2016

# Meta-Analytic Estimation Techniques for Non-Convergent Repeated-Measure Clustered Data

Aobo Wang
*Virginia Commonwealth University*

# Meta-Analytic Estimation Techniques for Non-Convergent Repeated-Measure Clustered Data

A dissertation submitted in partial fulfillment of the requirements for the Doctor of Philosophy in Biostatistics at Virginia Commonwealth University

By

Aobo Wang

Doctor of Philosophy

Director: Dr. Roy T. Sabo

Associate Professor

Department of Biostatistics

Virginia Commonwealth University

Richmond, Virginia

May, 2016

# Acknowledgment

Firstly, I would like to express my sincere gratitude to my advisor Dr. Sabo for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this dissertation. I could not have imagined having a better advisor and mentor for my Ph.D. study.

I also would like to thank my dissertation committee members Dr. Boone, Dr. Kang, Dr. Lu, and Dr. Sima for their advice with this research. Last but not least, I am grateful to the Department of Biostatistics at Virginia Commonwealth University for offering me this valuable opportunity of study.

# Table of Contents

# List of Tables

# List of Figures

# Abstract


META ANALYTIC ESTIMATION TECHNIQUES FOR NON-CONVERGENT REPEATED-MEASURE CLUSTERED DATA

By Aobo Wang, Ph.D.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biostatistics at Virginia Commonwealth University.

Virginia Commonwealth University, 2016.

Major Director: Dr. Roy T. Sabo, Associate Professor, Department of Biostatistics

Clustered data often feature nested structures and repeated measures. If coupled with binary outcomes and large samples (>10,000), this complexity can lead to non-convergence problems for the desired model especially if random effects are used to account for the clustering. One way to bypass the convergence problem is to split the dataset into small

enough sub-samples for which the desired model convergences, and then recombine results from those sub-samples through meta-analysis. We consider two ways to generate sub-samples: the *K* independent samples approach where the data are split into k mutually-exclusive sub-samples, and the cluster-based approach where naturally existing clusters serve as sub-samples. Estimates or test statistics from either of these sub-sampling approaches can then be recombined using a univariate or multivariate meta-analytic approach. We also provide an innovative approach for simulating clustered and dependent binary data by simulating parameter templates that yield the desired cluster behavior. This approach is used to conduct simulation studies comparing the performance of the *K* independent samples and cluster-based approaches to generating sub-samples, the results from which are combined either with univariate and multivariate meta-analytic techniques. These studies show that using natural clusters leaded to lower biased test statistics when the number of clusters and treatment effect were large, as compared to the *K* independent samples approach for both the univariate and multivariate meta-analytic approaches. And the independent samples approach was preferred when the number of clusters and treatment effect were small. We also apply these methods to data on cancer screening behaviors obtained from electronic health records of n=15,652 individuals and showed that these estimated results support the conclusions from the simulation studies.

KEY WORDS:
Non-Convergence; Clustered Design; Random Effects; Meta-Analysis

# Chapter 1 Introduction

## 1.1 Working Problem

Data are often collected from a number of different naturally existing groups, which are called clusters. Observations within these clusters are often more similar than are observations from different clusters. Clustered data can also feature nested random effects and repeated measures. For example, repeated measures are naturally clustered within individuals. Data from a study by Krist et al [1] exhibits a hierarchical nested cluster structure, where observations are nested within patients who are nested within physicians, who in turn are nested within practices. These cluster effects cannot be ignored in statistical analysis, since treatment effect may very across clusters and assuming independence of subjects within clusters is not valid [2][3]. If we omit cluster effects, we are omitting potential source of variability and overstate the power of the test[4][5]. The relatedness of clustered data is often measured by the intracluster correlation coefficient (ICC) [6], comparing the variance within clusters with the variance between clusters [7].

Mixed models are often used for analyzing clustered data, where clusters are modeled as random effects. If coupled with binary outcomes and large samples (>10,000) in addition

to repeated measures and nested clusters, this complexity can lead to non-convergence problems for the desired model. One possible explanation for the non-convergence issue is that the estimates for the three components of parameters (including fixed effects, random effects for clusters, and random effects for repeated measures) cannot be reached at the same time within the maximum number of iterations. To work around this computational challenge for the complicated and large data, we also tried to use HPMIXED [8] procedure and HPGLIMMIX [9] macro in SAS (version 9.4, Cary, NC, USA), and varied options for model estimation including the number of iterations, estimation method (ML, REML, QLS, etc.), and the convergence criterion. None of them could solve the non-convergence problem. At this point, some may be tempted to reconfigure the random cluster effects as fixed effects, a change which often allows the model to converge. However, this switch from random to fixed effects is not without consequences. When clusters are treated as fixed effects, only effects that vary within clusters can be estimated, while effects that vary among clusters can no longer be estimated [10]. Moreover, the model can become more complicated since these switched "fixed effects" are often categorical with multiple levels and could make the model difficult to estimate, interpret or test. Another approach is to fit a simpler model by omitting particular random effects. Naturally, this option could omit important information on the study design and lead to smaller standard errors, and biased estimates and effect sizes, all of which can lead to incorrect interpretation of associations between variables [4][5].

Rather than consider these less-than-desirable approaches, one could attempt to work around the non-convergence issue in order to fit the desired model. While still incorporating the desired hierarchical random structure, one way to bypass the convergence problem is to make the analyzing dataset small enough to ensure the convergence of the desired model. We could split the overall database into smaller components that are analyzable by the desired model, and then recombine results from those components through meta-analysis. It has been shown that results of meta-analyses of smaller trials are usually compatible with results from larger trials, though this compatibility can be compromised when treatment effect variability between different trials is not considered [11][12]. If we could appropriately split the overall dataset into smaller independent subsets, the variability of treatment effect among the overall database the smaller subsets would not be an issue due to subset independence of subsets and similarities between the subsets and the entire dataset. Therefore using meta-analytic approaches may produce reliable results in this situation.

### 1.2 Motivating Example

The motivating example for this research is from a study of an electronic medical record software program (My Preventive Care, or MPC) developed by Krist et al [1]. This program alerts patients when they are overdue for certain cancer screenings, such as prostate examinations in men and breast examinations in women. The MPC program was implemented at the practice level, and patients decided whether or not they wanted to register for and use the portal. Of interest was to determine whether MPC users were more likely to change their cancer screening behavior as compared to MPC non-users.

Using colon cancer screening as an example, a patient's status with respect to whether or not they had received any form of colon cancer screening at the most recent visit preceding the implementation of MPC into their practice, and this status was again updated one month, three months, and six months following that implementation.

A hierarchical linear mixed model was fit including the repeated-measure binary outcome (patient underwent colon cancer screening or they did not), fixed effects for time (four levels: baseline, one, three and six months), group (registered for MPC or no), and their interaction, as well as random effects for primary care physicians and practices, where the physicians are in turn nested with one of eight practices. It was this three-level nested design (patients within physicians within practices) that was desired, and was fit using the GLIMMIX procedure in SAS to test hypotheses of whether there were differences in the change in colon cancer screening rates between MPC program users and non-users from baseline to each of one, three and six months. Fitting the data to all n=110,029 subjects, the desired model did not converge.

In this case, we did not want to omit physician- and practice- level variability, as the changes in screening rates are likely to vary by practice. Nor did we want to include these as fixed effects, as there are too many levels, and the resulting model would lack parsimony. So our proposed idea to work around the non-convergence issue is to split the overall database into subsets that are analyzable by the desired model, and then recombine results from those subsets through meta-analytic approaches.

### 1.3 Approaches for Sub-Sample Generation

We consider two ways of splitting the overall dataset into more manageable subsets. First we split data into *K* independent, mutually-exclusive sub-samples, what we call the *K* independent samples approach, where the desired model is fit on each of the new sub-samples. The *K* independent, mutually-exclusive sub-samples are created by sampling without replacement for k-1 times. We sample out N/k subjects each time, where N is the total number of subjects in the overall database. The sampling process is balanced by group and cluster effects, so each sub-sample would have the same cluster structures and group allocation with the overall database. And the model fit to each sub-sample is the same desired model fit to the whole dataset. The sampling process can be easily implemented by statistical software. In this research, we use the SURVEYSELECT procedure [8] in SAS with the sampling method of simple random sampling and stratification of non-overlapping group and cluster effects.

The second proposal is to use natural existing clusters as sub-samples, where simplifications of the desired models – ignoring the clustering measure – are fit to each of the separate clusters. For example, the data from the study by Krist et al. [1] have nested cluster structures, where patients are nested within physicians and physicians are nested within practices. We used practices as the natural existing clusters as sub-samples in this data example.

Both splitting approaches have advantages and disadvantages. The cluster-based approach may not capture all inter-cluster variation when clusters vary in sizes. In addition, some clusters may be so large that the desired model would still not converge. In these circumstances, the *K* independent samples approach may be preferred. However, the *K* independent samples may be considered as arbitrary created. We did not try to pick different *K* independent samples or consider using boot-strapping due to increased computational time, the estimation process already takes a considerable amount of time for one sample.

### 1.4 Meta-Analytic Approaches

Test statistics from either of these sub-sampling approaches can then be recombined using random-effect-based meta-analytic approaches [13]. Meta-analysis is a statistical approach to combine results from different studies that are conducted to answer similar research questions [14]. It was first introduced by the British statistician Karl Pearson [15][16], who combined results from several studies of typhoid inoculation. In 1976, Gene Glass coined the term 'meta-analysis' to refer to integrate findings from individual studies [16][17]. Two common assumptions for meta-analysis are that the studies are independent and that the results and outcomes from studies are exchangeable [18][19]. Although we only have data from one study, the sub-samples created by either of the two approaches mentioned above should be viewed as independent and similar to the whole study. Thus they could be considered to have exchangeable outcomes and therefore satisfy the underlying assumptions of meta-analysis. We considered two meta-analytic approaches: a univariate meta-analytic approach and a multivariate meta-analytic

approach. Univariate meta-analysis is taking one test statistic and one standard error from each sample and combining them into one single estimate and standard error, while the multivariate approach takes multiple test statistics and multiple standard errors from each sample and combining them into a vector of estimates and their corresponding covariance structure.

The univariate meta-analytic approach we used in this research is a two-step random-effect estimate [13] starting with the DerSimonian and Laird approach [20]. Given the test statistics, $y_1, \ldots, y_k$, (k is the number of sub-samples) and the sampling variances $\sigma_1^2, \ldots, \sigma_k^2$, we need to calculate an estimate of the inter-study (or inter-cluster) variance $\tau^2$ and then estimate the overall effect μ and its standard error. Suppose $s_1^2, \ldots, s_k^2$ and $t^2$ are the estimates of $\sigma_1^2, \ldots, \sigma_k^2$ and $\tau^2$, a weighted estimator of μ and its standard error (SE) can be expressed as:

$$m_w = \frac{\sum_i w_i y_i}{\sum_i w_i}, \quad \mathrm{SE}(m_w) = \frac{1}{(\sum_i w_i)^{\frac{1}{2}}}, \text{ where } w_i = \frac{1}{t^2 + s_i^2}. \tag{1}$$

The expression for the standard error in Equation (1) is an underestimate of the true standard error of $m_w$ [13]. The first-step DerSimonian and Laird estimate for $\tau^2$ [20] is

$$t^2(DL) = \max\left\{0, \frac{[\sum_i w_{i0}(y_i - y_w(0))^2] - (k-1)}{[\sum_i w_{i0} - \sum_i w_{i0}^2 / \sum_i w_{i0}]}\right\}, \tag{2}$$

where $y_w(0) = \sum_i w_{i0} y_i / \sum_i w_{i0}$, and $w_{i0} = 1/s_i^2$. Substituting $t^2(DL)$ for $t^2$ in Equation (1) yields the corresponding DerSimonian and Laird estimate, $m_w(DL)$ and its standard

error SE($m_w(DL)$). Then substitute $w_{i0}$ and $y_w(0)$ with $w_{iD} = 1/(t^2(DL) + s_i^2)$ and $m_w(DL)$ respectively, we will get the two-step estimate, $t^2(DL2)$, where

$$t^2(DL2) = \max\left\{0, \frac{[\sum_i w_{iD}(y_i - m_w(DL))^2] - [\sum_i w_{iD} s_i^2 - \sum_i w_{iD}^2 s_i^2 / \sum_i w_{iD}]}{[\sum_i w_{iD} - \sum_i w_{iD}^2 / \sum_i w_{iD}]}\right\} \quad (3)$$

Substituting $t^2(DL2)$ for $t^2$ in Equation (1) yields the corresponding two-step estimate by DerSimonian and Kacker [13] $m_w(DL2)$ for μ and its approximate standard error, SE($m_w(DL2)$).

Other than taking one test statistic and one standard error from each sub-sample, we also want to consider extracting more available information from each sub-sample. By combining multiple test statistics and multiple standard errors from each sample through multivariate meta-analytic approach, we may capture more inter-study variability and have more accurate results.

The multivariate meta-analytic approach we used in this research is developed by Houwelingen *et al.* [21]. Instead of using one observed effect from each sub-sample, we use multiple observed effects from each sub-sample and combine them into a vector of overall effect. Given a vector of observed effects from each sub-sample, $\boldsymbol{y_1}, \ldots, \boldsymbol{y_k}$, (k is the number of sub-samples) and the corresponding sampling variances $\boldsymbol{s_1^2}, \ldots, \boldsymbol{s_k^2}$, we need to estimate the overall effect $\boldsymbol{\mu}$ and the between sub-samples covariance parameters. The random effect model is:

$$\boldsymbol{y_i} \sim N(\boldsymbol{\mu}, \Sigma + C_i)$$

with $\boldsymbol{y_i}$ the vector of observed effect from sub-sample $i$ , $C_i$ the diagonal matrix with the vector of $\boldsymbol{s_i^2}$, Σ the between sub-samples covariance matrix of the observed effects. Maximum likelihood estimation for this model can be carried out by linear mixed-effect models. This multivariate meta-analytic approach is applicable for large enough number of sub-samples, which is for cluster-based approach and *K* independent samples approach with large k.

### 1.5 Simulating Data

In order to evaluate the efficiency of our proposed approach, we need to conduct simulation studies in various situations. The type of data we are focusing on are clustered dependent and binary data. There are methods existing to simulate dependent binary data, so we need to extend these methods to incorporate clustered random effect structures.

Emrich and Piedmonte [22] developed a gold-standard method for simulating dependent binary outcomes based on the multivariate normal distribution. Kang and Jung [23] and Haynes, Sabo and Chaganty [24] introduced an approach based on the multinominal distribution of all possible combinations of the binary outcomes. We extend the multivariate normal- and multinomial sampling-based approaches for simulating dependent binary outcomes to also incorporate a desired cluster structure. This extension requires probabilistically generating parametric simulation templates for each of the desired cluster levels or combinations. Several simple probability distributions are used to exemplify the process of establishing the cluster-specific parameters and effect sizes,

including the normal, uniform and beta distributions. In this research, we incorporated one level of cluster structure. However, this approach extends naturally to more complicated scenarios, including cases of two or more clustering factors, or even nested factors. The unifying theme is that data are simulated uniquely for each combination of clusters. Further, we have discretion in selecting how those factors or levels affect the particular probability distribution and parameters used to simulate the simulation template for each cluster. The dependence levels between the binary outcomes can also be made to be cluster or level dependent, provided a distribution is selected that offers control in selecting the desired dependence while also ensuring the proper support.

### 1.6 Outline of Dissertation

The rest of this dissertation is outlined as follows. The method of simulating clustered dependent binary outcomes is described and examined in Chapter 2. We explain the procedures for splitting the existing database to ensure model convergence, and the meta-analytic approaches for recombining those sub-samples in Chapter 3. Chapter 4 focuses on the performance of multivariate meta-analytic approaches applied to the same settings in Chapter 3. A brief discussion follows in Chapter 5.

# Chapter 2 Simulating Clustered Dependent Binary Data

## 2.1 Introduction

Methods for simulating dependent binary outcomes are often required for the assessment of statistical methodologies suitable for repeated measure study designs with dichotomous outcomes. Such simulation techniques can also be useful in determining required sample sizes for longitudinal study designs featuring binary measurements. Emrich and Piedmonte [22] developed a gold-standard method for simulating dependent binary outcomes based on the multivariate normal distribution. Kang and Jung [23] introduced an approach based on the multinominal distribution of all possible combinations of the binary outcomes. Both of these approaches were extended to account for modeling dependencies with odds ratios in Sabo *et al.* [24].

While useful for repeated-measures or multiple-outcome studies, these methods require expansion if they are to be used in more complicated situations. For instance, certain research studies feature inherent clustering, where groups of subjects exist in natural clusters or groups. Examples include studies of school-age children attending various classrooms or schools [25], or primary care patients who attend one of several primary

care facilities [1], the latter of which also features patients nested within primary care physicians, who are in turn nested within primary care practices that are nested within larger health care systems. The previously mentioned simulation approaches cannot incorporate this type of complexity without amendment and are unsuitable as currently constructed to simulate clustered repeated measure data that would mimic such a scenario.

In this Chapter, we extend the multivariate normal- and multinomial sampling-based approaches for simulating dependent binary outcomes to also incorporate a desired cluster structure. This extension requires probabilistically generating parametric simulation templates for each of the desired cluster levels or combinations. Several simple probability distributions are used to exemplify the process of establishing the cluster-specific parameters and effect sizes, including the normal, uniform and beta distributions. The rest of this Chapter is outlined as follows. The two simulation methods are briefly described in the next Section, and are extended to account for a desired cluster structure. The performance of these extensions are then examined through simulation studies. A brief discussion concludes this Chapter.

## 2.2 Materials and Methods

### 2.2.1 Simulation Methodologies: Multivariate Normal Approach

The simulation approach by Emrich and Piedmonte [22] utilizes the multivariate normal distribution to generate vectors exhibiting desired dependence levels, which are then

categorized into binary observations. The process begins by using the desired pairwise correlations $\rho_{ij}$ between binary measures $Y_i$ and $Y_j$ with marginal probabilities $p_i = P(Y_i = 1)$ and $p_j = P(Y_j = 1)$ to solve for a bivariate correlation $r_{ij}$ using the bivariate normal cumulative distribution function (CDF)

$$\Phi\left[z(p_i), z(p_j), r_{ij}\right] = \rho_{ij}\left(p_i q_i p_j q_j\right)^{1/2} + p_i p_j, \quad (4)$$

where $z(p)$ is the $p^{th}$ percentile of the standard normal distribution and $q = 1 - p$. Odds ratios could be used in place of correlations by replacing the right-hand side of Equation (4) with the Plackett copula [26] $C\left(p_i, p_j, \psi_{ij}\right)$, where $\psi_{ij}$ is the desired odds ratio, as shown in Sabo *et al.*[24]. The values $r_{ij}\ \forall\ i \neq j$ are then placed into a correlation matrix **R** and used to simulate a $k \times 1$ multivariate normal vector $\underline{z} = (z_1, \ldots, z_k)^T \sim MVN\left(\underline{0}, \mathbf{R}\right)$. Binary observations are then created by classifying each element of $\underline{z}$ by letting $Y_i = 1$ if $z_i \leq z(p_i)$ and $Y_i = 0$ otherwise. This process can be repeated by generating and classifying $n$ such vectors to create the desired simulated sample.

### 2.2.2 Simulation Methodologies: Multinomial Approach

The multinomial-based simulation method introduced by Kang and Jung [23] uses a multinomial distribution of all possible combinations of dependent binary outcomes, which can be created through the joint and marginal probabilities, along with the desired correlation. Given a desired correlation $\rho_{ij}$ between binary variables $Y_i$ and $Y_j$ with desired marginal probabilities $p_i$ and $p_j$, we first calculate the joint probability $p_{ij}$ using the following expression.

$$p_{ij} = p_i p_j + \rho_{ij}\sqrt{p_i q_i}\sqrt{p_j q_j}. \quad (5)$$

Note that if odds ratios are used instead of correlations, then $p_{ij}$ can be solved for by inserting the desired odds ratio $\psi_{ij}$ and marginal probabilities $p_i$ and $p_j$ into the Plackett copula, as described in Sabo *et al.* [24]. Note that whether correlations or odds ratios are used to model dependence, the remainder of the multinomial-based approach is identical after the pair-wise joint probabilities $p_{ij}$ are calculated.

If three or more dependent binary measures are to be simulated, then higher order joint probabilities must be calculated. Let $p_{ijk}$ represent the joint probability $P(Y_i = 1, Y_j = 1, Y_k = 1)$, which is not uniquely defined by the marginal probabilities and the correlation. As shown in Chaganty and Joe [27], the minimum and maximum $p_{ijk}$ are defined as follows,

$$p_{ijk,L} = max\{0, p_{ij} + p_{ik} - p_i, p_{ij} + p_{jk} - p_j, p_{ik} + p_{jk} - p_k\} \quad (6)$$

$$p_{ijk,U} = max\{p_i, p_j, p_k, 1 - p_i - p_j - p_k + p_{ij} + p_{ik} + p_{jk}\}$$

where any value $p_{ijk} \in [p_{ijk,L}, p_{ijk,U}]$ leads to a valid probability density function with the desired marginal probabilities and dependence level. Though any value in this range is appropriate, we take the midpoint $p_{ijk} = (p_{ijk,L} + p_{ijk,U})/2$. Higher order joint probabilities in cases of four or more dependent binary observations can be determined in a similar manner, though the calculations become more tedious as the number of observations increases.

These quantities are used to calculate the multinomial probability density function (PDF) of all combinations of outcomes, which for the two-variable case are shown in the first two columns of Table 1. The CDF is created by progressively summing the values of the PDF, where the subscripts on $P$ indicate whether each binary outcome is successful, with 1 for success and 0 for failure. For example, $P_{01} = P(Y_1 = 0, Y_2 = 1)$. After the CDF is determined, a random number $u \sim Uniform[0, 1]$ is simulated, and the simulated observations are generated based on the decision rules based on the CDF, as shown in the last two columns of Table 1. For example, if $P_{11} < u \leq P_{11} + P_{10}$, then the observation is recorded as $Y_1 = 1$ and $Y_2 = 0$, or simply as 10. This process can be repeated to generate a sample of $n$ dependent binary outcomes. A similar approach – outlined in Kang and Jung [23] and in Haynes *et al.* [28] – can be used in cases of three or more dependent binary outcomes.

Table 1 Two-Variable PDF, CDF and Decision Rules for Multinomial Approach

| PDF | CDF | Decision Rule and Simulated Outcome | |
|---|---|---|---|
| $P_{11} = p_{12}$ | $P_{11}$ | $U \leq P_{11}$ | 11 |
| $P_{10} = p_1 - p_{12}$ | $P_{11} + P_{10}$ | $P_{11} < U \leq P_{11} + P_{10}$ | 10 |
| $P_{01} = p_2 - p_{12}$ | $P_{11} + P_{10} + P_{01}$ | $P_{11} + P_{10} < U \leq P_{11} + P_{10} + P_{01}$ | 01 |
| $P_{00} = 1 - p_1 - p_2 + p_{12}$ | $P_{11} + P_{10} + P_{01} + P_{00}$ | $U > P_{11} + P_{10} + P_{01}$ | 00 |

### 2.2.3 Accounting for Random Effects by Generating the Simulation Templates

For the two simulation approaches discussed in Section 2.2.1 and 2.2.2, we simulate a set of binary data representative of a single population by repeating either process $n$ times using a single simulation template, which consists of all desired marginal probabilities $p_i, i = 1, \dots, k$ and pairwise correlations $\rho_{ij}$ (or odds ratios $\psi_{ij}$). To generate two or more groups of simulated binary observations, where groups are differentiated by either different marginal probabilities, dependencies, or both, the simulation approach is repeated separately for each group with the desired simulation template.

Expanding the multivariate- and multinomial-based approaches to account for random effects (say from clustering) requires only a simple extension of the process used when simulating binary observations for multiple groups. As a motivating example, let's assume we want to simulate $M$ clusters of $n$ samples of two correlated binary measures. Let's further assume that those two measures have marginal probabilities that vary across the $M$ clusters in such a way that the averages are $p_1 = \pi_1$ and $p_2 = \pi_2$ and the corresponding cluster variances for those rates are $\sigma_1$ and $\sigma_2$.

First we assume that the marginal probability for each binary measure has some probability distribution $p_i \sim f(\theta_i)$, where $f(\circ)$ is some probability mass or density function and $\theta$ is some parameter (possibly vector-valued) selected such that $E(p_i) = \int p_i f(\theta_i)\, dy = \pi_i$ and $V(p_i) = \int \big(p_i - E(p_i)\big)^2 f(\theta_i) dy = \sigma_i$ for group $i = 1, 2$. If we desire $M$ clusters of simulated observations, then we simulate $M$ marginal probabilities $p_{1,m}$ and $p_{2,m}$ from $f(\theta_1)$ and $f(\theta_2)$, respectively, for $m = 1, \dots, M$. For cluster $m$, we

simulate the desired number of dependent binary observations using $p_{1,m}$ and $p_{2,m}$ and the desired dependence level $\rho_{12}$ (or $\psi_{12}$). This process is repeated for $m = 1, \ldots, M$, and the resulting $M$ clusters of simulated data will *on average* exhibit a distribution of marginal probabilities centered around $\pi_1$ and $\pi_2$, though the cluster-specific marginal means will vary according to $\sigma_1$ and $\sigma_2$, thus achieving the desired level of clustering.

In the previous scenario, the marginal probabilities were given probability distributions and themselves simulated $M$ times to achieve a clustering effect. An equivalent approach would be to simulate $p_1 \sim f(\theta_1)$ probabilistically to achieve a desired mean and variance for the first marginal mean across clusters, and then simulate some $\delta \sim g(\gamma)$ and define $p_2 = p_1 + \delta$, where $g(\circ)$ is some probability distribution not necessarily of the same form as $f(\circ)$, and $\gamma$ is some parameter (possibly vector-valued) such that $E(\delta) = \int \delta g(\gamma) d\delta$ yields the desired difference between $p_1$ and $p_2$ with some desired cluster variability $V(\delta) = \int \big(\delta - E(\delta)\big)^2 g(\gamma) d\delta$.

This approach extends naturally to more complicated scenarios, including cases of two or more clustering factors, or even nested factors. The unifying theme is that data are simulated uniquely for each combination of clusters, mainly through parametric templates that are probabilistically generated for each combination. For example, in the case of hierarchical clustering, where one factor is nested within the levels of another, the parameters $\theta_i$ used to simulate the parameter values used to simulate data for each cluster can itself be probabilistically determined. Further, the researcher has much discretion in

selecting how those factors or levels affect the particular probability distribution and parameters used to simulate the simulation template for each cluster. The dependence levels between the binary outcomes can also be made to be cluster- or level-dependent, provided a distribution is selected that offers control in selecting the desired dependence while also ensuring the proper support.

### 2.2.4 Distribution Examples

We will consider three examples of distributions that can be used in this simulation process, understanding that there are alternative and potentially more suitable options available. The only requirement is that the support of the distribution must either be equal to $[0, 1]$, be a proper subset of $[0, 1]$, or have a reasonably low probability of occurring outside $[0, 1]$. A simple choice would be to simulate the marginal probabilities from a uniform distribution such that $p_{i,m} \sim Uniform[\theta_{i1}, \theta_{i2}]$ for $m = 1, \dots, M$ clusters and $i = 1, \dots, k$ binary outcomes, where the midpoint of $\theta_{i1}$ and $\theta_{i2}$ yields the desired marginal mean $\pi_i$. In this case the inter-cluster variability in marginal probabilities can be controlled by increasing or decreasing the difference $\theta_{i2} - \theta_{i1}$, making it wider for greater variability and narrower for less variability. In this case the Uniform parameters can be selected such that $p_{i,m} \in [0, 1] \; \forall \; i, m$.

Another example would be to simulate the marginal probabilities from a beta distribution such that $p_{i,m} \sim Beta[\alpha_i, \beta_i]$ for $m = 1, \dots, M$ clusters and $i = 1, \dots, k$ binary outcomes, where shape parameters $\alpha_i$ and $\beta_i$ are selected so that the mode is equal to the desired

marginal probability (i.e. $(\alpha_i - 1)(\alpha_i + \beta_i - 2) = \pi_i$). We use mode instead of mean for the Beta distribution because we also want to investigate whether using mode has more accurate or biased results than using mean. There are infinite pairings of the shape parameters that give the same mode, so the inter-cluster variability in the marginal probabilities is controlled by making both $\alpha_i$ and $\beta_i$ larger (for less variability) or smaller (for more variability). Since the support of the distribution matches that of proportions and probabilities, we are assured that $p_{i,m} \in [0, 1] \; \forall \; i, m$.

The final example we consider is to simulate marginal probabilities from a normal distribution with low variance such that $p_{i,m} \sim Normal(\pi_i, \sigma_i^2)$, where $\pi_i$ is the desired $i^{th}$ marginal probability and $\sigma_i^2$ is the desired inter-cluster variation. While this choice of distribution has infinite support, the variance $\sigma_i^2$ can be made small enough to all but ensure values are greater than 0 and less than 1 while simultaneously providing the desired variability. The simulations can also be truncated so that $p = 0.001$ if the simulated value is less than 0 and $p = 0.999$ if it is greater than 1. Using the normal distribution also has the advantage of providing more direct control of the inter-cluster variability in marginal proportions as compared to the $Uniform$ and $Beta$ distributions.

### 2.2.5 Extension to Existing Approaches

For the multivariate normal-based approach, the simulated marginal probabilities $p_{1,m}, \ldots, p_{k,m}$ for clusters $m = 1, \ldots, M$ are matched with the desired dependence levels $\rho_{ij}$ (or $\psi_{ij}$) and are used in Equation 1 separately for each cluster. At this point, the process continues as stated in Section 2.1. Likewise, for the multinomial-based approach,

the simulated marginal probabilities are matched with the desired dependence levels and used to determine the joint pairwise probabilities in Equation 2. Thereafter, the multinomial approach continues as stated in Section 2.3.

## 2.3 Results and Discussion

### 2.3.1 Simulation Study

Here the performance of the multivariate normal and multinomial approaches to simulating dependent binary data with random effects is examined through simulation studies. The first case illustrates the simple situation where we simulate $k = 2$ dependent binary outcomes over $M = 20$ clusters. A second case looks at the situation where we simulate $k = 2$ dependent binary outcomes over $M = 20$ clusters, each consisting of both treatment and control subjects. For each case we assume the correlation between the two outcomes is $\rho_{12} = 0.2$, irrespective of group and cluster. Sample size was fixed at $n = 100$ subjects per cluster. For both cases we also investigate the use of the $Uniform$, $Beta$ and $Normal$ distributions for generating the simulation templates and incorporating the cluster-level variability. A total of 500 data sets were created for each combination of distribution and simulation method, and are used to estimate the average overall marginal probability for each measure, the standard deviation and distribution of those means, the average effect size (and standard deviation) for the case-specific hypothesis test, the empirical power for the case-specific hypothesis test, the mean and distribution of the inter-cluster variability, the mean estimated correlation, and the percentage of data sets for which the desired model converged. SAS (version 9.4, Cary,

NC, USA) was used to simulate data and fit generalized linear mixed models using the IML and GLIMMIX procedures, respectively.

### 2.3.2 Case One: Clustered, One-Group, Repeated-Measure Study

In this case we simulate $k = 2$ binary outcomes over $M = 20$ clusters, where the global marginal probabilities for the two outcomes are $p_1 = 0.25$ and $p_2 = 0.45$, indicating that the rate of our simulated outcome increases by $0.20$ after some time (possibly after an intervention). To incorporate inter-cluster variance we simulate the marginal probabilities according to the specifications listed in Table 2. Here we see that: the midpoints of the two $Uniform$ distributions are $(0.15 + 0.35)/2 = 0.25$ and $(0.34 + 0.55)/2 = 0.45$, respectively; the modes of the two *Beta* distributions are $(11 - 1)/(11 + 31 - 2) = 10/40 = 0.25$ and $(10 - 1)/(10 + 12 - 2) = 9/20 = 0.45$, respectively; and the means of the two normal distributions are $0.25$ and $0.45$; in each case matching the target levels. These values also imply that the inter-cluster variability in the marginal means is $0.0033$ for both measures with the $Uniform$ distribution, $0.0045$ and $0.0178$ for the $Beta$ distribution, and $0.01$ for both measures with the $Normal$ Distribution. The intended model is fit with a fixed two-level "time" effect, a cluster-level random effect to account for the inter-cluster variation, and a subject-level random effect to account for the correlation between the measures. The null hypothesis is no difference over time (i.e. $H_0: p_1 = p_2$) against a two-sided alternative.

Table 2 Simulation Template for Case One

| | Marginal | Distribution | | |
|---|---|---|---|---|
| Time | Mean | Uniform | Beta | Normal |
| 1 | $p_1$ | $U[0.15, 0.35]$ | $Beta(11, 31)$ | $N(0.25, 0.1^2)$ |
| 2 | $p_2$ | $U[0.35, 0.55]$ | $Beta(10,12)$ | $N(0.45, 0.1^2)$ |

The aggregate results over the $500$ simulations for Case One are found in Table 3. Here we see that both the MS and MVN simulation approaches were accurate in reproducing the marginal proportions $p_1$ and $p_2$, as well as the difference $\delta = p_2 - p_1$. We see that the MS and MVN approaches everywhere provided similar estimates and standard errors. The variability of these estimates is low and is also comparable between approaches. The cluster random effects averaged over all simulations are also provided; note these will not necessarily correspond to the theoretical inter-cluster variances stated earlier as these are model-derived and based on linked expectations in the generalized linear mixed model framework. The target correlation ($\rho = 0.2$) was also achieved by both methods, with reasonably small variance. The MS approach produced data sets that converged at least $99.0\%$, while the MVN approach always converged. The empirical powers for testing the null hypothesis of no difference in change over time between the two groups for the both the multivariate normal and multinomial approaches were $> 99.9\%$ for each of the three distributions (not shown in Table 3).

Table 3 Simulation Results for Case One

| Dist. | Approach | $p_1$ (SE) | $p_2$ (SE) | $p_2 - p_1$ (SE) | Cluster Random Effect (SE) | $\hat{\rho}$ (SE) | % Converged |
|---|---|---|---|---|---|---|---|
| Uniform | MS | 0.248 (0.015) | 0.449 (0.015) | 0.201 (0.023) | 0.034 (0.020) | 0.192 (0.021) | 99.0% |
| | MVN | 0.250 (0.016) | 0.449 (0.017) | 0.199 (0.022) | 0.033 (0.018) | 0.191 (0.023) | 100% |
| Beta | MS | 0.257 (0.017) | 0.453 (0.025) | 0.196 (0.030) | 0.075 (0.035) | 0.181 (0.023) | 99.8% |
| | MVN | 0.261 (0.019) | 0.457 (0.026) | 0.196 (0.031) | 0.076 (0.033) | 0.180 (0.022) | 100% |
| Normal | MS | 0.247 (0.026) | 0.447 (0.024) | 0.201 (0.034) | 0.102 (0.044) | 0.170 (0.025) | 100% |
| | MVN | 0.247 (0.025) | 0.450 (0.020) | 0.203 (0.035) | 0.100 (0.045) | 0.170 (0.023) | 100% |

The Beta distribution had more biased results than the other two distributions due to the use of mode being the mean marginal probabilities. Thus, we should use distribution mean as the mean marginal probabilities instead of mode. The inter-cluster variability estimates using the Normal distribution to generate the simulation template were $\hat{\sigma}_{IC}^2 = 0.102$ for the MS approach and $\hat{\sigma}_{IC}^2 = 0.100$ for the MVN approach. If these levels are deemed too large, then the variance assumed in the simulation template (here $\sigma = 0.1$) can be lowered. Likewise, the inter-cluster variability can be increased or decreased using the Uniform distribution by either increasing or decreasing the range about the desired proportions. While the process for the Beta distribution requires solving one equation for two unknowns (such that the given scale and shape parameters provided a desired mode), their sum can be increased or decreased to either decrease or increase the desired intra-cluster variability.

### 2.3.3 Case Two: Clustered, Two-Group, Repeated-Measure Study

In this case we simulate $k = 2$ binary outcomes over $m = 20$ clusters, where subjects in half the clusters belong to a treatment group and where subjects in the other half of the clusters belong to a control group. Assuming an effective treatment, the global marginal probabilities for the two outcomes in the treatment group are $p_{11} = 0.25$ and$p_{12} = 0.45$, while for an ineffective control the global marginal probabilities are $p_{21} = p_{22} = 0.25$; these values indicate that the difference in the changes over "time" is $(p_{12} - p_{11}) - (p_{22} - p_{21}) = 0.20$. To incorporate inter-cluster variance we simulate the marginal probabilities according to the specifications listed in Table 4. As in the previous case, we can easily show that that each distribution obtains the target marginal probability for that Group and time. The inter-cluster variabilities are similar to what was described before. The intended model is fit with a fixed two-level "time" effect, a fixed two-level "group" effect, a group-time interaction, a cluster-level random effect to account for the inter-cluster variation, and a subject-level random effect to account for the correlation between the measures. The null hypothesis is no difference in change over time between the two groups (i.e. $H_0$: $(p_{12} - p_{11}) = (p_{22} - p_{21})$) against a two-sided alternative.

Table 4 Simulation Template for Case Two

| | | Marginal | Distribution | | |
|---|---|---|---|---|---|
| Group | Time | Mean | Uniform | Beta | Normal |
| Treatment | 1 | $p_{11}$ | $U[0.15, 0.35]$ | $Beta(11, 31)$ | $N(0.25, 0.1^2)$ |
| | 2 | $p_{12}$ | $U[0.35, 0.55]$ | $Beta(10, 12)$ | $N(0.45, 0.1^2)$ |
| Control | 1 | $p_{21}$ | $U[0.15, 0.35]$ | $Beta(11,31)$ | $N(0.25, 0.1^2)$ |
| | 2 | $p_{22}$ | $U[0.15, 0.35]$ | $Beta(11,31)$ | $N(0.25, 0.1^2)$ |

The aggregate results over the 500 simulations for Case Two are found in Table 5. Here we see again that both approaches were effective in estimating the marginal means as well as the desired difference in the change in proportions over time ($\delta = 0.2$), and the efficiencies of these estimates were similar for both methods. The estimated inter-cluster variance and the correlation between the repeated measure outcomes were similar between the MS and MVN approaches, while the estimated correlations were also close to the desired level ($\rho = 0.2$). At least 98.8% of the data sets generated by the MS approach allowed models to converge, and at least 99.9% of the MVN-derived data sets allowed model convergence. The empirical powers for the testing the null hypothesis of no difference in change over time between the two groups for the multinomial approach were $>$ 99.9% (Uniform), 99.2% (Beta) and 96.8% (Normal), and were $>$ 99.9% (Uniform), 99.2% (Beta) and 96.6% (No

rmal) for the multivariate normal approach (not shown in Table 5).

Table 5 Simulation Results for Case Two

| | Uniform | | Beta | | Normal | |
|---|---|---|---|---|---|---|
| | MS Approach | MVN Approach | MS Approach | MVN Approach | MS Approach | MVN Approach |
| $p_{11}$ | 0.247 (0.022) | 0.246 (0.023) | 0.260 (0.026) | 0.259 (0.024) | 0.248 (0.036) | 0.246 (0.034) |
| $p_{12}$ | 0.450 (0.024) | 0.452 (0.023) | 0.452 (0.036) | 0.452 (0.037) | 0.449 (0.036) | 0.450 (0.035) |
| $p_{21}$ | 0.248 (0.023) | 0.247 (0.022) | 0.259 (0.025) | 0.260 (0.026) | 0.243 (0.035) | 0.244 (0.035) |
| $p_{22}$ | 0.249 (0.024) | 0.248 (0.023) | 0.261 (0.025) | 0.258 (0.026) | 0.245 (0.034) | 0.246 (0.033) |
| $(p_{12} - p_{11}) - (p_{22} - p_{21})$ | 0.203 (0.044) | 0.204 (0.044) | 0.190 (0.054) | 0.195 (0.058) | 0.199 (0.065) | 0.200 (0.066) |
| *Cluster* * | 0.034 (0.020) | 0.034 (0.020) | 0.060 (0.032) | 0.063 (0.030) | 0.115 (0.055) | 0.114 (0.057) |
| $\hat{\rho}$ | 0.192 (0.023) | 0.192 (0.023) | 0.186 (0.025) | 0.183 (0.024) | 0.169 (0.026) | 0.167 (0.025) |
| % Conv. | 98.8% | 100% | 99.8% | 100% | 100% | 100% |

* *R.E.* stands for random effect

## 2.4 Conclusion

We extended both the multinominal sampling approach and the multivariate normal approaches to simulating dependent binary data to account for desired random effect structures. The extensions for both methods are simple to implement and offer control of marginal probabilities, dependence between outcomes, and intra-cluster variability. Rather than being assigned constant values, the desired marginal probabilities are sampled from specified probability distributions, where a separate simulation template is simulated for each cluster. Simulation studies show that our extension to both approaches yields data that achieve the desired marginal probabilities with relatively low variability, and also exhibits the desired correlation between the binary repeated measures. The parameters for the distributions used in the simulation template can also be adjusted to achieve a desired inter-cluster variability.

One limitation in the presentation of this research is that the simulation templates used here are not exhaustive. In both examples offered we only considered cases of two repeated measures and we presented a limited selection of marginal means and correlations. However, extending this approach to account for more repeated measures or alternative simulation templates is straightforward. We also did not consider more complicated random effect structures, though the underlying principle remains the same: randomly generate a simulation template for each cluster or combination of clusters. This general idea can be applied to other simulation approaches for simulating dependent binary data (e.g., Qaqish [29]), and in principle can be adapted in simulation methodologies for other types of dependent outcomes.

An important statistical role in the preparation of clustered study designs is determining the sample size required to find a desired effect size. While equations or numerical procedures for estimating a required sample size are available for some situations (e.g., Donner, Birkett and Buck [30]), more complicated situations involving repeated measures and intricate clustering may require a simulation-based approach. Simulation templates can be designed to match the desired effect size and clustering structure, and empirical power can be estimated by repeatedly simulating such data. In a similar manner, new statistical methodologies suitable for repeated binary outcomes in clustered settings can be numerically assessed and compared with alternative procedures. Data can be simulated from a desired template and analyzed by the methodologies under consideration, and key features from that analysis (e.g., means, test statistics, confidence intervals, and hypothesis testing decisions) can be aggregated over repeated simulations and compared between competing models.

# Chapter 3 Using Univariate Meta-Analytic Approach for Analyzing Clustered Dependent Binary Data

## 3.1 Introduction

As we discussed in Chapter 1, the complex mixed models for analyzing clustered dependent binary data with large sample sizes may have non-convergence problem. Other than changing random effects into fixed effects or removing particular random effect, or solving the non-convergence problem directly, we figure out a way of "splitting and recombining" to analyze such type of data. That is to split the whole database into small subsets that are analyzable by the desired model and recombine results from subsets by meta-analytic approaches.

We consider two ways of splitting the overall dataset into more manageable subsets. First we split data into k independent, mutually-exclusive sub-samples, what we call the *independent samples approach*, where the desired model is fit on each of the new subsamples. The second proposal is to use natural existing clusters as sub-samples, where simplifications of the desired models – ignoring the clustering measure – are fit to each of the separate clusters. Estimates or test statistics from either of these sub-sampling approaches can then be recombined using random-effect-based meta-analytic approaches

[13]. In the first approach the clustering is taken into account by the model fit to each sub-sample, while in the second approach the clustering is taken into account by the meta-analysis.

The rest of this Chapter is outlined as follows. The sub-sample generating approaches and meta-analytic approach we used are presented in Section 3.2. These approaches are examined through simulation studies in Section 3.3 and an example of data on cancer screening behaviors in Section 3.4. A brief discussion follows in Section 3.5.

## 3.2 Method

### 3.2.1 Approaches for Creating Sub-Samples

We consider two ways of splitting the overall database into smaller components analyzable by desired model, which are called the *independent samples approach* and the *cluster-based approach*.

#### 3.2.1.1 Independent Samples Approach

The k independent, mutually exclusive sub-samples are sampled from the overall database without replacement k-1 times. Each time we sample out a sub-sample containing N/k subjects, where N is the total number of subjects in the overall database, and the remaining dataset will be the database for next sampling. We utilize the SURVEYSELECT procedure in SAS for sampling, which provides methods to select

probability-based random samples [8]. And we use the simple random sampling method with stratification of non-overlapping group and cluster effects. Running SURVEYSELECT procedure k-1 times, we will get the k independent mutually exclusive sub-samples.

Then each sub-sample is fitted by the desired model, which is the same model we applied to the overall database with non-convergence problem. If any of the k independent sub-samples cannot converge, we need to increase k to k+1 to get smaller sub-samples with higher chance to converge. If all k sub-samples can converge, we take one test statistic of the primary research question and one standard error out of each sub-sample, which is used as an observed effect for the meta-analysis, described in Section 3.2.2.

#### 3.2.1.2 Cluster-Based Approach

The cluster-based approach uses the natural existing clusters as sub-samples. If there are nested cluster structures in the dataset, we use the root clusters as sub-samples. For example, if the data has nested cluster design of patients within physicians within practices, practices will be used as sub-samples.

A simplified desired model – ignoring the clustering measure – is fit to each of the separate clusters. Though the model will not account for some amount of clustering, it will not be ignored in this case and is estimated by the meta-analytic approach when combining results from sub-samples, described in the next section.

### 3.2.2 Meta-analytic Approach

The meta-analytic approach we used in this research is the two-step estimate [13] which is based off of the DerSimonian and Laird estimate [20], as described in Chapter 1.4. Given the test statistics obtained from sub-samples, $y_1, \dots, y_k$, (k is the number of sub-samples) and the sampling variances $\sigma_1^2, \dots, \sigma_k^2$, we need to estimate the inter-study (or inter-cluster) variance $\tau^2$ and then estimate the overall test statistic μ and its standard error.

Supposing $s_1^2, \dots, s_k^2$ and $t^2$ are the estimates of $\sigma_1^2, \dots, \sigma_k^2$ and $\tau^2$, a weighted estimator of μ and its standard error (SE) can be expressed as:

$$m_w = \frac{\sum_i w_i y_i}{\sum_i w_i}, \quad \mathrm{SE}(m_w) = \frac{1}{(\sum_i w_i)^{\frac{1}{2}}}, \text{ where } w_i = \frac{1}{t^2 + s_i^2}. \tag{7}$$

Note that the expression for the standard error in Equation (7) is an underestimate of the true standard error of $m_w$ [13]. The DerSimonian and Laird estimate for $\tau^2$ [20] is

$$t^2(DL) = \max\left\{0, \frac{[\sum_i w_{i0}(y_i - y_w(0))^2] - (k-1)}{[\sum_i w_{i0} - \sum_i w_{i0}^2 / \sum_i w_{i0}]}\right\}, \tag{8}$$

where $y_w(0) = \sum_i w_{i0} y_i / \sum_i w_{i0}$, and $w_{i0} = 1/s_i^2$. Substituting $t^2(DL)$ for $t^2$ in Equation (7) yields the corresponding DerSimonian and Laird estimate, $m_w(DL)$ and its standard error $\mathrm{SE}(m_w(DL))$. Then we substitute $w_{i0}$ and $y_w(0)$ with $w_{iD} = 1/(t^2(DL) + s_i^2)$ and $m_w(DL)$ respectively, we will get the two-step estimate, $t^2(DL2)$, where

$$t^2(DL2) = \max\left\{0, \frac{[\sum_i w_{iD}(y_i - m_w(DL))^2] - [\sum_i w_{iD} s_i^2 - \sum_i w_{iD}^2 s_i^2 / \sum_i w_{iD}]}{[\sum_i w_{iD} - \sum_i w_{iD}^2 / \sum_i w_{iD}]}\right\} \tag{9}$$

Substituting $t^2(DL2)$ for $t^2$ in Equation (7) yields the corresponding two-step estimate $m_w(DL2)$ for μ and its approximate standard error, SE($m_w(DL2)$). The hypothesis test we are interested in will be conducted through a Wald test:

$$\frac{m_w(DL2)}{\mathrm{SE}(m_w(DL2))} \sim N(0,1) \text{ under the null hypothesis}$$

### 3.2.3 Hierarchical dependence

In the random-effect meta-analysis model, a traditional assumption is made that sampling error within sub-samples $\varepsilon \sim N(0, \mathbf{V})$ and the inter-cluster variation $\delta \sim N(0, \tau^2\mathbf{I})$ are independent so that

$$\boldsymbol{y} \sim N_k(\mu\mathbf{1}_{\boldsymbol{k}}, \mathbf{V} + \tau^2\mathbf{I}) \quad (10)$$

where $\boldsymbol{y}$ is a vector of *k* test statistics, $\mathbf{V}$ is the sampling variance matrix of $\varepsilon$, which is a diagonal matrix with diagonal elements equal to the squared standard errors of the test statistics, $\mathbf{I}$ is the identity matrix, and $\tau^2$ is the inter-cluster variance. However, an alternative hierarchical dependence structure may be considered [31]. For our *k* independent sub-samples approach, one may question about hierarchical dependence across sub-samples. Although these sub-samples are generated without replacement and are "independent" in the sense that subsets are independent, they may also be viewed as not being independent because they contain information from the same clusters between them. Therefore, we want to examine whether hierarchical dependence exists among the *k* independent sub-samples we created.

The hierarchical dependence between sub-samples can be modeled [31]

$$\boldsymbol{y} \sim N_k(\mu \mathbf{1}_{\boldsymbol{k}}, \mathbf{V} + \tau^2 \mathbf{I} + \eta \mathbf{M}) \quad (11)$$

where $\eta$ is the between-cluster covariance parameter and $\mathbf{M} = \mathbf{J} - \mathbf{I}$ is a matrix with of 0s on the diagonal and 1s on the off-diagonal and other elements are as described in formula (10). Estimation of μ and $\tau^2$, and $\eta$ is an iterative approach based on the method of moments [31] similar to that in DerSimonian and Laird [20]. Starting with $\tau^2$ and $\eta$ initialized to zero, the iterative process is based on evaluating the residual sum of squares (RSS) of model (11) and its expectation. If the estimate of $\eta$ is relatively small in magnitude and estimates of μ and $\tau^2$are similar to those without consideration of dependence across sub-samples, it suggests that we could ignore this hierarchical dependence across sub-samples when using k independent sub-samples based splitting approach.

### 3.3 Simulation Study

The performance of our proposed approach to work around the non-converging issue is examined through a simulation study. Clustered dependent binary data are simulated by using cluster-based templates [32] to simulate dependent binary data based on a multivariate normal approach [22]. We use a normal distribution with small variance to simulate cluster-level proportions for the simulation template. Using these proportions, we simulate dependent binary data using the multivariate normal sampling approach by Emrich and Piedmonte [22]. In this way, the simulated binary data will have cluster effect and repeated-measure structure at the same time.

We simulate 2 binary outcomes over M clusters, where the global marginal probabilities for the two outcomes are $p_1 = \pi_1$ and $p_2 = \pi_2$. To incorporate inter-cluster variability we simulate the marginal probabilities from $p_{i,m} \sim Normal(\pi_i, \sigma_i^2)$ for $m = 1, \dots, M$ clusters and $i = 1,2$ binary outcomes. The desired inter-cluster variation $\sigma_i$ we considered is 0.1. These marginal probabilities for clusters will be truncated so that $p = 0.001$ if the simulated value is less than 0 and $p = 0.999$ if it is greater than 1.

The first case illustrates the simple situation where we simulate 2 dependent binary outcomes (representing outcomes, for instance, measured before and after some intervention) over 20 clusters. The marginal probabilities for the two outcomes are $p_1 = 0.25$ and $p_2 = 0.45$, giving an effect size of 0.2. To incorporate inter-cluster variance we simulate the cluster specific marginal probabilities from $N(0.25, 0.01)$ and $N(0.45, 0.01)$ for the two outcomes respectively. We consider a balanced cluster sample size fixed at $n = 5000$. We also simulate an unbalanced cluster sample size where $n$ is sampled from a normal distribution $N(5000, 100^2)$. The generalized linear mixed model fitted to the $K$ independent sub-samples includes a fixed effect for time and random effects for repeated measures and cluster. For the cluster-based approach, the only difference in the model is without cluster random effect. Link function is logit link since we have binary data.

The second case is an extension of the first case to a more complicated situation with two groups (treatment and control) and two time points. Assuming an effective treatment, the global marginal probabilities for the two treatment group outcomes are $p_{11} = 0.25$

and $p_{12} = 0.45$, while for an ineffective control the global marginal probabilities are $p_{21} = p_{22} = 0.25$; these values indicate that the difference in the change over "time" is $(p_{12} - p_{11}) - (p_{22} - p_{21}) = 0.20$. In this case we also vary the number of clusters between 10, 20 and 40, and also consider other marginal probability values for the second outcome in the treatment group (0.35 and 0.25), which respectively yield overall effect sizes of 0.1 and 0.0. To incorporate inter-cluster variance we simulate the marginal probabilities from normal distribution with mean denoted as previous $\pi_i{}'_s$ and variance 0.01 for all probabilities except for $p_{12}$ with an inter-cluster variance of 0.04. For each case we assume that the correlation between the two outcomes is $\rho_{12} = 0.2$, irrespective of group and cluster. The generalized linear mixed model fitted to the K independent sub-samples includes fixed effects for time, group, and an interaction between time and group, and random effects for repeated measures and cluster. Again, the cluster-based approach does not have cluster random effect. And the link function is logit link.

A total of 1000 data sets were simulated for each case and are used to estimate the average overall inter-cluster variance, effect size (and standard error) for the case-specific hypothesis test, and the empirical power for the case-specific hypothesis test. The hypothesis test for the first case we are interested in is whether there are change in rates from time 1 to time 2, and for the second case is whether there is a difference in the change in rates between two groups from time 1 to time 2. For all cases, we split the overall dataset into $k = 2, \dots, 6$ independent sub-samples for the *k* independent samples approach and use the 20 clusters as sub-samples for the cluster-based approach. The *k* independent samples approach is also examined for existence of hierarchical dependence

for all cases. SAS was used to simulate data, split data, fit generalized linear mixed models, and do meta-analysis through the IML and GLIMMIX procedures [8]. The estimation method used in GLIMMIX procedure is the Maximum Likelihood with Laplace Approximation (MSPL).

### 3.3.1 Simulation Studies

#### 3.3.1.1 Case One: 20 Clusters, Balanced/Unbalanced, One-Group Data

The aggregate results over the 1000 simulations for Case One with/without considering possible dependence between sub-samples are found in Table 6 and Table 7, respectively. All tests correctly rejected the null hypothesis that there are no change in rates from time 1 to time 2, which was due to the large sample sizes we are assuming. The true test statistic of change in rates from time 1 to time 2, which is the parameter for time effect, is 0.9 given the marginal probabilities for the two outcomes are $p_1 = 0.25$ and $p_2 = 0.45$. The estimates of test statistics from both splitting approaches were higher than 0.9, and using natural clusters leaded to higher test statistic and standard error than independent samples approach. Therefore, the independent samples approach worked better than the natural existing cluster approach. The inter-cluster variances estimated by the independent samples approach were relatively small for all K values. The test statistics and standard errors stayed stable without showing a clear trend of change. . Comparing Table 6 and Table 7, we can see that estimates of hierarchical dependence across sub-samples for the independent samples approach were relatively small and all other estimates remained similar. Dependence between sub-samples can be ignored in this case.

Table 6 Simulation results for Case One without considering dependence across samples

| K | Balanced Cluster size | | | | Unbalanced Cluster Size | | | |
|---|---|---|---|---|---|---|---|---|
| | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 |
| 2 | 5.28E-05 | 0.9290 | 0.0105 | 1 | 3.40E-05 | 0.9291 | 0.0096 | 1 |
| 3 | 6.01E-05 | 0.9289 | 0.0098 | 1 | 8.92E-05 | 0.9290 | 0.0102 | 1 |
| 4 | 1.10E-04 | 0.9289 | 0.0102 | 1 | 6.32E-05 | 0.9289 | 0.0096 | 1 |
| 5 | 8.94E-05 | 0.9287 | 0.0096 | 1 | 8.48E-05 | 0.9288 | 0.0098 | 1 |
| 6 | 8.25E-05 | 0.9288 | 0.0097 | 1 | 6.23E-05 | 0.9287 | 0.0095 | 1 |
| cluster(20) | 0.6952 | 1.0081 | 0.1781 | 1 | 0.6893 | 1.0071 | 0.1776 | 1 |

Table 7 Simulation results for Case one after incorporating dependence across sub-samples for the independent samples approach

| K | Balanced Cluster size | | | | | Unbalanced Cluster Size | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\eta$ | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 | $\eta$ | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 |
| 2 | 2.20E-17 | 5.28E-05 | 0.9290 | 0.0105 | 1 | -1.35E-17 | 3.39E-05 | 0.9291 | 0.0096 | 1 |
| 3 | -1.90E-08 | 6.01E-05 | 0.9289 | 0.0098 | 1 | -2.01E-08 | 8.92E-05 | 0.9290 | 0.0102 | 1 |
| 4 | -5.76E-08 | 1.10E-04 | 0.9289 | 0.0102 | 1 | -3.51E-08 | 6.32E-05 | 0.9289 | 0.0096 | 1 |
| 5 | 4.73E-08 | 8.25E-05 | 0.9288 | 0.0097 | 1 | -1.41E-07 | 8.47E-05 | 0.9288 | 0.0097 | 1 |
| 6 | -1.85E-07 | 8.93E-05 | 0.9287 | 0.0096 | 1 | -4.35E-08 | 6.22E-05 | 0.9287 | 0.0095 | 1 |

#### 3.3.1.2 Case Two: 20 Clusters, Balanced/Unbalanced, Two-Group Data

The aggregate results over the 1000 simulations for Case Two including 2 groups with/without considering possible dependence between sub-samples are shown in Table 8 and Table 9, respectively. For the independent samples approach, all tests correctly rejected the null hypothesis that there are no difference in change in rates from time 1 to time 2 between treatment and control groups, while the cluster-based approach only had ~70% power for the corresponding test. Similarly to Case One, using the natural existing clusters leaded to larger standard errors of the requisite test statistic. The true test statistic of difference between the two groups in change in rates from time 1 to time 2, which is the parameter for the interaction term in the model, is 0.9 given the marginal probabilities are $p_{11} = 0.25$ and $p_{12} = 0.45$ for the treatment group, and $p_{21} = p_{22} = 0.25$ for the control group. Test statistic produced by the cluster-based approach had smaller bias than the independent samples approach, given the true test statistic of 0.9. The inter-cluster variances produced by the independent samples approach were all around 3E-04 ~ 4E-04. Again we do not need to consider dependence across sub-samples for *k* independent approach in this case as well since Table 8 and 9 presented similar estimates.

Table 8 Simulation results for Case Two without considering dependence across samples

| K | Balanced Cluster size | | | | Unbalanced Cluster Size | | | |
|---|---|---|---|---|---|---|---|---|
| | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 |
| 2 | 3.39E-04 | 0.9219 | 0.0226 | 1 | 3.62E-04 | 0.9593 | 0.0206 | 1 |
| 3 | 3.91E-04 | 0.9408 | 0.0214 | 1 | 4.33E-04 | 0.9490 | 0.0208 | 1 |
| 4 | 3.53E-04 | 0.9407 | 0.0209 | 1 | 3.55E-04 | 0.9173 | 0.0209 | 1 |
| 5 | 3.69E-04 | 0.9405 | 0.0206 | 1 | 4.54E-04 | 0.9530 | 0.0209 | 1 |
| 6 | 3.54E-04 | 0.9404 | 0.0202 | 1 | 3.54E-04 | 0.9449 | 0.0203 | 1 |
| cluster(20) | 2.6798 | 0.8927 | 0.3534 | 0.66 | 2.5828 | 0.9043 | 0.3490 | 0.69 |

Table 9 Simulation results for Case Two after incorporating dependence across sub-samples for the independent samples approach

| k | Balanced Cluster size | | | | | Unbalanced Cluster Size | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\eta$ | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 | $\eta$ | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 |
| 2 | 9.99E-18 | 3.39E-04 | 0.9219 | 0.0226 | 1 | 3.98E-18 | 3.62E-04 | 0.9593 | 0.0206 | 1 |
| 3 | 6.16E-07 | 3.91E-04 | 0.9408 | 0.0214 | 1 | -2.28E-07 | 4.33E-04 | 0.9490 | 0.0208 | 1 |
| 4 | 4.30E-07 | 3.53E-04 | 0.9407 | 0.0209 | 1 | 8.76E-08 | 3.56E-04 | 0.9173 | 0.0209 | 1 |
| 5 | -1.75E-07 | 3.69E-04 | 0.9405 | 0.0206 | 1 | -7.95E-07 | 4.63E-04 | 0.9530 | 0.0209 | 1 |
| 6 | 4.86E-07 | 3.54E-04 | 0.9404 | 0.0202 | 1 | 1.05E-06 | 3.51E-04 | 0.9463 | 0.0203 | 1 |

### 3.3.1.3 Case Three: Varying Other Study Parameters

We also performed simulations to investigate the effect of other factors which may affect the usefulness of these approaches. Based on the second case of two groups, two time points, 20 clusters, and unbalanced cluster size, we (1) change the number of clusters to 10; (2) change the number of clusters to 40; (3) lower the effect size between the two

treatment group outcomes change to 0.1 ($p_{11} = 0.25, p_{12} = 0.35$); and (4) reduce the effect size for the treatment group entirely by setting $p_{11} = p_{12} = 0.25$.

The aggregate results for 1000 simulated datasets for Case Two with the change of 10 and 40 unbalanced clusters are shown in Table 10 and Table 11, respectively. Compared to the results of Case Two in Table 8, we can see that for the independent samples approach, the standard errors of the hypothesis test decreased from ~0.03 to ~0.01 and the bias in test statistics (the true test statistic was the same with previous case since all marginal probabilities did not change) increased as the number of clusters increases from 10 to 40, although all tests were correctly significant. The cluster-based approach still had larger standard errors and lower bias in test statistic compared to the independent samples approach. However, as the number of clusters increased from 10 to 40, the standard errors decreased from 0.4883 to 0.2500 and the power of the tests increased from 53% to 91%. Therefore, the performance of the cluster-based approach increased as the number of clusters increases (the total sample sizes).

Table 10 Simulation results for Case Two with 10 unbalanced clusters

| | not consider dependence across sub-samples | | | | consider dependence across sub-samples | | | | |
|---|---|---|---|---|---|---|---|---|---|
| K | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 | $\eta$ | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 |
| 2 | 4.83E-04 | 0.9244 | 0.0300 | 1 | 2.62E-17 | 4.83E-04 | 0.9244 | 0.0300 | 1 |
| 3 | 6.46E-04 | 0.9264 | 0.0299 | 1 | -5.57E-07 | 6.45E-04 | 0.9264 | 0.0299 | 1 |
| 4 | 7.81E-04 | 0.9338 | 0.0296 | 1 | 5.82E-07 | 7.82E-04 | 0.9338 | 0.0296 | 1 |
| 5 | 1.13E-03 | 0.9379 | 0.0301 | 1 | 1.04E-06 | 1.13E-03 | 0.9379 | 0.0302 | 1 |
| 6 | 9.40E-04 | 0.9351 | 0.0291 | 1 | -3.00E-06 | 9.29E-04 | 0.9373 | 0.0291 | 1 |
| cluster (10) | 2.7215 | 0.8977 | 0.4883 | 0.53 | | | | | |

Table 11 Simulation results for Case Two with 40 unbalanced clusters

| | not consider dependence across sub-samples | | | | consider dependence across sub-samples | | | | |
|---|---|---|---|---|---|---|---|---|---|
| K | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 | $\eta$ | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 |
| 2 | 1.25E-04 | 0.9323 | 0.0150 | 1 | -4.82E-18 | 1.25E-04 | 0.9323 | 0.0150 | 1 |
| 3 | 1.26E-04 | 0.9399 | 0.0146 | 1 | 2.18E-07 | 1.26E-04 | 0.9399 | 0.0146 | 1 |
| 4 | 2.19E-04 | 0.9765 | 0.0151 | 1 | 8.41E-08 | 2.19E-04 | 0.9764 | 0.0151 | 1 |
| 5 | 1.86E-04 | 0.9414 | 0.0145 | 1 | -2.86E-07 | 1.09E-03 | 0.9414 | 0.0145 | 1 |
| 6 | 1.66E-04 | 0.9525 | 0.0142 | 1 | 6.57E-07 | 4.90E-04 | 0.9528 | 0.0152 | 1 |
| cluster (40) | 2.5989 | 0.9050 | 0.2500 | 0.91 | | | | | |

Table 12 and Table 13 show the aggregate results over 1000 simulation datasets for Case Two with the change of global marginal probability for treatment group outcome of time 2 to $p_{12} = 0.35$ and $p_{12} = 0.25$, respectively. Although we reduce the effect size in the treatment group to $p_{12} - p_{11} = 0.35 - 0.25 = 0.1$, the power of the corresponding tests was still high (about 96%) for the independent samples approach, while the power for the cluster-based approach decreased to 21%. The true test statistic is 0.48 in this case given the desired marginal probabilities for the two groups and time points. The independent samples approach had lower bias in test statistics and smaller standard errors than the cluster-based approach. And for ineffective treatment group which is the case shown in Table 13, the empirical error rate for the cluster-based approach was 22%, while the error rate was below 10% for the independent samples approach. The true test statistic is 0 given all the desired marginal probabilities are 0.25. Same with the less effective treatment case, the independent samples approach had smaller bias in test statistics and smaller standard errors than the cluster-based approach. Therefore, in the situation of

small number of clusters and small effect size, the independent samples approach worked better than the cluster-based approach.

Table 12 Simulation results for Case Two with treatment effect size of 0.1

| | not consider dependence across sub-samples | | | | consider dependence across sub-samples | | | | |
|---|---|---|---|---|---|---|---|---|---|
| k | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 | $\eta$ | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 |
| 2 | 2.17E-04 | 0.5386 | 0.0209 | 0.97 | 3.03E-18 | 2.17E-04 | 0.5386 | 0.0209 | 0.97 |
| 3 | 4.77E-04 | 0.5165 | 0.0222 | 0.95 | 2.22E-08 | 4.69E-04 | 0.5156 | 0.0221 | 0.95 |
| 4 | 4.95E-04 | 0.5299 | 0.0217 | 0.96 | 2.01E-07 | 4.95E-04 | 0.5299 | 0.0217 | 0.96 |
| 5 | 3.73E-04 | 0.5129 | 0.0207 | 0.97 | -7.72E-07 | 3.72E-04 | 0.5129 | 0.0207 | 0.97 |
| 6 | 3.14E-04 | 0.5193 | 0.0202 | 0.97 | 6.84E-07 | 3.42E-04 | 0.5211 | 0.0203 | 0.97 |
| cluster (20) | 3.2686 | 0.2597 | 0.3937 | 0.21 | | | | | |

Table 13 Simulation results for Case Two with treatment effect size of 0

| | not consider dependence across sub-samples | | | | consider dependence across sub-samples | | | | |
|---|---|---|---|---|---|---|---|---|---|
| k | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 | $\eta$ | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 |
| 2 | 2.42E-04 | 0.0856 | 0.0217 | 0.03 | -5.28E-19 | 2.42E-04 | 0.0856 | 0.0217 | 0.0 |
| 3 | 4.27E-04 | 0.0916 | 0.0221 | 0.08 | -4.48E-07 | 4.27E-04 | 0.0916 | 0.0221 | 0.08 |
| 4 | 5.46E-04 | 0.0452 | 0.0222 | 0.07 | -3.02E-07 | 5.46E-04 | 0.0452 | 0.0222 | 0.07 |
| 5 | 3.52E-04 | 0.0716 | 0.0209 | 0.08 | -7.64E-07 | 3.51E-04 | 0.0716 | 0.0209 | 0.08 |
| 6 | 2.95E-04 | 0.0660 | 0.0204 | 0.03 | -9.27E-06 | 7.27E-04 | 0.0682 | 0.0211 | 0.09 |
| cluster (20) | 4.2935 | -0.4824 | 0.4580 | 0.22 | | | | | |

### 3.4 Analysis of Cancer Screening Data

We also examined our approach on the motivating example data. As described in Section 1.2, data are from a study of an electronic medical record software program (My Preventive Care, or MPC) by Krist *et al* [1]. The MPC program was implemented at the practice level, and patients decided whether or not they wanted to register for and use the portal. We were interested in determine whether MPC users were more likely to change their cancer screening behavior as compared to MPC non-users. Using colon cancer screening as an example, the outcome was a patient's status with respect to whether or not they had received any form of colon cancer screening, measured at the most recent visit preceding the implementation of MPC into their practice and then updated one month, three months, and six months following that implementation.

A hierarchical linear mixed model was fit including the repeated-measure binary outcome (patient underwent colon cancer screening or they did not), fixed effects for time (four levels: baseline, one, three and six months), group (registered for MPC or not), and their interaction, as well as random effects for primary care physicians, who are in turn nested with one of eight practices. It was this three-level nested design (patients within physicians within practices) that was desired, and was fit using the GLIMMIX procedure in SAS to test hypotheses of whether there were differences in change of colon cancer screening rates between MPC program users and non-users from baseline to each of one, three and six months. Fitting the data to all n=110,029 subjects, the desired model did not converge. We chose a subset of the study with n=15,652 subjects where the desired model did converge as the overall database in this example. We split the overall database

into $k = 2, \ldots, 8$ independent sub-samples using the *k* independent samples approach and used the 8 practices as sub-samples for the cluster-based approach.

The original test results of differences of change in rates of colon cancer screening between users and non-users from baseline to month 1, 3, and 6 using the whole dataset are found in Table 14. Here we see that all the changes were significantly different between users and non-users using the whole dataset.

Table 14 The tests of differences of change in rates of colon cancer screening between users and non-users from baseline to month 1, 3, and 6 using the whole dataset

| User/Non-User | Test Statistic | SE | p-value |
|---|---|---|---|
| Baseline to Month1 | 0.1741 | 0.0317 | <.0001* |
| Baseline to Month3 | 0.1698 | 0.0429 | 0.0001* |
| Baseline to Month6 | 0.1235 | 0.0504 | 0.0130* |

* Z-statistic is significant at level $\alpha = 0.05$ (Here use z-statistic instead of t-statistic given in SAS output because we want to make it comparable to the results for sub-samples)

Table 15 shows the results of the same tests but applied to the sub-samples we created and recombined by meta-analysis. Estimates of inter-cluster variances, test statistics, standard errors and p-values shown in Table 15 are also plotted in Figure 1 through Figure 3. The test results of the whole dataset in Table 16 are also plotted at $k = 1$. Compared to results of the whole dataset shown in Table 16, we can see that $k = 2$ gave the most similar estimates, standard errors and p-values to those from the whole dataset. All test statistics for change from baseline to month 1 and month 3 were significant for the independent samples approach. And for change from baseline to month 6, $k =$ 4 and 8 yielded non-significant results. From Figures 1 through 3, we can see that smaller

k gave more similar results to the whole dataset and there was a trend of decreasing in test statistics, and a trend of increasing in both inter-cluster variance and standard error as $k$ increases. For the natural existing clusters approach (here based on the study practices), smaller test statistics and larger standard errors leaded to non-significant results which were different from those found using the whole dataset. Therefore, the natural existing cluster-based approach did not perform as well as the independent samples approach, and the smallest possible number of independent sub-samples, which was 2 in this situation, worked the best among all choices. The results were also consistent with the results in our simulation studies with small number of clusters and small treatment effect size,

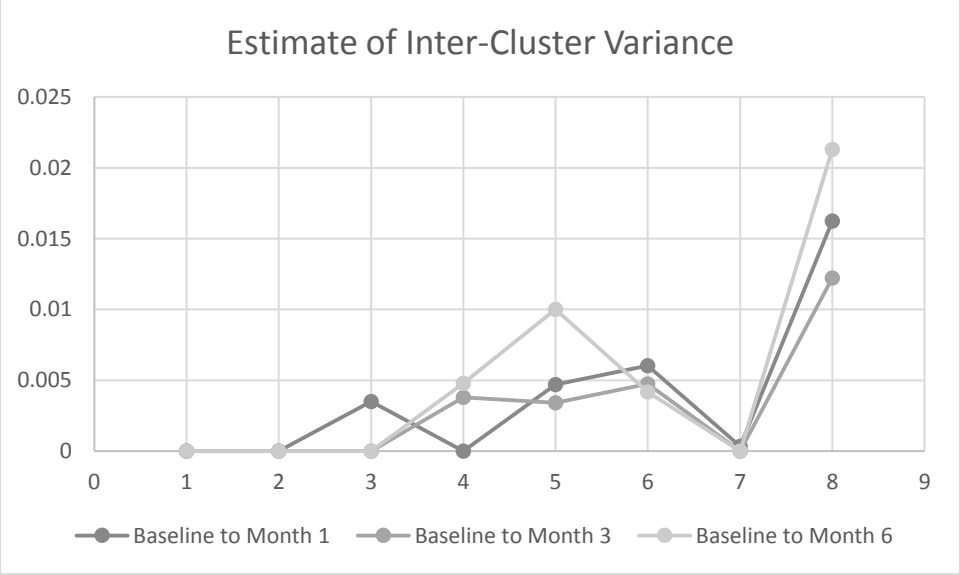
Table 15 Estimates of inter-cluster variance and the overall differences of change between users and non-users using sub-samples by univariate meta-analytic approach

| User/Non-User | k=2 | | | | k=3 | | | |
|---|---|---|---|---|---|---|---|---|
| | $\tau^2$ | Test Statistic | SE | p-value | $\tau^2$ | Test Statistic | SE | p-value |
| baseline to month1 | 0 | 0.1743 | 0.0315 | <.0001* | 0.0035 | 0.1717 | 0.0465 | 0.0002* |
| baseline to month3 | 0 | 0.1689 | 0.0425 | 0.0001* | 0 | 0.169 | 0.0426 | 0.0001* |
| baseline to month6 | 0 | 0.1232 | 0.05 | 0.0137* | 0 | 0.1206 | 0.0501 | 0.0161* |
| User/Non-User | k=4 | | | | k=5 | | | |
| | $\tau^2$ | Test Statistic | SE | p-value | $\tau^2$ | Test Statistic | SE | p-value |
| baseline to month1 | 0 | 0.174 | 0.0312 | <.0001* | 0.0047 | 0.2051 | 0.0435 | <.0001* |
| baseline to month3 | 0.0038 | 0.1686 | 0.0523 | 0.0013* | 0.0034 | 0.2065 | 0.0489 | <.0002* |
| baseline to month6 | 0.0048 | 0.1224 | 0.0605 | 0.0711 | 0.0100 | 0.1583 | 0.0662 | 0.0168* |
| User/Non-User | k=6 | | | | k=7 | | | |
| | $\tau^2$ | Test Statistic | SE | p-value | $\tau^2$ | Test Statistic | SE | p-value |
| baseline to month1 | 0.0060 | 0.1785 | 0.0434 | <.0001* | 0.0004 | 0.1787 | 0.0298 | <.0001* |
| baseline to month3 | 0.0048 | 0.1740 | 0.0489 | 0.0004* | 0 | 0.1658 | 0.0390 | <.0001* |
| baseline to month6 | 0.0042 | 0.1225 | 0.0536 | 0.0222* | 0 | 0.0920 | 0.0456 | 0.0437* |

| User/Non-User | k=8 | | | | 8 practices | | | |
|---|---|---|---|---|---|---|---|---|
| | $\tau^2$ | Test Statistic | SE | p-value | $\tau^2$ | Test Statistic | SE | p-value |
| baseline to month1 | 0.0162 | 0.1793 | 0.0542 | 0.0009* | 0.0203 | 0.0842 | 0.0619 | 0.1737 |
| baseline to month3 | 0.0122 | 0.1635 | 0.0559 | 0.0035* | 0.0138 | 0.0976 | 0.0627 | 0.1197 |
| baseline to month6 | 0.0213 | 0.1217 | 0.0701 | 0.0823 | 0.0064 | 0.1073 | 0.0604 | 0.0758 |

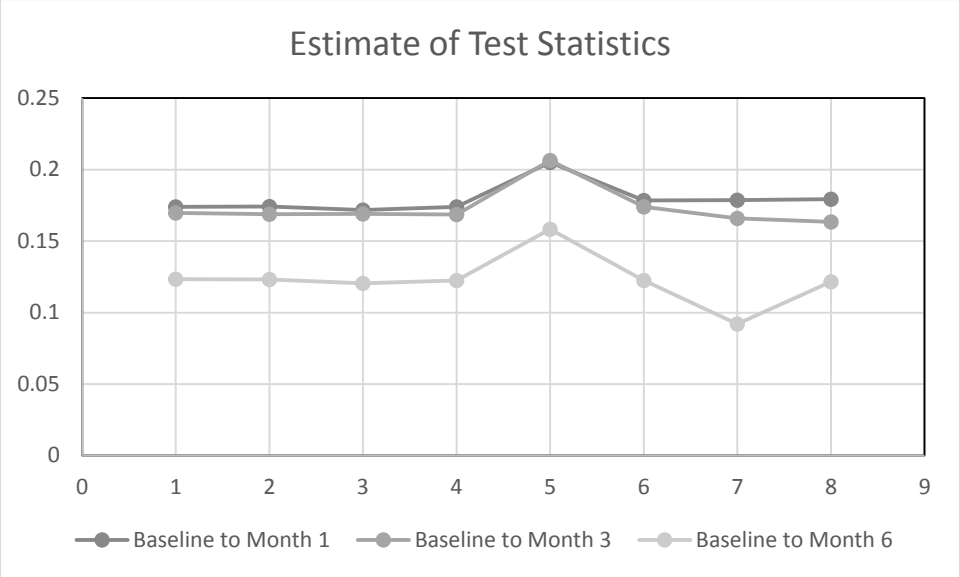
* Z-statistic is significant at level $\alpha = 0.05$

Figure 1 Plot of inter-cluster variances for the whole dataset* and sub-samples created by the k independent samples approach



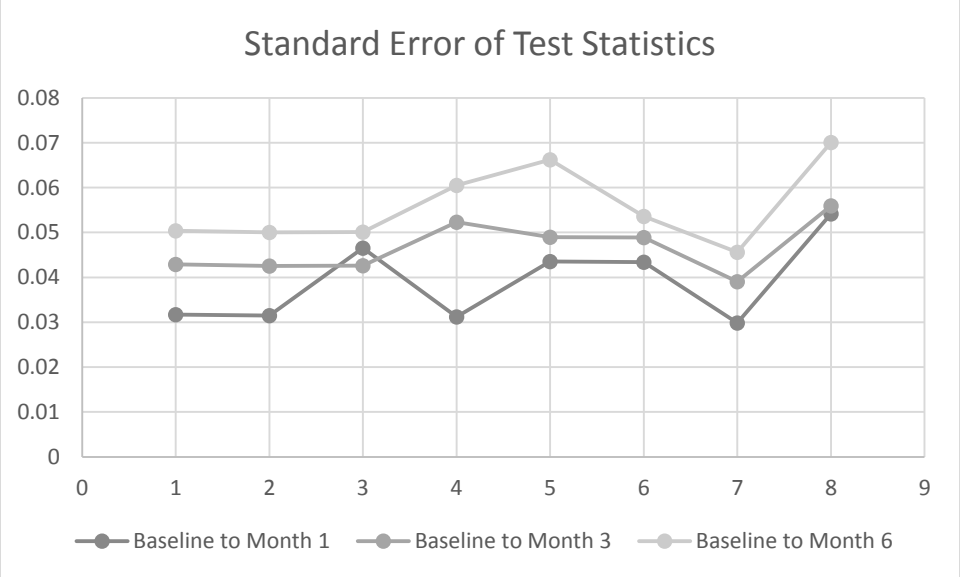

*k = 1 corresponds to the whole dataset and inter-cluster variance for the whole dataset is 0.

Figure 2 Plot of estimates of test statistics for the whole dataset* and sub-samples created by the k independent samples approach




*k = 1 corresponds to the whole dataset

Figure 3 Plot of standard errors of test statistics for the whole dataset* and sub-samples created by the k independent samples approach




*k = 1 corresponds to the whole dataset

We also present results after incorporating dependence between sub-samples for the independent samples approach in Table 16. The estimates of dependence across sub-samples $\eta$ were small and the test statistics and standard errors were stable compared to those without considering dependence across samples for small k values. The test results comparing change from baseline to month 6 did begin to change unpredictably for larger values of k, beginning with k=5, suggesting the dependence estimates become more unstable for smaller cluster sizes and may adversely affect the results. For small *k* values, there was no need to consider dependence between sub-samples. From the results listed in Tables 14 through 16, we conclude that smallest *k* works best for our motivating example, which supports the conclusion of our simulation studies.

Table 16 Test results after incorporating dependence across sub-samples for the independent samples approach

| User/Non-User | k=2 | | | | |
|---|---|---|---|---|---|
| | $\eta$ | $\tau^2$ | Test Statistic | SE | p-value |
| baseline to month1 | 0 | 0 | 0.1743 | 0.0314 | <.0001* |
| baseline to month3 | 0 | 0 | 0.1689 | 0.0425 | 0.0001* |
| baseline to month6 | 0 | 0 | 0.1232 | 0.05 | 0.0137* |
| User/Non-User | k=3 | | | | |
| | $\eta$ | $\tau^2$ | Test Statistic | SE | p-value |
| baseline to month1 | -0.0002 | 0.0033 | 0.1717 | 0.0447 | 0.0001* |
| baseline to month3 | 0.0001 | 4.70E-05 | 0.169 | 0.0432 | 0.0001* |
| baseline to month6 | 0.0055 | 0.0055 | 0.1206 | 0.0893 | 0.0177* |
| User/Non-User | k=4 | | | | |
| | $\eta$ | $\tau^2$ | Test Statistic | SE | p-value |
| baseline to month1 | 0.0001 | 0.0001 | 0.1741 | 0.0331 | <.0001* |
| baseline to month3 | 0.0002 | 0.0004 | 0.1686 | 0.0545 | 0.0020* |
| baseline to month6 | 0.0003 | 0.005 | 0.1224 | 0.0626 | 0.0507 |

| User/Non-User | | | k=5 | | |
|---|---|---|---|---|---|
| | $\eta$ | $\tau^2$ | Test Statistic | SE | p-value |
| baseline to month1 | 0.0003 | 0.0050 | 0.2051 | 0.0471 | <.0001* |
| baseline to month3 | 0.0004 | 0.0039 | 0.2066 | 0.0530 | <.0001* |
| baseline to month6 | -0.0008 | 0.0093 | 0.1583 | 0.0601 | 0.0084* |
| User/Non-User | | | k=6 | | |
| | $\eta$ | $\tau^2$ | Test Statistic | SE | p-value |
| baseline to month1 | 0.0000 | 0.0061 | 0.1785 | 0.0439 | <.0001* |
| baseline to month3 | -0.0004 | 0.0044 | 0.1740 | 0.0447 | <.0001* |
| baseline to month6 | -0.0004 | 0.0038 | 0.1226 | 0.0495 | 0.0132* |
| User/Non-User | | | k=7 | | |
| | $\eta$ | $\tau^2$ | Test Statistic | SE | p-value |
| baseline to month1 | 2.71E-06 | 0.0004 | 0.1787 | 0.0299 | <.0001* |
| baseline to month3 | 0 | 0 | 0.1658 | 0.0390 | <.0001* |
| baseline to month6 | 0.0011 | 0.0011 | 0.0920 | 0.0566 | 0.1039 |
| User/Non-User | | | k=8 | | |
| | $\eta$ | $\tau^2$ | Test Statistic | SE | p-value |
| baseline to month1 | -0.0003 | 0.0159 | 0.1793 | 0.0514 | 0.0005* |
| baseline to month3 | -3.39E-06 | 0.0122 | 0.1635 | 0.0559 | 0.0034* |
| baseline to month6 | -0.0011 | 0.0202 | 0.1217 | 0.0617 | 0.0485* |

## 3.5 Discussion

We offered an alternative approach to obtaining estimates when a desired model of clustered and dependent binary data does not converge. The idea is to split data into subsets that are analyzable by the desired model, and recombine results from subsets by meta-analytic approach. The two splitting methods we discussed in this paper are creating *k* independent sub-samples and using the natural existing clusters as sub-samples. Both

results from simulation studies and data on cancer screening show that we should avoid using natural existing clusters if we could use the *k* independent samples approach when the number of clusters and the treatment effect size are small. The smallest possible number of independent sub-samples (that is each sub-sample has converging result by the desired model) works best for the independent samples approach.. However, we suggest using the independent samples approach when the number of clusters and the treatment effect size are large due to the small bias in test statistics and high power of tests. Dependence between sub-samples by the *k* independent samples approach can be ignored.

One limitation in the presentation of this research is that the number of *k* independent sub-samples used here is not exhaustive. We only considered the number of independent samples up to 8. However, extending this approach to larger number of independent samples is straightforward. We have shown that the smallest possible number of independent sub-samples works best if the independent samples approach is preferred, therefore increasing the number of k should not alter our conclusion. Another limitation is that we only consider the meta-analytic approach by DerSimonian and Laird, which can be considered as a univariate meta-analysis. Rather than combining all sub-samples' results into one estimate and one standard error, there is also multivariate meta-analytic approach [21] combining sub-samples' results into multiple estimates and standard errors. The multivariate meta-analysis may work better than univariate meta-analysis since it uses more information from each of the sub-samples, though there may be a trade-off due to increased complexity of the meta-analytic model. However, this multivariate meta-

analysis may only be applicable to the natural existing cluster approach because the sample size of multivariate meta-analysis (*k*) will be too small to give valid inferences for the *k* independent samples approach. We will compare the results of univariate and multivariate meta-analytic approaches in Chapter 4.

Another limitation of this research is that we did not consider repeating the independent sample process by creating different sub-samples sets and re-estimating the meta-analysis outcomes for the independent samples approach. This in effect would create a process where the meta-analytic results would be re-estimated for each bootstrap sample, which can then be appropriately summarized. We didn't consider the bootstrap process due to long computational time. Since creating one independent sub-samples set worked well at providing stable estimates and low inter-cluster variance, we did not consider repeating the independent sample process in this research. We will evaluate the performance of repeating process in our future work.

# Chapter 4 Using Multivariate Meta-Analytic Approach for Analyzing Clustered Dependent Binary Data

## 4.1 Introduction

In our previous Chapter, we have shown that splitting the overall database into smaller components by either the *K* independent samples approach or the cluster-based approach and recombining results using univariate meta-analysis approach, can allow us to fit a desired model for clustered dependent binary data when that model won't converge on the entire dataset. In this approach the univariate meta-analysis was calculated on the single test statistic (and standard error) of interest. However, we can also try a multivariate meta-analytic approach incorporating multiple test statistics and covariance estimates from each sub-sample to combine into a vector of test statistics and standard errors. Through multivariate meta-analysis, we could base inferences based on more information and thus achieve more accurate results, though there may be a trade-off based the increased number of parameters. We want to determine whether multivariate meta-analytic approach could be incorporated in our approach and analyze how it performs when compared to univariate meta-analytic approach.

Multivariate meta-analysis is the synthesis analysis of estimates of several related parameters over several studies [33]. These related parameters often refer to multiple

outcomes within one study or to multiple comparisons between more than two groups [34]. For our purposes, since we are focusing on the clustered dependent binary measurements of a common outcome across treatment groups, we could parse at information from an overall test statistic by focusing on change with groups. These group-level estimates, along with the corresponding covariance structure, could then be captured within each sub-sample, and then combined using multivariate meta-analysis.

To work around the non-convergence issue of clustered dependent binary data, we are splitting the overall database by the same approaches mentioned in chapter 3, which are the *K* independent samples approach and clustered based approach. We then estimate group-specific statistics and their covariance, and recombine this information from the random-effect-based multivariate meta-analytic approach [21]. The rest of this chapter is outlined as follows. Section 4.2 presents the sub-sample-generating approaches and the multivariate meta-analytic approach we used in this research. These approaches are examined in Section 4.3 and Section 4.4 through the same simulation studies and cancer screening behaviors data used in Chapter 3, respectively. A brief discussion follows in Section 4.5.

## 4.2 Method

### 4.2.1 Splitting Approaches

We utilize the same sub-sample generating approaches presented in Chapter 3 to split the overall database into smaller subsets. They are the *K* independent samples approach and the cluster-based approach.

#### 4.2.1.1 *K* Independent Samples Approach

The k independent, mutually exclusive sub-samples are created by a process of sampling without replacement for k-1 times with the kth sub-sample consisting of the remaining N/k subjects, where N is the total number of subjects. Each time we select a sub-sample containing N/k subjects without replacement, and the remaining dataset will constitute the subject pool for the next sampling.

Sub-samples are generated through the SURVEYSELECT procedure in SAS, which provides methods to select probability-based random samples [8]. We use the simple random sampling method with stratification of non-overlapping group and cluster effects. Running SURVEYSELECT procedure k-1 times, we obtain the K independent mutually exclusive sub-samples.

Then we fit the desired model on each sub-sample. The desired model is the same model applied to the overall database that originally featured the non-convergence problem,

including all fixed and clustering random effects, and also accounting for repeated measurements. If any of the K independent sub-samples cannot converge, we need to increase K to k+1 to get smaller sub-samples with higher chance to converge. If all K sub-samples converge, we take the estimators of test statistics and standard errors out of each sub-sample, which are used as observed effects for the meta-analysis. Since we are focusing on binary outcomes with repeated measures and two treatment groups in this chapter, the estimators we obtain from each sub-sample are the group-specific test statistic and standard error of the change in rates over time on the log-odds scale.

#### 4.2.1.2 Cluster-Based Approach

The cluster-based approach splits the overall dataset based on the natural existing clusters. Each sub-sample then contains subjects from the same cluster. If there are nested cluster structures in the dataset (e.g., patients within physicians within practices), we use the root clusters as sub-samples (e.g., practices).

The model fit to each sub-sample/cluster is the desired model without the random effect of the clustering measure used to split the dataset. Although this clustering effect is ignored temporally during this phase, it will be estimated by the meta-analytic approach described in next section. The estimators we are taking out of each sub-sample are the same with those from the *K* independent samples approach, which are the group-specific test statistic and standard error of change in rates over time.

### 4.2.2 Multivariate Meta-Analytic Approach

#### 4.2.2.1 Multivariate Meta-Analysis

We utilize the random-effect based multivariate meta-analytic approach [21] to recombine estimates from our sub-samples. Suppose we have $k$ studies and $m$ effect size estimators from each study. The observed effect size estimators are $\boldsymbol{y_1}, \dots, \boldsymbol{y_k}$, and their corresponding variances are $\boldsymbol{s_1^2}, \dots, \boldsymbol{s_k^2}$, where $\boldsymbol{y_i}$ and $\boldsymbol{s_i^2}$ are both $m \times 1$ vectors for study $i = 1, \dots, k$.

The random-effect model for the multivariate meta-analysis [21] is:

$$\boldsymbol{y_i} \sim N_m(\boldsymbol{\mu}, V_i + \Sigma),$$

where $\boldsymbol{\mu}$ is a $m \times 1$ vector representing the overall effect size, $V_i$ is the diagonal matrix with diagonal elements equal to the vector of $\boldsymbol{s_i^2}$ representing the within study variances, $\Sigma$ is the between study covariance matrix of the observed effect. Given estimators of $\boldsymbol{y_i}$ and $V_i$, we need only estimate $\boldsymbol{\mu}$ and $\Sigma$. Maximum likelihood estimation for this model can be carried out by linear mixed-effect models. Since the sample size of this model formulation is the number of sub-samples generated, this model may not be appropriate when the number of sub-samples is small.

#### 4.2.2.2 Multivariate Meta-Analysis applied in our case

We are focusing on clustered dependent binary data in this research. Suppose we have two repeated (time 1, time 2) binary outcomes with two treatment groups (treatment, control). The hypothesis test we are interested in is whether there are differences between

the two groups in change of outcome rates from time 1 to time 2. Rather than taking the overall test statistic and standard error of this hypothesis test (as done in the previous Chapter), here we take the two test statistics and corresponding standard errors of the change in rates on the log-odds scale from time 1 to time 2 for the treatment group and the control group separately. Let $y_{T,i}$ denote the test statistic of change of outcome rates from time 1 and time 2 for treatment group of sub-sample $i$, and $y_{C,i}$ denote the same test statistic for control group of sub-sample $i, i = 1, \dots, k$. The corresponding standard errors for $y_{T,i}$ and $y_{C,i}$ are $s_{T,i}$ and $s_{C,i}$. The resulting random-effect model for multivariate meta-analysis in our case is:

$$\begin{pmatrix} y_{T,i} \\ y_{C,i} \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_T \\ \mu_C \end{pmatrix}, V_i + \Sigma\right), \text{ with } V_i = \begin{pmatrix} s_{T,i}^2 & 0 \\ 0 & s_{C,i}^2 \end{pmatrix}$$

where $\mu_T$ and $\mu_C$ are the overall test statistics for treatment group and control group, and $\Sigma$ is the between-sub-sample covariance matrix. $y_{T,i}$, $y_{C,i}$, $s_{T,i}$, and $s_{C,i}$ are the estimators taken from results of sub-sample $i$ fitted by the desired model without considering the root cluster random effect. Given the $k$ sets of ($y_{T,i}$, $y_{C,i}$, $s_{T,i}$, $s_{C,i}$), we need to estimate $\mu_T$, $\mu_C$, and $\Sigma$. Maximum likelihood estimation for this mixed model can be carried out by the MIXED procedure in SAS. However, the number of sub-samples $k$, which is the sample size in this analysis, should be large enough to get appropriate estimators.

The differences between the two groups in change of outcome rates from time 1 to time 2 can then be estimated through conducting a hypothesis test of the difference between $\mu_T$

and $\mu_C$ after the model is fitted. Let $\widehat{\boldsymbol{\mu}} = (\widehat{\mu_T}, \widehat{\mu_C})^T$ represent the vector of estimated overall test statistic on log-odds scale for treatment group and control group and $\widehat{\Phi} = \begin{pmatrix} \text{var}(\widehat{\mu_T}) & \text{cov}(\widehat{\mu_T}, \widehat{\mu_C}) \\ \text{cov}(\widehat{\mu_T}, \widehat{\mu_C}) & \text{var}(\widehat{\mu_T}) \end{pmatrix}$ represent the estimated covariance matrix for $\widehat{\boldsymbol{\mu}}$. The null hypothesis is that there is no difference between the two groups in change of log-odds ratios from time 1 to time 2 ($H_0$: $\boldsymbol{L}^T\boldsymbol{\mu} = [1 \quad -1][\mu_T \quad \mu_C]^T = \mu_T - \mu_C = 0$). We can conduct a Wald hypothesis of $\boldsymbol{L}^T\widehat{\boldsymbol{\mu}}$:

$$\frac{\boldsymbol{L}^T\widehat{\boldsymbol{\mu}}}{(\boldsymbol{L}^T\widehat{\boldsymbol{\Phi}}\boldsymbol{L})^{1/2}} \sim N(0,1) \text{ under } H_0$$

The test statistic $\boldsymbol{L}^T\widehat{\boldsymbol{\mu}}$ and its standard error $\left(\boldsymbol{L}^T\widehat{\Phi}\boldsymbol{L}\right)^{1/2}$ will be comparable to those of the same test generated by the univariate meta-analytic approach shown in Chapter 3.

And $\boldsymbol{L}^T\hat{\Sigma}\boldsymbol{L}$, where $\hat{\Sigma}$ is the estimated between-sub-sample variance matrix, represents the inter-study variance and is also comparable to inter-cluster variance estimator $\hat{\tau}^2$ produced by the univariate meta-analytic approach.

### 4.3 Simulation Study

The performance of using the multivariate meta-analytic approach is examined using the same parameter templates from the simulation studies conducted in Chapter 3, i.e. using the same 1000 simulation data sets and the same subsets generated by the *K* independent samples approach and the cluster-based approach. However, small k values from the *K* independent samples approach do not work for multivariate meta-analysis. The number of sub-samples k we considered for simulation studies in this Chapter are 5 and 6, excluding 2, 3, and 4 which are presented in Chapter 3 by univariate meta-analysis. Due

to poor performance in the previous Chapter, we do not consider dependence across sub-samples for the *K* independent approach.

The first case is the simple situation with one group and two time points over 20 clusters. The marginal probabilities for the two outcomes are $p_{11} = 0.25$ and $p_{12} = 0.45$, giving an effect size of 0.2. The cluster sample size has a balanced case where n is fixed at 5000 and also an unbalanced case where n is sampled from a normal distribution $N(5000, 100^2)$. The hypothesis test for the first case we are interested in is whether there are change in rates over time. Since there is only one group in this case, the time-specific mean log-odds ratios will be used as the input test statistic for the multivariate meta-analysis. The mixed linear model fitted to sub-samples remain the same as described in Chapter 3, including time as fixed effect, repeated measures and clusters as random effects (for cluster-based approach there will not be cluster random effect). We use logit link function because of binary data.

The second case extends the first case with two groups (treatment and control) and two time points over 20 clusters. Again the cluster sample size has a balanced case where n is fixed at 5000 and also an unbalanced case where n is sampled from a normal distribution $N(5000, 100^2)$. The global marginal probabilities are $p_{11} = 0.25$ and $p_{12} = 0.45$ for the two treatment outcomes and are $p_{21} = p_{22} = 0.25$ for the two control outcomes, indicating the difference between the two groups in the change over "time is $(p_{12} - p_{11}) - (p_{22} - p_{21}) = 0.20$. In this case we also vary the number of clusters

between 10, 20 and 40, and also consider other marginal probability values for the second outcome in the treatment group (0.35 and 0.25), which respectively yield overall effect sizes of 0.1 and 0.0. The hypothesis test for this two-group case is whether there is a difference in the change in rates between the two groups over time. Therefore, the group-specific log-odds ratios of change over time will be used as the input test statistic for the multivariate meta-analysis. The mixed linear model fitted to each sub-sample is the same with the first case except for including fixed effects for group and an interaction between time and group. Link function is the same as well.

### 4.3.1 Simulation Studies

#### 4.3.1.1 Case One: 20 Clusters, Balanced/Unbalanced, One-Group Data

The aggregate results over the 1000 simulations for Case One with balanced and unbalanced clusters using multivariate meta-analytic approach are shown in Table 17 and Table 18, respectively. The results using univariate meta-analytic approach shown in Chapter 3 are also provided for a comparison purpose (the univariate results are provided alongside all the multivariate results in Section 3.1). Here $\sigma_{11}$, $\sigma_{12}$, and $\sigma_{22}$ are the covariance parameters for the between sub-sample covariance matrix $\Sigma = \begin{bmatrix}\sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22}\end{bmatrix}$. The inter-cluster variance $\tau^2$ for the multivariate meta-analytic approach is calculated by $\tau^2 = \boldsymbol{C}^T\Sigma\boldsymbol{C} = [1 \quad -1]\begin{bmatrix}\sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22}\end{bmatrix}\begin{bmatrix}1 \\ -1\end{bmatrix} = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$.

From Table 17 and Table 18, we can see that for the *K* independent samples approach, multivariate meta-analysis had relatively larger $\tau^2$ standard errors compared to the

univariate meta-analysis. The true test statistic is still 0.9, both meta-analytic approaches had about 0.03 bias. This may be due to the relatively low convergence rate of multivariate meta-analysis (around 10% to 20%) applied to this one-group data (convergence rates were nearly 100% for all other cases of simulation studies). However, for the cluster-based approach, the multivariate approach had slightly smaller $\tau^2$, test statistics, and standard errors than the univariate approach. The bias in test statistic was small (below 0.01). The convergence rate of multivariate meta-analysis for the cluster-based approach was 100%. All tests correctly rejected the null hypothesis that there is no change over time due to the large sample sizes we are assuming.

Table 17 Case One with 20 balanced clusters using multivariate and univariate meta-analytic approaches

| Meta-Analysis | k | $\tau^2$ | Test Statistic | SE | proportion of p-value<.05 | $\sigma_{11}$ | $\sigma_{12}$ | $\sigma_{22}$ |
|---|---|---|---|---|---|---|---|---|
| Multivariate | 5 | 9.79E-04 | 0.8673 | 0.0140 | 1 | 0 | 2.12E-03 | 0 |
| | 6 | 4.54E-03 | 0.8725 | 0.0317 | 1 | 0 | 3.68E-04 | 0 |
| | Cluster (20) | 0.6720 | 1.0076 | 0.1745 | 1 | 0.507 | 3.91E-03 | 0.173 |
| Univariate | 5 | 8.94E-05 | 0.9287 | 0.0096 | 1 | | | |
| | 6 | 8.25E-05 | 0.9288 | 0.0097 | 1 | | | |
| | Cluster (20) | 0.6952 | 1.0081 | 0.1781 | 1 | | | |

Table 18 Case One with 20 unbalanced clusters using multivariate and univariate meta-analytic approaches

| Meta-Analysis | k | $\tau^2$ | Test Statistic | SE | proportion of p-value<.05 | $\sigma_{11}$ | $\sigma_{12}$ | $\sigma_{22}$ |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Multivariate | 5 | 5.51E-03 | 0.9160 | 0.0413 | 1 | 0 | -9.96E-04 | 0 |
| | 6 | 4.60E-03 | 0.8930 | 0.0307 | 1 | 0 | 7.86E-04 | 0 |
| | Cluster (20) | 0.6661 | 1.0064 | 0.1740 | 1 | 0.503 | 4.58E-03 | 0.172 |
| Univariate | 5 | 8.48E-05 | 0.9288 | 0.0098 | 1 | | | |
| | 6 | 6.23E-05 | 0.9287 | 0.0095 | 1 | | | |
| | Cluster (20) | 0.6893 | 1.0071 | 0.1776 | 1 | | | |

#### 4.3.1.2 Case Two: 20 Clusters, Balanced/Unbalanced, Two-Group Data

The aggregate results over the 1000 simulations for Case Two including 2 groups with balanced and unbalanced clusters using multivariate and univariate meta-analytic approaches are shown in Table 19 and Table 20, respectively. For the *K* independent samples approach, the multivariate meta-analysis gave larger $\tau^2$ and slightly smaller standard errors than the univariate meta-analysis, and all tests correctly rejected the null hypothesis that there is no difference between the two groups in change over time. For the cluster-based approach, the multivariate meta-analysis gave smaller $\tau^2$ and standard errors compared to the univariate meta-analysis. The power for the hypothesis test was almost the same for both meta-analyses (about 70%). Compared to the true test statistic 0.9, the cluster-based approach had lower bias in test statistic than the independent samples approach.

Table 19 Case Two with 20 balanced clusters using multivariate and univariate meta-analytic approaches

| Meta-Analysis | k | $\tau^2$ | Test Statistic | SE | proportion of p-value<.05 | $\sigma_{11}$ | $\sigma_{12}$ | $\sigma_{22}$ |
|---|---|---|---|---|---|---|---|---|
| Multivariate | 5 | 5.34E-04 | 0.9474 | 0.0191 | 1 | 3.78E-04 | 7.76E-05 | 3.11E-04 |
| | 6 | 4.91E-04 | 0.9483 | 0.0197 | 1 | 2.62E-04 | 1.31E-06 | 2.31E-04 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Cluster (20) | 2.5666 | 0.8917 | 0.3458 | 0.67 | 1.80 | 0.14 | 1.04 |
| Univariate | 5 | 3.69E-04 | 0.9405 | 0.0206 | 1 | | | |
| | 6 | 3.54E-04 | 0.9404 | 0.0202 | 1 | | | |
| | Cluster (20) | 2.6798 | 0.8927 | 0.3534 | 0.66 | | | |

Table 20 Case Two with 20 balanced clusters using multivariate and univariate meta-analytic approaches

| Meta-Analysis | k | $\tau^2$ | Test Statistic | SE | proportion of p-value<.05 | $\sigma_{11}$ | $\sigma_{12}$ | $\sigma_{22}$ |
|---|---|---|---|---|---|---|---|---|
| Multivariate | 5 | 6.71E-04 | 0.9557 | 0.0204 | 1 | 2.11E-04 | -1.09E-04 | 2.43E-04 |
| | 6 | 4.30E-04 | 0.9405 | 0.0196 | 1 | 1.54E-04 | 6.04E-05 | 3.97E-04 |
| | Cluster (20) | 2.4706 | 0.9044 | 0.3410 | 0.70 | 1.72 | 1.18E-01 | 0.99 |
| Univariate | 5 | 4.54E-04 | 0.9530 | 0.0209 | 1 | | | |
| | 6 | 3.54E-04 | 0.9449 | 0.0203 | 1 | | | |
| | Cluster (20) | 2.5828 | 0.9043 | 0.3490 | 0.69 | | | |

#### 4.3.1.3 Case Three: Varying Other Study Parameters

Tables 21 and Table 22 show the aggregate results for 1000 simulated datasets for Case Two with 10 and 40 unbalanced clusters, respectively. In both cases for both the *K* independent samples approach and the cluster-based approach, the multivariate meta-analysis gave smaller $\tau^2$ and standard errors compared to the univariate meta-analysis, which was similar to the results for 20 clusters shown in Table 20. Compared to the true test statistic 0.9, the independent samples approach had larger bias in test statistic than the cluster-based approach, and the bias increased as the number of cluster increased. For the cluster-based approach, the power for the hypothesis test decreased to about 50% for both meta-analysis approaches. However, if the number of clusters increased to 40, the power for the hypothesis test increased to 91% due to the larger total sample sizes in this

simulation study. Multivariate and univariate meta-analysis gave similar estimates of $\tau^2$, test statistics, and standard errors in this situation, and the standard errors were relatively smaller than those in the cases of 10 and 20 clusters. Based on Table 20 through 22, we conclude that the performance of multivariate meta-analysis and univariate meta-analysis for the cluster-based approach increased as the total sample size (number of clusters) increased. And we recommend the cluster-based approach over the independent samples approach for the situation with large number of clusters and large treatment effect size.

Table 21 Case Two with 10 unbalanced clusters using multivariate and univariate meta-analytic approaches

| Meta-Analysis | k | $\tau^2$ | Test Statistic | SE | proportion of p-value<.05 | $\sigma_{11}$ | $\sigma_{12}$ | $\sigma_{22}$ |
|---|---|---|---|---|---|---|---|---|
| Multivariate | 5 | 1.44E-03 | 0.9330 | 0.0291 | 0.99 | 6.57E-04 | -8.18E-05 | 6.17E-04 |
| | 6 | 3.14E-04 | 0.9367 | 0.0259 | 1 | 4.84E-04 | 3.09E-04 | 4.48E-04 |
| | Cluster (10) | 2.4731 | 0.8983 | 0.4650 | 0.54 | 1.50 | 8.63E-03 | 9.89E-01 |
| Univariate | 5 | 1.13E-03 | 0.9379 | 0.0301 | 1 | | | |
| | 6 | 9.40E-04 | 0.9351 | 0.0291 | 1 | | | |
| | Cluster (10) | 2.7215 | 0.8977 | 0.4883 | 0.53 | | | |

Table 22 Case Two with 40 unbalanced clusters using multivariate and univariate meta-analytic approaches

| Meta-Analysis | K | $\tau^2$ | Test Statistic | SE | proportion of p-value<.05 | $\sigma_{11}$ | $\sigma_{12}$ | $\sigma_{22}$ |
|---|---|---|---|---|---|---|---|---|
| Multivariate | 5 | 3.74E-04 | 0.9351 | 0.0148 | 1 | 1.22E-04 | -4.92E-05 | 1.54E-04 |
| | 6 | 2.79E-04 | 0.9530 | 0.0142 | 1 | 1.65E-04 | -3.07E-06 | 1.08E-04 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Cluster (40) | 2.5163 | 0.9055 | 0.2459 | 0.91 | 1.74 | 8.21E-02 | 9.37E-01 |
| | 5 | 1.86E-04 | 0.9414 | 0.0145 | 1 | | | |
| Univariate | 6 | 1.66E-04 | 0.9525 | 0.0142 | 1 | | | |
| | Cluster (40) | 2.5989 | 0.9050 | 0.2500 | 0.91 | | | |

Table 23 and Table 24 show the aggregate results for 1000 simulated datasets for Case Two with treatment effect sizes of 0.1 and 0, respectively. The multivariate meta-analysis and univariate meta-analysis behaved similarly in both cases. The power of the hypothesis test for the cluster-based approach was relatively low (about 20%) compared to the power for the *K* independent samples approach if the treatment was less effective (effect size 0.1). When the effect size between treatment groups was 0.0 (reflecting an ineffective treatment), the empirical error rates were all low, with the multivariate and univariate *K* samples approaches having ~10% error, and with the univariate and multivariate cluster-based approaches having ~20% error. When comes to the test statistics, the independent samples approach had lower bias than the cluster-based approach, compared to the true test statistic of 0.48 for treatment effect size of 0.1 and 0 for treatment effect size of 0.

In conclusion, the *K* independent samples approach seems to be out-performing the cluster-based approach when the number of clusters and treatment effect size are small. However, the cluster-based approach is better when the number of clusters and treatment effect size are large with respect to the low bias in test statistics. For both the *K* independent samples and the cluster-based approaches, the multivariate meta-analytic approach performs similarly with univariate meta-analytic approach when using large k

values, though the univariate approaches were better for small numbers of clusters and in the simple one-group, two-time point case.

Table 23 Case Two with 20 unbalanced clusters and a treatment effect size of 0.1 using multivariate and univariate meta-analytic approaches

| Meta-Analysis | k | $\tau^2$ | Test Statistic | SE | proportion of p-value<.05 | $\sigma_{11}$ | $\sigma_{12}$ | $\sigma_{22}$ |
|---|---|---|---|---|---|---|---|---|
| Multivariate | 5 | 7.62E-04 | 0.5128 | 0.0208 | 0.95 | 2.22E-04 | -8.10E-05 | 3.77E-04 |
| | 6 | 2.97E-04 | 0.5149 | 0.0192 | 0.97 | 1.57E-04 | 7.33E-05 | 2.86E-04 |
| | Cluster (20) | 3.1500 | 0.2617 | 0.3858 | 0.22 | 2.26 | 4.83E-02 | 9.86E-01 |
| Univariate | 5 | 3.73E-04 | 0.5129 | 0.0207 | 0.97 | | | |
| | 6 | 3.14E-04 | 0.5193 | 0.0202 | 0.97 | | | |
| | Cluster (20) | 3.2686 | 0.2597 | 0.3937 | 0.21 | | | |

Table 24 Case Two with 20 unbalanced clusters and an ineffective treatment effect using multivariate and univariate meta-analytic approaches

| Meta-Analysis | k | $\tau^2$ | Test Statistic | SE | Proportion of p-value<.05 | $\sigma_{11}$ | $\sigma_{12}$ | $\sigma_{22}$ |
|---|---|---|---|---|---|---|---|---|
| Multivariate | 5 | 1.03E-03 | 0.0715 | 0.0220 | 0.09 | 1.40E-04 | -1.95E-04 | 5.02E-04 |
| | 6 | 2.06E-04 | 0.0660 | 0.0190 | 0.08 | 1.27E-04 | 1.27E-04 | 3.33E-04 |
| | Cluster (20) | 4.1420 | -0.4779 | 0.4495 | 0.21 | 3.10 | -2.39E-02 | 9.93E-01 |
| Univariate | 5 | 3.52E-04 | 0.0716 | 0.0209 | 0.08 | | | |
| | 6 | 2.95E-04 | 0.0660 | 0.0204 | 0.09 | | | |
| | Cluster (20) | 4.2935 | -0.4824 | 0.4580 | 0.22 | | | |

### 4.4 Analysis of Cancer Screening Data

Here we apply the multivariate meta-analytic approach to the colon cancer screening data [1] used in Chapter 3. There are two treatment groups (the MPC program users and non-

users), 4 repeated measurements at baseline, one month, three months, and six months, and nested cluster structures (patients within physicians within practices) in the data.

The same hierarchical linear mixed model was fit including the repeated-measure binary outcome (patient underwent colon cancer screening or they did not), fixed effects for time, group, and their interaction, as well as random effects for primary care physicians, who are in turn nested with one of eight practices. This model was fit using the GLIMMIX procedure in SAS to test whether there were differences in the change in colon cancer screening rates between MPC program users and non-users from baseline to each of one, three and six months. The group-specific test statistic of change of colon cancer screening rates are used as observed effects for multivariate meta-analysis. We chose a subset of the study with n=15,652 subjects where the desired model did converge as the overall database in this example. We split the overall database into large k from 5 to 8 independent sub-samples using the k independent samples approach and used the 8 practices as 8 sub-samples for the cluster-based approach.

Table 25 shows the test results of differences in change in log-odds between users and non-users from baseline to month 1, 3, and 6 applied to the sub-samples we created and recombined by multivariate meta-analytic approach. The results of the same tests by univariate meta-analytic approach are also provided for comparison purpose. For the independent samples approach, all test statistics were correctly significant except for $k = 8$ from baseline to month 6 using univariate meta-analytic approach. The estimated inter-

cluster variances $\tau^2$ and the standard errors of the test statistic by multivariate meta-analysis were slightly lower than those by univariate meta-analysis, resulting in lower p-values of the hypothesis test. For the cluster-based approach, all test statistics were not significant, which did not match the results of whole dataset presented in Table 25 in Chapter 3. The inter-cluster variances and the standard errors from multivariate meta-analysis were slightly larger than those by univariate meta-analysis, resulting in larger p-values of the hypothesis test. However, since all tests were not significant, we conclude that the clustered based approach did not perform as well as the *K* independent samples approach when results were combined by either the univariate or multivariate meta-analysis. Compared to the results from the whole dataset, multivariate meta-analysis for large k values behaved similarly with, or at least as well as univariate meta-analysis, although there was one test was incorrectly non-significant.

Table 25 Test results applied to cancer screening data by multivariate and univariate meta-analytic approaches

| k | User/Non-User | Multivariate | | | | | | | Univariate | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\tau^2$ | Test Statistic | SE | p-value | $\sigma_{11}$ | $\sigma_{12}$ | $\sigma_{22}$ | $\tau^2$ | Test Statistic | SE | p-value |
| k=5 | baseline to month1 | 0.0029 | 0.2048 | 0.0390 | <.0001* | 4.33E-03 | 7.07E-04 | 0 | 0.0047 | 0.2051 | 0.0435 | <.0001* |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | baseline to month3 | 0.0018 | 0.2060 | 0.0452 | <.0001* | 3.62E-03 | 9.20E-04 | 0 | 0.0034 | 0.2065 | 0.0489 | <.0002* |
| | baseline to month6 | 0.0037 | 0.1549 | 0.0555 | 0.0053* | 5.08E-03 | 6.75E-04 | 0 | 0.0100 | 0.1583 | 0.0662 | 0.0168* |
| k=6 | baseline to month1 | 0.0045 | 0.1804 | 0.0403 | <.0001* | 4.28E-03 | -1.00E-04 | 0 | 0.0060 | 0.1785 | 0.0434 | <.0001* |
| | baseline to month3 | 0.0012 | 0.1749 | 0.0420 | <.0001* | 4.22E-03 | 2.27E-03 | 1.47E-03 | 0.0048 | 0.1740 | 0.0489 | 0.0004* |
| | baseline to month6 | 0 | 0.1242 | 0.0459 | 0.0068* | 0 | 1.13E-04 | 0 | 0.0042 | 0.1225 | 0.0536 | 0.0222* |
| k=7 | baseline to month1 | 0.0003 | 0.1755 | 0.0297 | <.0001* | 0 | -1.10E-04 | 1.20E-04 | 0.0004 | 0.1787 | 0.0298 | <.0001* |
| | baseline to month3 | 0 | 0.1663 | 0.0380 | <.0001* | 0 | 2.48E-04 | 0 | 0 | 0.1658 | 0.0390 | <.0001* |
| | baseline to month6 | 0 | 0.1018 | 0.0388 | 0.0087* | 0 | 1.85E-03 | 2.64E-04 | 0 | 0.0920 | 0.0456 | 0.0437* |
| k=8 | baseline to month1 | 0.0127 | 0.1836 | 0.0497 | 0.0002* | 1.52E-02 | 1.34E-03 | 1.69E-04 | 0.0162 | 0.1793 | 0.0542 | 0.0009* |
| | baseline to month3 | 0.0093 | 0.1642 | 0.0524 | 0.0017* | 9.27E-03 | 6.45E-04 | 1.28E-03 | 0.0122 | 0.1635 | 0.0559 | 0.0035* |
| | baseline to month6 | 0.0149 | 0.1252 | 0.0639 | 0.0499* | 1.93E-02 | 3.41E-03 | 2.44E-03 | 0.0213 | 0.1217 | 0.0701 | 0.0823 |
| Cluster (8) | baseline to month1 | 0.0224 | 0.0834 | 0.0637 | 0.1908 | 4.02E-02 | 1.37E-02 | 9.68E-03 | 0.0203 | 0.0842 | 0.0619 | 0.1737 |
| | baseline to month3 | 0.0220 | 0.0797 | 0.0696 | 0.2523 | 2.86E-02 | 3.30E-03 | 6.80E-05 | 0.0138 | 0.0976 | 0.0627 | 0.1197 |
| | baseline to month6 | 0.0066 | 0.0910 | 0.0601 | 0.1299 | 1.09E-02 | 3.11E-03 | 1.83E-03 | 0.0064 | 0.1073 | 0.0604 | 0.0758 |

## 4.5 Discussion

To work around the non-convergence issue for clustered dependent binary data, we split data into subsets either by the *K* independent samples approach or the cluster-based approach and recombine results from subsets by a meta-analytic approach. We have shown the efficiency of using univariate meta-analytic approach in Chapter 3, and we also investigated the performance of using multivariate meta-analytic approach to recombine results from subsets in this Chapter. Both results from simulation studies and data on cancer screening showed that multivariate meta-analytic approach behaved similarly with univariate meta-analytic approach. We recommend using the independent samples approach over the cluster-based approach when the number of clusters and treatment effect size are small, and in this case the smallest possible number of K works

best. However, the cluster-based approach is preferred if data has large number of clusters and treatment effect size.

One limitation in the presentation of this research is that multivariate meta-analysis is only suitable for large *k* values. We only considered the number of independent sub-samples $k \geq 5$ in this research. However, if k is small, using univariate meta-analytic approach is more straightforward than using multivariate meta-analytic approach since we have already shown that the smallest possible number of independent sub-samples works best if we use univariate meta-analysis. If k is small, our suggestion is to use the *K* independent samples approach to split data and recombine by univariate meta-analytic approach.

Another limitation in this research is that the independent sub-samples are generated arbitrarily for the independent samples approach, as described in Chapter 3. We did not consider repeating the independent sample generation process by creating different sub-samples sets and re-estimating the meta-analysis outcomes due to long computational time. This actually affects the multivariate meta-analysis success rate because we use a mixed model to do multivariate meta-analysis and these mixed model's convergence depends on the data which is the sets of sub-samples when the sample size is so small. We will evaluate the performance of bootstrapping the sets of independent sub-samples as well as increasing the number of independent sub-samples k in our future work.

# Chapter 5 Discussion

We provided an innovative idea to work around the non-convergence issue of using mixed models to analyze large, clustered dependent binary data. This approach splits the overall database into smaller components that are analyzable by the desired model and recombines the separate results through meta-analytic approaches. We showed several ways to split the data, either by generating *K* independent samples or by using the naturally existing clusters. We also investigated using either the univariate or multivariate meta-analytic approaches. For both meta-analytic approaches, splitting the data into independent samples worked better than using the natural clusters if data have small number of clusters and treatment effect size. And we recommend to use the smallest possible number of sub-samples for the *K* independent samples approach in this case, with respect to the most similar estimates with results from the whole dataset. If data have large number of clusters and treatment effect size, then the cluster-based approach is preferred.

There are several limitations in the presentation of this research. Firstly, this approach is only applicable to data with large sample size ($n \gg 1000$). However, non-convergence issue is rarely found in data with small sample size. If a desired model does not converge for a "small" data, we do not recommend our approach, and researchers should consider

other ways to work around this issue. We also only focused on binary outcomes in our research. The approaches here could be applied to cases with continuous outcomes where the desired model does not converge, though their effectiveness will need to be investigated. In addition, our simulation template was not exhaustive with respect to the number of repeated measures, clustering or nesting types, and the number of treatment groups. We will consider varying these factors in our simulation templates to evaluate the performance of our approach in future work.

Another limitation in our research is that we did not consider the situation when some clusters were so large that the desired model would still not converge using the cluster-based approach. One possible solution will be to split those non-converging clusters into smaller subsets that converge using the independent samples approach, and then combine all subsets using meta-analysis. Since we would be creating subsets out of individual clusters, we may need to account for that dependence in our meta-analytic technique when combining the results. We would also need to investigate whether this approach works better than just using the independent samples approach from the start. We will study this question in our future work.

Another limitation as mentioned in the previous Chapters is that we did not try bootstrapping the process of generating the $K$ independent sub-samples. The main reason for not trying bootstrapping is that the computational time is exponentially long. And the estimate of inter-cluster variance is low in all cases suggesting that splitting for more than

one time may not improve the results. If splitting the data once and the estimated inter-cluster variance is large, then we recommend to generate another set of independent sub-samples to see whether the inter-cluster variance decreases.

An important statistical value of our research is that the idea that we provided is not limited to apply on clustered data only. Theoretically, one may try to apply this "split and combine" process to other non-convergence models as long as smaller components of the overall database are analyzable by the desired model. The biggest advantage of our idea is that the hierarchical structure of the desired model is not reduced. We also did not omit any information the whole dataset provided.

# List of References


1. Krist AH, Aycock RA, Etz RS, Devoe JE, Sabo RT, Williams R, Stein KL, Iwamoto G, Puro J, Deshazo J, Kashiri PL, Arkind J, Romney C, Kano M, Nelson C, Longo DR, Wlover S, Woolf SH (2015). MyPreventiveCare: implementation and dissemination of an interactive preventive health record in three practice-based research networks serving disadvantaged patients. *Implementation Science* 9(181): epub.
2. Murray DM, Varnell SP, Blitstein JL(2003). Design and analysis of group-randomized trials: a review of recent methodological advances. *American Journal of Public Health*. 94(3):423-432
3. Galbraith S, Daniel JA, Vissel B (2010). A study of clustered data and approaches to its analysis. *The Journal of Neuroscience.* 30(32):10601-10608
4. Campbell MK, Grimshaw JM (1998). Cluster randomized trials: time for improvement. The implications of adopting a cluster design are still largely being ignored [editorial]. *BMJ*. 317:1171-1172
5. Zyzanski SJ, Flocke SA, Dickinson LM (2004). On the Nature and Analysis of Clustered Data. *Annals of Family Medicine*. 2(3): 199–200.
6. Kerry SM, Bland JM (1998). The intracluster correlation coefficient in cluster randomization. *BMJ*. 316(7142):1455-1460

7. Killip S, Mahfoud Z, Pearce K (2004). What is an intracluster correlation coefficient? crucial concepts for primary care researchers. *Annals of Family Medicine*. 2(3):204-208
8. SAS Institute Inc. 2008. SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc.
9. Liang X, Laurence VM (2014). %HPGLIMMIX: A high-performance SAS macro for GLMM estimation. *Journal of Statistical Software*. 58(8): 1-25
10. Fitzmaurice G.M (n.d.). Overview of Methods for Analyzing Cluster-Correlated Data [PowerPoint slides]. Retrieved April 15, 2016, from https://catalyst.harvard.edu/docs/biostatsseminar/Fitzmaurice_BSP-Workshop-Slides.pdf
11. Cappelleri JC, Loannidis JP, Schmid CH, Ferranti SD, Aubert M, Chalmers T, Lau J (1996). Large trials vs meta-analysis of smaller trials—how do their results compare? *The Journal of the American Medical Association* 276(16): 1332-1338.
12. The Institute of Medicine (2001). *Small Clinical Trials: Issues and Challenges*. Washington, DC: The National Academies Press.
13. DerSimonian R, Kacker R (2007). Random-effects model for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials* 28: 105-114.
14. Haidich AB (2010). Meta-analysis in medical research. *Hippokratia*. 14(Suppl 1): 29–37.
15. Pearson K (1904). Report on certain enteric fever inoculation statistics. *BMJ*. 3(2288): 1243-1246

16. O'Rourke K (2007). An historical perspective on meta-analysis: dealing quantitatively with varying study results. *Journal of the Royal Society of Medicine*. *100*(12): 579-582.
17. Glass GV(1976). Primary, secondary and meta-analysis of research. *Educ Researcher*.10: 3-8
18. Reinard, J. (n.d.). Chapter 12: Meta-Analysis. *Communication Research Statistics*. Retrieved April 15, 2016, from http://commfaculty.fullerton.edu/jreinard/stat_ch12.htm
19. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009). *Introduction to metaanalysis*. Chichester: Wiley.
20. DerSimonian R, Laird N (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* 7: 177-187
21. Houwelingen HC, Arends LR, Stinen T (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 21:589-624.
22. Emrich LJ, Piedmonte MR (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician* 45(4): 302-304.
23. Kang S-H, Jung S-H (2001). Generating correlated binary variables with complete specification of the joint distribution. *Biometrical Journal* 43(3): 263-269.
24. Sabo RT, Haynes ME, Chaganty NR (2015). Using odds ratios to simulate dependent binary outcomes. *Communications in Statistics: Simulation and Computation*; *submitted*.

25. Koning IM, Eijnden RJ, Verdurmen JE, Engels RC, Vollebergh WA (2011). Long-term effects of a parent and student intervention on alcohol use in adolescents:a cluster randomized controlled trial. *American Journal of Preventive Medicine* 40(5):541–547.
26. Joe H (1997). *Multivariate Models and Dependence Concepts*. London: Chapman and Hall.
27. Chaganty NR, Joe H (2006). Range of correlation matrices for dependent Bernoulli random variables. *Biometrika* 93(1): 197-206.
28. Haynes ME, Sabo RT, Chaganty NR (2016). Simulating dependent binary variables through multinomial sampling. *Journal of Statistical Computation and Simulation* 86(3): 510-523.
29. Qaqish BF(2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* 90: 455-463.
30. Donner A, Birkett N, Buck C (1981). Randomization by cluster: sample size requirements and analysis. *American Journal of Epidemiology* 114: 906-914.
31. Stevens, J. R., Taylor, A. M. (2009). Hierarchical dependence in meta-analysis. *Journal of Educational and Behavioral Statistics* 34: 46-73.
32. Wang A, Sabo RT (2015). Simulating clustered and dependent binary variables. *Austin Biometrics and Biostatistics* 2(2): e1020.
33. White IR (2009). Multivariate random-effects meta-analysis. *The Stata Journal* 9(1): 40–56.

34. Becker BJ (2000). Multivariate meta-analysis. In *Handbook of applied multivariate statistics and mathematical modeling*. Academic Press: 501-502.

# Appendix A. SAS code for simulating clustered dependent binary data

```
/* two time points case */ /* normal dist */
/* include group var: treatment vs. control */
libname lib "c:\d\bios\dissertation\simulation_new\sim2\norm";
proc iml;
repeat=500;
study=0;
seed=12345;
do ii=1 to repeat;
        study=study+1;
        seed=seed+1;
        k=20;  /* total number of clusters */
        n=100; /* number of subjects per cluster */
        rou12=0.2;
        c=j(k,1,1);
        s=j(k/2,1,seed);
        s1=j(k,1,seed);
        sigma=0.1;
        /* treatment group */
                z1=normal(s);
                p1_t=0.25+z1*sigma; /* mu1=0.25 */
                z2=normal(s);
                p2_t=0.45+z2*sigma; /* mu2=0.45 */
```

```
/* control group */
        z1=normal(s);
        p1_c=0.25+z1*sigma; /* mu1=0.25 */
        z2=normal(s);
        p2_c=0.25+z2*sigma; /* mu2=0.25 */
 p1=p1_t//p1_c;
 p2=p2_t//p2_c;
do j=1 to k;
        if p1[j]>1 then p1[j]=0.999;
        if p1[j]<0 then p1[j]=0.001;
        if p2[j]>1 then p2[j]=0.999;
        if p2[j]<0 then p2[j]=0.001;
end;

q1=c-p1;
q2=c-p2;
p12=p1#p2+rou12*sqrt(p1#q1)#sqrt(p2#q2); /* # : elementwise multiplication */
/* joint pdf */
p11=p12;
p10=p1-p12;
p01=p2-p12;
p00=c-p1-p2+p12;
/* cdf bounds */
b1=p11;
b2=b1+p10;
b3=b2+p01;
do i=1 to n;
   u=uniform(s1);
        out1 = (u<=b2);
        out2 = (u<=b1 | ( b2<u & u<=b3 ));
```

```
                y1=y1//out1;
                y2=y2//out2;
                cluster=cluster//(1:k)`;
                group=group//j(k/2,1,1)//j(k/2,1,2); /* group 1 means treatment, group 2
means control */
        end;
        mean1=mean(y1);
        mean2=mean(y2);
        *print mean1 mean2;
        id=id//(1:(n*k))`;
        study_ind=study_ind//j(n*k,1,study);
end;

create lib.t2_cluster2 var{ study_ind id y1 y2 cluster group };
append;
close lib.t2_cluster2;
quit;

proc sort data=lib.t2_cluster2; by group; run;
proc means data=lib.t2_cluster2; by group; var y1 y2; run;

data lib.t2_repeat_long_2; set lib.t2_cluster2;
y=y1; time=1;  output;
y=y2; time=2;  output;
drop y1 y2; run;

proc sort data=lib.t2_repeat_long_2; by study_ind id time; run;
proc glimmix data=lib.t2_repeat_long_2 maxopt=500 pconv=1e-6 method=MSPL;
by study_ind;
class time id cluster group;
```

```
model y(event='1') = time group time*group /dist=binary link=logit solution covb;
random intercept /subject=cluster;
random _residual_ /subject=id type=ar(1);
nloptions technique=quanew maxiter=400 gconv=1e-5;
lsmeans time*group / ilink e cov;
estimate "group1: time1 to time2" time -1 1 time*group -1  0  1  0 /ilink e;
estimate "group2: time1 to time2" time -1 1 time*group 0  -1  0  1/ilink e;
estimate "group 1-2 diff: time1 to time2"   time*group -1  1  1  -1/ilink e;
ods output lsmeans = lib.t2_repeat_long_2_lsmeans;
ods output CovParms = lib.t2_repeat_long_2_random;
ods output estimates = lib.t2_repeat_long_2_estm;
run;

proc sort data=lib.t2_repeat_long_2_lsmeans; by group time; run;
proc means data=lib.t2_repeat_long_2_lsmeans mean std;
by group time;
var mu stderrmu;
run;

/* p2-p1 */
data time_diff;
set lib.t2_repeat_long_2_lsmeans;
keep study_ind time group mu;
run;
proc sort data=time_diff; by  group study_ind  time; run;

proc transpose data=time_diff out=time_diff_1 prefix=p;
by group study_ind ;
id time;
var mu;
```

```
run;

data time_diff_1;
set time_diff_1;
diff=p2-p1;
run;

proc means data=time_diff_1 mean std;
by group;
var diff;
run;


proc sort data=lib.t2_repeat_long_2_estm; by statement; run;
proc means data=lib.t2_repeat_long_2_estm mean std;
by statement;
var probt ;
run;

proc sort data=lib.t2_repeat_long_2_random; by covparm; run;
proc means data=lib.t2_repeat_long_2_random mean;
by covparm;
var estimate;
run;
```

# Appendix B. SAS code for the independent samples approach and fitting desired model to sub-samples

```
/* This macro select random samples of specified size and run glimmix */
%macro select(study,k,seed,i,size,origname,origset,rootset,rootlib);
        /* study represents the study index */
        /* i indicates ith selection of samples */
        /* origset & origlib are baseline dataset and its folder*/
        /* rootset & rootlib are  whole   dataset and its folder */

        proc sort data=&origset; by group cluster ID time; run;
        /* select among baseline data */
        proc surveyselect data=&origset out=&origset.select
                method=srs sampsize=&size seed=&seed;
                strata group cluster /alloc=proportional;
        run;

        /* selected IDs */
        data IDs; set &origset.select; keep ID; run;

        proc sort data=IDs; by ID; run;

        /* selected samples */
        proc sort data=&rootset; by ID; run;
        data &rootset.select;
```

```
        merge IDs(in=b) &rootset (in=a);
        by ID;
        if b;
    run;
    /* rest baseline data */
    proc sort data=&origset; by ID; run;
    data &origname.&i.rest;
        merge &origset (in=a) IDs(in=b);
        by ID;
        if a & ~b;
    run;


    /*glm on selected samples */
    proc sort data=&rootset.select; by ID time; run;
    proc glimmix data=&rootset.select maxopt=500 pconv=1e-6 method=MSPL;
    class group cluster id time;
    model y(event='1') = time group time*group /dist=binary link=logit solution
covb;
    random intercept /subject=cluster;
    random _residual_ /subject=id type=ar(1);
    nloptions technique=quanew maxiter=400 gconv=1e-5;
    lsmeans time*group / ilink e cov;
    estimate "group1: time1 to time2" time -1 1 time*group -1  1  0  0 /ilink e;
    estimate "group2: time1 to time2" time -1 1 time*group 0  0  -1  1/ilink e;
    estimate "group 1-2 diff: time1 to time2"   time*group -1  1  1  -1/ilink e;
    ods output lsmeans = lsmeans;
    ods output CovParms = random;
    ods output estimates = estm;
    run;
    data lsmeans; set lsmeans; study_ind=&study; sample_ind=&i; run;
```

```
        data random;  set random;  study_ind=&study; sample_ind=&i; run;
        data estm;    set estm;    study_ind=&study; sample_ind=&i; run;

        /* add estimates */
        proc append base=&rootlib..lsmeans_k&k data=lsmeans ; run;
        proc append base=&rootlib..random_k&k data=random ; run;
        proc append base=&rootlib..estm_k&k data=estm ; run;
%mend select;

/* This macro generate simulation dataset */
/* and output summary statistics of mlm on each subsample*/
%macro meta(k,seed1,root_lib);
%let kminus1= %eval(&k - 1) ;
%let repeat=1000;
%do j=1 %to &repeat;
        proc iml;
                seed2= &seed1+ 2*&j;
                nk=20; /* # of clusters */
                rou12=0.2;
                c=j(nk*2,1,1);
                s1=j(nk,1,seed2);
                s2=j(nk*2,1,seed2);
                sigma=0.1;
                /* p1 is the same for both treatment and control group */
                z1=normal(s1);
                p1_tc=0.25+z1*sigma;/* mu1=0.25 */
                /* p2 for treatment group */
                del=normal(s1);
                sigma1=0.2;
                delta=0.2+del*sigma1;
```

```
p2_t=p1_tc+delta; /* mu2 = 0.45 */
/* p2 for control group */
z2=normal(s1);
p2_c=0.25+z2*sigma; /* mu2=0.25 */
p1=p1_tc//p1_tc;
p2=p2_t//p2_c; /* treatment then control */
do m=1 to nk*2;
        if p1[m]>1 then p1[m]=0.999;
        if p1[m]<0 then p1[m]=0.001;
        if p2[m]>1 then p2[m]=0.999;
        if p2[m]<0 then p2[m]=0.001;
end;
q1=c-p1;
q2=c-p2;
p12=p1#p2+rou12*sqrt(p1#q1)#sqrt(p2#q2); /* # --> elementwise
multiplication */
/* joint pdf */
p11=p12;
p10=p1-p12;
p01=p2-p12;
p00=c-p1-p2+p12;
/* cdf bounds */
b1=p11;
b2=b1+p10;
b3=b2+p01;
z3=normal(s2);
n=2500+z3*200;
n=round(n);/* # of subjects for each cluster of trt and control group*/
do mm=1 to nk*2;
s3=j(n[mm],1,seed2);
```

```
        u=uniform(s3);
                out1 = (u<=b2[mm]);
                out2 = (u<=b1[mm] | ( b2[mm]<u & u<=b3[mm] ));
                y1=y1//out1;
                y2=y2//out2;
                if mm<nk+1 then do;  clust=mm; gp=1; end;
                else do; clust=mm-nk; gp=2; end;
                cluster=cluster//j(n[mm],1,clust);
                group=group//j(n[mm],1,gp);
        end;

Ntotal=sum(n);
samplesize = round(Ntotal/&k);
create size var {samplesize};
append;
close;
id=(1:Ntotal)`;
study_ind=j(Ntotal,1,&j);
create root var{ study_ind id y1 y2 cluster group};
append;
close root;
quit;
data _NULL_;
        set size;
        call symput('smpsz',trim(left(put(samplesize,8.))));
run;
data root;
        set root;
        y=y1; time=1;  output;
        y=y2; time=2;  output;
```

```
        drop y1 y2;
    run;
    proc sort data=root; by id time; run;
    data base0rest; set root; if time=1;run;

    %let seed= %eval(&seed1+ &j);

    /* the first k-1 samples */
    %do jj=1 %to &kminus1;
        %let ii = %eval(&jj-1);
        %select(&j, &k, &seed, &jj, &smpsz, base, base&ii.rest, root, &root_lib);
    %end;
    /* the last sample */
    proc sort data=base&kminus1.rest; by ID; run;
    data rootselect;
        merge base&kminus1.rest(in=b) root(in=a);
        by ID;
        if b;
    run;
    proc sort data= rootselect; by ID time; run;
    /* run glimmix on last sample */
    proc glimmix data=rootselect maxopt=500 pconv=1e-6 method=MSPL;
    class group cluster id time;
    model y(event='1') = time group time*group /dist=binary link=logit solution
covb;
    random intercept /subject=cluster;
    random _residual_ /subject=id type=ar(1);
    nloptions technique=quanew maxiter=400 gconv=1e-5;
    lsmeans time*group / ilink e cov;
    estimate "group1: time1 to time2" time -1 1 time*group -1  1  0  0 /ilink e;
```

```
        estimate "group2: time1 to time2" time -1 1 time*group 0  0  -1  1/ilink e;
        estimate "group 1-2 diff: time1 to time2"   time*group -1  1  1  -1/ilink e;
        ods output lsmeans = lsmeans;
        ods output CovParms = random;
        ods output estimates = estm;
        run;

        data lsmeans; set lsmeans; study_ind=&j; sample_ind=&k; run;
        data random;  set random;  study_ind=&j; sample_ind=&k; run;
        data estm;    set estm;    study_ind=&j; sample_ind=&k; run;

        /* add estimates */
        proc append base=&root_lib..lsmeans_k&k data=lsmeans  ; run;
        proc append base=&root_lib..random_k&k data=random ; ; run;
        proc append base=&root_lib..estm_k&k data=estm  ; run;

%end;
%mend meta;
libname rootlib "C:\D\BIOS\Dissertation\simulation_new\sim6\k4\c20_02";
%meta(4,1234,rootlib);
```

# Appendix C. SAS code for univariate meta-analysis

```
libname lib1 "C:\D\BIOS\Dissertation\simulation_new\sim6\k6\c10_02";
data one; set lib1.Estm_k6; run;
data one; set one; if statement=3; run;

proc iml;
  k=6;
  do ii= 1 to 1000; /* ii is the statement number */
        use one;
        read all var {Estimate StdErr} where(Study_ind=ii);
        close one;
        var=StdErr##2;
        *** t2(DL) ***;
        Ai = 1/var;
        yw = sum(Ai#Estimate)/sum(Ai);
        y_w=j(k,1,yw);

        t2_DL = (sum(Ai#((Estimate-y_w)##2))- (k-1)) / (sum(Ai)-ssq(Ai)/sum(Ai));
        t2_DL = max(0,t2_DL);
        w_i = 1/(t2_DL + var);
        mw_DL = sum(w_i#Estimate)/sum(w_i);
        se_DL = 1/sqrt(sum(w_i));

        *** t2(DL2) ***;
        Ai = 1/(t2_DL + var);
```

```
        t2_DL2 = (sum(Ai#((Estimate-mw_DL)##2))-
sum(Ai#var)+sum((Ai##2)#var)/sum(Ai) ) / (sum(Ai)-ssq(Ai)/sum(Ai));
        t2_DL2 = max(0,t2_DL2);

        *** estimate mu ***;
        w_i = 1/(t2_DL2+var);
        mw_DL2 = sum(w_i#Estimate)/sum(w_i);
        se_DL2 = 1/sqrt(sum(w_i));

        t2_DL2_total = t2_DL2_total//t2_DL2;
        mw_DL2_total = mw_DL2_total//mw_DL2;
        se_DL2_total = se_DL2_total//se_DL2;
        study_ind=study_ind//ii;
 end;

create lib1.uni_result_k6 var {t2_DL2_total mw_DL2_total se_DL2_total study_ind};
append;
close lib1.uni_result_k6;
quit;

proc means data=lib1.uni_result_k6;
var t2_DL2_total mw_DL2_total se_DL2_total;
run;

data lib1.uni_result_k6; set lib1.uni_result_k6; z=mw_DL2_total/se_DL2_total;
  if abs(z)>1.96 then reject=1; else reject=0;
run;
proc means data=lib1.uni_result_k6; var reject; run;
```

# Appendix D. SAS code for multivariate meta-analysis

```
libname lib1 "C:\D\BIOS\Dissertation\simulation_new\sim6\k6\c10_02";
data data1; set lib1.estm_k6; if statement<3; run;
data data3; set data1;
if statement=1 then do exp=1; con=0; trial=sample_ind; var=StdErr**2; end;
if statement=2 then do exp=0; con=1; trial=sample_ind; var=StdErr**2; end;
keep statement study_ind exp con trial estimate StdErr var;
run;

data data4;  set data3; arm = mod(_N_,12); if arm=0 then arm=12; run;
data covvars; do study_ind= 1 to 1000; est=0;output; output; output; end; run;
data covvars1; set data4; keep study_ind var; rename var=est; run;
data covvars2; set covvars covvars1; by study_ind; run;

ods output CovParms=CovP Estimates=lib1.estmdiff;
proc mixed cl method=ml data=data4 asycov;
by study_ind;
class trial arm;
model estimate = exp con / noint s cl covb ;
random exp con/ subject=trial type=un s;
repeated /group=arm;
estimate 'difference' exp 1 con -1 /cl df=1000;
parms /parmsdata=covvars2 eqcons = 4 to 15 ; /*first three elements are zero */
run;
```

```
proc means data=lib1.estmdiff; var study_ind; run;
data CovP;
set CovP;
if Subject = 'trial';
keep study_ind CovParm Estimate;
run;

proc transpose data=CovP out=CovP1; by Study_ind; var estimate; id CovParm; run;

data lib1.CovP;
set CovP1;
rename UN_1_1_=UN11 UN_2_1_=UN21 UN_2_2_=UN22;
tau2 = UN_1_1_ + UN_2_2_ - 2*UN_2_1_;
run;

proc means data=lib1.CovP ; var UN11 UN21 UN22 tau2; run;

proc means data=lib1.estmdiff; var estimate stderr; run;

data lib1.estmdiff;
set lib1.estmdiff;
z=estimate/stderr;
if abs(z)>1.96 then reject=1; else reject=0;
run;
proc means data=lib1.estmdiff; var reject; run;
```

# Appendix E. SAS code for checking hierarchical dependence across sub-samples

```
libname lib1 "C:\D\BIOS\Dissertation\simulation_new\sim6\k5\c20_02";
data one; set lib1.estm_k5; run;
data one; set one; if statement=3; run;

proc iml;
k=5;
do ii = 1 to 1000; /* ii is the study number */
                use one;
                read all var {Estimate StdErr} where(Study_ind=ii);
                close one;

                var=StdErr##2;
                V=diag(var);
                M=j(k,k,1)-I(k);
                Ident=I(k);
                X=j(k,1,1);

                sig_k=0;
                tau_k=0;
                sig_k1=100;
                tau_k1=100;
                m=0;
```

```
do while( max(abs(tau_k-tau_k1), abs(sig_k-sig_k1)) > 1e-5 );
m=m+1;
if tau_k1=100 then tau_k=0; else tau_k=tau_k1; * check if initial value;
if sig_k1=100 then sig_k=0; else sig_k=sig_k1;

phi= V + (tau_k**2)*Ident + sig_k*M;
invp = inv(phi);
A = invp - invp*X*inv(t(X)*invp*X)*t(X)*invp;
RSS = t(estimate)*A*estimate;

tau_k1 = (RSS - trace(A*V) - sig_k*trace(A*M))/trace(A);
if tau_k1 > 0  then tau_k1 = sqrt(tau_k1); else tau_k1 = 0; *if tau2<0 then
set to 0;
phi_k1 = V + (tau_k1**2)*Ident + sig_k*M;
invp = inv(phi_k1);
A_k1 = invp - invp*X*inv(t(X)*invp*X)*t(X)*invp;
RSS_k1 = t(estimate)*A_k1*estimate;
sig_k1 = (RSS_k1 - trace(A_k1*V) - (tau_k1**2)*trace(A_k1)) /
trace(A_k1*M);
        if (sig_k1 > (tau_k1**2)) | (sig_k1 < (-tau_k1**2)/(k-1)) then do;
            if abs(sig_k1-(tau_k1**2))< abs(sig_k1+(tau_k1**2)/(k-1))
then sig_k1 = tau_k1**2;
            else sig_k1 = -(tau_k1**2)/(k-1);
        end;/*ensure phi is positive definite */
end;
phi_f = V + (tau_k1**2)*Ident + sig_k1*M;
se22=inv(t(X)*inv(phi_f)*X);

beta_i = inv(t(X)*inv(phi_f)*X)*t(X)*inv(phi_f)*estimate;
se_i = sqrt(inv(t(X)*inv(phi_f)*X));
```

```
                z_i = abs(beta_i)/se_i;
                p_i = 2*(1 - probnorm(z_i));

                beta = beta//beta_i;
                se = se //se_i;
                pvalue = pvalue//p_i;
                itera = itera//m;
                tau2 = tau2//(tau_k1**2);
                sig = sig//sig_k1;
                study_index = study_index//ii;
end;

create uni_dependence var {sig tau2 beta se pvalue study_index};
append;
close uni_dependence;
quit;

proc means data=uni_dependence_k5; run;
```