Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2023;15(3):161-70

ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA

DOI: 10.5336/biostatic.2023-98699

The Effect of Cluster Size for Model Performance in High-Dimensional Longitudinal Studies: A Simulation Study

Yüksek Boyutlu Boylamsal Çalışmalarda Küme Büyüklüğünün Model Performansına Etkisi: Bir Simülasyon Çalışması

Merve TÜRKEGÜN ŞENGÜL[a], Bahar TAŞDELEN[b], Saim YOLOĞLU[c]

[a]Department of Biostatistics and Medical Informatics, Alaaddin Keykubat University Faculty of Medicine, Antalya, Türkiye
[b]Department of Biostatistics and Medical Informatics, Mersin University Faculty of Medicine, Mersin, Türkiye
[c]Department of Biostatistics and Medical Informatics, İnönü University Faculty of Medicine, Malatya, Türkiye

ABSTRACT Objective: In order to prevent model estimation errors and deviations in high-dimensional longitudinal studies, risk models are established through penalized methods. The aim of this study is to examine the effect of small cluster effects on the generalized estimating equations (GEE) and penalized GEE (PGEE) model performances in high-dimensional longitudinal data. **Material and Methods:** A simulation study was designed to compare the GEE and PGEE model performances, Type I error rates, and power in two-period longitudinal data structures with different cluster sizes (n=20, 30, 50, 100, 200), different numbers of predictors (p=10, 20, 50) and different correlation levels between predictors (r=0.20, 0.50, 0.80). **Results:** It was observed that the GEE coefficient estimates were misleading and inconsistent, the Type I error rates were high, and the power of the test was weak at insufficient cluster sizes and high correlations between predictors. Even when the number of predictors and cluster size were in the balance (p=10, n=100, 200), Type I error rates were obtanied high for GEE. Increasing the cluster size was not enough to reduce the Type I error rate of GEE. The PGEE produced more successful results than GEE in all conditions. The power of PGEE increased to over 80% in all scenarios. **Conclusion:** The PGEE yielded more consistent results by controlling the relationships both within the cluster and between the predictors. In high-dimensional longitudinal studies, it was observed that the use of PGEE is more effective than GEE.

Keywords: Generalized estimating equations; penalized generalized estimating equations; model selection; penalized methods; high dimensional longitudinal data

ÖZET Amaç: Yüksek boyutlu boylamsal çalışmalardaki model tahmin hatalarının ve sapmaların önüne geçebilmek amacıyla risk modelleri, cezalı yöntemler aracılığı ile oluşturulur. Bu çalışmada amaç; yüksek boyutlu boylamsal veride küçük küme büyüklüğünün etkisinin, genelleştirilmiş tahmin eşitlikleri [generalized estimating equations (GEE)] ve cezalı genelleştirilmiş tahmin eşitlikleri [penalized generalized estimating equations (PGEE)] model performansları üzerine etkisini incelemektir. **Gereç ve Yöntemler:** Farklı küme büyüklüklerine (n=20, 30, 50, 100, 200), farklı açıklayıcı değişken sayılarına (P=10, 20, 50) ve açıklayıcı değişkenler arasında farklı korelasyon düzeylerine sahip (r=0.20, 0.50 ve 0.80) iki periyotlu boylamsal veri yapılarında GEE ve PGEE model performanslarını, Tip I hata oranlarını ve testin gücünü karşılaştırmak amacıyla simülasyon çalışması kurgulanmıştır. **Bulgular:** Yetersiz küme büyüklüklerinde ve açıklayıcı değişkenler arasındaki yüksek korelasyonlarda, GEE katsayı tahminlerinin yanıltıcı ve tutarsız olduğu, Tip I hata oranlarının yüksek ve testin gücünün ise zayıf olduğu gözlemlenmiştir. Değişken sayısı ile küme büyüklüğünün dengede olduğu durumlarda dahi (P=10, n=100, 200) GEE için Tip I hata oranları yüksek elde edilmiştir. Küme büyüklüğünü artırmak GEE'nin Tip I hata oranını düşürmek için yeterli olmamıştır. PGEE ise her koşulda GEE'den daha başarılı sonuçlar üretmiştir. PGEE'nin gücü tüm senaryolarda %80'in üzerine çıkmıştır. **Sonuç:** PGEE küme içi ve kümeler arası ilişkileri kontrol altında tutarak GEE'ye göre daha geçerli sonuçlar üretmiştir. Yüksek boyutlu boylamsal çalışmalarda GEE yerine PGEE'nin kullanımın daha etkili olduğu gözlemlenmiştir.

Anahtar kelimeler: Genelleştirilmiş tahmin eşitlikleri; cezalı genelleştirilmiş tahmin eşitlikleri, model seçimi; cezalı yöntemler; yüksek boyutlu boylamsal veri

Correspondence: Merve TÜRKEGÜN ŞENGÜL
Department of Biostatistics and Medical Informatics, Alaaddin Keykubat University Faculty of Medicine, Antalya, Türkiye
E-mail: merveturkegun@gmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 03 Jul 2023 ***Received in revised form:*** 15 Sep 2023 ***Accepted:*** 20 Sep 2023 ***Available online:*** 12 Oct 2023

High-dimensional longitudinal data (HDLD), which are gathered by measuring many variables (P) from a small group (n) of individuals over many time points, have been frequently seen in many areas of medical studies.[1] It causes inevitably "large P, small n" scenario and a high-dimensional data structure. If the cluster sizes (n) are insufficient, most of the statistical methods cannot control the Type-I error.[2] The generalized estimating equation (GEE) the most popular method to analyze longitudinal data, yields consistent and unbiased estimates of the regression coefficients if the cluster sizes are enough larger even though the working correlation matrix is misspecification among the responses.[3] But, the GEE estimate is biased under a finite sample, particularly for a small n. This may result in an incorrect inference and inflate the Type I error.[4]

The other important issue is separation or overfitting in HDLD for small cluster sizes if the outcome is binary. If a single covariate or a linear combination of covariates predicts the binary response exactly, this is considered "separation". Especially the large number of covariates causes separation in HDLD.[5]

Various variable selection methods were proposed, such as the quasi-likelihood information criterion (QIC), Mallows's C_p, and the Bayesian information criterion with quadratic inference function (QIF) for GEE in longitudinal applications.[6-8] On the other hand, the novel variable selection methods based on penalized functions, which exclude unrelated covariates by shrinking their coefficients to zero, estimate the parameters associated with selected covariates simultaneously, and minimize modeling bias, have been expanded with different penalty functions for HDLD applications. Fu introduced penalized generalized estimating equations (PGEE) with the bridge penalty and the least absolute shrinkage and selection operator (LASSO) penalty.[9] The smoothly clipped absolute deviation approach (SCAD) was expanded by Fan and Li for partially linear models using longitudinal data.[10] The SCAD penalty was generalized by Dziak and Dziak and Li for longitudinal generalized linear models; and the SCAD-penalized GEE was developed by Wang et al. for analyzing longitudinal data with high-dimensional covariates for variable selection and estimation.[11-13]

Researchers frequently have some challenges to obtain a complete and balanced data to test their hypothesis due to small target groups, rare disease studies, difficult reach, cost of data collection, or dropouts.[14] It is not always possible to encounter ideal data structures particularly in longitudinal studies. But some researchers want to avoid from missing data particularly if they study for rare diseases or cancer research. For example, death can occur in a short time in sickle cell anemia patients who have a painful vaso-occlusive crisis. Therefore, studies for these patients can generally be limited to two periods with the prescience of the physician because patients may die at the third or fourth period. So, they can work with a limited data structure instead of ignoring missing values and causing bias.

The aim of this simulation study is to investigate the effects of large P, small n, and different correlation levels on the Type I error, power, and model performances of the GEE and the PGEE with the SCAD penalty function.

MATERIAL AND METHODS

GENERALIZED ESTIMATING EQUATIONS

Suppose a sample of n subjects chosen at random. $Y_i=(Y_{i1}, \ldots, Y_{i_{mi}})^T$ is the m_i x p matrix of correlated responses for subject i at time t with t=1, 2, …, m_i where i=1, ..., n. $X_i=(Xi_1, \ldots, X_{i_{mi}})^T$ is a m_i x p vector of covariates measured at the same time as the responses for subject ith. Assume that observations from the same subject are correlated, while observations from different subjects are independent. We suppose that Y is generated from a distribution in the exponential family. Marginal mean of Y_{ij} is $E(Y_{it}|X_{it}) = \mu_{it} = (\mu_{i1}, \ldots, \mu_{in}) = g(X^T{}_{it}\beta)$ where $g(.)$ is the inverse of the known link-function, $\beta=(\beta_1, \ldots, \beta_p)^T$ is an unknown px1 vector of regression parameters. V_i denotes the variance of Y_{ij} $V_i = Var(Y_i|X_{it}) = \nu(\mu_{it})\phi$ with a variance function $\nu(\cdot)$ and an overdispersion parameter ϕ. According to Liang and Zeger, GEE uses a common working correlation matrix for correlated responses of each subject in longitudinal studies, and they recommend estimating V_i via a working correlation matrix.[15]

As a working covariance matrix $V_i = \phi A_i^{1/2} R_T A_i^{1/2}$, where $A_i = diag\{v(\mu_{i1}), \ldots, v(\mu_{i_{mi}})\}$ is a diagonal matrix for known variance function and, R_T is the true and unknown correlation matrix of Y_i. For this reason, Liang and Zeger proposed to use a working correlation matrix *Ri(α)*, which is completely specified by the parameter α vector, instead of R_T. The most commonly used working correlation structures are independence, autocorrelation, unstructured, exchangeable, symmetry, and fix. Even though *Ri(α)* is a mis-specified, parameter estimates are consistent and asymptotically normal. Finally, regression coefficient β is estimated by solving the following estimating equation as (1).[15]

$$U_\beta(\beta) = \sum_{i=1}^{n} A_i^{'} V_i^{-1} E\{Y_i - \mu_i(\beta)\} = 0 \quad (1)$$

PENALIZED GENERALIZED ESTIMATING EQUATIONS

Wang et al. proposed to estimate β, which solves the following set of penalized estimating equation as (2);[13]

$$U_n(\beta_n) = S_n(\beta_n) - q_{\lambda n}(|\beta_n|)\text{sign}(\beta_n) \quad (2)$$

where, $S_n(\beta_n) = n^{-1} \sum_{i=}^{n} X_i^T A_i^{1/2}(\beta_n) \widehat{R(\alpha)^{-1}} A_i^{-1/2}(\beta_n)(Y_i - \mu_i(\beta_n)) = 0$ are the estimating functions defining the GEE, $q_{\lambda_n}(|\beta_n|) = \left(q_{\lambda_n}(|\beta_{n1}|), \ldots, q_{\lambda_n}(|\beta_{n_{pn}}|)\right)^T$ is a p-n dimensional vector of penalty functions, and $\text{sign}(\beta_n) = \left(\text{sign}(\beta_{n1}), \ldots, \text{sign}(\beta_{n_{pn}})\right)^T$ with sign(t)=I(t>0)- I(t<0). The shrinking quantity is determined by the tuning parameter λ_n. The $q_\lambda(|\beta|)\text{sign}(\beta)$ stands for the component-wise product.

The penalty function $q_\lambda(|\beta_j|)$ is equal to zero if value of $|\beta_j|$ is large on the other hand, the $q_\lambda(|\beta_j|)$takes a large value if value of $|\beta_j|$ is a small value. So, high valued β_{nj}, the generalized estimating function $S_{nj}(\beta_n)$ which is the *j*th component of $S_n(\beta_n)$, is not penalized; as the penalty level is high, β_{nj} is approximately zero (but not equal). As a result, the penalty function $q_\lambda(|\beta_j|)$ aims to shrink estimates of small coefficients to zero. When an estimated coefficient is shrunken to zero, it is removed from the model that is ultimately chosen. Briefly, penalized estimation equations shrink small coefficients to zero, so they can perform variable selection while producing robust estimators of nonzero coefficients.[13]

According to Fan and Li, the three important characteristics of a successful penalty function estimator simultaneously are unbiasedness, sparsity, and continuity.[10] SCAD penalty has all these features. The nonconvex SCAD penalty allows selection of consistent variables avoiding over penalizing large coefficients and shrinks the coefficients of redundant covariates to exactly zero.[11] Because of these features, in this study, we considered the nonconvex SCAD penalty, which was given by (3),

$$q_{\lambda_n}(\theta) = \lambda_n \left\{ I(\theta \leq \lambda_n) + \frac{(\alpha\lambda_n - \theta)_+}{(\alpha-1)\lambda_n} I(\theta > \lambda_n) \right\} \quad (3)$$

Where $\theta \geq 0$ and some $\alpha > 2$ and I(.) is an indicator function. if c is true, I(c)=1 otherwise is zero $b_+ = bI(b>0)$ for real number *b*. Fan and Li suggested taking α=3.7.[11]

TUNING PARAMETER SELECTION

In penalized regression methods, both the robustness and consistency of the obtained estimators and the determination of important variables depend on the selection of the tuning parameter in the penalty function. According to tuning parameter, over and under penalization problems are possible and cause deviations in the estimated penalty functions.[12] K-fold cross-validation (CV) has been suggested by Cantoni et al. to select the tuning parameter.[7] Since there is no likelihood function in the PGEE framework, it has been extended by including the working correlation structure in the CV.[16]

SIMULATION STUDY

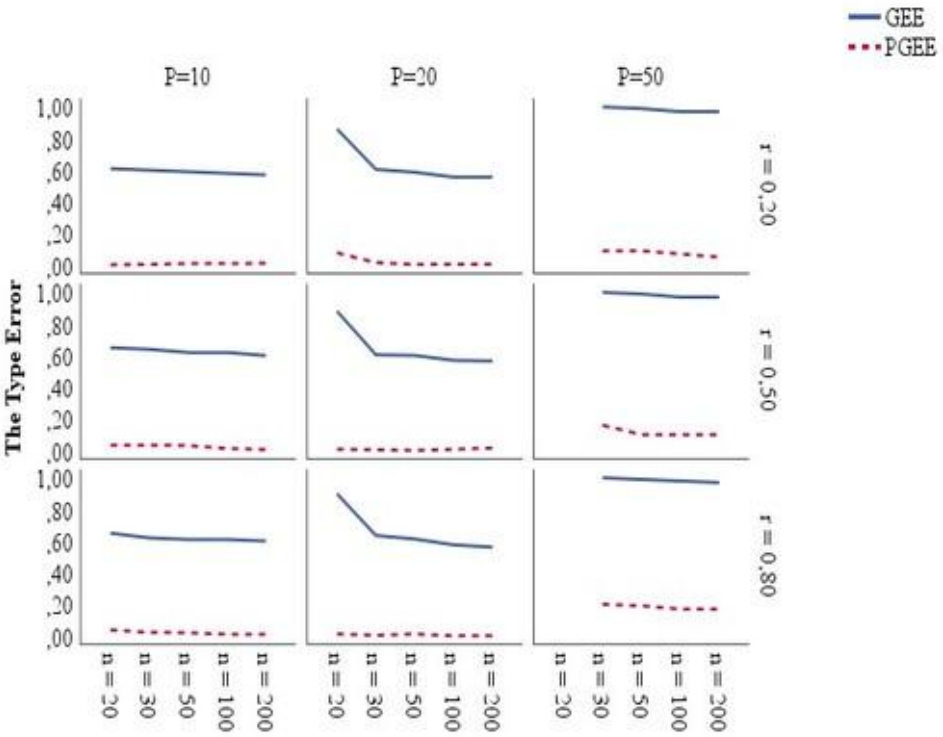
To evaluate the effects of cluster sizes, number of periods and variables, and correlation between variables on Type I error rates, power of the test, and model performances for GEE and PGEE. R (Version 4.2.2) programming language was used. We generated correlated binary response data from binom distribution by *rbin(…)* function in the *"SimCorMultR"* package with cluster sizes were n=20, 30, 50, 100, and 200. The explanatory variables were derived from the multivariate normal distribution with 0 mean, 1 standard deviation and correlations ρ=0.20 (weak), 0.50 (medium), and 0.80 (high). The number of explanatory variables in each cluster was taken as P=10, 20, and 50. This package simulates correlated binary responses assuming a regression model for the marginal probabilities. For this, the working correlation structure was selected only as "*exchangeable*" and ρ=0.40 since we studied only two periods. Also, for simulated binary response we determined initial beta coefficients as β_i=($\beta_1, \dots, \beta_{(10,20,and,50)}$) for $\beta_0 = 0$, $\beta_1 = -0.50$, $\beta_2 = 0.01$, and $\beta_3 = 0.50$. These coefficients were determined according to the results of our doctorate thesis data in order to take into account the potential difficulties encountered in real application data. 45 different scenarios with 500 replications were created for each cluster sizes. *"geepack"* and *"PGEE"* packages were used for GEE and PGEE analysis. We selected only the non-convex SCAD penalty for PGEE. Type I error (α) rates and the power of the test were recorded. In order to evaluate the model performances of the GEE and PGEE, the median squared error (MEDSE) and BIAS statistics were calculated using the *"mlr3measures"* package. MEDSE values of regression coefficients are calculated for *s.* simulation as MEDSE $= \text{median}_s[(\hat{\beta}_s - \beta_s)^2]$. MEDSE is more resistant to outliers and is used like the mean squared error (MSE). MEDSE values must be close to zero for model accuracy.[17] Also, Bias is defined for *s.* simulation as $[\hat{\beta}_s - \beta_s]$. We calculated median of bias values due to outliers.

RESULTS

The Type I error rates are shown in Figure 1 for GEE and PGEE according to number of variables (P=10, 20, and 50), cluster sizes (n=20, 30, 50, 100, and 200), and correlation levels (r=0.20, 0.50, and 0.80). The results of the power for GEE and PGEE are given in Figure 2. Generally, it was observed that Type I error rates for the PGEE were lower than for the GEE in all scenarios. While the Type I error rates for the GEE were negatively affected by the high correlation, the Type I error rates for the PGEE were quite low in all scenarios under the same conditions.

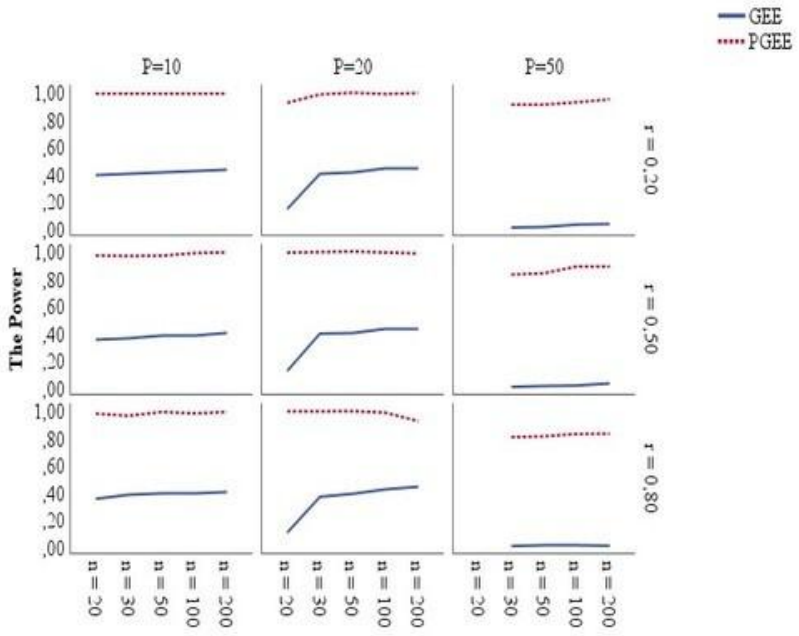
In weak, medium, and high correlations, although it was observed that Type I errors decreased when the cluster sizes were balance according to the number of variables, this was not effective for the GEE results. The Type I errors for the GEE could not reach the statistical significance level (Figure 1). For P=10 and 20, they were calculated as 0.50 or above, and the maximum power was 44%. It was observed that PGEE protected the statistical significance level and power was obtained at 90% or above even though the cluster sizes were equal to the number of variables and insufficient (n=20, P=20). The same results were observed in other scenarios for the PGEE. It was seen that the high correlation did not affect the PGEE (Figure 2).

Type I errors of GEE could not be calculated for small cluster sizes (n=20) and large variables (P=50) due to a convergence problem. Although increasing the cluster sizes solved the convergence problem in GEE, it was observed that the Type I error rates were close to 1 at all correlations. For this situation, the power of the GEE was calculated quite low. On the other hand, in small cluster sizes (n=20, 30, and 50) and high dimensionality (P=50), all Type I errors of PGEE were lower than the GEE, and the power of PGEE was 90% and above.



PGEE: Penalized generalized estimating equation.

FIGURE 1: Type I error rates for PGEE and GEE.



PGEE: Penalized generalized estimating equation.

FIGURE 2: The power for PGEE and GEE.

PGEE has derived more successful results than GEE for all scenarios. When both small cluster sizes and increasing correlation for P=50, Type I errors for PGEE were calculated above 0.05, and it has been observed that they increase to 10-20% levels. In this case, although the power decreases, it has been observed that it maintains 80% or above power. While PGEE has not been affected by small cluster sizes, the GEE has been negatively affected by both insufficient cluster sizes and high correlation.

The comparisons of model performances for the GEE and PGEE were presented in Table 1, Table 2, and Table 3. The model performances of PGEE and GEE produced successful results for P=10. However, the GEE was more affected by the increase in correlation in small clusters (n=20, 30) than in large clusters (n=50, 100, 200). As the cluster sizes increased, MEDSE and BIAS values approached 0, and GEE's model performance improved. Otherwise, the PGEE was not affected by the low cluster number and high correlation. It was observed that PGEE gave more consistent results than the GEE in all scenarios (Table 1).

TABLE 1: Model performances measurements of PGEE and GEE for P=10.

P=10		r=0.20		r=0.50		r=0.80	
		MEDSE	BIAS	MEDSE	BIAS	MEDSE	BIAS
n=20	PGEE	<0.001	0.001	<0.001	0.001	<0.001	0.001
		(<0.001)	(0.000)	(<0.001)	(0.000)	(<0.001)	(0.000)
	GEE	0.168	0.006	0.229	-0.003	0.496	-0.002
		(0.346)	(0.157)	(0.474)	(0.100)	(0.858)	(0.079)
n=30	PGEE	<0.001	0.001	<0.001	0.001	<0.001	0.001
		(<0.001)	(0.000)	(<0.001)	(0.000)	(<0.001)	(0.000)
	GEE	0.062	0.003	0.100	-0.003	0.217	0.001
		(0.077)	(0.084)	(0.113)	(0.063)	(0.260)	(0.052)
n=50	PGEE	<0.001	0.001	<0.001	0.001	<0.001	0.001
		(<0.001)	(0.000)	(<0.001)	(0.000)	(<0.001)	(0.000)
	GEE	0.030	-0.001	0.043	0.000	0.098	-0.001
		(0.033)	(0.055)	(0.043)	(0.041)	(0.116)	(0.034)
n=100	PGEE	0.000	0.001	<0.001	0.001	<0.001	0.001
		(0.013)	(0.000)	(<0.001)	(0.000)	(<0.001)	(0.000)
	GEE	0.013	0.001	0.019	0.002	0.047	-0.002
		(0.012)	(0.040)	(0.020)	(0.028)	(0.051)	(0.020)
n=200	PGEE	0.000	0.001	<0.001	0.001	<0.001	0.001
		(0.007)	(0.003)	(<0.001)	(0.000)	(<0.001)	(0.000)
	GEE	0.006	0.002	0.010	0.000	0.023	0.000
		(0.006)	(0.027)	(0.009)	(0.019)	(0.022)	(0.014)

All MEDSE and BIAS values were summarized by median (inter quantile range); PGEE: Penalized generalized estimating equation; MEDSE: Median squared error.

The GEE produced very large MEDSE values when cluster sizes and number of variables were unbalanced (P=20; n=20, 30). GEE was insufficient in these clusters at all correlation levels. On the other hand, when cluster sizes increased, the MEDSE values decreased below 0.05. MEDSE values were estimated near

zero for medium and high correlation with increasing cluster sizes, even though GEE was impacted by correlation in small cluster sizes. The GEE model was weak compared to the PGEE model in all correlations where the number of variables was equal to or close to the number of clusters (Table 2).

TABLE 2: Model performances measurements of PGEE and GEE for P=20.

P=20		r=0.20		r=0.50		r=0.80	
		MEDSE	BIAS	MEDSE	BIAS	MEDSE	BIAS
n=20	PGEE	0.000	0.001	0.000	0.001	0.000	0.001
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	GEE	2908.340	0.148	5568.850	0.022	14783.790	-0.047
		(50147.700)	(10.088)	(5.66E+28)	(7.624)	(2.53E+29)	(7.858)
n=30	PGEE	0.000	0.001	0.000	0.001	0.000	0.001
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	GEE	0.244	-0.005	0.499	-0.003	1.105	0.000
		(3.39E+28)	(0.135)	(7.50E+28)	(0.117)	(1.84E+29)	(0.076)
n=50	PGEE	0.000	0.001	0.000	0.001	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	GEE	0.041	0.002	0.043	-0.001	0.140	0.000
		(0.040)	(0.037)	(0.043)	(0.024)	(0.123)	(0.019)
n=100	PGEE	0.000	0.000	0.000	0.000	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	GEE	0.015	0.000	0.022	0.001	0.050	0.000
		(0.010)	(0.024)	(0.017)	(0.016)	(0.038)	(0.012)
n=200	PGEE	0.000	0.001	0.000	0.000	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	GEE	0.006	0.001	0.010	0.000	0.023	0.000
		(0.004)	(0.015)	(0.007)	(0.009)	(0.015)	(0.008)

All MEDSE and BIAS values were summarized by median (inter quantile range); PGEE: Penalized generalized estimating equation; MEDSE: Median squared error.

Low cluster sizes and high correlation had no impact on the PGEE model when P=50. The problem of convergence arose in GEE due to an increase in the number of variables at small cluster sizes (n=20), and GEE did not work. In all cluster sizes up to n=200, the MEDSE and BIAS values of GEE were far from 0. We saw that the PGEE model gave more consistent results than the GEE (Table 3). As a result, it was observed that the PGEE model was not affected by low cluster sizes and high correlation when P=10, 20, and 50. Also, we know that that an insufficient cluster size, an unbalanced distribution of occurrence and non-occurrence for the outcome variable, a large number of explanatory variables, and high within-cluster correlations cause overfitting. The robust sandwich variance estimator of GEE generally provides poor estimators due to small cluster size and over-fitting.[5] We have observed that the model performances results of GEE were inconsistent and unreliable because of overfitting for n=20, 30, and 50, while P=20, and 50. Also, the same BIAS values were calculated negatively because the predicted values were so small due to overfitting (Table 2 and Table 3).These models were not useful.

TABLE 3: Model performances measurements of PGEE and GEE for P=50.

P=50		*r=0.20*		*r=0.50*		*r=0.80*	
		MEDSE	**BIAS**	**MEDSE**	**BIAS**	**MEDSE**	**BIAS**
n=20	PGEE	0.000	-0.097	0.000	-0.128	0.000	-0.113
		(0.000)	(0.339)	(0.000)	(0.185)	(0.000)	(0.208)
	GEE	--	--	--	---	--	--
		--	--	--	--	--	--
n=30	PGEE	0.000	0.128	0.000	0.041	0.000	-0.012
		(0.000)	(0.307)	(0.000)	(0.273)	(0.000)	(0.124)
	GEE	104.820	-0.813	166.640	-0.817	415.030	-0.826
		(128.670)	(0.296)	(171.500)	(0.165)	(467.390)	(0.107)
n=50	PGEE	0.000	0.128	0.000	-0.014	0.000	-0.002
		(0.000)	(0.262)	(0.000)	(0.083)	(0.000)	(0.063)
	GEE	30.160	-1.407	46.280	-1.411	118.520	-1.413
		(16.270)	(0.482)	(25.540)	(0.288)	(73.580)	(0.265)
n=100	PGEE	0.000	-0.030	0.000	0.001	0.000	-0.002
		(0.000)	(0.214)	(0.000)	(0.034)	(0.000)	(0.037)
	GEE	17.740	-3.888	27.350	-3.958	69.790	-3.991
		(7.740)	(1.723)	(12.820)	(1.520)	(31.910)	(1.635)
n=200	PGEE	0.000	-0.012	0.000	0.001	0.000	0.000
		(0.000)	(0.030)	(0.000)	(0.027)	(0.000)	(0.024)
	GEE	0.290	-0.218	0.450	-0.210	0.860	-0.173
		(796.920)	(16.648)	(1684.910)	(17.088)	(915.020)	(8.842)

All MEDSE and BIAS values were summarized by median (inter quantile range); PGEE: Penalized generalized estimating equation; MEDSE: Median squared error.

DISCUSSION

In this study, we evaluated the GEE and PGEE for P>n in low periods and cluster sizes. The sandwich estimators of beta coefficients in the GEE approach are asymptotically consistent and unbiased. However, preserving these asymptotic properties depends on the cluster sizes.[18] It is recommended that the number of clusters should be at least 30, especially if the response variable is binary, and at least 50 for the continuous or discrete response variable. Otherwise, the asymptotic unbiasedness of the estimators lost and the Type I error rates of the coefficients increases for GEE.[19] In cases where the cluster sizes cannot be increased, the sandwich estimators such as Morel et al., and Firth were recommended.[18,20-26] According to our results, although the number of variables and cluster sizes were not only balanced (P=10, n=100, 200), but also close to each other (P=10, 20, 50; n=20, 30, 50), the Type I error rates were much higher, and the power of the test was much lower for the GEE results than for the PGEE results. Also, the MEDSE and BIAS values of the GEE coefficients were calculated to be very large for the GEE. In these scenarios, while increasing the cluster sizes improved the model performance of GEE, the MEDSE and BIAS values for PGEE were calculated close to zero or smaller in all scenarios. Although there has been a rich literature on variable selection or small cluster sizes for longitudinal data, our simulation study differs from these in that the cluster sizes, and the number of periods is very small.

The number of periods is a factor that affects the power of the GEE. Ma et al. showed that the increase in the number of periods has eliminated the negative effects of the low cluster sizes.[16] In our study, since we consider the challenges of study design for rare diseases, the number of periods was kept constant at two (k=2). On the contrary GEE, PGEE has protected Type I error, power, and model performance successfully even if low number of periods. In order to see the effect of the number of periods on the power in the case of P>n, this study is planned with different period numbers in the future.

The overfitting is another negative effect of the imbalance between the variable and the cluster sizes on the binary response variables. Although "bias correction" or "bias reduction" methods were used to control the increase in bias due to the low cluster sizes, they were insufficient to solve the overfitting problem for P>n. Mondol and Rahman's only examined the effect of overfitting of small cluster sizes and GEE with Firth's type penalty was proposed to reduce the effect of low cluster sizes in this study.[3] Although GEE is frequently insufficient, Mondol and Rahman found that PGEE achieves convergence even in the presence of complete or quasi-complete separation.[3] Additionally, they showed how PGEE is superior to GEE in terms of the bias of the regression coefficients under near-to-complete separation. Similarly, Gosho et al. showed that in small samples, PGEE will be a better choice than GEE and bias-corrected GEE for assessing sparse binary data.[4] But the impact of P>n in small sample was not investigated in these studies. We did not set a simulation for separation and sparse, but we observed that complete separation occurred at P>n in GEE results and GEE didn't work particularly at *P=50*, and *n=20*. Similar to these studies we observed that the PGEE was superior than the GEE in the presence of complete separation. No study has been found in the literature that compares the GEE and PGEE model performances in the case of P>n, where there is overfitting, highly correlated covariates, small *cluster* sizes, and low periods. For this reason, this study is the first in this field.

CONCLUSION

Increasing the cluster size was not sufficient to decrease the Type I error rate and increase the power of GEE. High correlations between variables negatively affected the reliability of GEE coefficients and created difficulties in their interpretation in small cluster sizes. The PGEE generated more reliable and reasonable results than GEE, despite the low number of clusters and high correlation between the variables.

The model performances of the GEE were too much low, also. We observed major deviations in the GEE estimations. The MEDSE and BIAS values were calculated close to 0 or smaller in all scenarios for PGEE. The PGEE yielded more consistent results by controlling the relationships both within the cluster and between the variables despite lower cluster sizes, periods, and complete separation.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Bahar Taşdelen, Saim Yoloğlu, Merve Türkegün Şengül;* ***Design:*** *Bahar Taşdelen, Saim Yoloğlu, Merve Türkegün Şengül;* ***Control/Supervision:*** *Bahar Taşdelen, Saim Yoloğlu, Merve Türkegün Şengül;* ***Data Collection and/or Processing:*** *Merve Türkegün Şengül, Bahar Taşdelen;* ***Analysis and/or Interpretation:*** *Merve Türkegün Şengül, Bahar Taşdelen, Saim Yoloğlu;* ***Literature Review:*** *Merve Türkegün Şengül;* ***Writing the Article:*** *Merve Türkegün Şengül, Bahar Taşdelen;* ***Critical Review:*** *Bahar Taşdelen, Saim Yoloğlu, Merve Türkegün Şengül.*

REFERENCES

1. Zhong P-S, Li J, Kokoszka P. Multivariate analysis of variance and change points estimation for high-dimensional longitudinal data. Scand J Statist. 2021;48(2):375-405. [Crossref]
2. Konietschke F, Schwab K, Pauly M. Small sample sizes: a big data problem in high-dimensional data analysis. Stat Methods Med Res. 2021;30(3):687-701. [Crossref] [PubMed] [PMC]
3. Mondol MH, Rahman MS. Bias-reduced and separation-proof GEE with small or sparse longitudinal binary data. Stat Med. 2019;38(14):2544-60. [Crossref] [PubMed]
4. Gosho M, Ishii R, Noma H, Maruo K. A comparison of bias-adjusted generalized estimating equations for sparse binary data in small-sample longitudinal studies. Stat Med. 2023;42(15):2711-27. [Crossref] [PubMed]
5. Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. Stat Med. 2006;25(24):4216-26. [Crossref] [PubMed]
6. Pan W. Akaike's information criterion in generalized estimating equations. Biometrics. 2001;57(1):120-5. [Crossref] [PubMed]
7. Cantoni E, Flemming JM, Ronchetti E. Variable selection for marginal longitudinal generalized linear models. Biometrics. 2005;61(2):507-14. [Crossref] [PubMed]
8. Wang L, Qu A. Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2009;71(1):177-90. [Crossref]
9. Fu WJ. Penalized estimating equations. Biometrics. 2003;59(1):126-32. [Crossref] [PubMed]
10. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001;96(456):1348-60. [Crossref]
11. Dziak JJ. Penalized quadratic inference functions for variable selection in longitudinal research [Doctorate thesis]. Pennsylvania: The Pennsylvania State University; 2006. [Cited: June 23, 2023]. Available from: [Link]
12. Dziak JJ, Li R. An overview on variable selection for longitudinal data. In: Hong D, Shyr Y, eds. Quantitative Medical Data Analysis Using Mathematical Tools and Statistical Techniques. Default Book Series. 1st ed. Singapore: World Scientific; 2007. p.3-24. [Crossref]
13. Wang L, Zhou J, Qu A. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. Biometrics. 2012;68(2):353-60. [Crossref] [PubMed]
14. Ren VS, Miočević, M. Introduction. Small Sample size Solutions: A Guide for Applied Researchers and Practitioners. 1st ed. New York: Routledge; 2020. p.viii.
15. Liang KY, Zeger, SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13-22. [Crossref]
16. Ma Y, Mazumdar M, Memtsoudis SG. Beyond repeated-measures analysis of variance: advanced statistical methods for the analysis of longitudinal data in anesthesia research. Reg Anesth Pain Med. 2012;37(1):99-105. [Crossref] [PubMed] [PMC]
17. Rousseeuw PJ. Least median of squares regression. Journal of the American Statistical Association. 1984;79(388):871-80. [Crossref]
18. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics. 2001;57(1):126-34. [Crossref] [PubMed]
19. McNeish DM, Harring JR. Clustered data with small sample sizes: comparing the performance of model-based and design-based approaches. Communications in Statistics-Simulation and Computation. 2017;46(2):855-69. [Crossref]
20. Westgate PM, Burchett WW. Improving power in small-sample longitudinal studies when using generalized estimating equations. Stat Med. 2016;35(21):3733-44. [Crossref] [PubMed] [PMC]
21. Kauermann G, Carroll RJ. A Note on the efficiency of sandwich covariance matrix estimation. Journal of the American Statistical Association. 2001;96(456):1387-96. [Crossref]
22. Fay MP, Graubard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. Biometrics. 2001;57(4):1198-206. [Crossref] [PubMed]
23. Morel JG, Bokossa MC, Neerchal NK. Small sample correction for the variance of GEE estimators. Biom. J. 2003;45(4):395-409. [Crossref]
24. McCaffrey DF, Bell RM. Improved hypothesis testing for coefficients in generalized estimating equations with small samples of clusters. Stat Med. 2006;25(23):4081-98. [Crossref] [PubMed]
25. Wang M, Long Q. Modified robust variance estimator for generalized estimating equations with improved small-sample performance. Stat Med. 2011;30(11):1278-91. [Crossref] [PubMed]
26. Westgate PM. A bias correction for covariance estimators to improve inference with generalized estimating equations that use an unstructured correlation matrix. Stat Med. 2013;32(16):2850-8. [Crossref] [PubMed]