Received: 5 June 2019 | Revised: 12 May 2020 | Accepted: 8 July 2020

DOI: 10.1111/biom.13333

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Poststratification fusion learning in longitudinal data analysis

Lu Tang[1] | **Peter X.-K. Song**[2]

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

[2] Department of Biostatistics, University of Michigan, Ann Arbor, Michigan

Correspondence
Lu Tang, Department of Biostatistics, University of Pittsburgh, 130 DeSoto St, Pittsburgh, PA 15261.
Email: lutang@pitt.edu

Funding information
National Institute of Environmental Health Sciences, Grant/Award Number: R01ES024732; National Science Foundation, Division of Mathematical Sciences, Grant/Award Number: DMS1811734

Abstract

Stratification is a very commonly used approach in biomedical studies to handle sample heterogeneity arising from, for examples, clinical units, patient subgroups, or missing-data. A key rationale behind such approach is to overcome potential sampling biases in statistical inference. Two issues of such stratification-based strategy are (i) whether individual strata are sufficiently distinctive to warrant stratification, and (ii) sample size attrition resulted from the stratification may potentially lead to loss of statistical power. To address these issues, we propose a penalized generalized estimating equations approach to reducing the complexity of parametric model structures due to excessive stratification. Specifically, we develop a data-driven fusion learning approach for longitudinal data that improves estimation efficiency by integrating information across similar strata, yet still allows necessary separation for stratum-specific conclusions. The proposed method is evaluated by simulation studies and applied to a motivating example of psychiatric study to demonstrate its usefulness in real world settings.

KEYWORDS
GEE, pattern-mixture model, regularization, stratification, variable selection

1 | INTRODUCTION

In biomedical studies, stratification has been commonly undertaken in design and analytical phases by investigators to gain more targeted insight. Due to heterogeneity of populations from observational studies, patients may be partitioned into multiple strata so that those within a stratum have similar characteristics and more comparable association effect sizes with respect to an outcome. For examples, unsupervised clustering and partitioning by treatment propensity scores are two commonly used methods to stratify populations for estimating conditional treatment effects of these subpopulations. A key technical concern is that data-driven approaches to stratification may be subject to errors. It is unclear if the resulting stratification is representative of the true underlying subgroup structure of the effects of interest. The examples above aim to group subjects into homogeneous strata so that the resulting statistical inferences are better powered within each specific stratum. Moreover, they may benefit from a poststratification group merging procedure, termed as *fusion learning*, for better efficiency and interpretation.

Throughout this paper, we assume that a data sample comprised many strata, within each subjects are homogeneous and sampled from a common parametric distribution. The development is motivated by the situation where erroneous stratification often occurs in practice and excessive stratification may potentially do more harm than good due to data attrition. Arguably, the more strata used, the more ad hoc noise is introduced to subsequent analyses, so that results may be overly specific and lack reproducibility.

A direct consequence of stratification is that we often have limited amount of data in each stratum. It is natural to hope we could borrow information from other strata, if possible, to increase statistical power. A strategy of information integration, whose form is to be unfolded, has to account for a desirable trade-off between specificity and generalizability. We propose a new approach to address this problem via the technique of fusion learning (e.g., Tang and Song, 2016; Wang *et al.*, 2016; Ma and Huang, 2017), in the context of generalized estimating equations (GEE) for longitudinal data.

GEE is one of the two primary methods of choice in longitudinal or clustered data analyses. In comparison to its competitor, the generalized linear mixed-effects model (GLMM), GEE is a quasi-likelihood approach based only on the first two moments of data distributions and provides estimates of population-average effects of covariates. Ollier *et al.* (2016) proposed a fused LASSO (Tibshirani *et al.*, 2005) approach for the GLMM where parameter fusion is operated under given random cluster effects. Their method is different from ours in the following ways. As pointed in Song (2007), the likelihood estimation in the GLMM with nonnormal data is subject to numerical approximation errors arising from the integral of augmented likelihoods with respect to the distribution of random effects (e.g., Laplace approximation), whereas GEE yields consistent estimates of the parameters in the marginal model. These two models have different interpretations of the model parameters, namely conditional effects in the GLMM versus population-average effects in the GEE (Neuhaus *et al.*, 1991). Fusion learning in the GEE framework is preferred when certain regularization is used in the analysis as the GLMM may become unstable and unreliable due to both numerical errors and estimation bias from the penalty-led shrinkage.

Consider a longitudinal study of N individuals, each with T_i repeated measurements. For the sake of exposition, in the rest of the paper, we assume an equal number of repeated measures, namely, $T_i = T, i = 1, \dots, N$, are collected at prespecified follow-up times. For complete data, the design matrix of an individual i, $i \in \{1, \dots, N\}$, is denoted as $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{iT})^T$, where $\boldsymbol{x}_{it}$ is a p-element covariate vector measured at visit t, $t = 1, \dots, T$. Response vector $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{iT})^T$ denotes the longitudinally correlated responses of subject i. Denote the first two conditional marginal moments of Y_{it} by $\mu_{it} = E(Y_{it}|\boldsymbol{X}_{it})$ and $\sigma^2_{it} = \text{var}(Y_{it}|\boldsymbol{X}_{it})$, where the marginal density of Y_{it} is a member in the family of exponential dispersion (ED) models (Jorgensen, 1997). The mean μ_{it} follows a generalized linear model, $g(\mu_{it}) = \boldsymbol{x}_{it}^T\boldsymbol{\beta}$, where g is a known link function and $\boldsymbol{\beta}$ is the regression parameters of interest, and variance σ^2_{it} of the ED model takes the form $\sigma^2_{it} = \phi\upsilon(\mu_{it})$, where $\upsilon(\cdot)$ is the known unit variance function and ϕ is the dispersion parameter. See more details of the ED models in Jorgensen (1997). In this paper, we consider canonical link function g, satisfying $\dot{g}(\mu) = \upsilon^{-1}(\mu)$. With an invocation of sample stratification, the N subjects are stratified into K strata, each having a sample size n_k, $k = 1, \dots, K$ and $N = \sum_k n_k$. Accordingly, we denote stratum-specific quantities within, say, stratum k, via suitable subscripts, that is, as $\boldsymbol{X}_{ki}, \boldsymbol{Y}_{ki}, \mu_{kit}, \sigma^2_{kit}, \phi_k, \boldsymbol{\beta}_k$, and so on, where $i = 1, \dots, n_k$. For ease of presentation, we package the parameters of interest $\{\beta_{jk} : j = 1, \dots, p, k = 1, \dots, K\}$ into a covariate-major vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_{1\cdot}^T, \dots, \boldsymbol{\beta}_{p\cdot}^T)^T$, where $\boldsymbol{\beta}_{j\cdot} = (\beta_{j1}, \dots, \beta_{jK})^T$ denotes regression coefficients associated with the jth covariate $\boldsymbol{X}_j$, for $j = 1, \dots, p$. Similarly, we denote $\boldsymbol{\beta}_{\cdot k} = (\beta_{j1}, \dots, \beta_{jK})^T$ the vector of regression coefficients associated with stratum k, for $k = 1, \dots, K$.

It follows that a stratified GEE solves the following aggregated estimating equations

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k} \boldsymbol{S}_{ki}(\boldsymbol{\beta}_{\cdot k}) = \boldsymbol{0}, \tag{1}$$

where each of the K components $\sum_{i=1}^{n_k} \boldsymbol{S}_{ki}(\boldsymbol{\beta}_{\cdot k}) = \boldsymbol{0}$ corresponds to the GEE of a stratum. By Liang and Zeger (1986), the individual estimating equations $\boldsymbol{S}_{ki}(\boldsymbol{\beta}_{\cdot k})$ take the forms $\boldsymbol{S}_{ki}(\boldsymbol{\beta}_{\cdot k}) = \frac{\partial \boldsymbol{\mu}_{ki}(\boldsymbol{\beta}_{\cdot k})}{\partial \boldsymbol{\beta}_{\cdot k}^T}\boldsymbol{V}_{ki}^{-1}\{\boldsymbol{Y}_{ki} - \boldsymbol{\mu}_{ki}(\boldsymbol{\beta}_{\cdot k})\}$, where $\frac{\partial \boldsymbol{\mu}_{ki}(\boldsymbol{\beta}_{\cdot k})}{\partial \boldsymbol{\beta}_{\cdot k}^T} = \boldsymbol{X}_{ki}^T\boldsymbol{A}_{ki}(\boldsymbol{\beta}_{\cdot k})$ with $\boldsymbol{A}_{ki}(\boldsymbol{\beta}_{\cdot k}) = \text{diag}\{\sigma^2_{ki1}(\boldsymbol{\beta}_{\cdot k}), \dots, \sigma^2_{kiT}(\boldsymbol{\beta}_{\cdot k})\}$ (ϕ will be cancelled with zero on the right-hand side of (1), thus omitted), and variance matrix $\boldsymbol{V}_{ki} = \boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}_{\cdot k})\boldsymbol{R}_k(\tau_k)\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}_{\cdot k})$ with $\boldsymbol{R}_k(\tau_k)$ being a working correlation matrix parameterized by correlation parameter τ_k, for $k = 1, \dots, K$, $i = 1, \dots, n_k$. As a result, $\boldsymbol{S}_{ki}(\boldsymbol{\beta}_{\cdot k})$ may be written as

$$\boldsymbol{S}_{ki}(\boldsymbol{\beta}_{\cdot k}) = \boldsymbol{X}_{ki}^T\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}_{\cdot k})\boldsymbol{R}_k^{-1}(\tau_k)\boldsymbol{A}_{ki}^{-1/2}(\boldsymbol{\beta}_{\cdot k})\{\boldsymbol{Y}_{ki} - \boldsymbol{\mu}_{ki}(\boldsymbol{\beta}_{\cdot k})\}. \tag{2}$$

If $\boldsymbol{\beta}_{\cdot k}$'s are treated as all different across K strata, the aggregated analysis in (1) is equivalent to performing stratum-specific GEE analyses separately. Our approach attempts to identify common effects among some of the K strata and fuse them together to achieve higher statistical power. In addition, sparsity is induced to allow detection of grouped zero effects. In short, we develop a method of aggregated GEE with structural regularization to refine the grouping structure on the basis of individual covariates across strata. We establish asymptotic results on consistency and normality when both K and p go to infinity.

The paper is organized as follows. The motivating scientific problem from a longitudinal study of depression is introduced in Section 2. A review of fusion learning and penalized GEE is given in Section 3. The proposed approach is presented in details in Section 4. In Section 5, we apply our method to pattern-mixture models in

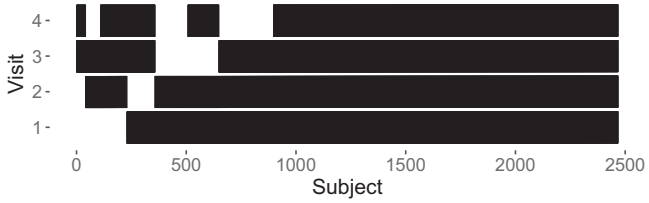

FIGURE 1 Response availability across the four longitudinal follow-up visits for IHS participants, grouped into eight response patterns. Black color indicates responses, and white color indicates nonresponses.

handling longitudinal missing-data. We evaluate the method via simulation experiments in Section 6 and present a real data analysis of predicting suicidal ideation in Section 7. We conclude with discussion in Section 8. Additional supporting results are included in the Web Appendices.

2 | MOTIVATING EXAMPLE: THE INTERN HEALTH STUDY

This paper is motivated by the Intern Health Study (IHS), an ongoing longitudinal cohort study that investigates depression in medical interns at institutions in the United States (Sen *et al.*, 2010). Medical interns are trainees in their first year of residency. This population is considered one of the highest risk groups to develop depression. The broader aim of this study is to detect important risk factors in connection to the development of depression, and subsequently, to implement constructive measures and policy interventions that may foster a healthier and more educational environment for medical trainees.

Participants are recruited at baseline before their official start of residency, and followed quarterly for up to four times in their internship year. Pre-internship risk factors, including demographic factors and depression history, are collected at the baseline visit. Time-varying internship risk factors, including mental health, anxiety level, work stress, work performance, and other questionnaires, are assessed at each follow-up visit. A specific outcome of interest, suicidal ideation (SI), is repeatedly measured at follow-ups. Focusing on predicting SI, an early alarming signal of depression, rather than actual suicides or attempts, allows us to take timely intervention measures. Also, suicidal ideation is observed more frequently than suicidal action thus more appealing from a modeling perspective.

Due to nonresponses in follow-ups, we either observe no information or complete response at any given visit for each individual. Figure 1 shows response availability for participants across the four visits, based on our study data set that consists of subjects recruited from 2012 to 2014. In the analysis of this longitudinal behavioral data, we take into account potential estimation biases due to potential nonignorable missingness, as we expect the decision of responding to a follow-up survey prompt is associated with their mental state (e.g., participant might ignore a follow-up because of heavy work load or loss of interest in engaging social activities that they committed previously). Such outcome-dependent missingness can in turn bias the main association of interest between SI and risk factors due to sampling bias associated with nonresponses. A common approach taken is to stratify subjects by missing-data patterns (Dawson, 1994), which enables the investigation of temporal heterogeneity on the associations. However, the method of stratification by data availability patterns may generate excessive strata, which can lead to over-attrition in stratum sample size, loss of statistical power, and even misleading results. These consequences are especially concerned when the number of follow-up visits is large, because the number of unique strata grows with the number of visits. This motivates a poststratification fusion strategy that aims to utilize shared information across strata to improve statistical efficiency in the context of GEEs.

3 | A REVIEW OF FUSION LEARNING

Fusion learning exploits similarity of parameters in a statistical model by aggregation, for examples, marginally by K-means clustering or conditionally by regularized regression, to produce subgrouping structures of parameters. Earlier work of fusion learning penalizes the distance between parameter estimates of biologically ordered genes on chromosomes to achieve piecewise constancy in estimation. This idea is extended to network-type fusion as considered in OSCAR (Bondell and Reich, 2008), grouping pursuit (Shen and Huang, 2010), and homogeneity pursuit (Ke *et al.*, 2015) by relaxing the ordering constraint. Penalties are further relaxed to learn more complicated parametric structures (Tibshirani and Taylor, 2011; Bach *et al.*, 2012). All of the above consider fusing similar parameter estimates within the scope of single data sets.

Some other work consider clustering parameter of similar meaning, but from various sample strata or data sets. The idea appears in the ANOVA setting to fuse the differences among the levels within each factor to achieve grouped estimation (Bondell and Reich, 2009). It is also studied in data integration problems to evaluate the comparative effectiveness of a same factor across different data sets (Tang and Song, 2016; Wang *et al.*, 2016), and to recover the structural differences and commonalities in heterogeneous graphical models (Hao *et al.*, 2018). Fusion of subject level estimates is proposed to achieve individualized

treatment effect clustering where each individual forms its own stratum (Ma and Huang, 2017). Most of the previous work ignores the serial dependency of longitudinal or clustered data, except Wang *et al.* (2016) that considers a penalized quadratic inference function for longitudinal data with no missing values. However, none of these studies consider when K diverges according to N, which is an important feature pertaining to sample stratification.

Our approach of poststratification fusion learning is developed in the framework of penalized GEE to account for correlations of repeated measurements. Penalized GEE has been studied by several groups. The penalized estimating equations is first studied by Fu (2003) to address collinearity issue through the bridge penalty. Johnson *et al.* (2008) relaxes the smoothness requirement in Fu (2003) to a more general discrete case with the consideration of variable selection, and establishes the oracle properties under a family of convex and nonconvex penalties. Wang *et al.* (2012) derives the oracle properties under a diverging number of covariates. All of these methods concern variable selection, but none consider fusion penalties for the need of parameter clustering across sample strata.

A key contribution of our approach is the ability to fuse parameter estimates from different sample strata to learn the underlying similarity among stratum-specific parameters in an integrative longitudinal data analysis. In a way, it improves statistical efficiency and simplifies the model. This procedure is particularly appealing when overstratification is suspected. The extension to nonconvex fusion learning in GEE brings forth new numerical algorithms; and the asymptotic rate of K provides appropriate guidance towards application practices.

4 | GEE-BASED FUSION LEARNING

4.1 | Fusion method

The proposed method takes initial estimates from stratum-specific GEE estimates from respective marginal longitudinal models (1) and proceeds to fuse similar estimates across strata via suitable regularization. To encourage parameter fusion and sparsity, we propose to solve for the model parameters jointly using the following penalized GEE

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{1}{N}\left(\boldsymbol{S}_1^T(\boldsymbol{\beta}_{\cdot 1}), \ldots, \boldsymbol{S}_K^T(\boldsymbol{\beta}_{\cdot K})\right)^T - \boldsymbol{q}_{\lambda,\alpha}(|\boldsymbol{D}\boldsymbol{\beta}|)\text{sign}(\boldsymbol{D}\boldsymbol{\beta}) = \boldsymbol{0}, \tag{3}$$

with stratum-specific estimating equations $\boldsymbol{S}_k(\boldsymbol{\beta}_{\cdot k}) = \sum_{i=1}^{n_k} \boldsymbol{S}_{ki}(\boldsymbol{\beta}_{\cdot k})$ for $\boldsymbol{S}_{ki}(\boldsymbol{\beta}_{\cdot k})$ given in (2). By definition of $\boldsymbol{\beta} = (\boldsymbol{\beta}_{1\cdot}^T, \ldots, \boldsymbol{\beta}_{p\cdot}^T)^T$, $\boldsymbol{\beta}_{\cdot k} = \boldsymbol{C}_k\boldsymbol{\beta}$ where $\boldsymbol{C}_k = (\boldsymbol{e}_k, \boldsymbol{e}_{K+k}, \ldots, \boldsymbol{e}_{(p-1)K+k})^T$ is an extraction matrix with $\boldsymbol{e}_i$ being the ith unit vector of length pK. We use a block-diagonal matrix $\boldsymbol{D} = \text{block-diag}(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_p)$ that consists of matrices $\boldsymbol{D}_j$ to specify the contrasts of coefficients in $\boldsymbol{\beta}_{j\cdot} = (\beta_{j1}, \ldots, \beta_{jK})$ for covariate X_j, $j = 1, \ldots, p$. The penalty term $\boldsymbol{q}_{\lambda,\alpha}(\boldsymbol{x})$ with tuning parameters $\lambda \geq 0, \alpha \geq 0$ is a pK-dimensional element-wise function vector whose specific form is dependent on $\boldsymbol{D}$ and will be defined later, and $\text{sign}(\boldsymbol{\beta}) = (\text{sign}(\beta_{11}), \ldots, \text{sign}(\beta_{pK}))^T$ is pK-dimensional element-wise sign function vector. Letting $\boldsymbol{D}_j = \boldsymbol{I}_K$ reduces to variable selection for GEE (Fu, 2003; Johnson *et al.*, 2008; Wang *et al.*, 2012). Deviating from the focus of variable selection in most of the previous work, we consider identifying homogeneous structures of similar parameters within $\boldsymbol{\beta}_{j\cdot}$ via matrix $\boldsymbol{D}_j$ that characterizes contrasts or differences between pairs of elements in $\boldsymbol{\beta}_{j\cdot}$'s for fusion, $j = 1, \ldots, p$. We consider nonconvex penalties that have been shown to be numerically more robust than convex ones for preserving strong signals.

Next, we provide a practical formulation of an order-dependent $\boldsymbol{D}$ matrix that allows both parameter fusion and variable selection. Previous work (Ke *et al.*, 2015; Wang *et al.*, 2016) has shown great computational efficiency gain by simplifying the contrast structures without loss of estimation precision. Hence, we reduce the number of contrasts by assuming a certain ordering, and provide theoretical assurance of such reduction when ordering is based on stratified GEE estimates. For a simple example of $K = 4$, if the ordering of elements $\boldsymbol{\beta}_{j\cdot}$ for covariate X_j is $0 < \beta_{j2} = \beta_{j4} < \beta_{j3} = \beta_{j1}$, we would define $\boldsymbol{D}_j$ as

$$\boldsymbol{D}_j = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & -1 \\ 1 & 0 & -1 & 0 \end{pmatrix} \begin{matrix} \} \text{selection} \\ \\ \} \text{fusion} \\ \\ \end{matrix}.$$

Rows 2 to $K = 4$ of $\boldsymbol{D}_j$ specify the penalized contrast of adjacent parameters. In addition, we let the first row be the reference parameter, or the "anchor" parameter, and allow it to be shrunk toward zero to enjoy sparsity. This reference parameter is chosen to be the one with the smallest distance to zero, that is, β_{j2} in this example. Similarly, we can derive $\boldsymbol{D}_j$ for each individual $\boldsymbol{X}_j$, $j = 1, \ldots, p$, each may have different structure. The true ordering matrix is not unique due to clustered coefficient values, and is discussed in Section 4.3. Denote $\boldsymbol{\theta} = (\boldsymbol{\theta}_{1\cdot}^T, \ldots, \boldsymbol{\theta}_{p\cdot}^T)^T = \boldsymbol{D}\boldsymbol{\beta}$, where $\boldsymbol{\theta}_{j\cdot} = (\theta_{j1}, \ldots, \theta_{jk})^T = \boldsymbol{D}_j\boldsymbol{\beta}_{j\cdot}, j = 1, \ldots, p$. To separately control sparsity and fusion shrinkage, we introduce an additional tuning parameter so that sparsity may be explained in the rate of fusion. For constants $\lambda \geq 0, \alpha \geq 0$, we let

$$\boldsymbol{q}_{\lambda,\alpha}(|\boldsymbol{\theta}|) = (q_{\alpha\lambda}(\theta_{11}), q_\lambda(\theta_{12}), \ldots, q_\lambda(\theta_{1K}), \ldots, q_{\alpha\lambda}(\theta_{p1}), q_\lambda(\theta_{p2}), \ldots, q_\lambda(\theta_{pK}))^T.$$

Tuning parameter λ controls the magnitude of fusion and a larger value corresponds to stronger fusion. The choice of α depends on the goal of estimation; more emphasis is given to fusion when α is small, and vice versa for larger α. When variable selection (estimating zeros) is not of interest, we set $\alpha = 0$ so that $q_{\alpha\lambda}(\theta_{j1}) = 0, j = 1, \dots, p$. When $\alpha = 1$, there is equal strength of selection and fusion, and we simply write the penalty term of (3) as $\boldsymbol{q}_\lambda(|\boldsymbol{D\beta}|)$. Function $q_\lambda(\cdot)$ is the subdifferential of any penalty function. Some popular $q_\lambda(\cdot)$ functions include (a) LASSO: $q_\lambda(x) = \lambda,\ x > 0$, (b) SCAD (Fan and Li, 2001): $q_\lambda(x) = \lambda\{I(x \le \lambda) + \frac{(a\lambda - x)_+}{(a-1)\lambda} I(x > \lambda)\},\ x > 0,\ a > 2$, and $t_+ = \max(0, t)$, and (c) MCP (Zhang, 2010): $q_\lambda(x) = \lambda \frac{(a\lambda - x)_+}{a\lambda},\ x > 0,\ a > 1$. A comparison of the three is shown in Web Appendix F. Although LASSO is simpler and guarantees a global optimal solution, its regularized solution tends to over-shrink large coefficients, resulting in too many subgroups (Ma and Huang, 2017). We choose MCP as it provides the most distinguishable clustering pattern in its solution paths and fix $a = 1.5$ as similar to Wang *et al.* (2012).

4.2 | Estimation via iterative algorithm

We propose an efficient algorithm for obtaining an approximate regularized solutions to (3). Equivalently, we attempt to solve the following system of equations:

$$\boldsymbol{S}(\boldsymbol{\beta}) - N\boldsymbol{q}_\lambda(|\boldsymbol{D\beta}|)\text{sign}(\boldsymbol{D\beta}) = \boldsymbol{0}, \tag{4}$$

where $\boldsymbol{S}(\boldsymbol{\beta}) = (\boldsymbol{S}_1^T(\boldsymbol{\beta}_{\cdot 1}), \dots, \boldsymbol{S}_K^T(\boldsymbol{\beta}_{\cdot K}))^T$. The algorithm concerns a general $m \times pK$ contrast matrix $\boldsymbol{D}$, with $m \le pK$ and $\text{rank}(\boldsymbol{D}) = m$, which includes the case considered in Section 4.1.

To begin, we consider $m = pK$, thus, $\boldsymbol{D}$ is full-rank invertible. Denote $\theta = (\theta_{1\cdot}^T, \dots, \theta_{p\cdot}^T)^T = \boldsymbol{D\beta}$, where $\theta_{j\cdot} = (\theta_{j1}, \dots, \theta_{jK})^T$ such that $\theta_{j\cdot} = \boldsymbol{D}_j\boldsymbol{\beta}_{j\cdot}, j = 1, \dots, p$. Thus $\boldsymbol{\beta} = \boldsymbol{D}^{-1}\theta$. We write (4) as a function of θ, and effectively solve the following system of equations:

$$\widetilde{\boldsymbol{U}}(\theta) = \widetilde{\boldsymbol{S}}(\theta) - N\boldsymbol{q}_\lambda(|\theta|)\text{sign}(\theta) = \boldsymbol{0}, \tag{5}$$

where $\widetilde{\boldsymbol{S}}(\theta) = \boldsymbol{S}(\boldsymbol{D}^{-1}\theta) = (\boldsymbol{S}_1^T(\boldsymbol{C}_1\boldsymbol{D}^{-1}\theta), \dots, \boldsymbol{S}_K^T(\boldsymbol{C}_K\boldsymbol{D}^{-1}\theta))^T$, and $\boldsymbol{C}_k\boldsymbol{D}^{-1}\theta = \boldsymbol{C}_k\boldsymbol{\beta} = \boldsymbol{\beta}_{\cdot k},\ k = 1, \dots, K$. In fact, $\widetilde{\boldsymbol{U}}(\theta)$ is a penalized GEE whose solution $\widehat{\theta}$ may be efficiently obtained by an iterative algorithm. Following Wang *et al.* (2012), we estimate $\widehat{\theta}$ in (5) by coupling the Newton-Raphson iterative algorithm and the minorization-maximization (MM) algorithm for nonconvex penalty (Hunter and Li, 2005). The key idea in the originally studied penalized likelihood problems is to perturb the penalty function slightly so it becomes differentiable (Fan and Li, 2001). In GEE, the element-wise derivative of the perturbed penalty corresponds to $Nq_\lambda(|\theta_{jk}|)\text{sign}(\theta_{jk})\frac{|\theta_{jk}|}{\epsilon + |\theta_{jk}|}, j = 1, \dots, p, k = 1, \dots, K$, for some $\epsilon > 0$ ensuring the denominator is well defined. To proceed, we select a small $\epsilon = 10^{-6}$ and obtain the penalized GEE estimate $\widehat{\theta}$ for θ, as the solution that approximately satisfies

$$S_{kj}(\boldsymbol{C}_k\boldsymbol{D}^{-1}\theta) - Nq_\lambda(|\widehat{\theta}_{jk}|)\text{sign}(\widehat{\theta}_{jk})\frac{|\widehat{\theta}_{jk}|}{\epsilon + |\widehat{\theta}_{jk}|},$$
$$j = 1, \dots, p,\ k = 1, \dots, K,$$

where $S_{kj}(\cdot)$ denotes the jth element of $\boldsymbol{S}_k(\cdot)$. The updating step for $\widehat{\theta}$, at iteration b, is

$$\begin{aligned}
\widehat{\theta}^b &= \widehat{\theta}^{b-1} + \{\widetilde{\boldsymbol{H}}(\widehat{\theta}^{b-1}) + N\boldsymbol{J}(\widehat{\theta}^{b-1})\}^{-1}\{\widetilde{\boldsymbol{S}}(\widehat{\theta}^{b-1}) \\
&\quad - N\boldsymbol{J}(\widehat{\theta}^{b-1})\widehat{\theta}^{b-1}\}, \text{with} \\
\widetilde{\boldsymbol{H}}(\widehat{\theta}^{b-1}) &= \text{block-diag}\{\boldsymbol{H}_1(\boldsymbol{C}_1\boldsymbol{D}^{-1}\widehat{\theta}^{b-1}), \dots, \\
&\quad \boldsymbol{H}_K(\boldsymbol{C}_K\boldsymbol{D}^{-1}\widehat{\theta}^{b-1})\}\boldsymbol{C}\boldsymbol{D}^{-1}, \\
\boldsymbol{J}(\widehat{\theta}^{b-1}) &= \text{diag}\left\{\frac{q_{\lambda_{11}}(|\widehat{\theta}_{11}^{b-1}|_+)}{\epsilon + |\widehat{\theta}_{11}^{b-1}|}, \dots, \frac{q_{\lambda_{1K}}(|\widehat{\theta}_{1K}^{b-1}|_+)}{\epsilon + |\widehat{\theta}_{1K}^{b-1}|}, \dots, \right. \\
&\quad \left. \frac{q_{\lambda_{p1}}(|\widehat{\theta}_{p1}^{b-1}|_+)}{\epsilon + |\widehat{\theta}_{p1}^{b-1}|}, \dots, \frac{q_{\lambda_{pK}}(|\widehat{\theta}_{pK}^{b-1}|_+)}{\epsilon + |\widehat{\theta}_{pK}^{b-1}|}\right\}, \qquad (6) \\
\boldsymbol{H}_k(\boldsymbol{C}_k\boldsymbol{D}^{-1}\widehat{\theta}^{b-1}) &= \sum_{i=1}^{n_k} w_{ki}\boldsymbol{X}_{ki}^T\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{C}_k\boldsymbol{D}^{-1}\widehat{\theta}^{b-1}) \\
&\quad \times \boldsymbol{R}_k^{-1}(\widehat{\tau}_k^{b-1})\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{C}_k\boldsymbol{D}^{-1}\widehat{\theta}^{b-1})\boldsymbol{X}_{ki},
\end{aligned}$$

and $\boldsymbol{C} = (\boldsymbol{C}_1^T, \dots, \boldsymbol{C}_K^T)^T$. The update of $\boldsymbol{J}(\widehat{\theta})$ in (6) takes $\lambda_{jk} = \alpha\lambda$ if $k = 1$ and $\lambda_{jk} = \lambda$ if $k \ne 1, \forall j$, to achieve both fusion and variable selection. It can be further extended using the idea of adaptive penalties (Zou, 2006) to avoid the tuning of α. In this case, $\boldsymbol{J}(\widehat{\theta})$ will be updated by taking $\lambda_{jk} = \frac{1}{|\widehat{\theta}_{jk}^0|^\gamma}\lambda, \gamma \ge 0$, where $\widehat{\theta}_{jk}^0$ is the unpenalized estimate, for all j, k. We recommend γ to be less than 1 to avoid extreme values in the denominators and set $\gamma = 0.5$ throughout this paper.

Both $\widehat{\tau}_k$ and $\widehat{\phi}_k$ (involved through $\boldsymbol{A}_{ki}$) can be updated sequentially using the method of moments as suggested in the standard GEE (see section 3.3 of Liang and Zeger (1986)). For simplicity, we assume $\boldsymbol{R}_k$'s, τ_k's, and ϕ_k's are common across strata, thus τ_k's and ϕ_k's can be omitted in the iteration. Note that these parameters do not affect the mean-zero assumption of the GEE, that is, $E\boldsymbol{S}_{ki}(\boldsymbol{\beta}_{\cdot k}) = \boldsymbol{0}$ for any $\boldsymbol{R}_k$, τ_k, ϕ_k. After convergence, we obtain

$\widehat{\boldsymbol{\beta}} = \boldsymbol{D}^{-1}\widehat{\boldsymbol{\theta}}^b$ for $\widehat{\boldsymbol{\theta}}^b$ from the last iteration as our final estimate. In practice, ordering of coefficients across strata is usually unknown without prior knowledge assistance. We propose to use estimated ordering from the stratified estimates by solving (1). Using the initial unpenalized coefficient estimates, we estimate the structure of $\boldsymbol{D}$, denoted as $\widehat{\boldsymbol{D}}$, and proceed with fusion. Refer to Wang *et al.* (2016, Proposition 1) for the theoretical justification of using $\widehat{\boldsymbol{D}}$ in the fused LASSO. Model selection is done by the tunning of λ with the extended Bayesian information criterion (EBIC) (Chen and Chen, 2012), with details given in Web Appendix A.

If $m < pK$, the above algorithm may still be applied by augmenting matrix $\boldsymbol{D}$ into a full rank matrix $\widetilde{\boldsymbol{D}}$ by appending $(pK - m)$ rows that are orthogonal to those in $\boldsymbol{D}$. Subsequently, the corresponding $(pK - m)$ elements in resulting $\theta = \widetilde{\boldsymbol{D}}\boldsymbol{\beta}$ will not be penalized, and can be achieved by setting the corresponding λ_{jk} to zero in (6). On the other hand, contrasts that over-identify the model such that $m > pK$ where $\text{rank}(\boldsymbol{D}) = pK$ create redundancy in optimization. Numerical solutions can still be obtained by the alternating direction method of multipliers, which, however, requires extra iteration steps.

4.3 | Theoretical properties

Main results in Theorem 2 shows the asymptotic properties of the proposed estimator when both p and K diverge. The key depends on Theorem 1 for estimator $\widetilde{\boldsymbol{\theta}}$ from (5) when the true parameter ordering is given, that is, $\boldsymbol{D} = \boldsymbol{D}_*$, and Lemma 1 for requirements on the estimated ordering, $\widehat{\boldsymbol{D}}$. Denote $n_{\text{inf}} = \inf_{k\in\{1,\dots,K\}} n_k$. Let the number of true parameter clusters for each covariate j (ie, number of nonzero adjacent differences) be s_j, and $s = \sum_{j=1}^{p} s_j$, which may diverge along with K. We use $\boldsymbol{D}_*$ to denote the full-rank invertible contrast matrix constructed based on the true parameter values $\boldsymbol{\beta}_*$ as described in Section 4.2, and $\widehat{\boldsymbol{D}}$ to denote the data-dependent estimate of $\boldsymbol{D}_*$. Correspondingly, $\theta_* = \boldsymbol{D}_*\boldsymbol{\beta}_*$ is a sparse vector with support at the set $\mathcal{A} = \{j : \theta_{*j} \neq 0, j = 1, \dots, pK\}$, and subscript $\mathcal{A}$ is used to denote the corresponding subvector or submatrix corresponding to the indices in $\mathcal{A}$. We use $g_j(k), k = 1, \dots, K$, to denote the cluster membership of true parameter $\beta_{*j\cdot}$ of covariate $\boldsymbol{X}_j$, such that $g_j(k) \in \{1, \dots, s_j\}$ and $\beta_{*jk} = \beta_{*jk'}$ only if $g_j(k) = g_j(k')$. Denote $\boldsymbol{R}_{k*}$ the true correlation matrix of stratum k, and $\overline{\boldsymbol{R}}_k$ a constant positive-definite limiting matrix of $\widehat{\boldsymbol{R}}(\tau_k)$. In the following discussion, we assume that $s = o(N^{1/3})$, $||\boldsymbol{R}(\widehat{\tau}_k)^{-1} - \overline{\boldsymbol{R}}_k^{-1}||_2 = O(\sqrt{s/n_{\text{inf}}K})$, $k = 1, \dots, K$. Regularity conditions (C1)-(C5), Lemma 1, and proofs are given in Web Appendices B and C.

Theorem 1. *Under conditions (C1)-(C5), as $n_{\text{inf}} \to \infty$, there exist a solution $\widetilde{\boldsymbol{\theta}}$ to (5) with $\boldsymbol{D} = \boldsymbol{D}_*$ that satisfies (a) $P(\widetilde{\theta}_{\mathcal{A}^c} = \boldsymbol{0}) \to 1$, and (b) $\forall \boldsymbol{a} \in \mathbb{R}^s$ such that $||\boldsymbol{a}|| = 1$, $\boldsymbol{a}^T \boldsymbol{M}^*_{\mathcal{A}}(\boldsymbol{\beta}_*)^{-1/2}\widetilde{\boldsymbol{H}}^*_{\mathcal{A}}(\boldsymbol{\beta}_*)(\widetilde{\boldsymbol{\theta}}_{\mathcal{A}} - \theta_{*\mathcal{A}}) \xrightarrow{d} \mathcal{N}(0, 1)$, where*

$$\widetilde{\boldsymbol{H}}^*_{\mathcal{A}\cup\mathcal{A}^c}(\boldsymbol{\beta}_*) = \text{block-diag}\{\boldsymbol{H}_1(\boldsymbol{\beta}_{*\cdot 1}), \dots, \boldsymbol{H}_K(\boldsymbol{\beta}_{*\cdot K})\}\boldsymbol{C}\boldsymbol{D}_*^{-1};$$

$$\boldsymbol{M}^*_{\mathcal{A}\cup\mathcal{A}^c}(\boldsymbol{\beta}_*) = \text{block-diag}\{\boldsymbol{M}_1(\boldsymbol{\beta}_{*\cdot 1}), \dots, \boldsymbol{M}_K(\boldsymbol{\beta}_{*\cdot K})\};$$

$$\boldsymbol{H}_k(\boldsymbol{\beta}_{*\cdot k}) = \sum_{i=1}^{n_k} \boldsymbol{X}_{ki}^T \boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}_{*\cdot k})\overline{\boldsymbol{R}}_k^{-1}\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}_{*\cdot k})\boldsymbol{X}_{ki} \text{ and}$$

$$\boldsymbol{M}_k(\boldsymbol{\beta}_{*\cdot k}) = \sum_{i=1}^{n_k} \boldsymbol{X}_{ki}^T \boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}_{*\cdot k})\overline{\boldsymbol{R}}_k^{-1}\boldsymbol{R}_{k*}\overline{\boldsymbol{R}}_k^{-1}\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}_{*\cdot k})\boldsymbol{X}_{ki},$$

$$k = 1, \dots, K.$$

Theorem 1 follows from the asymptotic results of penalized GEE with diverging number of covariates (Wang *et al.*, 2012) after reparameterization. The rates in (C5) imply that when the number of unique true parameters s is fixed and $\tau_k = \tau$, the total number of parameters pK is allowed to diverge in the sense that $pK = O(N)$. This reduces to the result in Wang *et al.* (2012, Theorem 1) when K is held fixed. On the other hand, if p is fixed, intuitively, consistency would require that sample size increases faster than the number of strata, that is, $K = o(n_{\text{inf}})$. However, due to information sharing by fusion, K can be of order $O(N)$.

Corollary 1. *Under the same conditions as Theorem 1 but with s fixed such that the dimension of $\widetilde{\boldsymbol{\theta}}_{\mathcal{A}}$ is finite, as $n_{\text{inf}} \to \infty$, there exists a solution $\widetilde{\boldsymbol{\theta}}$ to (5) with $\boldsymbol{D} = \boldsymbol{D}_*$, such that $\widetilde{\boldsymbol{\theta}}_{\mathcal{A}}$ follows an asymptotic normal distribution with mean $\theta_{*\mathcal{A}}$ and variance $N^{-1}\boldsymbol{\Sigma}_{\mathcal{A}}$, where $\boldsymbol{\Sigma}^{-1}_{\mathcal{A}} = \widetilde{\boldsymbol{H}}^*_{\mathcal{A}}(\boldsymbol{\beta}_*)^T \boldsymbol{M}^*_{\mathcal{A}}(\boldsymbol{\beta}_*)^{-1}\widetilde{\boldsymbol{H}}^*_{\mathcal{A}}(\boldsymbol{\beta}_*)$.*

Corollary 1 directly results from Theorem 1, yet they both rely on knowing $\boldsymbol{D}_*$. The dependency of $\widetilde{\boldsymbol{\theta}}$ on $\boldsymbol{D}_*$ is restrictive in practice because the true parameter ordering is usually unknown. We will need more stringent requirement on the speed of divergence of K when the initial ordering $\boldsymbol{D}_*$ is unknown to guarantee similar results. Theorem 2 relaxes the dependence of $\boldsymbol{D}_*$ required above for the estimator $\widehat{\boldsymbol{\theta}}$ under $\boldsymbol{D} = \widehat{\boldsymbol{D}}$, where $\widehat{\boldsymbol{D}}$ is the contrast matrix based on the unpenalized estimates from (1). For simplicity, we assume s is fixed in Theorem 2 and Corollary 2, while still allow both p and K to diverge.

Theorem 2. *Suppose that conditions (C1)-(C5) hold, $N^{-1}p^2 = o(1)$, $K = O(N^{1/4-\xi})$ with $\xi \in (0, 1/4]$, $n_{\text{inf}} \to \infty$, then the solution $\widehat{\boldsymbol{\theta}}$ to the penalized GEE in (5) under $\boldsymbol{D} = \widehat{\boldsymbol{D}}$ satisfies (a) $P(\widehat{\boldsymbol{\theta}}_{\mathcal{A}^c} = \boldsymbol{0}) \to 1$, and (b) $\widehat{\boldsymbol{\theta}}_{\mathcal{A}}$ follows an*

asymptotic normal distribution with mean $\theta_{*\mathcal{A}}$ *and variance* $N^{-1}\Sigma_{\mathcal{A}}$.

Corollary 2. *Under the same conditions as Theorem 2, for each* $j \in \{1, \ldots, p\}$, *the estimator* $\widehat{\boldsymbol{\beta}}$ *obtained from the penalized GEE (4) with* $\boldsymbol{D} = \widehat{\boldsymbol{D}}$ *satisfies* $P(\widehat{\beta}_{jk} = \widehat{\beta}_{jk'}) \to 1$, $\forall k, k' \in \{k, k' | g_j(k) = g_j(k')\}$.

The rate established for K in Theorem 2 is dependent on both p and N. In the case of fixed-dimensional covariate p, the number of strata $K = O(N^{1/2-\xi})$, $\xi \in (0, 1/2]$. In regards to $\boldsymbol{\beta}$, Corollary 2 states that the clustering structure in $\boldsymbol{\beta}$ can be estimated consistently. However, the joint asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ cannot be directly obtained from Theorem 2 because it involves elements of $\widehat{\theta}$ in $\mathcal{A}^c$ whose asymptotic distribution is unknown. Bias-correction approaches (Van de Geer *et al.*, 2014; Zhang and Zhang, 2014) may be adopted, however, challenges remain as the size of $\widehat{\boldsymbol{\beta}}$ diverges.

5 | FUSION LEARNING IN PATTERN-MIXTURE MODELS

We apply the proposed poststratification fusion learning method for the analysis of incomplete longitudinal data using pattern-mixture models. In the missing-data literature, pattern-mixture model is one of the primary modeling methods to deal with not missing at random mechanisms (NMAR), where stratification according to missing-data patterns is taken. We propose to perform poststratification fusion learning in this framework to investigate the heterogeneous interaction of missing-data pattern and time varying covariates. After stratification by missing-data patterns, separate complete data analyses can be conducted.

Due to dropouts or intermittent nonresponses, suppose we observe K distinct missing-data patterns. By stratifying the N subjects according to the patterns, each stratum has a sample size n_k, $k = 1, \ldots, K$. Subjects in the kth response pattern are observed at certain visits as indexed by $\mathcal{L}_k = \{t : r_{kt} = 1, t = 1, \ldots, T\}$, where $r_{kt} = 0$ if no response at visit t and $r_{kt} = 1$ *vice versa*. In one extreme case where missing data are missing completely at random (MCAR), we can estimate the common $\boldsymbol{\beta}$ using the sum of GEE across all patterns

$$\sum_{k=1}^{K} \sum_{i=1}^{n_k} \boldsymbol{S}_{ki}(\boldsymbol{\beta}) = \boldsymbol{0}, \tag{7}$$

where $\boldsymbol{S}_{ki}(\boldsymbol{\beta})$ is defined as (2). Note that the number of repeated measurements may be different for different k, and are uniquely determined by the missing-data patterns outlined in $\mathcal{L}_k$. Equation (7) is the GEE under the assumption of MCAR, or ignorable missingness defined in Little (1993), in which the underlying relationships between outcomes and covariates are not affected by missing-data patterns. When such MCAR assumption is violated, GEE (7) may be invalid due to biased sampling across missing-data patterns (Song, 2007).

To test the MCAR assumption, Chen and Little (1999) proposes a Wald-type test, where under the null, there exists a common true parameter vector $\boldsymbol{\beta}_*$ such that $E\{\boldsymbol{S}_k(\boldsymbol{\beta}_*)\} = \boldsymbol{0}$ for all K strata. Assuming parameters of interest are identifiable in every stratum, and letting $\widehat{\boldsymbol{\beta}}_k$ and $\widehat{\boldsymbol{\Sigma}}_k$ be the GEE estimator and sandwich variance estimator from stratum k, we can obtain a meta estimator $\widehat{\boldsymbol{\beta}}_c = (\sum_{k=1}^{K} \widehat{\boldsymbol{\Sigma}}_k^{-1})^{-1} \sum_{k=1}^{K} \widehat{\boldsymbol{\Sigma}}_k^{-1} \widehat{\boldsymbol{\beta}}_k$. Then, Chen and Little (1999)'s test statistic takes the form:

$$d = \sum_{k=1}^{K} (\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_c)^T \boldsymbol{\Sigma}_k^{-1} (\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_c), \tag{8}$$

which asymptotically follows a χ_r^2 distribution with degrees of freedom $r = p(K-1)$. Similar tests are developed to test the same null hypothesis under different scenarios (Diggle, 1989; Qu and Song, 2002; Qu *et al.*, 2011). When the null is rejected, a stratified analysis is required to generate interpretable results. As pointed above, any fully stratified model based on missing-data patterns specifies model parameters to be distinct across strata, leading to potential overparameterization. This is again an issue caused by overstratification that has not been addressed in the current literature. Arguably when the number of repeated measurements increases, the number of missing-data patterns may become very large, so the sample size of each stratum would be too small to provide reliable and meaningful results.

Our poststratification fusion learning developed in this paper can greatly relax simultaneously the MCAR assumption and the NMAR assumption with the flexibility on departure from the full homogeneous parameterization under MCAR and/or from the full heterogeneous parameter under NMAR. A legitimate decision may be reached in between the two extreme cases. The above test (8) is only able to assess the MCAR assumption, and is not designed to understand partially homogeneous structures.

In lieu of an exclusive conditional distribution given each missing pattern, a more realistic situation is that there exist some strata that may share some common parameter structures due to various reasons, such as over stratification. In other words, two data generation machines under missing-data patterns k and k', $P(\boldsymbol{Y}|k, \boldsymbol{X})$ and $P(\boldsymbol{Y}|k', \boldsymbol{X})$, respectively, may share a common parameter for covariate

$\boldsymbol{X}_j$ when $g_j(k) = g_j(k')$. Thus, conducting poststratification analysis is needed in the context of pattern-mixture model based analysis of longitudinal data with nonignorable missing values.

6 | MONTE-CARLO SIMULATION EXPERIMENTS

Multiple methods are applied to pattern-mixture models for studying simulated nonignorable missing-data problems. Both linear and logistic models are presented, in which we simulate longitudinal data with $T = 4$ repeated measurements. Samples are generated from $K = 8$ strata corresponding to missing-data patterns $\{0011, 0110, 0111, 1011, 1101, 1100, 1110, 1111\}$, with 0 means missing and 1 otherwise. For simplicity, we let $n_k = n = 100$ for all k.

For the linear model with continuous outcomes, the following model is used to simulate experimental data for the (k, i, t)-trios, $k = 1, \ldots, K, i = 1, \ldots, n$, and $t \in \mathcal{L}_k$:

$$Y_{kit} = \beta_{k1}X_{kit1} + \beta_{k2}X_{kit2} + \beta_{k3}X_{kit3} + \beta_{k4}X_{ki4} + \beta_{k5}X_{ki5} + \epsilon_{kit},$$

where $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{k5})^T$ is the model coefficient vector under stratified pattern k, and ϵ_{kit} is the temporally correlated marginal error. For subject i in stratum k, X_{kit1} and X_{kit2} are drawn independently from Bernoulli distributions each with 0.5 success rate, and X_{kit3}, X_{ki4}, and X_{ki5} are drawn independently from the standard normal distribution, for $t \in \mathcal{L}_K$. Covariates X_1, X_2, and X_3 are time-varying, whereas X_4 and X_5 are constant. To induce correlation among repeated responses, $\boldsymbol{\epsilon}_{ki} = (\epsilon_{ki1}, \epsilon_{ki2}, \epsilon_{ki3}, \epsilon_{ki4})^T$ is drawn from a multivariate normal distribution, $\mathcal{N}_4(\mathbf{0}, \phi\boldsymbol{R}(\tau))$, and only the elements in $\mathcal{L}_k$ are kept. In the simulation experiments, we let $\boldsymbol{R}(\tau)$ be exchangeable with true correlation coefficient $\tau = 0.6$. The true regression coefficients of covariates X_1 (binary), X_3 (continuous), and X_5 (continuous) are set to be homogeneous across strata, and the true coefficients of X_2 (binary) and X_4 (continuous) are heterogeneous. Specifically, $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1K})^T = (0.5, \ldots, 0.5)^T$, $\boldsymbol{\beta}_3 = (\beta_{31}, \ldots, \beta_{3K})^T = (0.3, \ldots, 0.3)^T$, and $\boldsymbol{\beta}_5 = (\beta_{51}, \ldots, \beta_{5K})^T = (0, \ldots, 0)^T$. The true coefficients of X_2 and X_4 are each divided into G distinct-valued groups, $G \leq K$, with δ controlling the gap between values of distinct groups (ie, signal strength) and $G \in \{1, 2, 3\}$. Note that $G = 1$ corresponds to the MCAR setting, where the true responses-covariate relationships are homogeneous and independent of the missing-data patterns; in this case, we let $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2K})^T = (0.3, \ldots, 0.3)^T$ and $\boldsymbol{\beta}_4 = (\beta_{41}, \ldots, \beta_{4K})^T = (0, \ldots, 0)^T$. When $G = 2$, we split elements in $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_4$ into two clusters based on binomial draws $\text{Binom}(K, 1/G)$ and discard draws with only one group; the coefficients are then set so that one cluster is larger than the other by δ. For example, we set $\boldsymbol{\beta}_2 = (0.3, 0.3 + \delta, 0.3, 0.3, 0.3, 0.3 + \delta, 0.3, 0.3 + \delta)^T$ and $\boldsymbol{\beta}_4 = (0, \delta, 0, 0, 0, \delta, 0, \delta)^T$. The clusterings of $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_4$ are aligned for ease of presentation but is not required. Similarly, we create heterogeneous groups for $G = 3$ by subtracting δ from coefficients in the third cluster.

Our method is compared with (a) the homogeneous GEE assuming MCAR, (b) the stratified GEE, and (c) a two-stage GEE approach described as follows: based on a test for parameter homogeneity (8), use (a) if fail to reject the null (MCAR) and (b) otherwise. The metrics we use to evaluate the methods include mean squared errors of the estimated coefficients and the average number of groups estimated.

Table 1 summarizes results for the linear model with the independent working correlation $\boldsymbol{R}(\cdot) = \boldsymbol{I}$, averaged from 100 replications. Results for the proposed fusion learning method are reported for sparsity tuning levels $\alpha = 0, 0.5, 1$, and adaptive, where α is the ratio between variable selection and fusion penalties. When $G = 1$ and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_5$ are all homogeneous, the proposed method ($\alpha = 0$) has consistently smaller MSE than stratified and two-stage GEE approaches, and close to the benchmark homogeneous GEE method. As δ becomes larger in $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_4$ when $G = 2$ or 3, the homogeneous method becomes worse in estimating $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_4$. In comparison, the stratified and two-stage GEE methods yield smaller MSE for $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_4$ when $\delta = 0.3$, but have very large MSE for estimating $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_5$. The fusion only method ($\alpha = 0$) produces small MSE for all $\boldsymbol{\beta}$ when $G > 1$. Additional results for the combined fusion and sparsity method ($\alpha = 0.5, 1$) suggest improved estimation of $\boldsymbol{\beta}_4$ and $\boldsymbol{\beta}_5$, which contains zero coefficient clusters, but also introduces bias to $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\beta}_3$ as expected for all shrinkage estimators. The adaptive tuning method provides balanced performance for all coefficients.

Table 2 summarizes results for the linear model using exchangeable working correlation matrix, averaged from 100 replications. By considering the dependency between longitudinal outcomes, we show general improvement in estimation efficiency reflected in MSE for all methods when compared to Table 1. Comparing between methods, all conclusions remain similar as to those drawn from Table 1. The distributions of estimates across all simulation replicates for all settings are plotted in Figure 2. The histograms show better separation of peaks for $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_4$ as δ increases, and better concentration at zero for truly zero coefficients as α increases. The distributions of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are wider because $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are binary covariates thus have larger variance estimates. In general, the proposed method is able to adequately detect parameter

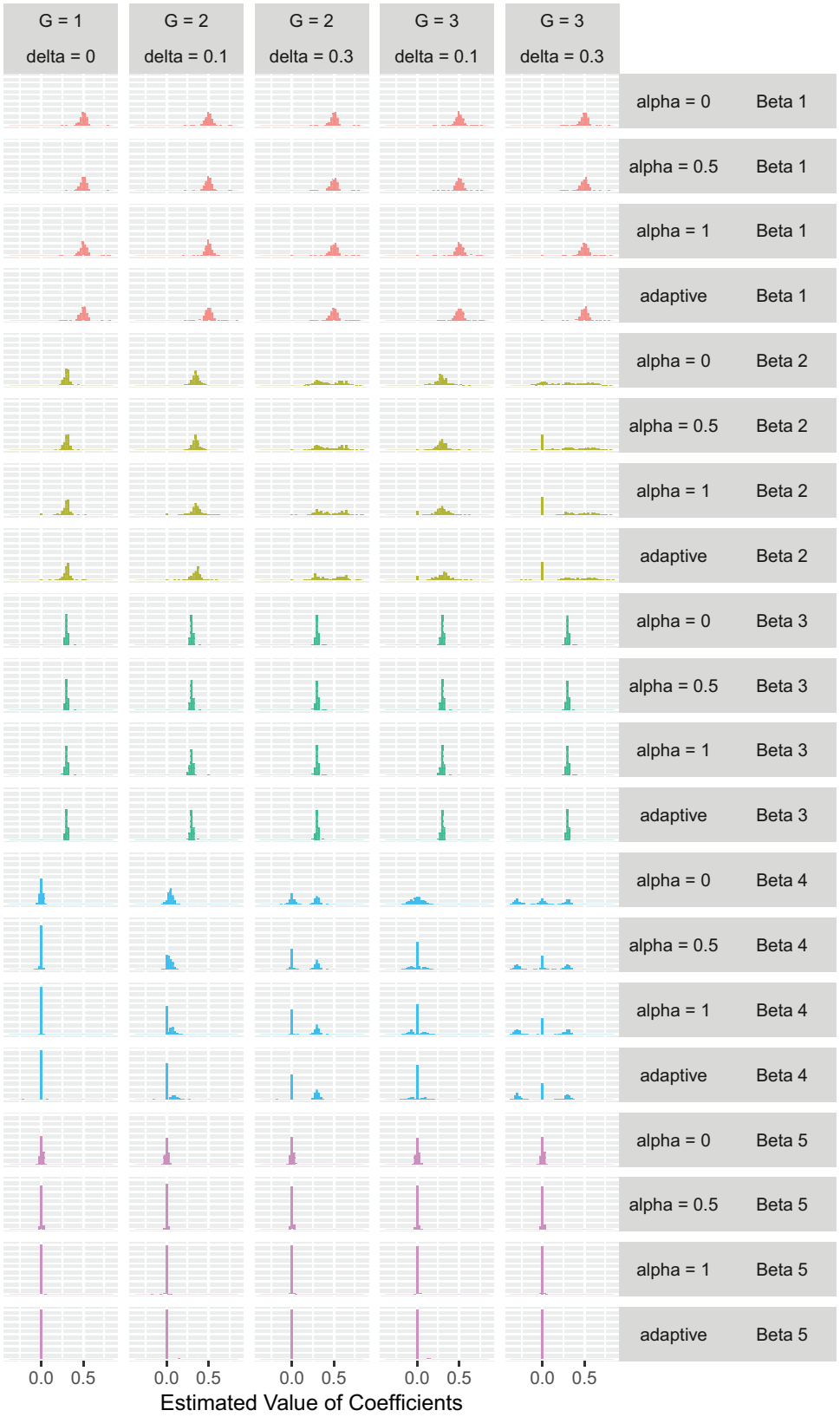

FIGURE 2 Histograms of coefficient estimates in the linear model with exchangeable working correlation matrix. This figure appears in color in the electronic version of this article.

TABLE 1 Mean squared error (×100) of GEE coefficient estimates in the linear model with independent working correlation matrix (ie, working correlation matrix misspecified). The true values of the coefficients across strata are specified in bold

G	δ	Method	β_1	β_2	β_3	β_4	β_5
			$\beta_1 = \mathbf{0.5}$	$\beta_2 = \mathbf{0.3}$	$\beta_3 = \mathbf{0.3}$	$\beta_4 = \mathbf{0}$	$\beta_5 = \mathbf{0}$
1	0.0	Homogeneous	0.158	0.144	0.041	0.056	0.044
1	0.0	Stratified	1.215	1.185	0.357	0.425	0.380
1	0.0	Two-stage	0.321	0.353	0.089	0.113	0.108
1	0.0	Proposed ($\alpha = 0$)	0.300	0.278	0.058	0.102	0.067
1	0.0	Proposed ($\alpha = 0.5$)	0.328	0.312	0.064	0.046	0.033
1	0.0	Proposed ($\alpha = 1$)	0.553	0.808	0.088	0.032	0.013
1	0.0	Proposed (adaptive)	0.549	0.725	0.069	0.045	0.015
			$\beta_1 = \mathbf{0.5}$	$\beta_2 \in \{\mathbf{0.3}, \mathbf{0.3} + \delta\}^8$	$\beta_3 = \mathbf{0.3}$	$\beta_4 \in \{\mathbf{0}, \delta\}^8$	$\beta_5 = \mathbf{0}$
2	0.1	Homogeneous	0.156	0.361	0.041	0.273	0.045
2	0.1	Stratified	1.215	1.185	0.357	0.425	0.380
2	0.1	Two-stage	0.544	0.655	0.136	0.325	0.159
2	0.1	Proposed ($\alpha = 0$)	0.362	0.521	0.062	0.318	0.060
2	0.1	Proposed ($\alpha = 0.5$)	0.430	0.562	0.061	0.355	0.031
2	0.1	Proposed ($\alpha = 1$)	0.619	0.989	0.101	0.380	0.031
2	0.1	Proposed (adaptive)	0.612	0.874	0.073	0.432	0.021
2	0.3	Homogeneous	0.155	2.097	0.042	2.011	0.049
2	0.3	Stratified	1.215	1.185	0.357	0.425	0.380
2	0.3	Two-stage	1.215	1.185	0.357	0.425	0.380
2	0.3	Proposed ($\alpha = 0$)	0.535	1.171	0.057	0.244	0.062
2	0.3	Proposed ($\alpha = 0.5$)	0.539	1.216	0.049	0.203	0.026
2	0.3	Proposed ($\alpha = 1$)	0.691	1.459	0.080	0.191	0.022
2	0.3	Proposed (adaptive)	0.644	1.398	0.072	0.203	0.004
			$\beta_1 = \mathbf{0.5}$	$\beta_2 \in \{\mathbf{0.3}, \mathbf{0.3} \pm \delta\}^8$	$\beta_3 = \mathbf{0.3}$	$\beta_4 \in \{\mathbf{0}, \pm\delta\}^8$	$\beta_5 = \mathbf{0}$
3	0.1	Homogeneous	0.154	0.726	0.041	0.642	0.045
3	0.1	Stratified	1.215	1.185	0.357	0.425	0.380
3	0.1	Two-stage	1.015	1.102	0.292	0.489	0.298
3	0.1	Proposed ($\alpha = 0$)	0.558	0.900	0.071	0.553	0.079
3	0.1	Proposed ($\alpha = 0.5$)	0.538	0.948	0.065	0.551	0.038
3	0.1	Proposed ($\alpha = 1$)	0.691	1.310	0.095	0.556	0.029
3	0.1	Proposed (adaptive)	0.682	1.286	0.084	0.565	0.032
3	0.3	Homogeneous	0.153	5.434	0.043	5.352	0.051
3	0.3	Stratified	1.215	1.185	0.357	0.425	0.380
3	0.3	Two-stage	1.215	1.185	0.357	0.425	0.380
3	0.3	Proposed ($\alpha = 0$)	0.619	1.463	0.051	0.457	0.080
3	0.3	Proposed ($\alpha = 0.5$)	0.574	1.423	0.067	0.382	0.024
3	0.3	Proposed ($\alpha = 1$)	0.686	1.453	0.089	0.295	0.014
3	0.3	Proposed (adaptive)	0.631	1.423	0.069	0.357	0.006

heterogeneity and cluster estimates into sensible groups. Web Appendix D illustrates through boxplots the performance of fusion learning procedure in selecting the right number of clusters.

Similarly, simulation experiments for the logistic model with binary outcomes, which mimics the actual real data analysis in Section 7, are performed and given in Web Appendix D. Results remain largely consistent with findings in the linear model in that we see generally smaller MSE in our method than other stratified methods; and we observe comparable MSE in our method for the truly homogeneous covariates and smaller MSE for truly heterogeneous covariates, in comparison to the homogeneous method.

TABLE 2 Mean squared error (×100) of GEE coefficient estimates in the linear model with exchangeable working correlation matrix (ie, working correlation matrix correctly specified). The true values of the coefficients across strata are specified in bold

G	δ	Method	β_1	β_2	β_3	β_4	β_5
			$\boldsymbol{\beta_1 = 0.5}$	$\boldsymbol{\beta_2 = 0.3}$	$\boldsymbol{\beta_3 = 0.3}$	$\boldsymbol{\beta_4 = 0}$	$\boldsymbol{\beta_5 = 0}$
1	0.0	Homogeneous	0.101	0.081	0.021	0.030	0.026
1	0.0	Stratified	0.787	0.741	0.208	0.257	0.225
1	0.0	Two-stage	0.202	0.173	0.047	0.068	0.055
1	0.0	Proposed ($\alpha = 0$)	0.142	0.085	0.023	0.029	0.026
1	0.0	Proposed ($\alpha = 0.5$)	0.132	0.090	0.023	0.009	0.005
1	0.0	Proposed ($\alpha = 1$)	0.173	0.346	0.028	0.002	0.001
1	0.0	Proposed (adaptive)	0.163	0.288	0.022	0.010	0.000
			$\boldsymbol{\beta_1 = 0.5}$	$\boldsymbol{\beta_2 \in \{0.3, 0.3 + \delta\}^8}$	$\boldsymbol{\beta_3 = 0.3}$	$\boldsymbol{\beta_4 \in \{0, \delta\}^8}$	$\boldsymbol{\beta_5 = 0}$
2	0.1	Homogeneous	0.102	0.300	0.021	0.249	0.027
2	0.1	Stratified	0.787	0.741	0.208	0.257	0.225
2	0.1	Two-stage	0.451	0.519	0.114	0.269	0.122
2	0.1	Proposed ($\alpha = 0$)	0.155	0.317	0.023	0.243	0.027
2	0.1	Proposed ($\alpha = 0.5$)	0.136	0.318	0.023	0.259	0.004
2	0.1	Proposed ($\alpha = 1$)	0.126	0.644	0.037	0.239	0.008
2	0.1	Proposed (adaptive)	0.205	0.500	0.026	0.323	0.003
2	0.3	Homogeneous	0.107	2.039	0.023	1.993	0.029
2	0.3	Stratified	0.787	0.741	0.208	0.257	0.225
2	0.3	Two-stage	0.787	0.741	0.208	0.257	0.225
2	0.3	Proposed ($\alpha = 0$)	0.195	0.820	0.027	0.100	0.026
2	0.3	Proposed ($\alpha = 0.5$)	0.202	0.859	0.028	0.074	0.006
2	0.3	Proposed ($\alpha = 1$)	0.240	1.004	0.027	0.064	0.002
2	0.3	Proposed (adaptive)	0.222	0.914	0.027	0.056	0.000
			$\boldsymbol{\beta_1 = 0.5}$	$\boldsymbol{\beta_2 \in \{0.3, 0.3 \pm \delta\}^8}$	$\boldsymbol{\beta_3 = 0.3}$	$\boldsymbol{\beta_4 \in \{0, \pm\delta\}^8}$	$\boldsymbol{\beta_5 = 0}$
3	0.1	Homogeneous	0.099	0.67	0.021	0.623	0.027
3	0.1	Stratified	0.787	0.741	0.208	0.257	0.225
3	0.1	Two-stage	0.716	0.75	0.191	0.302	0.204
3	0.1	Proposed ($\alpha = 0$)	0.213	0.667	0.022	0.438	0.027
3	0.1	Proposed ($\alpha = 0.5$)	0.203	0.711	0.024	0.453	0.004
3	0.1	Proposed ($\alpha = 1$)	0.277	1.021	0.031	0.414	0.003
3	0.1	Proposed (adaptive)	0.229	0.986	0.023	0.464	0.008
3	0.3	Homogeneous	0.105	5.360	0.024	5.338	0.030
3	0.3	Stratified	0.787	0.741	0.208	0.257	0.225
3	0.3	Two-stage	0.787	0.741	0.208	0.257	0.225
3	0.3	Proposed ($\alpha = 0$)	0.207	1.048	0.026	0.21	0.026
3	0.3	Proposed ($\alpha = 0.5$)	0.196	1.044	0.025	0.18	0.006
3	0.3	Proposed ($\alpha = 1$)	0.232	1.023	0.027	0.155	0.003
3	0.3	Proposed (adaptive)	0.190	1.039	0.023	0.171	0.000

7 | APPLICATION: INTERN HEALTH STUDY

As discussed in Section 2, the goal of the IHS is to identify risk factors that predict SI. Our sample consists of 2467 qualified subjects recruited across medical institutes in the United States between 2012 and 2014, who have at least two consecutive study responses. Four baseline covariates of interest are age, gender, baseline SI, and score of psychological health from Patient Health Questionnaire (PHQ) (Kroenke *et al.*, 2001), and four time-dependent risk factors are PHQ score, anxiety score from General Anxiety Disorder questionnaire (GAD) (Spitzer *et al.*, 2006), binary indicator of whether conducted medical error in the past 3

months (MEDERR), and average work hours in the past 3 months (HOUR). Over 30% of the sample have at least one nonresponse, and exhibits $K = 8$ distinct missing-data patterns as shown in Figure 1. Continuous covariates are standardized in subsequent data analysis. See Web Appendix E for a summary of variables in their original scales, stratified by missing-data patterns.

Originally proposed analysis stratifies subjects by missing-data patterns due to the concern that missingness may be nonignorable, and that the predictive effects of risk factors on suicidal ideation are dependent on the missing-data patterns. We examine this hypothesis by performing the test given in (8). It rejects the hypothesis that data are MCAR with high confidence ($p < 10^{-9}$), indicating a strong deviation from homogeneity and suggesting the need of stratification via pattern-mixture structure. Thus, the application of stratification seems to be a reasonable choice of method to analyze effects of the risk factors when MCAR is proved to be invalid missing-data mechanism in the longitudinal data analysis.

The stratified GEE is applied to fit stratum-wise models for the binary suicidal ideation outcome $E(SI_{k,ij}) = \mu_{k,ij}$ of the following form:

$$\begin{aligned}\text{logit}(\mu_{kij}) = \beta_0 + \beta_1 AGE_i + \beta_2 SEX_i + \beta_3 SI_{i0} + \beta_4 PHQ_{i0} \\ + \beta_{5k} PHQ_{kij} + \beta_{6k} GAD_{kij} + \beta_{7k} MEDERR_{kij} \\ + \beta_{8k} HOUR_{kij},\end{aligned}$$

$k = 1, \ldots, 8$, where with consultation with our collaborators, in this analysis, the baseline coefficients are set to be homogeneous, and the effects of time-dependent risk factors are allowed to be different on the missing-data patterns. We are interested in assessing if the pattern-mixture model (7) pertains to an over-stratification and the effective number of heterogeneous risk groups may be smaller than 8. In particular, we want to learn if there exist some shared effects of the risk factors in the time-varying covariates, that is, $\boldsymbol{\beta}_5, \ldots, \boldsymbol{\beta}_8$. Stratum-specific estimating equations are weighted by the inverse of their respective sample sizes. We set $\alpha = 0.5$ to induce a moderate amount of sparsity and use EBIC to select λ.

We compare fused GEE estimates with stratified GEE estimates. Figure 3A-D overlays the estimated values from the two methods for time-dependent covariates PHQ, GAD, MEDERR, and HOUR, respectively. The 95% confidence intervals from stratified GEE, plotted by vertical bars, overlaps each other and do not provide informative knowledge about coefficient clustering. On the other hand, the proposed method identifies coefficient clusters in *PHQ*, *MEDERR*, and *HOUR* that differ among others. Due to the sparsity penalty, some coefficients are estimated exactly as zero, such as in *GAD*. The values of our estimates are strikingly consistent with the coverage of zero by the 95% confidence intervals from the stratified analysis. For example, the 95% confidence intervals of PHQ effects exclude zero except for the "0011" pattern, which are in agreement with the findings from fusion learning. As fusion learning selects a model that is quite close to the homogeneous model assumed under MCAR, we further examine their prediction performance by five-fold cross-validation. Samples from each missing-data pattern are split separately to ensure each stratum is represented in both the training and testing sets. Prediction performance is evaluated by the cross-validated area under curve (AUC) from 50 replications, and the results are summarized as boxplots in Figure 3E-G for fusion learning, homogeneous, and stratified models, respectively. Results indicate that the homogeneous GEE predicts as good as the selected model, that is, with comparable AUC and uncertainty. Both models have consistently higher AUC and smaller uncertainty than the stratified GEE across patterns. The difference between the selected model and the homogeneous model may be due to the uncertainty in tuning and the relatively small sample sizes in some of the strata. Nevertheless, our method provides additional protection against model misspecification with little extra computational cost and no loss of predictability.

In summary, regression coefficients with a simplified structure is easier to interpret and more informative than those obtained by commonly used pattern-mixture models. The discovered grouping results may support tailored decisions being made when implementing training and caring programs across medical institutes. The results also warrant further validation in a future cohort with a larger sample size.

8 | CONCLUDING REMARKS

Stratification is an important and widely used method in statistical analysis to address data heterogeneity and complex interaction effects. However, it is often taken for granted with no systematic follow-up assessment, for example, to check overstratification and to identify shared information or hidden homogeneity. The latter may be utilized to improve statistical power and provide better interpretation of response-covariate relationships. The proposed poststratification fusion learning overcomes these shortcomings in stratification-based analyses. Although pattern-mixture models is one of the primary methods used to handle incomplete longitudinal data under NMAR missing-data mechanism, there is no sensitivity analysis available in the literature to assess the influence of stratification on stratum-specific analysis. The proposed poststratification fusion learning fills in this

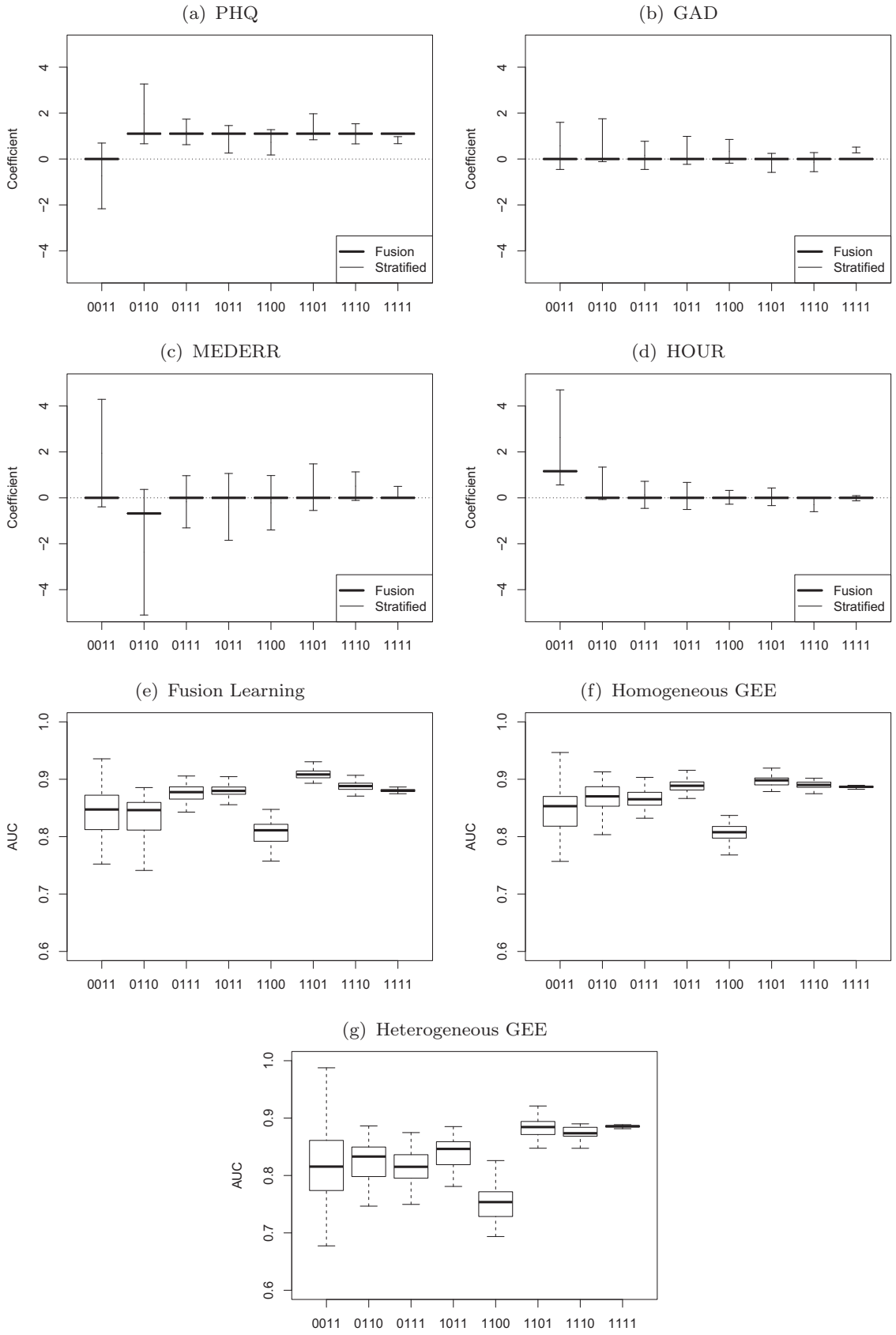

FIGURE 3 Coefficient estimates from fusion learning and stratified GEE (with 95% confidence intervals), by missing-data patterns, of time-dependent covariates PHQ (A), GAD (B), MEDERR (C), and HOUR (D). Boxplots of AUC from cross-validation repeated 50 times by missing-data patterns for fusion learning (E), homogeneous (F), and heterogeneous GEE (G).

critical methodology gap. Moreover, through simulation experiments and real data analysis, we demonstrate the usefulness of the proposed methodology in the application of pattern-mixture models.

In regard to technical advances, we generalize fusion learning to the GEE framework, allowing modeling of correlated data, in particular, longitudinal data, and to handle nonignorable missing-data issues. It can be readily applied to other correlated cases, such as spatial and spatiotemporal data. Second, we derive asymptotically the rates permitted for the number of strata and variables in terms of sample size to guarantee theoretical properties, which is useful in guiding applications. Cluster memberships produced from fusion learning may be modified to derive individualized treatment rules in longitudinal randomized trials.

ACKNOWLEDGMENTS

The authors thank co-editor, associate editor, and anonymous reviewer for constructive comments that help improve the manuscript.

DATA AVAILABILITY STATEMENT

The Intern Health Study data are not shared. Simulated data similar to the real data are provided through data generating code available in the Supporting Information of this article.

ORCID

Lu Tang https://orcid.org/0000-0001-6143-9314
Peter X.-K. Song https://orcid.org/0000-0001-7881-7182

REFERENCES

Bach, F., Jenatton, R., Mairal, J. and Obozinski, G. (2012) Structured sparsity through convex optimization. *Statistical Science*, 27, 450–468.

Bondell, H.D. and Reich, B.J. (2008) Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64, 115–123.

Bondell, H.D. and Reich, B.J. (2009) Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, 65, 169–177.

Chen, J. and Chen, Z. (2012) Extended BIC for small-n-large-p sparse GLM. *Statistica Sinica*, 22, 555–574.

Chen, H.Y. and Little, R. (1999) A test of missing completely at random for generalised estimating equations with missing data. *Biometrika*, 86, 1–13.

Dawson, J.D. (1994) Stratification of summary statistic tests according to missing data patterns. *Statistics in Medicine*, 13, 1853–1863.

Diggle, P.J. (1989) Testing for random dropouts in repeated measurement data. *Biometrics*, 45, 1255–1258.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Fu, W.J. (2003) Penalized estimating equations. *Biometrics*, 59, 126–132.

Hao, B., Sun, W.W., Liu, Y. and Cheng, G. (2018) Simultaneous clustering and estimation of heterogeneous graphical models. *Journal of Machine Learning Research*, 18, 1–58.

Hunter, D.R. and Li, R. (2005) Variable selection using MM algorithms. *The Annals of Statistics*, 33, 1617.

Johnson, B.A., Lin, D. and Zeng, D. (2008) Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*, 103, 672–680.

Jorgensen, B. (1997). *The Theory of Dispersion Models*. Boca Raton, FL: CRC Press.

Ke, Z.T., Fan, J. and Wu, Y. (2015) Homogeneity pursuit. *Journal of the American Statistical Association*, 110, 175–194.

Kroenke, K., Spitzer, R.L. and Williams, J.B. (2001) The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16, 606–613.

Liang, K.-Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.

Little, R.J. (1993) Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88, 125–134.

Ma, S. and Huang, J. (2017) A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112, 410–423.

Neuhaus, J.M., Kalbfleisch, J.D. and Hauck, W.W. (1991) A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, 59, 25–35.

Ollier, E., Samson, A., Delavenne, X. and Viallon, V. (2016) A SAEM algorithm for fused lasso penalized nonlinear mixed effect models: application to group comparison in pharmacokinetics. *Computational Statistics & Data Analysis*, 95, 207–221.

Qu, A. and Song, P.X.-K. (2002) Testing ignorable missingness in estimating equation approaches for longitudinal data. *Biometrika*, 89, 841–850.

Qu, A., Yi, G., Song, P. X.-K. and Wang, P. (2011) Assessing the validity of weighted generalized estimating equations. *Biometrika*, 98, 215–224.

Sen, S., Kranzler, H.R., Krystal, J.H., Speller, H., Chan, G., Gelernter, J. and Guille, C. (2010) A prospective cohort study investigating factors associated with depression during medical internship. *Archives of General Psychiatry*, 67, 557–565.

Shen, X. and Huang, H.-C. (2010) Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, 105, 727–739.

Song, P. X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. New York, NY: Springer.

Spitzer, R.L., Kroenke, K., Williams, J.B. and Löwe, B. (2006) A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, 166, 1092–1097.

Tang, L. and Song, P.X. (2016) Fused lasso approach in regression coefficients clustering: learning parameter heterogeneity in data integration. *The Journal of Machine Learning Research*, 17, 3915–3937.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 91–108.

Tibshirani, R.J. and Taylor, J. (2011) The solution path of the generalized lasso. *The Annals of Statistics*, 39, 1335–1371.

Van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014) On asymptotically optimal confidence regions and tests for

high-dimensional models. *The Annals of Statistics*, 42, 1166–1202.

Wang, F., Wang, L. and Song, P. X.-K. (2016) Fused lasso with the adaptation of parameter ordering in combining multiple studies with repeated measurements. *Biometrics*, 72, 1184–1193.

Wang, L., Zhou, J. and Qu, A. (2012) Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68, 353–360.

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38, 894–942.

Zhang, C.-H. and Zhang, S.S. (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 217–242.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

SUPPORTING INFORMATION

Web Appendices referenced in Sections 4, 6, 7 and R code implementing the method are available with this paper at the Biometrics website on Wiley Online Library.

How to cite this article: Tang L, Song PXK. Poststratification fusion learning in longitudinal data analysis. *Biometrics*. 2021;77:914–928. https://doi.org/10.1111/biom.13333