Appendix for 'Analytical Considerations for Repeated Measures of Estimated Glomerular Filtration Rate in Cohort Studies of Chronic Kidney Disease"

Haochang Shou, PhD[a,b], Jesse Y. Hsu, PhD[a,b], Dawei Xie, PhD[a,b], Wei Yang, PhD[a,b], Jason Roy, PhD[a,b], Amanda H. Anderson, PhD[a,b], J. Richard Landis, PhD[a,b], Harold I. Feldman, MD, MSCE[a,b,] Afshin Parsa, MD, MPH[c,d], Christopher Jepson, PhD[a,b] on behalf of the Chronic Renal Insufficiency Cohort (CRIC) Study Investigators

[a]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[b]Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[c]Department of Medicine and Division of Nephrology, University of Maryland School of Medicine, Baltimore, MD, USA

[d]Department of Medicine, Baltimore Veterans Affairs Medical Center, Baltimore, MD, USA

1. Data Organization and Visualization

For longitudinal study with moderate size, the dataset are often prepared in either of the two ways: wide format and long format. Wide format is suitable when the repeated measures are collected over discrete visits and subjects have roughly the same number of total visits. In a wide format dataset, a subject's observations from multiple variables and visits (e.g. eGFR1, eGFR2,…,eGFRm) fit into one single row and different columns. The long format, on the other hand, is more efficient and suitable when the visit numbers are large and vary across subjects. In a long format dataset, the repeated measures of each subject from different time points are stacked in multiple rows. For variables that were measured at baseline (e.g., gender, age of enrollment and baseline hemoglobin levels), or those do not change over time such as genotype, their values are copied across all rows within each subject. A dataset that is arranged in the long format will contain, in addition to the subject IDs, the time variable (or variables) denoting when each observation was collected. In our motivating example, the variable 'YEARS' is the continuous time variable indicating when eGFRs were measured. Table A1 illustrates the wide and long format organization for the repeated eGFR values of two hypothetical subjects (01001 and 01002). The first subject (01001) has two eGFR measures and the second subject (01002) was followed for 3 years with 4 repeated eGFR values.

Long format is in general more preferred for advanced longitudinal data analysis since it is more flexible to handle subjects with different numbers of clinical visits or irregular time points of measurement, and is also easier for dynamically updating the dataset with future follow-ups.' For example in Table A1, in order to shape a wide format data frame, we need to create 4 columns of eGFR variables (eGFR0, eGFR1, eGFR2 and

eGFR3) to accommodate the second subject who has the maximum number of clinical visits. This leaves many cells with NA values for the first subject who had fewer eGFR measures. While in long format data frame, eGFR and Year are separately into two variables, and the repeated eGFR values over time are stacked on multiple rows. No extra space with NA needs to be created. Furthermore, as both subjects stay enrolled in the study, it is also easier to update the long-format data with future eGFR values by simply merging the new rows of observations with the old data frame without changing the contents of the old data frame.

The simple analytic functions such as rANOVA in most of the standarad softwares (e.g., SAS, R and Stata) often accept wide format dataset. While long format data would be the default data format for longitudinal data analysis such GEE and linear mixed effects model. Many of them also provide functions to convert wide format to long format, or vice versa. When using the long-format data in the analysis, however, one should always remember to identify and specify the variable name for subject IDs, so that the program could recognize which rows of repeated measures belonging to the same subject.

A. Wide format

ID	eGFR0	eGFR1	eGFR2	eGFR3
01001	78.59	62.78	NA	NA
01002	55.28	47.32	47.22	31.54

B. Long format

ID	YEARS	eGFR
01001	0	78.59
01001	1	62.78
01002	0	55.28
01002	1	47.32
01002	2	47.22

01002	3	31.54

Table A1: Wide-format (A) and Long-format (B) for the repeated eGFR values of two hypothetical subjects (01001 and 01002) with different number of clinical visits.

2. Correlation Structures

As emphasized in the paper, a unique and crucial characteristic of the longitudinal data is that the repeated measures within the same subject are potentially correlated. Here we introduce the concepts and the mathematical formulation of a correlation matrix using the *APOL1* example. Suppose that eGFR is measured three times for each subject, eGFR1, eGFR2 and eGFR3. The correlation between eGFR values at any two visits specifies the extent to which the values of eGFR covary within subject. To represent this, we use a correlation matrix, as follows

	eGFR1	eGFR2	eGFR3
eGFR1	1	ρ_{12}	ρ_{13}
eGFR2	ρ_{12}	1	ρ_{23}
eGFR3	ρ_{13}	ρ_{23}	1

The correlation on the diagonal is always 1 as it is perfectly correlated with itself. Those off diagonal elements ρ_{kl} (k, l=1,2,3) indicate the amount of correlations in eGFR values between visit k and visit l and take value between 0 and 1, and need to be estimated from the model fit. The matrix is symmetric and the values in the lower triangular below diagonal are a mirror image of the values above it. As mentioned in the

paper, many longitudinal modeling techniques require one to specify the nature of these correlations, generally referred to as correlation structures. Typical choices of correlation structures include independence, exchangeable, autoregressive (AR), m-dependent and unstructured. Under independent structure assumption, all the off diagonal elements in the correlation matrix are assumed to be 0 indicating eGFR values from the two different visits are uncorrelated and are treated as measures obtained from two different subjects.

Independence correlation structure:

	eGFR1	eGFR2	eGFR3
eGFR1	1	0	0
eGFR2	0	1	0
eGFR3	0	0	1

The exchangeable structure (also called compound symmetry) assumes that correlations between any pair of the observations are identical to be value ρ, where ρ needs to be estimated from the data. In the following example matrix, any pair of eGFR values are estimated to have a correlation of 0.51.

Exchangeble correlation structure:

	eGFR1	eGFR2	eGFR3
eGFR1	1	0.51	0.51
eGFR2	0.51	1	0.51
eGFR3	0.51	0.51	1

Auto-regressive (AR) structure is often used in longitudinal analysis and refers to when the magnitude of the within-subject correlation decreases exponentially as the two visits get farther apart; in the example below, correlations of eGFR values from one visit apart such Year 1 and 2, 2 and 3 are equal and estimated to be 0.69, while eGFR values from two visits apart between Year 1 and 3 are less correlated to be 0.48 which is equal to 0.69^2.

Auto-regressive correlation structure:

	eGFR1	eGFR2	eGFR3
eGFR1	1	0.69	0.48
eGFR2	0.69	1	0.69
eGFR3	0.48	0.69	1

All the aforementioned structures only require the analysis to estimate one parameter. m-dependent structure is similar to AR but is more general in that the observations from different visits are correlated if they are within m visits from each other. The following matrix shows an example of 1-dependent structure where correlation is 0.70 only when the visits are one year apart (Year 1 and 2, 2 and 3). Year 1 and 3 have distance of two years and hence their eGFR values are assumed to be uncorrelated.

1-dependent correlation structure:

	eGFR1	eGFR2	eGFR3
eGFR1	1	0.70	0
eGFR2	0.70	1	0.70
eGFR3	0	0.70	1

At the opposite extreme is the unstructured covariance structure, in which the correlations between any pair of visits are assumed to be unique with no pre-specified patterns. As a result, more unknown parameters need to be estimated from data and a larger sample size is required. The following example shows an estimated unstructured correlation matrix where no specific patterns were assumed except that the matrix is symmetric.

unstructured correlation structure:

	eGFR1	eGFR2	eGFR3
eGFR1	1	0.67	0.34
eGFR2	0.67	1	0.57
eGFR3	0.34	0.57	1

3. Statistical Methods for Repeated Measures as Outcome

GEE model

Model (1) in the paper only include repeated eGFR as an outcome and the time is expressed in Year and the exposure variable is $APOL1$ risk groups. $APOL1$ variable contains three categories: Caucasian as reference, *APOL1* low risk and $APOL1$ high risk. The detailed mathematical formula of model (1) is specified in below

$$eGFR_{ij} = \beta_0 + \boldsymbol{\beta}_1 APOL1_i + \gamma_1 Year_{ij} + \boldsymbol{\gamma}_2 Year_{ij} * APOL1_i + e_{ij} \quad \text{(A.1)}$$

The mean response model $eGFR = \beta_0 + \boldsymbol{\beta}_1 APOL1 + \gamma_1 Year + \boldsymbol{\gamma}_2 Year * APOL1$ describes the average relationship between the outcome and covariates in the population. β_0 and $\boldsymbol{\beta}_1$ account for the associations between baseline eGFR and *APOL1* risk groups. The coefficients in front of the time varying variables (γ_1 and $\boldsymbol{\gamma}_2$) depict the relationship between eGFR slope and the corresponding risk factors. The coefficients in bold represent a vector of length two. Since *APOL1* is coded as a three-level categorical variable, $\boldsymbol{\beta}_1$ is a vector of coefficients with two values (β_{1L}, β_{1H}). That is, β_0 represents the average baseline eGFR for the reference group (Caucasian), $\beta_0 + \beta_{1L}$ stands for the baseline eGFR for *APOL1* low risk groups and $\beta_0 + \beta_{1H}$ is the baseline eGFR for *APOL1* high risk groups. The coefficients in front of the time-varying variables quantify the association of eGFR slope and the variables. In particular, γ_1 represents the average eGFR slope for the reference Caucasian group. $\boldsymbol{\gamma}_2$ contains two coefficients that correspond to the differences of eGFR slopes between the two *APOL1* risk groups and Caucasian, respectively. e_{ij} is the error term indicated in model (1) of the paper and we typically assume e_{ij} satisfies one of the working correlation structures.

Mixed effects model

In our example, the subject-specific random intercept characterizes the difference between an individual's baseline eGFR and the population average, and a random slope in front of a time variable describes the deviation of individual slope from the population average. Thus, the mixed effects model for our example looks like the following:

$$eGFR_{ij} = \beta_0 + \boldsymbol{\beta}_1 APOL1_i + \gamma_1 Year_{ij} + \boldsymbol{\gamma}_2 Year_{ij} * APOL1_i + b_{0i} + b_{1i} Year_{ij} + e_{ij} \quad \text{(A.2)}$$

The fixed effects coefficients β_0, $\boldsymbol{\beta}_1$, γ_1, $\boldsymbol{\gamma}_2$ share the same interpretations as in the GEE model (A.1) for the population average trend. b_{0i} and b_{1i} are random intercept and random slope that are unique for subject i. The estimated coefficients can be used to construct the individual trajectory. Suppose subject i is Caucasian, his/her eGFR trajectory was estimated to start with baseline value $\beta_0 + b_{0i}$ at the entry of the study, and changes $\gamma_1 + b_{1i}$ per year. The estimated eGFR trajectories from two hypothetical subjects are illustrated in Figure 2 in the paper.

4. Missing Data Mechanism

In general, GEE assumes that data are covariate-dependent missing completely at random (MCAR) [1]. This is saying that whether an outcome variable, such as eGFR, is missing at a particular visit does not depend on the renal function itself, but might be related with other observed covariates such as the time of visit or baseline characteristics. This assumption could be practically valid in longitudinal studies in that it allows the chance of missing eGFR measurements becomes larger in the later phase of the study, but is unrelated to other observed eGFR values at a given time point. Mixed-effects model handles both MCAR and missing at random (MAR), where MAR allows the chance of missing eGFR to be also related with eGFRs from the past. When their corresponding missing assumptions are satisfied, since both GEE and mixed effects models allow for the outcome data being observed on a different set of time points for different subjects, no data imputation is necessary. However, if the assumptions are violated, then simply apply

the regular GEE or mixed effects model on longitudinal data with missing values will result in biased estimates and invalid inference.

Another important scenario that is often encountered in CKD is missing not at random (MNAR) [1] or informative censoring [2], where neither mixed effects model or data imputation is suitable. For instance, patients drop out of the study due to some competing events that are related with CKD progression such as initiation of dialysis, renal transplantation or death. In these scenarios, the chance of missing eGFR could truly depend on the unobserved eGFR values and kidney function. Alternative methods discussed in the paper that jointly combine the missing mechanism and repeated measurements are more appropriate.

5. Statistical softwares

GEE model

GEE has been implemented in many statistical softwares such as PROC GENMOD in SAS and 'geepack' package in R. Model selection in GEE cannot be achieved using likelihood ratio test. Instead, one can choose the best model that minimizes the quasi-likelihood under the independence model criterion (QIC) [3], or QICu that penalizes the model complexity when too many covariates are included. This can be done using PROC IML in SAS or 'MuMIn' in R. Alternatively, to select an appropriate working correlation structure for GEE, in SAS one can choose to output either the 'empirical variance' that refers to the robust variance estimate, or the 'model-based' estimate that assumes the

working correlation is true. The closer the two estimates are, the more likely that the working correlation is appropriately selected.

To fit a weighted GEE (WEE) model when data are missing at random (MAR), we can specify WEIGHT statement when using PROC GENMOD, or MISSMODEL in PROC GEE in SAS.

Mixed effects model

The linear mixed effects model can be implemented in SAS using PROC MIXED, and in R using package 'lme4' or 'lmerTest'. For generalized linear mixed effects model with non-normal outcome, one can use command PROC GLIMMIX in SAS, or 'glmm' package in R. Model selection in GEE can be achieved using likelihood ratio test and compare Aikake Information Criterion (AIC) and Bayesian Information Criterion (BIC).

A detailed summary of available statistical softwares and packages can be found in Table 4 of Boucquemont et al.[4]

Reference:

1. Rubin DB. Inference and Missing Data. *Biometrika* 1976; **63:** 581-592.

2. Schluchter MD, Greene T, Beck GJ. Analysis of change in the presence of informative censoring: application to a longitudinal clinical trial of progressive renal disease. *Statistics in medicine* 2001; **20**(7)**:** 989-1007.

3. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2001; **57**(1)**:** 120-125.

4. Boucquemont J, Heinze G, Jager KJ, Oberbauer R, Leffondre K. Regression methods for investigating risk factors of chronic kidney disease outcomes: the state of the art. *BMC nephrology* 2014; **15:** 45.