 2 Simulation results—Colors represent the SLaPMEG (Green), separate LMM (Red), and pairwise (Blue) methods. Correlation of the metabolites and subjects are both low for the plot on the left (A) and high for the plot on the right (B). All methods have type I error rates close to 5%. SLaPMEG, shared latent process mixed effects modeling within global test [Colour figure can be viewed at wileyonlinelibrary.com]

5.2 | Simulation study results

Type I error. Both SLaPMEG and the naive method preserve the type I error close to nominal level (Figure 2). This is true for all correlation structures. Low- and high-correlation structures are presented in Figure 2 and the other two intermediate structures are presented in the Supplementary Figure 1. More specifically, SLaPMEG is slightly anticonservative for testing a large pathway with large sample size and few time points.

Power. In general, SLaPMEG improves in power over the naive method, especially for small sample sizes (Figure 3). Under low-correlation setting with large effect and sample size, both methods have high power to detect active pathways. For small sample sizes, the naive method has very little power, this can be improved up to twofold using SLaPMEG. Under the high-correlation setting and large sample size, both methods still have relatively high power, but the power of naive method decreases in some conditions even up to 73%. With highly correlated metabolites and subjects, the power of SLaPMEG is up to seven times more than the naive method. Power plots for other conditions can be found in Supplementary Figures 2 to 5.

Pairwise method. The type I error of the pairwise approach does not depend on the correlation structure and is very close to the standard SLaPMEG. In case of a pathway of size 40, for which only pairwise modeling was efficient, the type I error of the method is well preserved as is the case for the naive method. As for the power, the pairwise approach closely resembles the full SLaPMEG under different scenarios. Even with pairwise fitting, SLaPMEG improves the power up to two times in the low-correlation setting and up to seven times in the high-correlation setting for large pathways (40 metabolites).

Differential progression. SLaPMEG can detect the difference in progression of the longitudinal metabolites. In fact, when only a time-dependent group effect is present the method has more power. Results for a small sample size are presented in Figure 4, other results are presented in Supplementary Figure 6. For this set of simulations, we have conducted two global tests in step II, one with all random effects, and one only with the random slope (specified in figure with *slope* in parentheses). The second variation has an even higher power, hence random slope can sufficiently capture the differential progression of longitudinal metabolites.

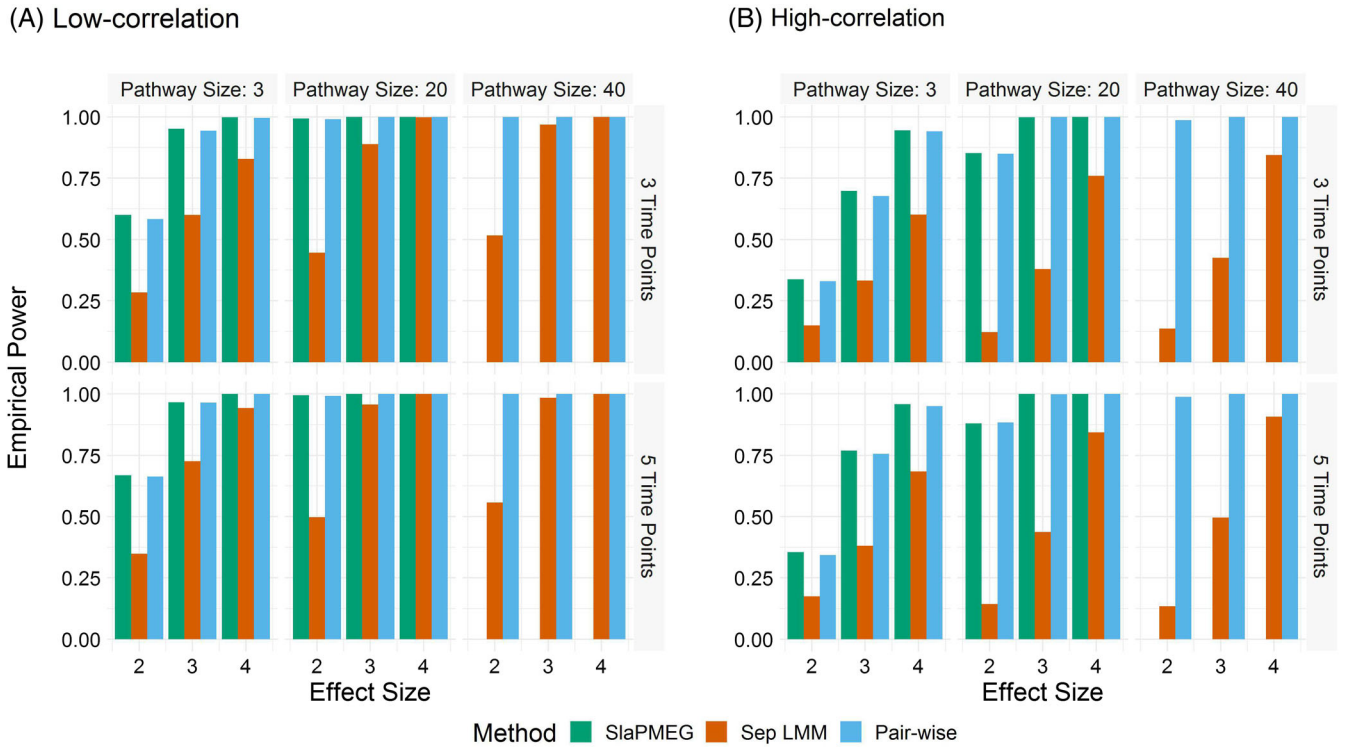

FIGURE 3 Simulation results—Colors represent the SLaPMEG (Green), separate LMM (Red), and pairwise (Blue) methods. The plots show a scenario with a small sample size (26) where only 1/3 of the metabolites are associated with the phenotype. Correlation of the metabolites and subjects are both low for the plot on the left (A) and high for the plot on the right (B). Comparing the two plots, it is clear that Sep. LMM has less power under the high-correlation scenario. The SLaPMEG model has more power in general and especially for a small sample and/or effect sizes. A similar pattern holds for larger sample size (100) and pathways with more active genes (2/3). SLaPMEG, shared latent process mixed effects modeling within global test [Colour figure can be viewed at wileyonlinelibrary.com]

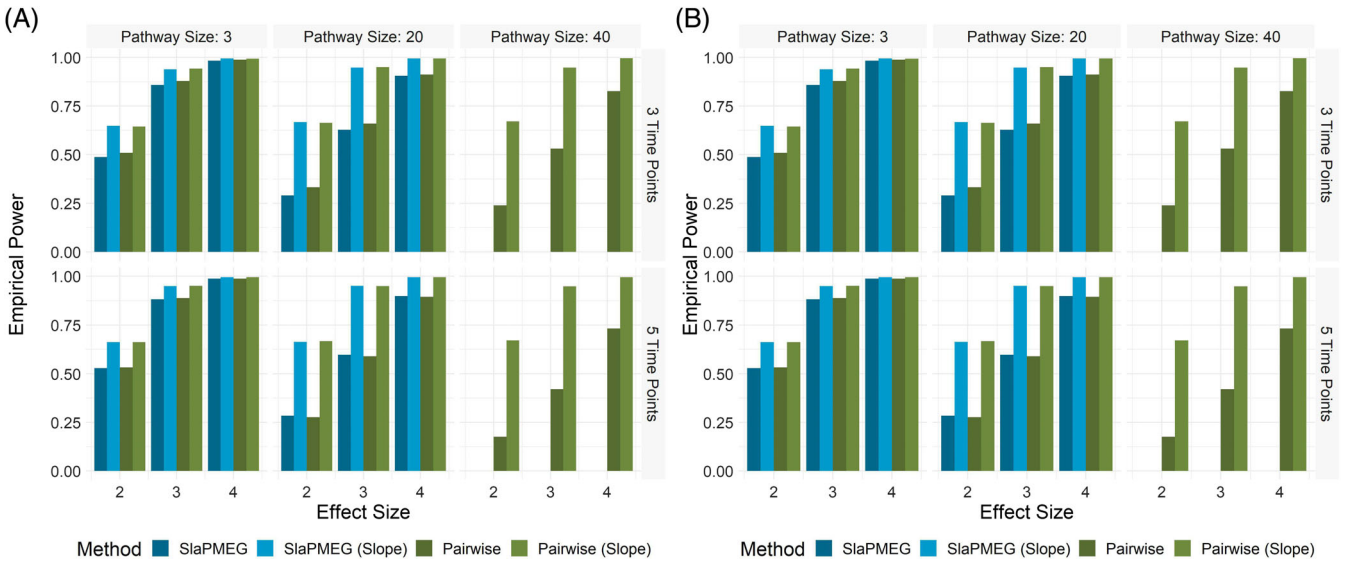

FIGURE 4 Simulation results—Colors represent the SLaPMEG (Blue) and pairwise (Green) methods. The bar with darker colors are the results of Globaltest using all random effects, the lighter colors are the results of Globaltest using only the random slope. These plots show a scenario with an small sample size (n = 26) where (A) 1/3 and (B) 2/3 of the Metabolites are associated with a time-dependent effect. SLaPMEG method has sufficient power to detect differential progression and the pairwise method follows closely. A similar pattern holds for larger sample size (n = 100). SLaPMEG, shared latent process mixed effects modeling within global test [Colour figure can be viewed at wileyonlinelibrary.com]

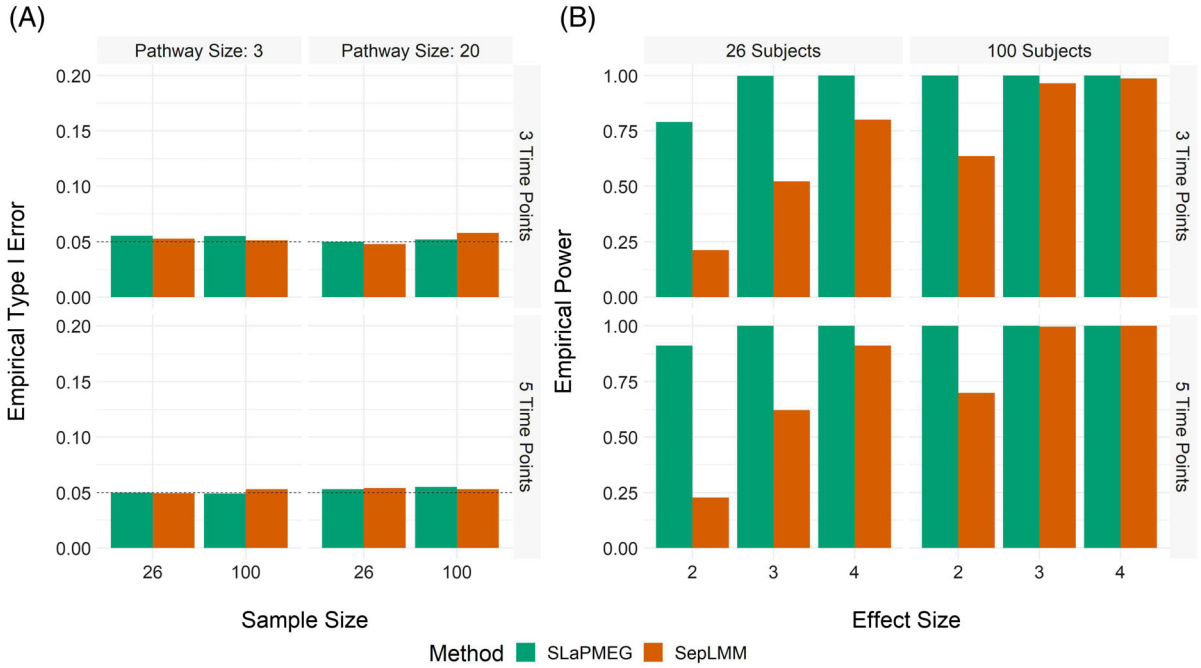

FIGURE 5 Simulation results—Colors represent the SLaPMEG (Green) and separate LMM (Red) methods. The plot on the left (A) shows type I error, and the plot on the right (B) shows power for a scenario with a small sample size (26) and 1/3 of the Metabolites in each pathway are associated with the phenotype. For this plot the data are simulated according to separate LMM models. Type I error rate is controlled even under misspecification of the correlation structure. Power follows a similar pattern to that of jointly simulated data. SLaPMEG, shared latent process mixed effects modeling within global test [Colour figure can be viewed at wileyonlinelibrary.com]

Robustness to misspecification. The results of the SLaPMEG approach are robust to misspecification of the underlying correlation structure in the data. To show this, we performed the same procedure with a new dataset where the data were generated from separate LMMs, hence assuming independence among metabolites. As depicted in Figure 5, type I error is controlled, and our proposed method has more power in detecting active pathways compared with the naive method. Results of power comparison with larger sample size and a larger proportion of active features are presented in Supplementary Figure 6. It is worth noting that when the metabolites are independent, especially with larger pathways, fitting a shared latent process is challenging simply because the shared effect is very small. SLaPMEG may encounter convergence issues in such a case and thus the pairwise approach has been adopted instead. In our misspecified simulations the joint model failed to converge about 14% of the time on average over different scenarios.

6 | CASE STUDY

6.1 | DMD study

We illustrate the application of SLaPMEG by analyzing data from a longitudinal experiment, designed to study disease progression in a mouse model of *DMD*, briefly introduced in Section 1. In particular, 21 dystrophin-lacking (*mdx*) mice (5 mdx mice, 5 mdx utrn+/+, 6 mdx utrn+/-, and 5 mdx utrn-/- mice) and five WT mice were followed up in time and blood samples were taken at five time points, that is, at 6, 12, 18, 24, and 30 weeks. Plasma samples were obtained in heparin lithium tubes after fasting the mice for 4 to 6 hours. Details of the data acquisition along with comprehensive analysis and results are reported separately in Reference 14.

For the illustration of our method, we will use here longitudinal data on 132 metabolites from two mice groups mdx (21 mice) and WT (5 mice). Metabolites provide a close perspective on the biologic system's response to mutations in the DMD gene leading to lack of dystrophin. Our goal is to investigate the metabolic signature between *mdx* and WT groups. To this end, we have considered 11 pathways from the Wikipathways database, as indicated in the first column of Table 1. These pathways were selected as relevant according to the results of a previous study.[14]

TABLE 1 DMD study results—Testing association of the selected metabolic pathways and disease status based on SLaPMEG and Naive methods

Pathway name	Pathway size	LMM conv.	Sep. LMM	SLaPMG
Alanine and aspartate metabolism	5	3	<0.001	0.036
Vitamin B12 metabolism	3	1	0.75	0.89
Transport of bile salts and organic acids, metal ions and amine compounds	5	1	0.65	0.43
Transport of inorganic cations/anions and amino acids/oligopeptides	5	2	0.005	0.34
Transport of vitamins, nucleosides, and related molecules	10	2	0.34	0.48[a]
Metabolism of polyamines	8	4	0.001	0.001
Histidine, lysine, phenylalanine, tyrosine, proline, and tryptophan catabolism	5	1	<0.001	0.019
Human metabolism overview	7	3	0.005	0.070
Biochemical pathways	55	15	0.302	0.042[a]
Urea cycle and metabolism of amino groups	12	5	<0.001	0.0.023[a]
Glucose homeostasis	13	7	0.54	0.29[a]

Abbreviations: DMD, Duchenne muscular dystrophy; SLaPMEG, shared latent process mixed effects modeling within global test.
[a]For these pathways the joint SLaPMEG failed to converge and the *P*-values are calculated based on the pairwise approach.

The data were first scaled and then our proposed method, as discussed in Section 2 was fitted:

$$y_{kij} = \beta_0 + \beta_1 \cdot t_{kij} + u_{i0} + u_{i1} \cdot t_{kij} + b_{ki} + \epsilon_{kij}. \tag{11}$$

The model includes a fixed effect of time (β), a subject-specific random intercept u_{i0}, a subject-specific random slope u_{i1} and a number of metabolite-specific random terms b_{ki} depending on the size of the pathway. The size of the pathways varies from 3 to 55. The estimation is based on the full model likelihood as described in Section 3 for small or moderate-sized pathways. The pairwise solution is utilized for larger pathways (ie, >12 metabolites) or in case the standard SLaPMEG failed to converge. As in the simulation, we also present results using the naive approach, which models each longitudinal metabolite separately. For these individual LMMs, a similar design was adopted excluding the metabolite-specific random terms. For both methods, *P*-values at the second stage were calculated using 10 000 permutations. The reported *P*-values have been adjusted for multiple testing across the 11 pathways using the Benjamini-Hochberg procedure. Furthermore, we have explored the density plots of the estimated random-effects to gain more insight into the nature of the pathway-level effects.

6.2 | DMD study results

In our analysis, we found that six out of 11 metabolic pathways had a significantly different progression for *mdx* mice compared with WT. The list of selected pathways along with the *P*-values are presented in the last column of Table 1. To further investigate our findings, we explored the density plots of the estimated random-effects for *Metabolism of Polyamines* pathway as shown in Figure 6. The creatine-specific random-effects for the WT mice are smaller than those for the *mdx* mice (Figure 6B-top). Interestingly, this is confirmed by the spaghetti plot of creatine (Figure 6A-top). Conversely, the estimated ornithine-specific random-effects are lower in *mdx* group than that of the WT group (Figure 6B-bottom), which is also confirmed by the spaghetti plot for the ornithine progression over time Figure 6B-bottom. This implies that SLaPMEG can identify the heterogeneous effects of pathways. As shown by Table 1, the two methods select similar pathways. Nevertheless, the results of the naive approach is not directly comparable to SLaPMEG. Fitting separate LMM for each metabolite did not always converge. For these cases the global test in the second step was only based on a small portion of metabolites for most pathways. The convergence rate within each pathway was 20% to 60%. A complete list of convergence rates and *P*-values for the naive method are also presented in the third and fourth column of Table 1. For instance, for the fourth raw, the *Transport of inorganic cations/anions and amino acids* pathway, there are five metabolites in the

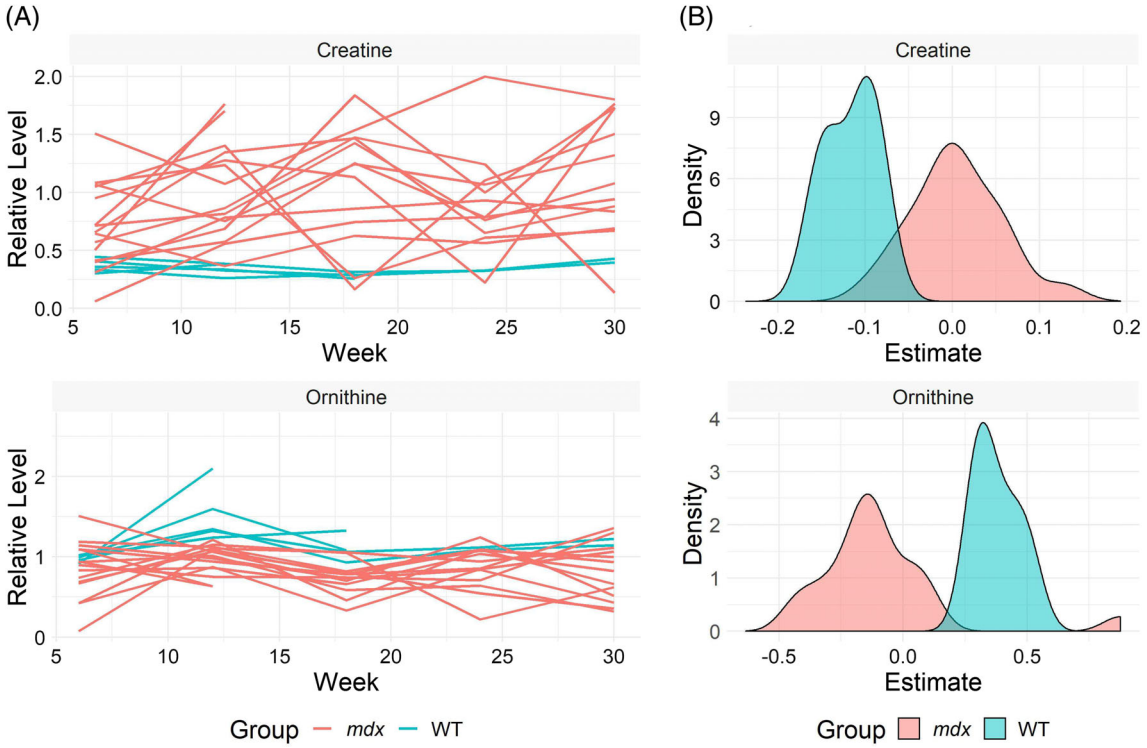

FIGURE 6 DMD study results—A, Time course of two metabolites from *metabolism of polyamines* pathway with nine metabolites and B, estimated b_{ki} values (metabolite-specific random-effects) for the corresponding metabolites. Colors represent the WT (blue) and *mdx* (red) groups. The direction of effect is different for the two metabolites. Creatine values (Top) lie below the shared longitudinal trajectory in the WT group compared with *mdx*. The opposite is true for the ornithine (Bottom) values. Despite the heterogeneity of effects, SLaPMEG identified the differential expression of this pathway between two groups. DMD, Duchenne muscular dystrophy; SLaPMEG, shared latent process mixed effects modeling within global test; WT, wild type [Colour figure can be viewed at wileyonlinelibrary.com]

pathway. The LMM convergence for this pathway is 2 (40%), meaning that we could fit an LMM for two of the metabolites and the model did not converge for the other three metabolites. This consequently means that the *P*-value (.0039) is calculated based on the random effects estimated based on only two metabolites. A reason for the failure of the separate linear mixed models may be that the missingness rate was about 30%.

7 | DISCUSSION

We have proposed a top-down approach to test whether a set of longitudinal metabolites in a pathway are associated with a clinical outcome of interest. The strength of our approach is that it exploits all sources of correlation in the data levering thereby the power of the test while remaining computationally efficient even for big pathway sizes. For small studies, which are common in genomic research, SLaPMEG shows good power properties. Furthermore, we observed that the random-effects can efficiently summarize the longitudinal metabolites. As stated in the results, a higher level of creatine was observed in *dmx* mice which is expected as they do not metabolize creatine due to loss of muscle mass.[15] On the other hand, the direction of difference for ornithine was the opposite. Reduced levels of amino acids such as ornithine in *dmx* mice are also in accordance with literature.[16] These two amino acids are involved in the same differentially expressed pathway. Based on this and evaluation of the density plots of the random-effects from other pathways, we can conclude that SLaPMEG can detect pathways with heterogeneous differential effects. This may be partly due to the fact that the global test can also detect heterogeneous differential effects.

We have built the pseudo-likelihood function using pairs of metabolites to provide a faster solution for large pathways. However, it should be noted that the pseudo-likelihood function in (9) could also be written in terms of triplets instead of pairs of metabolites. This would involve fitting some additional models, and according to the simulations, does not gain in term of power (results presented as Supplementary Figure 8). Note that these results are true when the data

are generated from a multivariate normal distribution and may change when certain correlation structures are present in the data. Moreover, optimizing the pseudo-likelihood instead of the joint likelihood ignores the correlations among estimates. Despite this fact, according to the empirical evaluations in Section 5, it can sufficiently approximate the unknown parameters of the joint model.

We have adopted a two-step procedure to avoid joint modeling of phenotype data and multiple longitudinal measures. The two step procedure does not require complex estimation techniques and is frequently used to analyze time-to-even data, for example, References 17-20. A disadvantage of these procedures is the potential bias in parameter estimates as a consequence of ignoring the phenotypic information when modeling the latent process.[18,20] Furthermore, the result of the second step depends on the quality of the estimates in the first step. The model in (2) should capture the shared latent trajectories of the data at hand. Therefore, it is recommended to perform formal model building and tests such as likelihood ratio test to select the best fitting random effects structure. Linear progressions are rather simple and a more elaborate structure may be more robust, for example, using polynomial or natural cubic splines depending on the length of the follow-up. According to the simulations, SlaPMEG may lose power when the model does not appropriately capture the longitudinal trend, specially when the sample size is low and the follow-up time is long (Supplementary Figure 9).

Our proposed procedure can be extended in different ways. In the case study, we have considered balanced data and a binary clinical outcome. However, the method can be successfully adopted for unbalanced datasets, with missing values and different types of outcomes including survival. In fact, the second step of our proposed method can be easily replaced by a different linear model, with some minor adjustments. In addition, in theory, SLaPMEG is not limited to metabolites and can be used with other omics, for example, RNA-seq. To analyze such overdispersed count data, instead of a linear model depending on normality, a generalized linear mixed model must be adopted in step I. This will greatly complicate the estimation of the parameters in Equation (2), and is currently not possible based on the SLaPMEG package. Nevertheless, the current method can be employed, as is, after normalizing the values or for example using a method similar to Reference 21, where the mean-variance trend is estimated and incorporated into a linear modeling procedure. However, the effect of these procedures on power properties and type I error control needs to be further studied.

In general, score tests are locally most powerful,[11] suggesting that our proposed procedure is also more powerful in detecting pathways that include many metabolites with small effect sizes. This is advantageous as the test is designed for detecting pathways with many relevant features. Finally, note that SLaPMEG analyzes the pathways independently, in case multiple pathways are of interest, the pathway-level *P*-values require multiple comparisons adjustments.

ACKNOWLEDGEMENTS

This work is financed by the Dutch Research Council (NWO) grant 639.072.412 and also the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements 305121 Integrated European Project on Omics Research of Rare Neuromuscular and Neurodegenerative Diseases (NEUROMICS).

DATA AVAILABILITY STATEMENT

The data of DMD study are available through Tsonaka et al (https://doi.org/10.1093/hmg/ddz309).

ORCID

Mitra Ebrahimpoor https://orcid.org/0000-0002-2299-876X
Pietro Spitali https://orcid.org/0000-0003-2783-688X

REFERENCES

1. Guo X, Qi H, Verfaillie CM, Pan W. Statistical significance analysis of longitudinal gene expression data. *Bioinformatics*. 2003;19(13):1628-1635. https://doi.org/10.1093/bioinformatics/btg206.
2. Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. Significance analysis of time course microarray experiments. *Proc Nat Acad Sci*. 2005;102(36):12837-12842. https://doi.org/10.1073/pnas.0504609102.
3. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*. 2012;8(2):e1002375. https://doi.org/10.1371/journal.pcbi.1002375.
4. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Nat Acad Sci U S America*. 2005;102:15545-15550. https://doi.org/10.1073/pnas.0506580102.
5. Wang L, Chen X, Wolfinger RD, Franklin JL, Coffey RJ, Zhang B. A unified mixed effects model for gene set analysis of time course microarray experiments. *Stat Appl Genet Molecul Biol*. 2009;8(1):1-18. https://doi.org/10.2202/1544-6115.1484.

6. Goeman JJ, Van De Geer SA, De Kort F, Van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*. 2003;20(1):93-99. https://doi.org/10.1093/bioinformatics/btg382.
7. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene sets net correlations analysis (GSNCA): a multivariate differential coexpression test for gene sets. *Bioinformatics*. 2014;30:360-368. https://doi.org/10.1093/bioinformatics/btt687.
8. Proust-Lima C, Amieva H, Jacqmin-Gadda H. Analysis of multivariate mixed longitudinal data: a flexible latent process approach. *Brit J Math Stat Psychol*. 2012;66(3):470-487. https://doi.org/10.1111/bmsp.12000.
9. Tsonaka R, van der Helm-van Mil AH, Houwing-Duistermaat JJ. A two-stage mixed-effects model approach for gene-set analyses in candidate gene studies. *Stat Med*. 2011;31(11-12):1190-1202. https://doi.org/10.1002/sim.4370.
10. Proust C, Jacqmin-Gadda H, Taylor JMG, Ganiayre J, Commenges D. A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. *Biometrics*. 2006;62(4):1014-1024. https://doi.org/10.1111/j.1541-0420.2006.00573.x.
11. Goeman JJ, Van De Geer SA, Van Houwelingen HC. Testing against a high dimensional alternative. *J Royal Stat Soc Ser B (Stat Methodol)*. 2006;68(3):477-493. https://doi.org/10.1111/j.1467-9868.2006.00551.x.
12. Fieuws S, Verbeke G. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*. 2006;62(2):424-431. https://doi.org/10.1111/j.1541-0420.2006.00507.x.
13. Renard D, Molenberghs G, Geys H. A pairwise likelihood approach to estimation in multilevel probit models. *Comput Stat Data Anal*. 2004;44(4):649-667. https://doi.org/10.1016/s0167-9473(02)00263-3.
14. Tsonaka R, Signorelli M, Sabir E, et al. Longitudinal metabolomic analysis of plasma enables modelling disease progression in Duchenne muscular dystrophy mouse models. *Human Molecul Genet*. 2020;29(5):745-755. https://doi.org/10.1093/hmg/ddz309.
15. Boca SM, Nishida M, Harris M, et al. Discovery of metabolic biomarkers for duchenne muscular dystrophy within a natural history study. *PLoS One*. 2016;11(4):e0153461. https://doi.org/10.1371/journal.pone.0153461.
16. Srivastava NK, Annarao S, Sinha N. Metabolic status of patients with muscular dystrophy in early phase of the disease: in vitro, high resolution NMR spectroscopy based metabolomics analysis of serum. *Life Sci*. 2016;151:122-129. https://doi.org/10.1016/j.lfs.2016.01.032.
17. Albert PS, Shih JH. On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. *Biometrics*. 2010;66(3):983-987. https://doi.org/10.1111/j.1541-0420.2009.01324.x.
18. Sweeting MJ, Thompson SG. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometr J*. 2011;53(5):750-763. https://doi.org/10.1002/bimj.201100052.
19. He B, Luo S. Joint modeling of multivariate longitudinal measurements and survival data with applications to Parkinson's disease. *Stat Methods Med Res*. 2016;25(4):1346-1358. https://doi.org/10.1177/0962280213480877.
20. Yang L, Yu M, Gao S. Joint models for multiple longitudinal processes and time-to-event outcome. *J Stat Comput Simul*. 2016;86(18):3682-3700. https://doi.org/10.1080/00949655.2016.1181760.
21. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. 2014;15(2):R29. https://doi.org/10.1186/gb-2014-15-2-r29.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Ebrahimpoor M, Spitali P, Goeman JJ, Tsonaka R. Pathway testing for longitudinal metabolomics. *Statistics in Medicine*. 2021;40:3053–3065. https://doi.org/10.1002/sim.8957