Comparing PMSE and QIC for Working Correlation Selection in GEE: A Simulation Study with FEV1 Data

1 Introduction

In many biomedical and longitudinal studies, repeated observations of the same subject are *correlated*, which violates the independence assumption of standard generalized linear models (GLMs). The **Generalized Estimating Equation (GEE)** approach Liang and Zeger (1986) handles such correlated data by specifying a marginal mean model and a *working* correlation structure for the repeated measurements.

Even though GEE yields consistent estimates of the regression coefficients $\boldsymbol{\beta}$ under mild conditions, its *efficiency* (variance) and potential bias can depend on how well the chosen correlation structure approximates the true correlation (see, e.g., Pan (2001a)). This leads to a model–selection problem: *How do we choose among correlation structures such as Independence, Compound Symmetry (CS), or AR-1* One popular approach is to compare the **predictive mean squared error (PMSE)** of models fitted under each structure and select the one with the lowest estimated PMSE.

The remainder of this document first reviews the exponential–family formulation of GLMs and the corresponding score equations; we then show how these ideas extend to GEEs and discuss data–driven criteria for choosing a working correlation structure.

2 The Background of GEE

For clustered/correlated data, the difficulties in interpreting the parameter estimates from generalized linear mixed effects models with non-identity link functions suggest that other approaches are needed. One of the most popular of these is generalized estimating equations (GEE). GEE extends generalized linear models to correlated data but differs from mixed effects models in that GEE explicitly fits a marginal model to data.

3 Method: From the GLM to the GEE

3.1 The GLM

The density (or mass) function of any member of the exponential family takes the following form:

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi)\right)$$

The functions a, b, and c in the formula will vary from distribution to distribution. The parameter θ is called the canonical parameter and is a function of μ. This function of μ is referred to as the canonical link function.

Suppose we have a random sample of size n from a member of the exponential family of distributions. The likelihood is given by

$$\log L(y) = \sum_{i=1}^{n} \left(\frac{y_i\theta - b(\theta)}{a(\phi)} - c(y_i,\phi)\right)$$

Estimating equations of Equation (2) can be written as the following single matrix equation.

$$\mathbf{D} = \begin{bmatrix} \frac{\partial\mu_1}{\partial\beta_0} & \frac{\partial\mu_1}{\partial\beta_1} & \cdots & \frac{\partial\mu_1}{\partial\beta_p} \\ \frac{\partial\mu_2}{\partial\beta_0} & \frac{\partial\mu_2}{\partial\beta_1} & \cdots & \frac{\partial\mu_2}{\partial\beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\mu_n}{\partial\beta_0} & \frac{\partial\mu_n}{\partial\beta_1} & \cdots & \frac{\partial\mu_n}{\partial\beta_p} \end{bmatrix} \quad , \quad \mathbf{V} = a(\phi)\begin{bmatrix} \mathrm{Var}(\mu_1) & 0 & \cdots & 0 \\ 0 & \mathrm{Var}(\mu_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{Var}(\mu_n) \end{bmatrix}$$

$$\mathbf{y} - \boldsymbol{\mu} = \begin{bmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ \vdots \\ y_n - \mu_n \end{bmatrix} \quad , \quad \frac{d}{d\boldsymbol{\beta}} \log L(y) = \begin{bmatrix} \frac{\partial \log L}{\partial\beta_0} \\ \frac{\partial \log L}{\partial\beta_1} \\ \vdots \\ \frac{\partial \log L}{\partial\beta_p} \end{bmatrix} = \mathbf{D}^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = 0$$

3.2 The GEE

Given longitudinal data on K subjects, let Y_{ij} denote the jth response for subject i, where $i = 1, \ldots, K$, $j = 1, \ldots, n_i$. Let X_{ij} be a $p \times 1$ vector of covariates. Write

$$Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})', \quad \mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})',$$

where $\mu_{ij} = \mathbb{E}(Y_{ij} \mid X_{ij})$. Within-subject correlation is allowed, but observations across subjects are independent.

The marginal model is specified by

$$g(\mu_{ij}) = X_{ij}'\beta, \tag{1}$$

where g is a known link function and β an unknown $p \times 1$ vector. The conditional variance is

$$\mathrm{Var}(Y_{ij} \mid X_{ij}) = v(\mu_{ij})\,\phi,$$

with $v(\cdot)$ a known variance function and ϕ a scale parameter. For example, if Y_{ij} is continuous then $v(\mu) = 1$; if Y_{ij} is a count then $v(\mu) = \mu$.

Define

$$A_i = \mathrm{diag}(v(\mu_{i1}), \ldots, v(\mu_{in_i})), \quad V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2},$$

where $R_i(\alpha)$ is the matrix working correlation indexed by parameter α. Common choices for $R_i(\alpha)$ include the following. In the independence structure, $\alpha = 0$, and $\mathrm{Corr}(Y_{ij}, Y_{ik}) = 1$ if $j = k$, and 0 otherwise. In the exchangeable structure, $\alpha = \rho$, and $\mathrm{Corr}(Y_{ij}, Y_{ik}) = 1$ if $j = k$, and ρ otherwise. In the autoregressive (AR(1)) structure, $\alpha = \rho$, and $\mathrm{Corr}(Y_{ij}, Y_{i,j+m}) = \rho^m$ for $j + m \leq n_i$. In the Toeplitz structure, $\alpha = (\theta_0, \theta_1, \ldots, \theta_{n_i-1})^\top$, and $\mathrm{Corr}(Y_{ij}, Y_{i,j+m}) = 1$ if $m = 0$, and θ_m if $1 \leq m \leq n_i - j$. Finally, in the unstructured case, $\alpha = (\theta_{12}, \theta_{13}, \ldots, \theta_{n_i-1,n_i})^\top$, and $\mathrm{Corr}(Y_{ij}, Y_{ik}) = 1$ if $j = k$, and θ_{jk} otherwise.

The GEE approach estimates $\boldsymbol{\beta}$ by solving the estimating equations Liang and Zeger (1986):

$$\sum_{i=1}^{N} \mathbf{D}_i^\top \mathbf{V}_i^{-1} \left(\mathbf{Y}_i - \boldsymbol{\mu}_i \right) = 0, \tag{2}$$

where $\mathbf{D}_i = \dfrac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$ and $\mathbf{V}_i$ is the *working* covariance matrix of $\mathbf{Y}_i$.

$\mathbf{V}_i$ can be expressed in terms of a correlation matrix $R(\alpha)$:

$$\mathbf{V}_i = \mathbf{A}_i^{1/2} \, R(\alpha) \, \mathbf{A}_i^{1/2},$$

where $\mathbf{A}_i$ is a diagonal matrix whose elements are $\mathrm{Var}(Y_{it}) = V(\mu_{it})$, specified as functions of the means μ_{it}, and α is an unknown parameter. The parameter α can be estimated through moment methods or another set of estimating equations.

However, when dealing with correlated or repeated-measurement data, the GEE approach is typically used. In this case, only the first two moments of the data—the mean and the variance are specified, without fully modelling the joint distribution of the data. This methodology is therefore based on the concept of *quasi-likelihood* rather than full likelihood.

The more accurately the working correlation structure reflects the true correlation, the smaller the variance of the estimator, and thus the greater the efficiency.

4 Common Methods for Selecting a GEE Working Correlation Structure

In Generalized Estimating Equations (GEE), the *working correlation structure* refers to the assumed or approximated pattern of correlation among repeated measurements within the same subject (or observational unit) that is specified in the estimating equations. It is termed "working" because, even if this structure does not perfectly match the true underlying correlation, the regression coefficient estimates from GEE remain **consistent** in large samples. However, when the working correlation more closely resembles the true correlation, estimation becomes more **efficient** (i.e., the variance of the estimates is reduced). There are three recommended methods for choosing a correlation structure for GEE. (1)Choose a correlation structure that reflects the manner in which the data were collected. For instance with temporal data a sensible correlation structure is one that includes time dependence, such as AR(1). (2)Choose a correlation structure that minimizes Pan's QIC. QIC is a statistic that generalizes AIC to GEE but is used only for comparing models that are identical except for their different correlation structures (Pan, 2001a).(3) Choose a correlation structure for which the sandwich estimates of the variance most closely approximate the estimate of the variance.

4.1 The Traditional QIC Approach

QIC was proposed by Pan (2001) and comes in two flavors: one for selecting a correlation structure and a second for choosing models all of which were fit with the same correlation structure. Here we focus on the version used for selecting a correlation structure.

Recall that for a parametric model, AIC is defined by

$$\mathrm{AIC} = -2 \log L + 2K,$$

where L is the likelihood and K the number of parameters.

Analogously, define the quasi-likelihood under the independence model as

$$Q(\mathbf{y}; \hat{\mu}) = \int^{\hat{\mu}} \frac{y_i - \mu}{a(\phi)\,\mathrm{Var}(\mu)}\, d\mu, \tag{3}$$

where $\hat{\mu}$ denotes the fitted means under the independence structure.

Then QIC is given by

$$\mathrm{QIC} = -2\, Q\big(\widehat{\boldsymbol{\beta}}(R),\, I,\, \mathrm{Data}\big) + 2\,\mathrm{trace}\big(\widehat{\Omega}\, \widehat{V}_R\big),$$

in which the first term is the quasi-likelihood evaluated at the working-correlation fit, and the second term is a penalty based on the robust sandwich variance $\widehat{\Omega}$ and model-based variance $\widehat{V}_R$.

An alternative expression of the quasi-likelihood (over the full data range) is

$$Q(\mu; y) = \int_y^{\mu} \frac{y - t}{\phi(t)}\, dt, \tag{4}$$

which reduces to (3) when $a(\phi)\,\mathrm{Var}(\mu) = \phi(\mu)$.

4.2 Predictive Mean Squared Error (PMSE)

4.2.1 Background of PMSE

In the GEE framework we consider the conditional expectation

$$\mathbb{E}\big(Y_{it} \mid x_{it}\big) \;=\; \mathbb{E}\big(Y_{it} \mid x_{ij},\, j = 1, \ldots, N_i\big), \tag{5}$$

for subjects $i = 1, \ldots, N$ and times $t = 1, \ldots, N_i$.

If past covariate information has additional effects on the current response, and these effects are not fully captured in the model, then using non-diagonal working correlation structures (such as Compound Symmetry [CS] or AR-1) may introduce bias.

4.2.2 Practical Implications of Condition (5)

In practice, Condition (5), the assumption of the marginal model may not hold. Its indicates it may violet the condition that does not include individual-level random effects. On the one hand, when a diagonal working correlation matrix (i.e. independence) is used, the resulting estimate of $\boldsymbol{\beta}$ still enjoys the desirable common properties of the GEE framework(Zeger (1988), McDonald (1993)). Therefore, if Condition (5) is violated, one may reasonably opt for the independence working correlation structure.

Case 1: Condition (5) **holds** If the covariate set is properly specified—e.g. by incorporating historical covariate information—then Condition (5) may hold. In such cases, using a non-diagonal working correlation structure like Compound Symmetry (CS) or AR(1) is appropriate. These structures account for existing correlation in repeated measurements and improve estimation efficiency by reducing variance.

Case 2: Condition (5) **fails** If prior covariate values affect the response but are not fully captured—i.e. Condition (5) does not hold—then non-diagonal structures (CS or AR(1)) may introduce bias. In this scenario, the independence structure is preferred to preserve unbiasedness, even at the cost of efficiency. Thanks to the robust (sandwich) variance estimator in the GEE framework, consistent and unbiased estimates of $\boldsymbol{\beta}$ are still achieved.

In longitudinal or clustered data, repeated or within-cluster observations are generally correlated. Assuming independence may yield consistent estimates of regression coefficients, but often with reduced efficiency. GEE addresses this by specifying a working correlation matrix that captures within-subject correlation and combining it with the marginal mean and variance structure in the estimating equations. If the chosen working correlation approximates the true correlation well, more precise (lower-variance) estimates follow; if not, estimates remain consistent but may be less efficient—or even biased if key assumptions (e.g. as discussed in Pepe and Anderson, 1994) are violated.

4.2.3 Methodology of PMSE

We denote our observed data by

$$D = \big\{(Y_i, X_i),\ i = 1, \dots, N\big\},$$

a random sample from some unknown distribution F. To assess predictive performance, we draw two independent replicates of size N each:

$$D^{(1)} = \big\{(Y_i^{(1)}, X_i),\ i = 1, \dots, N\big\}, \quad D^{(2)} = \big\{(Y_i^{(2)}, X_i),\ i = 1, \dots, N\big\}.$$

Under a candidate working correlation R, fit the GEE to $D^{(1)}$ to obtain $\hat{\beta}\big(D^{(1)}, R\big)$. Use these parameter estimates to predict the means on $D^{(2)}$:

$$\hat{\mu}_{ij}^{(2)} = g^{-1}\big(X_{ij}^{\top}\,\hat{\beta}(D^{(1)}, R)\big), \quad j = 1, \ldots, N_i.$$

Define the predictive mean-squared error as

$$\mathrm{PMSE}(R) = \mathbb{E}_{D^{(1)},D^{(2)}}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{\big(Y_{ij}^{(2)} - \hat{\mu}_{ij}^{(2)}\big)^2}{V\big(\hat{\beta}(D^{(1)}, R)\big)}\right], \tag{6}$$

where $V(\hat{\beta})$ denotes the model-based variance function under R.

In practice, the expectation in (6) is approximated by repeated sampling (e.g. bootstrap or Jacknife), and one selects the working correlation R that minimizes the empirical $\widehat{\mathrm{PMSE}}(R)$.

Here, $D^{(1)}$ is used to estimate $\hat{\beta}$ and hence $\hat{\mu}^{(2)}$, while $D^{(2)}$ provides the observed $Y^{(2)}$ values for computing squared errors. This two-stage procedure isolates prediction error under each candidate correlation structure.

Bootstrap Approximation A practical way to estimate the expectation in (6) is via bootstrap sampling. One proceeds as follows:

1. For $b = 1, \ldots, B$, draw a bootstrap sample $D^{*(b)}$ of size N with replacement from the original data D.
2. Fit the GEE under working correlation R to $D^{*(b)}$ to obtain $\hat{\beta}^{*(b)}$ and predicted means $\hat{\mu}_{ij}^{*(b)}$.
3. Compute the within-sample squared prediction error
$$\mathrm{SE}^{*(b)}(R) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_i}\sum_{j=1}^{N_i}\big(Y_{ij} - \hat{\mu}_{ij}^{*(b)}\big)^2.$$
4. The bootstrap PMSE is then
$$\widehat{\mathrm{PMSE}}_{\mathrm{boot}}(R) = \frac{1}{B}\sum_{b=1}^{B}\mathrm{SE}^{*(b)}(R).$$

We then select the structure with the lowest PMSE in that replication

4.3 Rotnitzky–Jewell's Criterion

Another approach, due to Rotnitzky and Jewell (1990, Sec. 4), assesses the adequacy of the working correlation by comparing two matrices. Define

$$Q_0 = n^{-1}\, D^{\top}V^{-1}D, \tag{7a}$$

$$Q_1 = n^{-1}\, D^{\top}V^{-1}(Y - \hat{\mu})(Y - \hat{\mu})^{\top}V^{-1}D, \tag{7b}$$

and set

$$Q \;=\; Q_0^{-1}Q_1. \tag{8}$$

If the working correlation is correctly specified, $Q \approx I_p$. Hence define

$$C_1 = \frac{\mathrm{tr}(Q)}{p}, \qquad C_2 = \frac{\mathrm{tr}(Q^2)}{p},$$

where p is the dimension of β. Values of (C_1, C_2) close to $(1, 1)$ indicate a good fit. Finally, the Rotnitzky–Jewell measure is

$$\mathrm{RJ}(R) = \sqrt{(1 - C_1)^2 + (1 - C_2)^2}. \tag{9}$$

One selects the working correlation R that minimizes $\mathrm{RJ}(R)$, i.e. that yields (C_1, C_2) nearest to $(1, 1)$ in Euclidean distance.

4.4 Correlation Information Criterion, CIC(R)

Hin and Wang (2009) proposed the *Correlation Information Criterion*, CIC(R), as a simplified version of QIC(R) by retaining only the penalty term:

$$\mathrm{CIC}(R) \; = \; \mathrm{tr}\big(\widehat{\Omega}_I \, V_{\hat{\beta}_{\mathrm{GEE}}}\big).$$

The idea is that the quasi-likelihood term in QIC(R) does not depend on the working correlation structure (though it must be estimated with error), so removing it yields a criterion focused purely on the variance penalty. One then selects the working correlation R that minimizes CIC(R) among the candidates.

5 Simulations

We first generate simulated data under three models and initially analyze the results using GEE. We will calculate the mean-squared error (MSE), based on empirical variance, under different working correlation structures to identify the best structure. Then, the predictive mean-squared error (PMSE) method will be applied to select the optimal structure. First, we use Pan's QIC for correlation-structure selection.

5.1 Data-Generating Models

Model 1 (Lagged Model)

$$Y_{it} \; = \; \alpha \, Y_{i,t-1} \; + \; \beta \, x_{it} \; + \; \varepsilon_{it},$$

where $Y_{i,0} = 0$ (or another initial value), and x_{it}, ε_{it} are independently drawn from the standard normal distribution. Although the marginal model $\mathbb{E}(Y_{it} \mid x_{it}) = \beta \, x_{it}$ holds, Condition (5) is violated because the past value $Y_{i,t-1}$ directly affects Y_{it}.

Model 2 (Multiplicative Lagged Model)

$$Y_{it} \; = \; Y_{i,t-1} \, \beta \, x_{it} \; + \; \varepsilon_{it},$$

with $Y_{i,0} = 1$ and $\beta = 1$. Here, even though the marginal model $\mathbb{E}(Y_{it} \mid x_{it}) = \beta \, x_{it}$ still holds, Condition (5) is not satisfied due to the dependence on previous Y values.

Model 3 (Random Effects Model)

$$Y_{it} = b_i + \beta x_{it} + \varepsilon_{it},$$

where b_i is drawn from a normal distribution. In this setting, the marginal model is correctly specified, and the true within-subject correlation structure is compound symmetry (CS).

In the next subsections, we present the results of fitting each GEE model under various working correlation structures, compare their empirical MSEs, and then assess their predictive performance via PMSE.

5.2 The Response Variables and Bias

For Models 1 and 2, the marginal model is $\mathbb{E}(Y \mid x) = \beta x$, but the data-generating process depends on past values of Y. Therefore, using non-independence structures (such as CS or AR(1)) may result in biased estimates, whereas the independence structure yields more stable results. For Model 3, the marginal model is also $\mathbb{E}(Y \mid x) = \beta x$, and the true correlation structure is CS. Therefore, CS is expected to be most efficient. To compare methods, we perform multiple replications (e.g. 1,000 simulations) and calculate bias, variance, and mean squared error (MSE) of the β estimates under each working correlation structure, as well as the frequency with which each structure is selected.

Table 1: Simulation Results for Models 1–3

Model	Correlation	Mean MSE	MSE(VAR)	Freq. Selected
	Independence	0.00242	7.61×10^{-6}	86
Model 1	Exchangeable	0.01219	1.10×10^{-4}	14
	AR(1)	0.03393	2.68×10^{-4}	0
	Independence	0.116	0.0435	71
Model 2	Exchangeable	0.193	0.0312	18
	AR(1)	0.199	0.0252	11
	Independence	0.00364	1.66×10^{-5}	22
Model 3	Exchangeable	0.00227	6.94×10^{-6}	40
	AR(1)	0.00260	8.86×10^{-6}	38

5.3 Selecting by QIC

For each simulated dataset, we computed Pan's QIC under the three candidate working correlation structures and recorded which structure minimized QIC. Table 2 reports, for each data-generating model, the frequency (out of 100 replicates) that each structure was selected.

Table 2: Frequency of working correlation structure selected by QIC

Model	Independence	Exchangeable	AR(1)
Model 1	99	0	1
Model 2	100	0	0
Model 3	100	0	0

Note: For Model 3, QIC always picked the independence structure, which contradicts the MSE-based results (where CS was most efficient). This highlights a potential drawback of relying solely on QIC for working-correlation selection.

5.4 Selecting by PMSE

We also assessed structure-selection performance using the bootstrap (BOOT) method over 100 replications. Table 3 gives the frequency with which each working correlation structure was chosen for Models 1–3.

Table 3: Bootstrap selection of working correlation structures (100 replications)

Model	Independence	Exchangeable	AR(1)
Model 1	100	0	0
Model 2	81	8	11
Model 3	42	43	15

These results align with the theoretical predictions (Pan's PMSE guide) and are consistent with the earlier findings of MSE and MSE (VAR). The selecting results showed even after it is detected that condition (5) is violated, PMSE can still to find a working correlation structure that works best.

6 Comparison of PMSE and QIC: Simulation on Generated FEV1 Data

FEV1 Data Generation We simulated a synthetic dataset for $N_{\text{all}} = 1000$ subjects, each measured annually over 5 years. For each subject, baseline covariates including age, gender, BMI, weight, pack-years, cigarettes per day, and systolic blood pressure (SBP) were generated. The smoking status was assigned independently at each time point with a 40% chance of being a smoker. This setup allowed for various realistic smoking trajectories across the 5 years, including transitions from smoking to non-smoking and vice versa.

To reflect real-world findings, we incorporated the effect of smoking cessation by adding a positive effect of 0.09–0.19 liters to FEV1 after quitting smoking. This effect was drawn from a uniform distribution

Table 4: Sample Simulated Longitudinal Data (Subjects 193–194)

id	time	smoking	treatment	age	gender	BMI	weight	pack_years	cigs_per_day
193	1	0	1	50.95	1	21.15	63.83	22	0
193	2	1	1	51.95	1	20.41	65.25	23	18
193	3	1	1	52.95	1	21.22	65.16	24	18
193	4	0	1	53.95	1	21.17	64.44	24	0
193	5	1	1	54.95	1	21.41	64.33	25	18
194	1	1	1	41.05	1	30.25	62.54	23	19
194	2	0	1	42.05	1	30.40	63.08	23	0
194	3	0	1	43.05	1	30.06	63.84	23	0
194	4	1	1	44.05	1	30.48	61.10	23	19
194	5	0	1	45.05	1	30.05	62.04	23	0

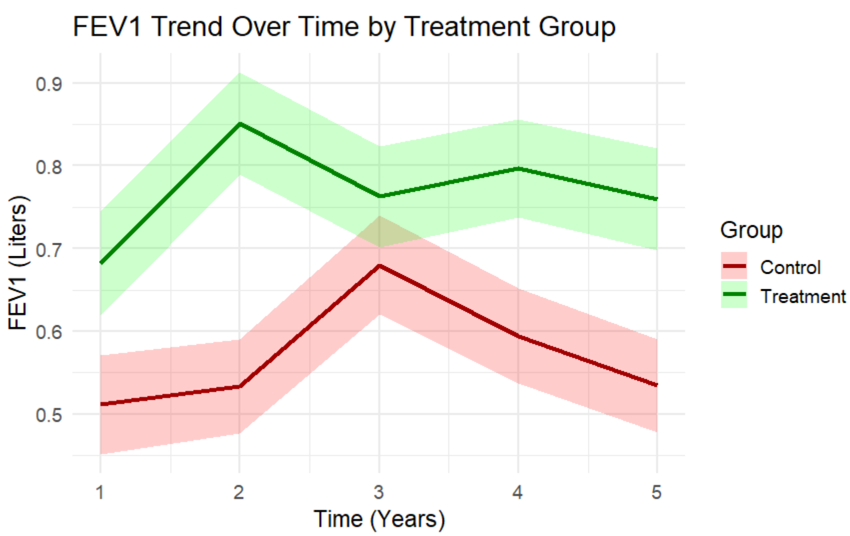

Figure 1: Mean FEV1 Trend Over Time by Treatment Group. Shaded areas indicate 95% confidence intervals.

$U(0.09, 0.19)$ and applied only during years following a transition from smoking to non-smoking.

6.1 GEE Modeling and Working Correlation Structures

To account for within-subject correlation across repeated measures, we fit marginal models using Generalized Estimating Equations (GEE) under three working correlation structures: independence, exchangeable, and autoregressive of order one (AR-1). The primary predictors of interest were current smoking status and years since quitting.

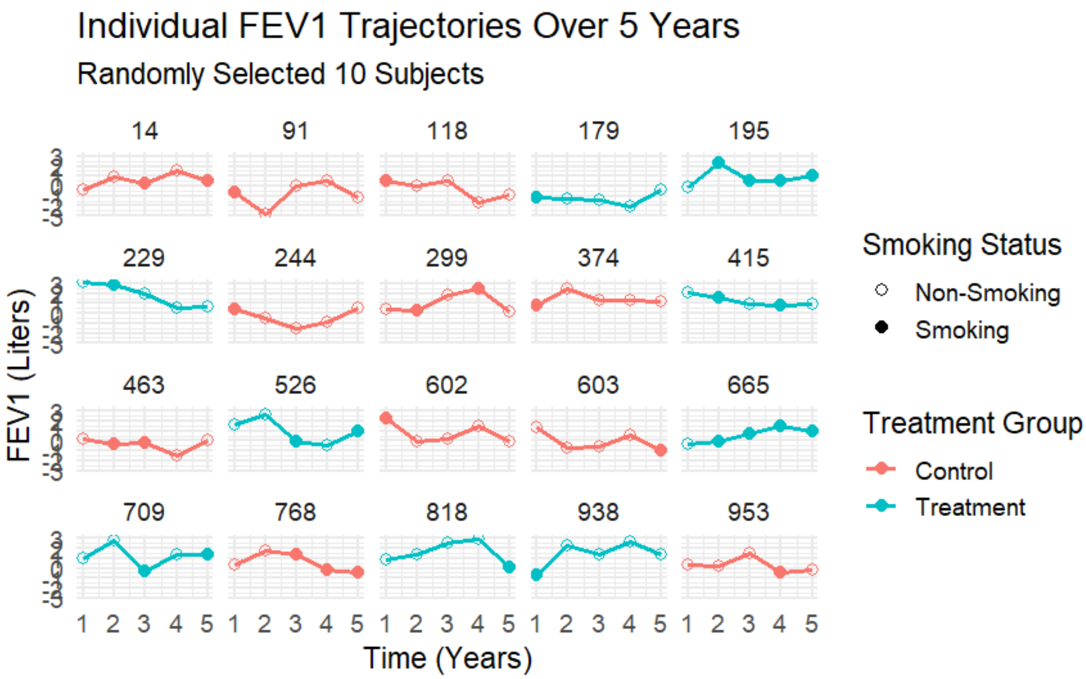

Figure 2: Individual FEV1 Trajectories Over 5 Years for 10 Randomly Selected Subjects. Marker shape indicates smoking status; color indicates treatment group.

Table 5: GEE Estimates Under Different Working Correlation Structures

Structure	Term	Estimate	Std. Error	Statistic	p-value
Independence	Smoking	-0.1827	0.0796	2.2969	0.1296
	Treatment	0.2051	0.0588	12.1792	**0.0005**
	Years Quit	0.0229	0.0174	1.7292	0.1885
Exchangeable	Smoking	-0.1209	0.0987	1.4998	0.2207
	Years Quit	**0.0397**	0.0148	7.2108	**0.0072**
AR-1	Smoking	**-0.1804**	0.1044	2.9853	**0.0084**
	Years Quit	**0.0419**	0.0167	6.2875	**0.0122**

Table 6: Predictive Mean Squared Error (PMSE) for GEE Models

Working Correlation Structure	PMSE
GEE Independence	**1.6669**
GEE Exchangeable	1.6678
GEE AR-1	1.6777

The GEE models under exchangeable and AR-1 correlation structures produced statistically significant estimates for the cumulative effect of smoking cessation (years_quit). In contrast, the independence structure failed to detect this effect, likely due to its assumption of no within-subject correlation, which reduces efficiency when true correlation exists.

Selecting the Optimal Correlation Structure via PMSE

To evaluate model predictive performance, we computed the predictive mean squared error (PMSE) for each correlation structure. As shown in Figure X, the independence structure yielded the lowest PMSE (1.6669), closely followed by exchangeable (1.6678) and AR-1 (1.6777).

Despite the lower PMSE for independence, this structure may fail to efficiently estimate time-varying cumulative effects such as years_quit, especially when those effects accumulate over time. This tradeoff highlights the distinction between **predictive accuracy** and **estimation efficiency**. If the scientific interest lies in recovering the true average population effect (e.g., long-term impact of smoking cessation), exchangeable or AR-1 structures may be preferable. However, if prediction of future FEV1 at individual time points is the target response, the independence structure can still be a reasonable choice under certain conditions.

Overall, the PMSE results emphasize that model choice should be guided by analytic objectives: whether for robust inference or predictive accuracy.

6.2 Model Selection Based on QIC

To further evaluate model performance with respect to both fit and complexity, we computed the quasi-likelihood under the independence model criterion (QIC) for three GEE models with different working correlation structures. The results are summarized in Table 7.

Table 7: QIC Model Comparison Across Correlation Structures

Model	(Int)	age	BMI	cigs/day	dgn	pck_yrs	sbp	smk	trt	yrs_quit	QIC
fm_ex	-0.0130	-0.0104	0.0163	0.0097	-0.00239	-0.1177	0.355	0.486	0.0486	0.0378	**536**
fm_ind	-0.0134	-0.0101	0.0144	0.0126	-0.0013	-0.0981	0.357	0.478	0.0454	0.0361	537
fm_ar1	-0.0132	-0.0105	0.0128	0.0132	-0.0030	-0.2382	0.378	0.127	0.0459	0.0372	539

Based on the QIC values, the exchangeable model (QIC = 536) outperformed the independence (QIC = 537) and AR-1 (QIC = 539) structures. The difference in QIC (DeltaIC) compared to the best model was 0.45 for independence and 2.68 for AR-1. According to standard thresholds, delta IC > 2 is generally considered meaningful, suggesting AR-1 is statistically worse.

Considering both model fit (QIC) and power to detect key long-term effects (e.g., years_quit), the exchangeable structure emerges as the most appropriate working correlation specification for GEE in this context.

Conclusion

Our simulation study demonstrates the complementary strengths of PMSE and QIC in evaluating GEE models for longitudinal FEV1 outcomes. PMSE is designed to select models that minimize prediction error on the response variable (FEV1) and effectively eliminates noisy or weakly informative predictors—such as current smoking status, which is often unstable and varies across time. However, PMSE may fail to detect the impact of cumulative risk factors, such as the duration of smoking cessation, which manifest over multiple time points.

In contrast, the QIC criterion favored the exchangeable working correlation structure and consistently selected models in which the cumulative effect of quitting smoking (years_quit) remained statistically significant. This suggests that QIC better accommodates the intra-subject correlation inherent in repeated measures and is more sensitive to long-term covariate effects.

These findings align with existing literature. As Pan (2001b) emphasized in their analysis of the Lung Health Study, "even though FEV1 tends to increase if one quits smoking in the current year, it may also depend on when the participant stopped smoking"—for instance, a five-year quitter may have significantly different lung function from a one-year quitter.

In summary, QIC is better suited for identifying structurally important covariates with cumulative effects, while PMSE is more effective in optimizing predictive performance. Together, they offer a comprehensive strategy for variable selection and model evaluation in longitudinal health data.

References

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

Pan, W. (2001a). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57, 120–125.

Pan, W. (2001b). Model selection in estimating equations. *Statistics in Medicine*, 20, 89–104.