Distribution Agreement

Signature:

______________________________	______________
Guandong Yang	Date

Evaluation of Efficiency in Pooling Strategies for Some Mixed Linear and Nonlinear Models in Longitudinal Studies

By

Guandong Yang
MPH

Department of Biostatistics and Bioinformatics

____ [Chair's signature]
Robert Lyles
Committee Chair

____ [Member's signature]
Amita Manatunga
Committee Member

Abstract

Design Strategies Under Some Random Effects Models for Repeated Measures or
Longitudinal Data Subject to Outcome Pooling
By Guandong Yang

For repeated measures or longitudinal studies, the cost and the effort of applying expensive laboratory assays to assess biomarker levels can be prohibitive. To mitigate assay costs, it is relatively common to pool samples prior to performing a laboratory test. In certain settings, an optimal pooling design has been shown to minimize information loss. Pooling laboratory samples can also preserve biospecimens and avoid issues with limits of detection. In this thesis, we explore the efficiency of pooling strategies for estimating fixed effects and variance components under mixed linear models for normally distributed outcomes as well as a mixed nonlinear model assuming a gamma-distributed outcome. We conducted a series of simulations to assess and compare the pooling strategies and models discussed in section 2. We evaluate the efficiencies of different pooling design strategies with initial simulation studies under standard Gaussian assumption-based linear mixed models. We also examine the efficiency of within-individual pooling a right-skewed outcome under the gamma model. The design strategies are presented with simulations inspired by The HIV Epidemiology Research Study, with assumed parameters mimicking a prior longitudinal study that estimated average trajectories of HIV ribonucleic acid (RNA) over time.

Design Strategies Under Some Random Effects Models for Repeated Measures or Longitudinal Data Subject to Outcome Pooling

By

Guandong Yang

Bachelor of Science
Nanjing Agricultural University
2016

Thesis Committee Chair: Robert Lyles, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Biostatistics
2018

Evaluation of Efficiency in Pooling Strategies for Some Mixed Linear and Nonlinear Models in Longitudinal Studies

1. Introduction

The basic nature of longitudinal data consists of repeated measures on each of a number of subjects. We must use more sophisticated statistical models to take account of the dependency within each subject, which could allow various correlation structure (McCulloch and Searle, 2005). However, when you have multiple measures for each subject, the cost and the effort of testing the assay are prohibitive. Longitudinal studies involve resource-intensive measurements obtained form laboratory assays, which motivates researchers to consider study designs to mitigate those costs. Strategies such as randomly sampling a part of biospecimens or combining samples prior to performing laboratory assays is often deployed. These techniques, while effective in reducing cost, are often accompanied by a considerable loss of statistical efficiency. An optimal pooling design prior to performing these lab assays has been shown to minimize the information loss, respective to a given model, parameters of interests in a logistic regression and linear regression settings. At the same time, pooling laboratory samples can also preserve biospecimens and avoid limits of detection issue.

There are numerous articles that investigate the advantages of deploying pooling strategies to different scenarios. Dorfman (1943) showed that combining samples or group testing can reduce expense while archiving elimination of rare disease. Vansteelandt et al. (2000) and Chen et al. (2009) considered regression models with fixed and/or random effects for group testing with binary outcomes. McMahan et al. (2012) investigated the technical violation of dilution effects when performing group testing. The pooling idea also been extended to continuous outcomes. For example, Schisterman and Vexler (2008) presented a cost-efficient pooling design for biomarker studies. Mitchell et al. (2014) took consideration of the skewness of the outcome and proposed three different models for pooling: lognormal, Gamma with constant shape and Gamma with constant scale. Malinovsky et al. (2012) investigated pooling designs for estimating the interclass correlation and variance components under a Gaussian random intercept model.

Biospecimen data are often continuous but right skewed and positive. One of the common ways to analyze such data is to log-transform outcome variable, in an effort to maintain the nice properties of normality that permits the use of a standard mixed linear model. There are alternatives in the generalized linear model framework, which might provide better model interpretation and avoid model assumptions violations. For example, through the log-link function, we can use Gamma distribution to model positive, continuous and right-skewed outcome data. For a pooled outcome, they share a useful summation property with the normal distribution; that is, if an individual level measure follows normal or Gamma distribution, then a pooled measure will also follow a normal or Gamma distribution.

In this article, we explore the efficiency of pooling strategies for fixed effects and random effects under mixed linear models and Gamma models for repeated measures or longitudinal data. In sections 2 and 3, we introduce the basic properties of mixed effect models and Gamma models with constant scale and their corresponding notation. In section 4, we describe three types of pooling strategy: pooling within subjects, pooling across subjects, and mixed pooling. The pooling strategy is illustrated by introducing a transformation matrix **Q**. In section 5, we summarized simulation studies motivated by the data from HIV Epidemiology Research Study (HERS) and investigate the mean bias and relative efficiency of different pooling design in estimating fixed effects and variance components. In section 6, we summarize different pooling strategies under different models and discuss potential future research topics.

2. Linear Mixed Effect Model

Mixed effect linear models are widely used for repeated measures and longitudinal data. Let Y_{ij} denote a continuous outcome, where $i = 1, ...I$ indicates individual and $j = 1, ..., J_i$ indicates repeated measurements on a given individual. When we do not pool, we assume that each measurement Y_{ij} is obtained from a separate assay. A common model is a randomized regression, whereby each subject is allowed to have their own random intercept and slope. For right-skewed outcomes, which are relatively common, are frequently applied a log transformation and works with a variant of the following models:

$$Y_{ij} \quad = \quad \beta_0 + b_{0i} + (\beta_1 + b_{1i})t_{ij} + \epsilon_{ij}, \tag{1}$$

where b_{0i} denotes the individual random effect on the intercept, b_{1i} denotes the individual random slope for the effect time and ϵ_{ij} denotes the random within-subject error. Typically, one assume the random intercept deviations, random slope deviations and random errors (b_{0i}, b_{1i}, ϵ_{ij}) follows a trivariate normal distribution:

$$\begin{bmatrix} b_{0i} \\ b_{1i} \\ \epsilon_{ij} \end{bmatrix} \sim MVN(0, \begin{bmatrix} \sigma_0^2 & \sigma_{01} & 0 \\ \sigma_{01} & \sigma_1^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix})$$

If β_1 and b_{1i} are all equal to 0, then Model (1) reduce to the one-way random effects analysis of variance(ANOVA) model. Then we have:

$$Y_{ij} \quad = \quad \mu + \alpha_i + \epsilon_{ij}, \tag{2}$$

The notation remains similar, α_i denotes the individual random effect on the intercept, ϵ_{ij} denotes the random within-individual error. One typically assumes that α_i and ϵ_{ij} are mutually indeoendent, where $\alpha_i \overset{iid}{\sim} N(0, \sigma_\alpha^2)$, $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

There is a more general expression for linear mixed model (McCulloch and Searle (2005)). Consider that the i_{th} individual, and consider $\mathbf{X_i}$ and $\mathbf{Z_i}$ as corresponding design matrices of fixed effects and random effects. In matrix notation, it expressed as:

$$\mathbf{y_i} = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b_i} + \boldsymbol{\epsilon_i}$$

$$\text{where, } \mathbf{b_i} = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim MVN(0, \mathbf{D} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix})$$

$$\boldsymbol{\epsilon_i} \sim MVN(0, \sigma_\epsilon \mathbf{I}_{Ji})$$

Therefore for each individual, we have

$$\mathbf{y_i} \sim MVN(\mathbf{X_i}\boldsymbol{\beta}, \mathbf{H_i} = \mathbf{Z_i}\mathbf{D}\mathbf{Z_i}^T + \sigma_\epsilon^2 \mathbf{I}_{Ji})$$

Then we build on this to write the model for all the individuals as:

$$\mathbf{Y} \quad = \quad \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon} \tag{3}$$

where

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \\ \vdots \\ \mathbf{y_I} \end{bmatrix} \mathbf{X} = \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \\ \vdots \\ \mathbf{X_I} \end{bmatrix} \mathbf{Z} = \begin{bmatrix} \mathbf{Z_1} & 0 & \cdots & 0 \\ 0 & \mathbf{Z_2} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{Z_I} \end{bmatrix} \mathbf{b} = \begin{bmatrix} \mathbf{b_1} \\ \mathbf{b_2} \\ \vdots \\ \mathbf{b_i} \end{bmatrix} \epsilon = \begin{bmatrix} \boldsymbol{\epsilon_1} \\ \boldsymbol{\epsilon_2} \\ \vdots \\ \boldsymbol{\epsilon_I} \end{bmatrix}$$

Such that,

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) \tag{4}$$

$$\text{where,}\mathbf{V} = \begin{bmatrix} \mathbf{H_1} & 0 & \cdots & 0 \\ 0 & \mathbf{H_2} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{H_I} \end{bmatrix}$$

3. Gamma Model

Biomarker data are often right-skewed so that it is not always realistic to make a Gaussian distributional assumption. Gamma regression is more suitable for modeling continuous right-skewed outcomes while still offering convenient summation characteristics that are appealing to pooling analysis[Emily 2014]. Gamma regression with a constant shape parameter is closely associated with quasi-likelihood models, which lends it some robustness to model misspecification[26]. In our application, the observed data likelihood can be maximized using PROC NLMIXED procedure in SAS (Institute (2011)). We propose a Gamma model with a random intercept in the linear predictor for the scale parameter; in this way, we can directly model outcome without a non-linear transformation (i.e., log, square-root). We assume the shape parameter θ is constant across individuals, and the scale terms β_i are constant within the i_{th} subject.

Let Y_{ij} denote the observation from the j_{th} measures in the i_{th} individual, such that

$$\begin{aligned} Y_{ij} &\sim Gamma(\theta, \beta_i), where\ \theta > 0, \beta_i > 0 &(5)\\ ln(\beta_i) &= \beta + b_i &(6)\\ where\ b_i &\sim N(0, \sigma^2) &(7) \end{aligned}$$

In particular, the overall mean and variance for the outcome follows rules involving conditional moments. $\mu_i = E(Y_{ij}) = E[E(Y_{ij}|\beta_{ij})] = E[\theta\beta_{ij}] = \theta E(\beta_i) = \theta e^{\beta+\sigma^2/2}$ and $Var(Y_{ij}) = E[Var(Y_{ij}|\beta_{ij})] + Var[E(Y_{ij}|\beta_{ij})] = \theta e^{2\beta+\sigma^2}[(\theta+1)e^{\sigma^2} - \theta]$. Under this model, our interest is in valid and efficient estimation of the parameter(θ, β, σ^2) subjected to various pooling designs.

4. Pooling Strategies

In repeated measures or longitudinal data, pooling of biospecimens could be used to reduce the cost of assays. We will compare different pooling strategies under different scenarios with the goal of obtaining the estimators of fixed effects and variance components with as little information loss as possible. These strategies include combining samples within individual and/or across individuals. In all the pooling strategies that we considered, we do not allow samples to be pooled more than once. In other words, each biospecimen will only be used in one pool as an individual sample. For Gaussian mixed linear model, because of the properties of the multivariate normal distribution, we can express the pooling strategies conveniently with a transformation matrix $\mathbf{Q}$ that convert full data $\mathbf{Y}$ to pooled data $\mathbf{Y}^*$

$$\mathbf{Y}^* = \mathbf{QY} \sim MVN(\mathbf{QX}\boldsymbol{\beta}, \mathbf{QVQ}^T = \mathbf{V}^*) \qquad (8)$$

For Gamma model with constant shape, it is difficult to express the pooled distribution with Q matrix.

Pooling within Individuals, Type I

We define pooling performed within each individual or across time as Type I pooling. To maintain consistent notation, here we use $\mathbf{Y}$ and $\mathbf{Y}^*$ to denote full data and pooled data. We demonstrate the type I pool matrix $\mathbf{Q}$ with $I = 2$, $J = 4$, and pool size = 2.

$$\mathbf{Y}^* = \mathbf{QY} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \end{bmatrix} = \begin{bmatrix} (Y_{11} + Y_{12})/2 \\ (Y_{13} + Y_{14})/2 \\ (Y_{21} + Y_{22})/2 \\ (Y_{23} + Y_{24})/2 \end{bmatrix}$$

In a general form, if the pool size is n and J is divisible by n, then we define the p_th poolwise outcome for subject i as:

$$Y^*_{ip} = \frac{1}{n} \sum_{j=(p-1)*n+1}^{pn} (y_{ij}) \text{ where } i = 1, ..., I, p = 1, ..\frac{J}{n}$$

For the random intercept/slope model,

$$Y^*_{ip} \;\; = \;\; \beta_0 + b_{0i} + (\beta_1 + b_{1i})t^*_{ip} + \epsilon^*_{ip}, \tag{9}$$

where

$$t^*_{ip} = \frac{1}{n} \sum_{j=(p-1)*n+1}^{pn} (t_{ij})$$
$$\epsilon_{ip} \overset{iid}{\sim} N(0, \sigma^2/n)$$

For the one-way ANOVA(random intercept) model,

$$Y^*_{ip} \;=\; \mu + \alpha_i + \epsilon^*_{ip} \tag{10}$$

where

$$\epsilon^*_{ip} \overset{iid}{\sim} N(0, \sigma_\epsilon/n)$$

For the Gamma model, we work with the poolwise sum ($Y^*_{ip} = \sum_{j=(p-1)+1}^{pn} y_{ij}$)

$$Y^*_{ip} \sim Gamma(n\theta, \beta_i) \tag{11}$$

Pooling across Individuals, Type II

Pooling performed across individuals is defined as Type II pooling. We demonstrate the type II pool matrix $\mathbf{Q}$ with $I = 2$, $J = 4$, and pool size n = 2.

$$\mathbf{Y}^* = \mathbf{QY} = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \end{bmatrix} = \begin{bmatrix} (Y_{11} + Y_{21})/2 \\ (Y_{12} + Y_{22})/2 \\ (Y_{13} + Y_{23})/2 \\ (Y_{14} + Y_{24})/2 \end{bmatrix}$$

In a general form, if the pool size is n and I is divisible by n, then we have

$$Y^*_{pj} = \tfrac{1}{n} \textstyle\sum_{i=(p-1)*n+1}^{pn} y_{ij}, \;\; where\ j = 1, ..., J, p = 1, ..\tfrac{I}{n}$$

For random slope model,

$$Y^*_{pj} \quad = \quad \beta_0 + b^*_{0p} + (\beta_1 + b^*_{1p})t^*_{pj} + \epsilon^*_{pj} \tag{12}$$

where

$$t^*_{pj} = \tfrac{1}{n}\textstyle\sum_{i=(p-1)*n+1}^{pn} t_{ij}$$
$$\begin{bmatrix} b^*_{0p} \\ b^*_{1p} \end{bmatrix} \sim MVN(\mathbf{0}, \mathbf{D} = \tfrac{1}{n}\begin{bmatrix} \sigma^2_0 & \sigma_{01} \\ \sigma_{01} & \sigma^2_1 \end{bmatrix})$$
$$\epsilon^*_{pj} \sim N(0, \sigma^2/n)$$

For the one way ANOVA model with random intercepts,

$$Y^*_{pj} \quad = \quad \mu + \alpha^*_p + \epsilon^*_{pj} \tag{13}$$

where,

$$\alpha^*_p \sim N(0, \sigma^2_\alpha/n),\ \epsilon_{pj} \sim N(0, \sigma^2/n)$$

Mixed Pooling, Type III

Type III pooling is considered as the mixture of Type I and Type II pooling. We also demonstrate Type III pool matrix **Q** with $I = 4$, $J = 2$, and pool size n = 2. In this case we assume that half of the biospecimens are pooled within individual and half of them are pooled across individual.

$$\mathbf{Y}^* = \mathbf{QY} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \\ Y_{41} \\ Y_{42} \end{bmatrix} = \begin{bmatrix} (Y_{11} + Y_{12})/2 \\ (Y_{21} + Y_{22})/2 \\ (Y_{31} + Y_{41})/2 \\ (Y_{33} + Y_{42})/2 \end{bmatrix}$$

There is no simple closed-form expression for the mixed linear model in the case. However, the mixed linear model still dictates that $\mathbf{Y}^* \sim MVN(\mathbf{QX}\boldsymbol{\beta}, \mathbf{QVQ}^T = \mathbf{V}^*)$. This type of pooling is not easily accommodated by a standard procedure like SAS PROC MIXED for mixed linear model. Instead, we used the general likelihood facility in SAS PROC NLMIXED to specify and to maximize the observed data likelihood.

5. Simulation Study

The HIV Epidemiology Research Study (HERS) conducted from 1993 to 1995 is a perspective, large-scale cohort study of HIV infection. The study rationale, organization, and methods have been described in detail elsewhere (Smith et al. 1997). In brief, from 1993 to 1995, 871 HIV-infected women aged 16 to 55 years, and 439 demographically matched women at risk of HIV through either self-reported injecting drug use or sexual contact, were enrolled at four US cities (Baltimore, Detroit, New York City and Providence). Semiannual visits consisted of an extensive interview, a physical examination and specimen collection. For monitoring the progression of diseases, at each visit, a viral load cell lymphocyte count was determined and HIV RNA was quantified in heparinized plasma specimens by using a branched chain DNA signal amplification assay (Chiron Corp, Emeryville, California), with a quantification limit of 500 copies per milliliter (Todd et al. 1995).

Results from the analysis of longitudinal HIV RNA data in the HERS motivated (Lyles et al. (2000)) the simulations to test the pooling strategy and models were mentioned in section 2. To simplify the analysis, we simulate data such that each visit is at equally spaced fixed time points and without censoring of outcomes. 12 observations from I participants were generated under a multivariate normal distribution or multivariate Gamma distribution, in concordance with section 2. The outcome mimicking log(HIV RNA) was generated through the linear mixed with random intercepts and slopes model, or the Gamma model with random intercepts based roughly on the parameter estimator provided from the analysis on the full HERS dataset.

In this simulation study, we consider results for three values of the patient number $(I = 50, 100, 500)$ with 5000 iterations, under the model assumptions:

$$1)\ log(Viral\ load) = Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, j = 1, ..., 12, i = 1, ..., I$$

$$\text{where } \mu = 2.88, \alpha_i \sim N(0, \sigma_\alpha^2 = 0.718), \epsilon_i j \sim N(0, \sigma^2 = 0.382)$$

$$2)\ log(Viral\ load) = Y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})t_{ij} + \epsilon_{ij}, j = 1...12, i = 1, ..., I$$

$$\text{where, } \beta_0 = 2.88, \beta_1 = 0.062$$

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim MVN(0, \mathbf{D} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix}) = \begin{bmatrix} 0.718 & -0.06 \\ -0.06 & 0.039 \end{bmatrix})$$

$$\epsilon_{ij} \sim N(0, \sigma^2 = 0.382)$$

$$t = \{0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5\}$$

$$3)\ Viral\ load = Y_{ij} \sim Gamma(\theta = 0.626, \beta_i), j = 1...12$$

$$\ln(\beta_i) = \beta + b_i$$

$$\text{where } \beta = 0.346,\ b_i \sim N(0, \sigma^2 = 0.534)$$

Scenario 1 and 3 represent the repeated measures that are not affected by time, Scenario 2 represents longitudinal data, in which each subject is assumed to have a linear trend in log (Viral load) over time.

To compare the simulation results for pooling design, we also consider random sampling. We randomly sample n observation for each participant to maintain what would be the same total number of assumptions for the outcome variable. When considering mixed pooling, we only consider the mix of half Type I and half Type II pooling.

We used the SAS PROC MIXED procedure to perform the regression analysis for type I and type II pooling under the mixed linear model. We used the PROC NLMIXED procedure to perform analysis for type I pooling under Gamma model and mixed pooling under the mixed linear effect. In the mixed linear model analysis, we used both the Restricted Maximum Likelihood (REML) and Maximum Likelihood (ML) approaches. However, Type III pooling and Gamma model could only accommodate the ML approach. All pooling under the linear mixed model use pools of size 2, and for Type III pooling, we only consider the mixture of exact half Type I and half Type II pools. Pooling strategies subject to the Gamma model consider pool sizes 2, 3, 4 and 6 to explore the effect of sample size on the information loss.

In the mixed linear models, we compared the accuracy and precision for estimating fixed effects and variance component separately. In terms of fixed effects, we calculated the mean bias, relative efficiency and 95% confidence interval (CI) coverage. The relative efficiency is defined as the ratio of the empirical variance of the fixed effect estimate calculated from the pooled data, and the one calculated from the complete data ($\frac{Var_P(\hat{\beta})}{Var_F(\hat{\beta})}$). Since REML and ML produce different estimated standard errors, the relative efficiency is calculated separated for each approach. For 95% CI coverage on fixed effect, we rely on parameters approximate T reference distribution with degree of freedom calculated according to default rules conformed to by the SAS PROC MIXED and NLMIXED procedure. In terms of variance component, since the ML approach obtains biased but consistent estimators, we consider mean bias and empirical relative efficiency for each of the variance components.

In terms of fixed effects, the Type I and II pooling strategies obtain identical unbiased estimates as do the full data regressions (Table 1). There is no information loss for these parameters through performing pooling. All the methods provide close to 95% coverage rates, except for Mixed pooling when the sample size is small.

For one way ANOVA model, researchers are often interested in the ICC as well, due to its value for assessing the reproducibility of certain biomarkers (e.g Malinovsky et al. 2012). Thus, we also show the empirical efficiency for the ICC as well as the variance components. As seen in Table 2, the relative efficiency for Type I is less than 2 and closer to 1 (1.003, if rounded to the third digit) for the between individual variance (i.e. σ_α^2) and greater than 2 for the within-individual variance (i.e. σ^2). The relative efficiency of Type II pooling for both σ_α^2 and σ^2 is close to 2. It is much less efficient to pool across as opposed to within subjects for estimating σ_α^2, but more efficient for estimating σ^2. The results from the ANOVA model are consistent with known asymptotic results (Searle et al. 2009; Malinovsky et al. 2012) and the relative efficiency is invariant to the number of participants. Type I is a highly efficient strategy for estimating ICC compared to Type II, Type III and random pooling. Compared to random sampling, Type I pooling has better precision in both between individual variation and the ICC. However, they have similar efficiency for estimating within individual variation. As expected, Type III mixed pooling provides efficiency of variance components and the ICC between that of Type I and Type II. It loses some precision in estimating the fixed effect when the sample size is small, as seen in Table 1.

Table 1: Fixed effect for random intercept models

		μ[†]		
Approach	Strategy	$I = 50$	$I = 100$	$I = 500$
ML	Type I [§]	0.0 (1.00) *95.2*	0.0 (1.00) *95.3*	0.0 (1.00) *95.1*
	Type II [‖]	0.0 (1.00) *95.0*	0.0 (1.00) *95.2*	0.0 (1.00) *95.0*
	Type III [‡]	0.0 (1.08) *89.2*	0.0 (1.00) *95.2*	0.0 (1.00) *95.2*
	Random Sampling	0.0 (1.07) *95.2*	0.0 (1.09) *94.9*	0.0 (1.10) *95.3*
REML	Type I	0.0 (1.00) *95.5*	0.0 (1.00) *95.3*	0.0 (1.00) *95.2*
	Type II	0.0 (1.00) *95.5*	0.0 (1.00) *95.4*	0.0 (1.00) *95.0*
	Random Sampling	0.0 (1.04) *95.4*	0.0 (1.09) *95.0*	0.0 (1.09) *95.3*

[†] Bias (Relative Efficiency) *95% Confidence Interval Coverage Rate.*
[§] Within individuals pooling
[‖] Across individuals pooling
[‡] Mixed pooling

From Table 3, it appeared once again that all the methods obtain unbiased and efficient estimators of the fixed effect and CI coverages are close to 95% for Mixed linear models, excepts for Mixed pooling with sample size 50. As expected, random sampling is the least efficient methods. Even though the total number of the observations is the same as for the pooled data strategies, it suffers from relatively significant information loss.

For variance components in Mixed linear models, especially the between individual variation, REML obtains unbiased estimators even when the sample is small. While the ML estimator shows larger bias, it has better precision. These two approaches are approximately similar when the sample size is large enough ($n = 500$). In terms of design methods, we can rule out the random sampling. Random sampling shows the largest empirical relative efficiency, which by our definition indicates the most information loss with the same number of assays. Type I and Type II pooling provides nearly identical estimators as in the full data regression so that one can expect almost

Table 2: Variance components and interclass correlation of random intercept models(bias(relative efficiency)).

		$I = 50$			$I = 100$			$I = 500$		
	Strategy	σ^2_α [†]	σ^2 [†]	ICC[†]	σ^2_α	σ^2	ICC	σ^2_α	σ^2	ICC
ML	Type I[§]	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00
		(1.00)	(2.49)	(1.23)	(1.00)	(2.46)	(1.21)	(1.00)	(2.52)	(1.26)
	Type II[‖]	0.03	0.00	0.02	0.01	0.00	0.01	0.00	0.00	0.00
		(1.99)	(2.05)	(2.17)	(2.01)	(1.91)	(2.06)	(1.97)	(2.02)	(1.99)
	Type III [‡]	0.02	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.00
		(1.23)	(2.46)	(1.47)	(1.33)	(2.14)	(1.46)	(1.33)	(2.19)	(1.49)
	Random Sampling	0.02	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.00
		(1.19)	(2.46)	(1.50)	(1.19)	(2.46)	(1.49)	(1.19)	(2.57)	(1.48)
REML	Type I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		(1.00)	(2.49)	(1.23)	(1.00)	(2.46)	(1.21)	(1.00)	(2.52)	(1.26)
	Type II	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00
		(2.07)	(2.05)	(2.14)	(2.05)	(1.91)	(2.05)	(1.98)	(2.02)	(1.98)
	Random Sampling	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
		(1.11)	(2.52)	(1.42)	(1.19)	(2.46)	(1.49)	(1.16)	(2.52)	(1.45)

[†] Bias (Relative Efficiency)
[§] Within individuals pooling
[‖] Across individuals pooling
[‡] Mixed pooling

no information loss for estimating the fixed effect coefficient under the within- and across- subject pooling designs.

In Table 4, we examine the performance of different pooling strategies for estimating the variance components in the random/slope intercept model. Type I pooling showed better empirical relative efficiency than Type II pooling for estimating the between-subject variance components. Furthermore, the relative efficiency remains the same with the increase of the sample size. Type I pooling provides a relative efficiency close to 1, and Type II pooling has a relative efficiency close to 2, which indicates Type I pooling has a noticeably smaller information loss than Type II when estimating the between individual variance components.

Table 3: Fixed effect for mixed linear models (bias (relative efficiency) confidence interval coverage rate)

		β_0 [†]			β_1 [†]		
	Strategy	$I = 50$	$I = 100$	$I = 500$	$I = 50$	$I = 100$	$I = 500$
ML	Type I[§]	0.0 (1.00) *94.9*	0.0 (1.00) *94.9*	0.0 (1.00) *95.3*	0.0 (1.00) *95.1*	0.0 (1.00) *94.7*	0.0 (1.00) *95.5*
	Type II [‖]	0.0 (1.00) *95.1*	0.0 (1.00) *94.7*	0.0 (1.00) *95.4*	0.0 (1.00) *94.8*	0.0 (1.00) *94.7*	0.0 (1.00) *95.6*
	Type III [‡]	0.0 (1.09) *90.1*	0.0 (1.00) *94.8*	0.0 (1.00) *95.3*	0.0 (1.09) *90.1*	0.0 (1.00) *94.9*	0.0 (1.00) *95.3*
	Random Sampling	0.0 (1.61) *94.7*	0.0 (1.66) *94.6*	0.0 (1.66) *94.7*	0.0 (2.07) *94.5*	0.0 (2.18) *94.3*	0.0 (2.13) *95.5*
REML	Type I	0.0 (1.00) *95.1*	0.0 (1.00) *94.9*	0.0 (1.00) *95.4*	0.0 (1.00) *95.5*	0.0 (1.00) *94.9*	0.0 (1.00) *95.4*
	Type II	0.0 (1.00) *95.1*	0.0 (1.00) *94.9*	0.0 (1.00) *95.4*	0.0 (1.00) *94.8*	0.0 (1.00) *94.9*	0.0 (1.00) *95.6*
	Random Sampling	0.0 (1.61) *94.8*	0.0 (1.66) *94.7*	0.0 (1.66) *94.7*	0.0 (2.07) *94.7*	0.0 (2.18) *94.3*	0.0 (2.14) *95.0*

[†] Bias (Relative Efficiency) *95% Confidence Interval Coverage Rate*
[§] Within individuals pooling
[‖] Across individuals pooling
[‡] Mixed pooling

With respect to within individual variance; Type II pooling provides empirical relative efficiency close to 2. However, Type I provided empirical relative efficiency greater than 2, which indicates Type II pooling yields relatively smaller information loss than Type I in for estimation the within-individual variance. As expected, mixed pooling yielded precision in between Type I and Type II pooling methods. These results are consistent with one finding under the one-way ANOVA model.

Although one main goal is the access the performance of different pooling methods, it is also worthwhile to compare the performance of the REML and ML approaches in both the ANOVA model and mixed linear regression models for pooled data. When comparing the REML and ML within the same pooling strategy (full data regres-

Table 4: Variance components and interclass correlation of mixed linear models.

		$I = 50$				$I = 100$				$I = 500$			
	Strategy	σ_1^2 †	$\sigma_0^{2\dagger}$	σ_{01}^2 †	$\sigma^{2\dagger}$	σ_1^2	σ_0^2	σ_{01}^2	σ^2	σ_1^2	σ_0^2	σ_{01}^2	σ^2
ML	Type I§	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		(1.01)	(1.03)	(1.02)	(2.51)	(1.01)	(1.01)	(1.01)	(2.49)	(1.01)	(1.02)	(1.01)	(2.47)
	Type II‖	0.03	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		(2.01)	(2.04)	(2.01)	(1.99)	(2.03)	(1.98)	(2.00)	(2.00)	(2.05)	(2.00)	(2.02)	(2.01)
	Type III ‡	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		(1.27)	(1.33)	(1.31)	(2.55)	(1.31)	(1.31)	(1.32)	(2.21)	(1.34)	(1.35)	(1.36)	(2.24)
	Random	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Sampling	(2.62)	(4.92)	(4.23)	(8.42)	(2.67)	(4.56)	(3.92)	(7.91)	(2.65)	(4.60)	(3.94)	(8.24)
REML	Type I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		(1.01)	(1.03)	(1.02)	(2.51)	(1.01)	(1.01)	(1.01)	(2.49)	(1.01)	(1.02)	(1.01)	(2.47)
	Type II	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		(1.93)	(1.96)	(1.93)	(1.99)	(2.07)	(2.02)	(2.04)	(2.00)	(2.06)	(2.01)	(2.02)	(2.01)
	Random	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Sampling	(2.57)	(4.84)	(4.15)	(8.44)	(2.66)	(4.52)	(3.89)	(7.91)	(2.65)	(4.6)	(3.94)	(8.24)

† Bias (Relative Efficiency)

§ Within individuals pooling

‖ Across individuals pooling

‡ Mixed pooling

sion, Type I and Type II pooling), we note that, REML estimators provide unbiased estimators of variance components, especially between the individual variation, even when the sample size is small (n = 50 and 100). Even though the ML estimators are biased, they showed smaller empirical standard deviation than the REML estimator. As with the fixed effect, REML and ML estimators provide unbiased estimators with similar efficiency, when the sample size is large (n = 500).

Table 5 examines estimation under the Gamma model with random intercepts in the linear predictor for the scale parameter. With $I = 500$, ML yields nearly unbiased estimators. Indeed, we know from ML theory that for large enough sample size, all the estimators will converge in probability to the true value. From Table 5, we can see the Maximum Likelihood estimator (MLE) for the scale parameter θ and the fixed effect β appear to converge faster in probability than does to MLE variance σ^2. In the full data Gamma regression we note that the MLE for σ^2 still exhibited small bias.

With an increase in the pool size, the loss of efficiency for estimating θ is even more pronounced for within-subject (Type I) pooling than for random sampling. When pool size equals to 2, the empirical standard deviation from the Type I pooling is greater than random sampling with the same total number of assays. Pooling exhibits more efficiency in terms of estimating β and σ^2 and performs much better than corresponding random sampling for the purpose. Especially we have that σ^2 remains as precisely estimated as with the full data Gamma regression even when the pool size increases to 6.

In summary, our study support that the Type I pooling under Gamma model yields almost no information loss in terms the variance of the random effect. However, the efficiency loss increases with the increase of the pool size when it comes to estimating the scale parameter. On the contrary, Type I pooling under the ANOVA model showed no information loss in fixed effect estimation and only a small amount of loss of efficiency for estimating the variance component.

Table 5: parameter comparison among different pool sizes Gamma model

	θ †		β†		σ†	
	Type I§	Random Sampling	Type I	Random Sampling	Type I	Random Sampling
Pool Size 2	0.00 (2.47)	0.00 (2.09)	0.00 (1.20)	0.00 (1.36)	0.01 (1.00)	0.03 (1.36)
Pool Size 3	0.00 (4.43)	0.00 (3.42)	0.00 (1.46)	0.01 (1.82)	0.01 (1.01)	0.03 (1.73)
Pool Size 4	0.00 (6.92)	0.00 (5.08)	0.00 (1.77)	0.00 (2.36)	0.01 (1.00)	0.01 (2.67)
Pool Size 6	0.00 (14.57)	0.00 (8.39)	0.00 (2.78)	0.00 (3.81)	0.01 (1.04)	0.01 (4.43)

† Bias (Relative Efficiency)

§ Within individuals pooling

6. Discussion

We have generalized the settings consider by work of Malinovsky et al. (2012) to extend to considering a random slope model and a Gamma model with random intercepts. We considered three pooling strategy and have compared REML and ML estimates under different sample sizes. In a study with repeated measures, it is possible to cut assay costs by half even when performing pooling with pool size 2. Because biospecimen data are often continuous right-skewed and violating the Gaussian assumption, we proposed two types of models for pooling. One is the traditional mixed linear model, and the other is Gamma model with a constant shape.

Through the simulation study, we showed in the random intercept model, Type I pooling showed superior performance to that of Type II pooling for estimating between individual variation while Type II pooling was more efficient in terms of estimating within individual variation. Both the pooling strategies were efficient for estimating fixed effect and more efficient than the corresponding random sampling.

Again, the motivation for proposing the Gamma model is that the Gaussian assumption may not be realistic for biospecimens data, while a typical logarithm transformation complicates model fitting for pooled outcomes (Mitchell et al. 2014). In this article, we considered Type I pooling subject to a Gamma model with constant shape and random intercepts in the linear predictor for the scales, which provides better efficiency in estimating the parameters used to model the scale than does random pooling. In order to perform Type II pooling, we need a new way to parameterize the standard Gamma model. This parameterization can be referred as an "alternate" Gamma model, which assume constant scales and models the mean through a linear predictor of the shape term (Mitchell et al. 2015). This "alternate" Gamma model can also be estimated through PROC NLMIXED by maximizing the likelihood, and is a subject for our future work. Another interest of longitudinal data is marginal mean or overall population mean. Because of the nonlinear assumption, the marginal mean is different than the individual or conditional mean. It will be worthy to investigate the efficiency of different pooling strategies for estimating the marginal mean as a future topic.

Goodness-of-fit tests is of great interest in accessing model fitting and variable selection (Zheng 2000). In future work, we could develop goodness-of-fit measures to access that whether a particular model (e.g., mixed linear model or mixed nonlinear model with Gamma assumptions) fits better based pooled data.

In Type I pooling for the longitudinal data, if we have the linearity assumption on the effect of time, we can show that pooling specimens taken closer in time yields the most efficiency and it can provide conditional mean estimator for each individual.

In the random effect model, we did not incorporate covariate information of each individual. We assumed the baseline characteristics (i.e., gender, age) would not affect the virus load, which is not true for most of the studies. Further work with Type II pooling can seek optimization by forming pools based on covariate information. In Mitchell et al. 2014 it was shown that a k-means clustering strategy can reduce the information loss and provide more precision than random pooling or sampling. This pooling strategy could also be deployed in the random effect models to obtain the most efficiency. The model could have more complex hierarchical random effect structure. Future research could be to develop closed-form expressions for relative efficiency for both fixed effects and variance components in the longitudinal setting.

All the pooling considered above is under the assumption that there is no technical violation (i.e., dilution effects and measurement error). However, pooling strategies are not only used as a technique to save laboratory assay costs but also as a way to protect data confidentiality (Saha-Chaudhuri and Weinberg 2013). If researchers cannot access the clinic raw data, then pooling can be used for data ”encryption”. In this way, the clinic or medical center will perform data ”pooling”, the data we have is the average or the sum of two or more sample without no technical violation.

In summary, different pooling strategies showed advantages in different scenarios under the models considered herein. Choosing optimal strategies tailored to specify models and parameters of interest can save laboratory assay costs while maintaining minimal information loss. Next steps would include developing information matrix for our three types of pooling design and illustrating the strategies by the analysis of real data that involves biospecimens or the need to preserve data privacy.

References

Chen, P., J. M. Tebbs, and C. R. Bilder (2009). Group testing regression models with fixed and random effects. *Biometrics 65*(4), 1270–1278.

Dorfman, R. (1943). The detection of defective members of large populations. *The Annals of Mathematical Statistics 14*(4), 436–440.

Institute, S. (2011). *SAS/IML 9.3 user's guide.* Sas Institute.

Lyles, R. H., C. M. Lyles, and D. J. Taylor (2000). Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative drop-outs. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 49*(4), 485–497.

Malinovsky, Y., P. S. Albert, and E. F. Schisterman (2012). Pooling designs for outcomes under a gaussian random effects model. *Biometrics 68*(1), 45–52.

McCulloch, C. E. and S. R. Searle (2005). Linear mixed models (lmms). *Generalized, linear, and mixed models*, 156–186.

McMahan, C. S., J. M. Tebbs, and C. R. Bilder (2012). Regression models for group testing data with pool dilution effects. *Biostatistics 14*(2), 284–298.

Mitchell, E. M., R. H. Lyles, A. K. Manatunga, M. Danaher, N. J. Perkins, and E. F. Schisterman (2014). Regression for skewed biomarker outcomes subject to pooling. *Biometrics 70*(1), 202–211.

Mitchell, E. M., R. H. Lyles, A. K. Manatunga, N. J. Perkins, and E. F. Schisterman (2014). A highly efficient design strategy for regression with outcome pooling. *Statistics in medicine 33*(28), 5028–5040.

Mitchell, E. M., R. H. Lyles, and E. F. Schisterman (2015). Positing, fitting, and selecting regression models for pooled biomarker data. *Statistics in medicine 34*(17), 2544–2558.

Saha-Chaudhuri, P. and C. R. Weinberg (2013). Specimen pooling for efficient use of biospecimens in studies of time to a common event. *American journal of epidemiology 178*(1), 126–135.

Schisterman, E. F. and A. Vexler (2008). To pool or not to pool, from whether to when: applications of pooling to biospecimens subject to a limit of detection. *Paediatric and perinatal epidemiology 22*(5), 486–496.

Searle, S. R., G. Casella, and C. E. McCulloch (2009). *Variance components*, Volume 391. John Wiley & Sons.

Smith, D. K., D. L. Warren, D. Vlahov, P. Schuman, M. D. Stein, B. L. Greenberg, S. D. Holmberg, and H. I. V. E. R. S. Group (1997). Design and baseline participant characteristics of the human immunodeficiency virus epidemiology research (her) study: a prospective cohort study of human immunodeficiency virus infection in us women. *American Journal of Epidemiology 146*(6), 459–469.

Todd, J., C. Pachl, R. White, T. Yeghiazarian, P. Johnson, B. Taylor, M. Holodniy, D. Kern, S. Hamren, and D. Chernoff (1995). Performance characteristics for the quantitation of plasma hiv-1 rna using branched dna signal amplification technology. *Journal of acquired immune deficiency syndromes and human retrovirology: official publication of the International Retrovirology Association 10*, S35–44.

Vansteelandt, S., E. Goetghebeur, and T. Verstraeten (2000). Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics 56*(4), 1126–1133.

Zheng, B. (2000). Summarizing the goodness of fit of generalized linear models for longitudinal data. *Statistics in medicine 19*(10), 1265–1275.